RAIRO-Oper. Res. 51 (2017) 719–728
DOI: 10.1051/ro/2016059

RAIRO Operations Research
www.rairo-ro.org

A HYBRID ARIMA-ANN APPROACH FOR OPTIMUM ESTIMATION AND FORECASTING OF GASOLINE CONSUMPTION

Reza Babazadeh[1]

Abstract. Accurate estimation and forecasting of gasoline is vital for policy and decision-making process in energy sector. This paper presents a hybrid data-driven model based on Artificial Neural Network (ANN) and autoregressive integrated moving average (ARIMA) approach for optimum estimation and forecasting of gasoline consumption. The proposed hybrid ARIMA-ANN approach considers six lagged variables and one forecasted values provided by ARIMA process. The ANN trains and tests data with Multi Layer Perceptron (MLP) approach which has the lowest Mean Absolute Percentage Error (MAPE). To show the applicability and superiority of the proposed hybrid approach, daily available data were collected for 7 years (2005–2011) in Iran. Although eliminating subside from gasoline price has led to appearing noisy data in gasoline consumption in Iran the acquired results show high accuracy of about 9427% by using the proposed hybrid ARIMA-ANN method. The results of the proposed model are compared respect to regression's models and ARIMA process. The outcome of this paper justifies the capability of the proposed hybrid ARIMA-ANN approach in accurate forecasting gasoline consumption.

Mathematics Subject Classification. 60G25, 62M10, 91B84, 92B20, 62P30.

Received June 7, 2015. Accepted August 30, 2016.

1. Introduction

Time series forecasting is an interesting research area which has attracted many practitioners and researchers over the past several decades. In time series forecasting approaches, historical observations of the same variable are analyzed to extract the most appropriate model describing the relationship between current data and the past observed data. Time series modeling approach is used when there is no exhaustive model linking the prediction variable to other explanatory variables with high accuracy or there is little knowledge about the underlying data generating the process [1]. Autoregressive integrated moving average (ARIMA) is one of the prevalent linear models which has been used in forecasting energy, engineering, exchange rate and stock problems Moving average and exponential smoothing are other tools used in linear forecasting.

ARIMA models are composed from the pure autoregressive (AR), the pure moving average (MA) and combination of the AR and MA (ARMA). ARIMA models are able to represent different types of time series, *i.e.*, AR, MA, and ARMA series. Nevertheless, since the complex real-world problems usually have nonlinear structure,

Keywords. Gasoline consumption, artificial neural networks, ARIMA, forecasting, multi layer perceptron.

[1] Faculty of Engineering, Urmia University, Urmia, West Azerbaijan Province, Iran. r.babazadeh@ut.ac.ir

the pre-assumed linear correlation structure among the time series values is their major drawback in real-world applications [2]

Using ANN has proved its efficiency as an estimation tool for predicting factors through other input parameters which have no any specified relationship. Some examples of this work are provided in references [3,4]. Also the capabilities of ANN methods help us to gain more reliable results (see Refs. [5,6]). In this study we have introduced seven parameters including six lagged variables and one forecasted value of gasoline consumption specified by ARIMA model. The output is daily gasoline consumption. We have applied these input parameters in the framework of ANN and data have been tested and trained by Multi Layer Perceptron (MLP). Comparison the acquired results of this study with regression prediction models and ARIMA model shows a considerable improvement in error amount and accuracy of prediction. As an instance case study, we collected daily data for 7 years (2005–2011) in Iran. In the present work we provide a prediction with a believable amount of error which is obtained with regard to more available input data. The aim of the proposed hybrid ARIMA-ANN model is to reduce the risk of using an inappropriate model by combining several models to decrease the risk of failure and obtain results that are more accurate. The main difference of the proposed approach respect to those existing in the literature (see Ref. [7]) is that this paper uses autocorrelation function (ACF) and the partial autocorrelation function (PACF) for suitable recognition of inputs of the ANN model.

2. The ARIMA and ANN forecasting models

Totally, an ARIMA model is specified *via* three components including order of autoregressive process (p), order of moving average process (q), and order of differencing (d). In an ARIMA (p, d, q) model, the future value of a variable is assumed to be a linear function of several past observations and random errors. That is, the underlying process that generates the time series with the mean μ has the form [8]:

$$\phi(B)\nabla^d(y_t - \mu) = \theta(B)a_t \tag{2.1}$$

where, y_t and a_t are the actual values and random error at time period t, respectively.

$$\phi(B) = 1 - \sum_{i=1}^{p} \varphi_i B^i \tag{2.2}$$

$$\theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j \tag{2.3}$$

$\varphi(B)$ and $\theta(B)$ are polynomials in B of degree p and q, $\phi_i(i = 1, 2, \ldots, p)$ and $\theta_j(j = 1, 2, \ldots, q)$ are model parameters, $\nabla = (1 - B)$, B is the backward shift operator, p and q are integers and often referred to as orders of the model, and d is an integer and often referred to as order of differencing. Random errors, ε_t, are assumed to be independently and identically distributed with a mean of zero and a constant variance of σ^2.

The Box-Jenkins [9] methodology encompasses three iterative steps including model identification, parameter estimation, and diagnostic checking. Box and Jenkins [9] used ACF and PACF of the sample data as the basic tools to identify the order of the ARIMA model. We also use these functions to identify the preliminary components of the considered time series and then the most appropriate components are specified using suitable diagnostic statistical test. Recently other approaches based on intelligent paradigms, such as neural networks [10], genetic algorithms [11] or fuzzy systems [12] have been presented to improve the accuracy of order selection of ARIMA models.

In the identification step, data transformation is often required to make the time series stationary. Stationarity is a necessary condition in building an ARIMA model used for forecasting. In other words, the model estimation step is performed only after stationarity of a time series is confirmed. A stationary time series is characterized by statistical characteristics such as the mean and the autocorrelation structure being constant over time. When the observed time series show trend and heteroscedasticity, differencing and power transformation are applied

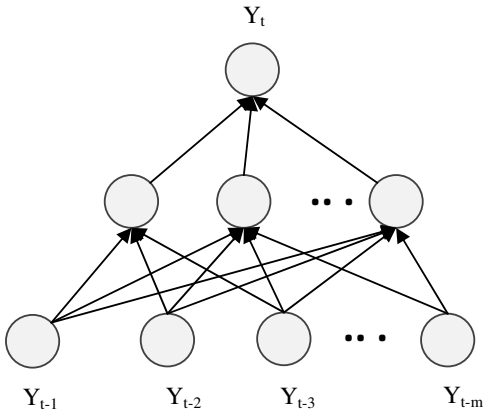

FIGURE 1. A three-layer MLP network.

to the data to remove the trend and to stabilize the variance before an ARIMA model can be fitted. Once a tentative model is identified, estimation of the model's parameters is straightforward. The parameters are estimated such that an overall measure of errors is minimized. This can be accomplished using a nonlinear optimization procedure. The last step in model building is the diagnostic checking of model adequacy. This is basically to check if the model assumptions about the errors, ε_t, are satisfied. Several diagnostic statistics and plots of the residuals can be used to examine the goodness of fit of the tentatively entertained model to the historical data. If the model is not adequate, a new tentative model should be identified which is followed by the steps of parameter estimation and model verification. Diagnostic information may suggest alternative model(s). This three-step model building process is typically repeated several times until a satisfactory model is finally selected. The final selected model can then be used for prediction purposes.

ANNs are composed of attributes that lead to perfect solutions in applications where we need to learn a linear or nonlinear mapping. Some of these attributes are: learning ability, generalization, parallel processing and error endurance. These attributes would cause the ANNs solve complex problem methods precisely and flexibly [13]. ANNs consists of an inter-connection of a number of neurons. There are many varieties of connections under study, however here we will discuss only one type of network which is called the MLP. In this network the data flows forward to the output continuously without any feedback. The MLP uses a supervised learning technique called backpropagation for training the network [14] Figure 1 shows a typical three-layer feed forward model used for forecasting purposes. The input nodes are the previous lagged observations while the output provides the forecast for the future value. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes. The model can be written as:

$$y_t = \alpha_0 + \sum_{j=1}^{n} \alpha_j f\left(\sum_{i=1}^{m} \beta_{ij} y_{t-1} + \beta_{0j}\right) + \varepsilon_t \tag{2.4}$$

where m is the number of input nodes, n is the number of hidden nodes, f is a sigmoid transfer function such as the logistic: $f(x) = \frac{1}{1+\exp(-x)}$. $\{\alpha_j, j = 0, 1, \ldots, n\}$ is a vector of weights from the hidden to output nodes and $\{\beta_{ij}, i = 1, 2, \ldots, m;\ j = 0, 1, \ldots, n\}$ are weights from the input to hidden nodes. α_0 and β_{0j} are weights of arcs leading from the bias terms which have values always equal to 1. Note that equation (2.4) indicates a linear transfer function employed in the output node as desired for forecasting problems. The MLP's most popular learning rule is the error back propagation algorithm. Back Propagation learning is a kind of supervised learning introduced by Werbos [15] and later developed by Rumelhart and McClelland [16]. At the beginning of the learning stage all weights in the network are initialized to small random values. The algorithm uses

a learning set, which consists of input–desired output pattern pairs. Each input–output pair is obtained by the offline processing of historical data. These pairs are used to adjust the weights in the network to minimize the sum squared error (SSE) which measures the difference between the real and the desired values over all output neurons and all learning patterns.

After computing SSE, the back propagation step computes the corrections to be applied to the weights. Most of the suggested models use MLP networks references [17, 18]. The attraction of MLP has been explained by the ability of the network to learn complex relationships between input and output patterns, which would be difficult to model with conventional algorithmic methods.

2.1. The proposed hybrid ARIMA-ANN method

The hybrid models for forecasting time series often decompose a time series into its linear and nonlinear form (Zhang 2003). Khashei and Bijari [1] and Wang *et al.* [19] investigated the effectiveness of hybrid ARIMA-ANN methods in forecasting respect to existing forecasting methods. In hybrid models, a time series can be considered to be composed of a linear autocorrelation structure and a nonlinear component. In the present work, we consider the gasoline consumption time series as function of a linear and a nonlinear component. The amount of consumption (y_t) is a function of linear component (L_t) and nonlinear component (N_t).

$$y_t = f(L_t, N_t). \tag{2.5}$$

In order to estimate L_t and N_t, we use the forecasted values predicted by ARIMA process. Indeed, first the valid components are chosen from the component list specified by ACF and PACF. Second, the estimated model is constructed *via* the valid components and used for forecasting the considered stationary time series. This phase constructs the linear component of the proposed hybrid ARIMA-ANN approach. Then, the forecasted values and specified valid components by ARIMA process are utilized to be used as input parameters of the ANN model. This phase constructs nonlinear component of the proposed hybrid ARIMA-ANN approach.

In this paper, data is collected for a robust period and is further divided to train and test groups. Train data is used to train the MLP models. Test data is used to be compared with actual data (Validation). Moreover, the best fitted MLP is identified by the lowest MAPE. In addition, the selected MLP is compared with different regression's models. Figure 2 presents the overall description of the model. Figure 3 presents the ANN pictorial of the proposed model.

Where Z_{t-i} is the lagged observations representing valid component specified by ARIMA process, and L_t represents the forecasted values of gasoline consumption provided by ARIMA process. The process of extracting

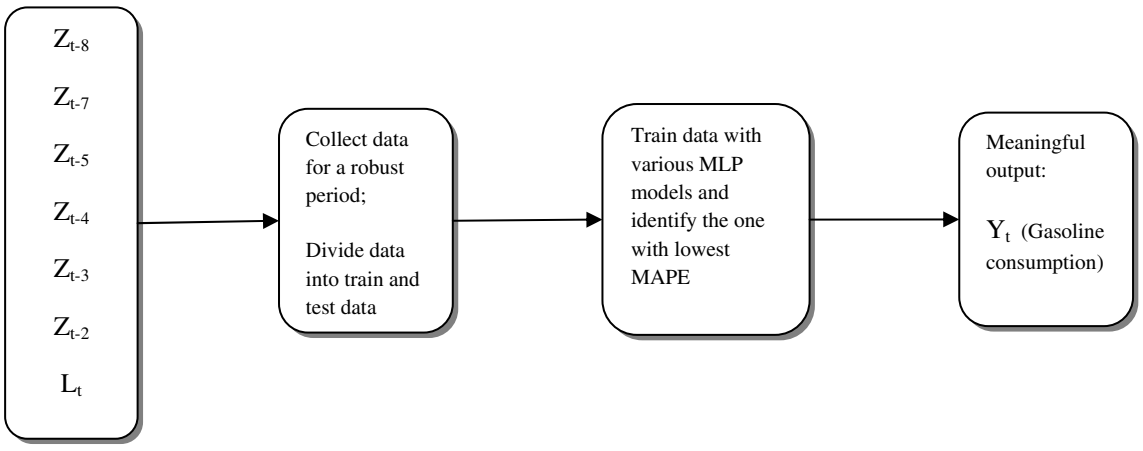

FIGURE 2. Description of the ANN model.

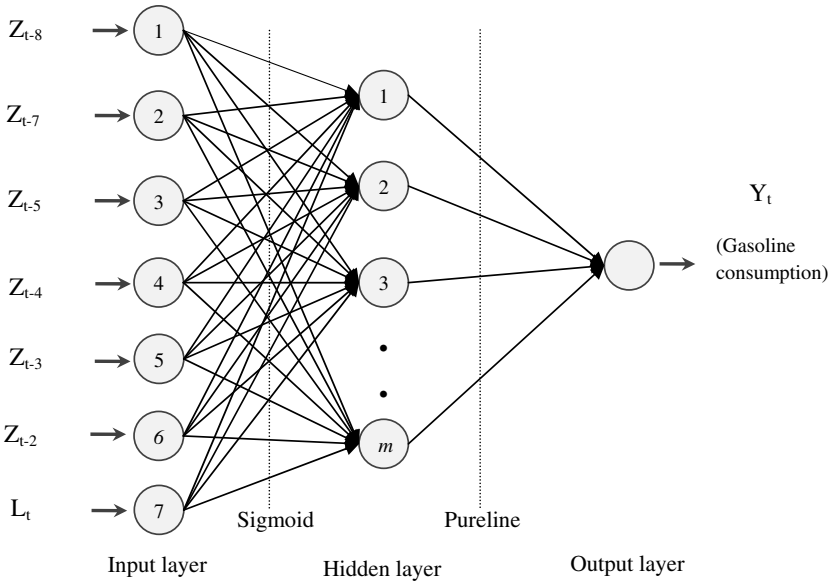

FIGURE 3. The integrated ANN-MLP model.

the most suitable lagged observations would be described in Section 3 Among different error estimation methods we use the MAPE method to assess the performance of employed forecasting models. The MAPE calculation is as follows:

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{x_t - x'}{x_t} \right|}{n}. \tag{2.6}$$

3. Experiment: The case study

The proposed model was applied in Iran. Data for these parameters are provided from Institute for International Energy Studies (IIES) in Iran and Energy Information Administration (EIA) website http://www.eia.doe.gov/emeu/international/contents. Before going ahead, we shortly explain the most important motivations to forecast gasoline consumption in Iran.

According to "Iranian targeted subsidy plan", The Iranian government presented an energy price reform in 2008. The major aim of the policy was to slow down the increasing trend of energy consumption in Iran by removing the energy subsidies. According to the plan, all energy prices were to increase by 20 percent annually. In 2006, daily gasoline consumption stood at 74 million liters and the country paid $5 billion for gasoline imports. The overall consumption of gasoline after the reform decreased from about 65 million liters per day to about 54 million liters per day. Therefore, accurate forecasting of gasoline consumption leads to creating insights in planning for imports, price reform, and etc.

The required data were collected daily for seven years from 2005 to 2011. We divided the dataset into two groups: the training subset and the test subset. For the training subset related data of 1851 periods (days) were considered and used for learning the model and for the test subset relevant data of 20 days were used to test the capability of the model.

Table 1. Different MLP specifications and MAPE results.

Model number	1	2	3	4	5	6	7	8	9	10
Number of neurons in first hidden layer	6	6	4	7	7	7	8	8	9	9
Number of neurons in second hidden layer	2	4	2	1	2	3	1	2	1	2
Learning method	BP	BP	BP	BP	BP	BP	BP	BP	BP	BP
Relative error (MAPE %)	6.74	5.73	6.33	6.32	6.07	6.30	5.76	6.6	6.07	6.22

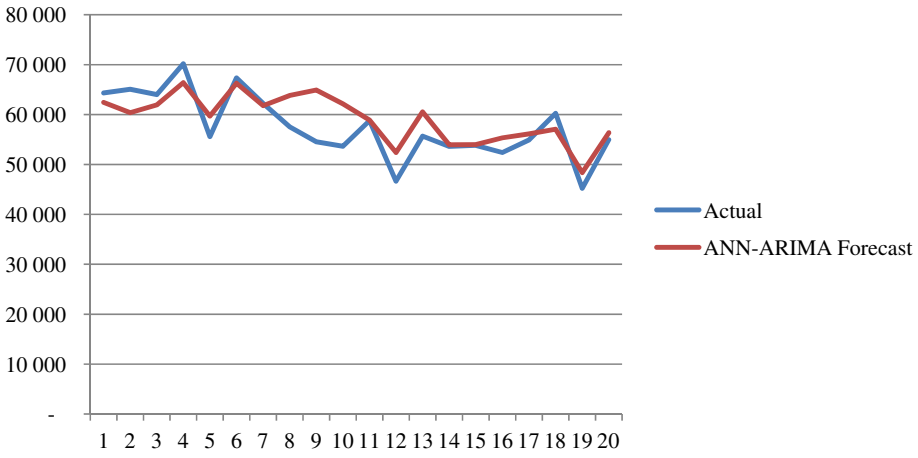

Figure 4. Actual *vs.* hybrid ARIMA-ANN Forecast.

3.1. The best structure of the proposed hybrid ARIMA-ANN model

Several MLP networks are generated and tested. The transfer function for the first layer and all hidden layers are sigmoid and for last one is linear. Back propagation algorithm is used to adjust the learning procedure.

The results of the ten best models and their errors are shown in Table 1. The acquired errors in the last row of Table 1 are derived for test data. According to Table 1 the best ANN structure including six neurons in the first hidden layer and four neurons in the second hidden layer is selected. This ANN structure is trained with back propagation (BP) learning algorithm. Figure 4 illustrates the actual values against forecasted values by hybrid ARIMA-ANN approach.

3.2. Implementing ARIMA process

To implement ARIMA process, we first specify the AR and MA components using autocorrelation function (ACF) and partial autocorrelation function (PACF). Variation of the considered time series has been shown in Figure 5. Stationary test on the considered time series, which is performed by diagnostic statistical test, confirms that the time series is stationary.

The ACF and PACF have been shown in Figures 6 and 7 for 100 lagged observations.

From the ACF and PACF illustrations, the following factors are extracted to be considered as lagged observations and errors.

"ar(1)ar(2)ar(3)ar(4)ar(5)ar(7)ar(8)ar(9)ar(12)ar(14)ma(1)ma(2)ma(3)ma(5)ma(7)ma(8)ma(14)ma(15)".

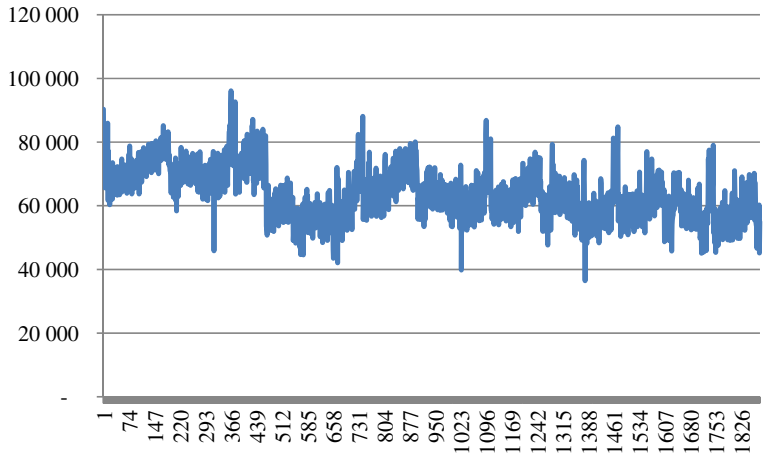

FIGURE 5. Variation of gasoline consumption.

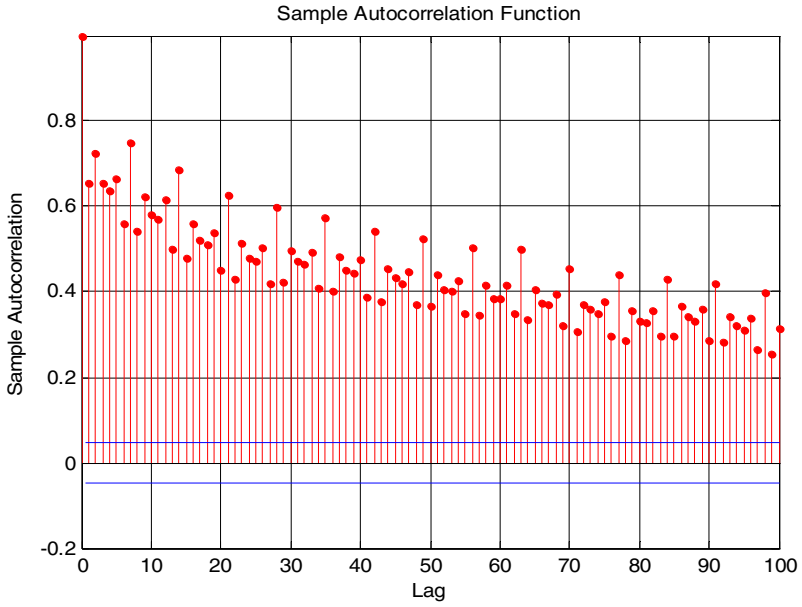

FIGURE 6. Autocorrelation function.

In the estimated model for ARIMA model, from the above-mentioned factors, those factors would be valid that their p_value be less than 0.05 in 95% confidence level. After try and error process, the following valid components by using ARCH (1, 1) method are found:

"ar(1)ar(2)ar(7)ar(8)ar(9)ar(14)ma(1)ma(3)ma(7)".

The acquired model is nonlinear due to using ARCH method. These specified factors are recognized as lagged observations which are used as input parameters of the ANN model to create hybrid ARIMA-ANN structure.

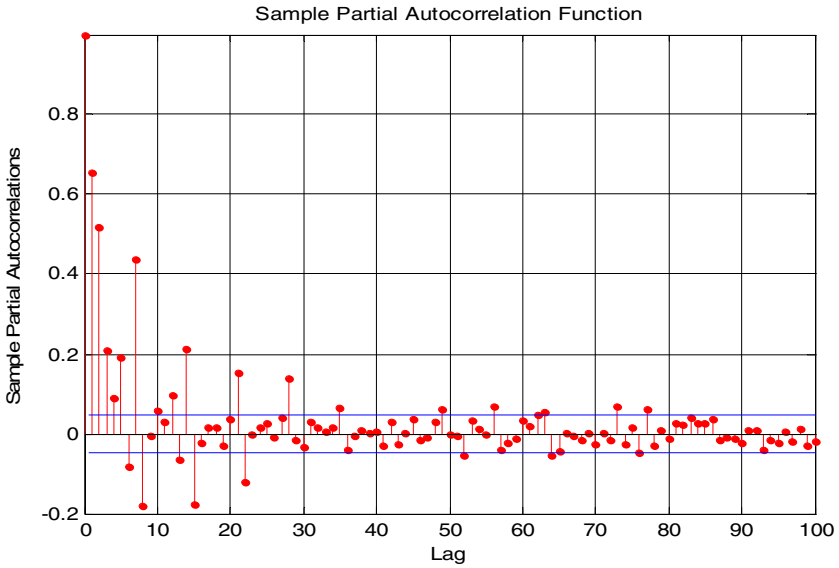

FIGURE 7. Partial autocorrelation.

3.3. Verification and validation

In this section, the proposed hybrid ARIMA-ANN model is compared respect to ARIMA model and well-known regression's models. The specified lag variables by ARIMA model are used as X-values in regression's models. Indeed, the inputs of ANN are considered as X-values in regression's models. The used regression's models are described as follows:

$$\text{Model (I)} \qquad y = \alpha_0 + \sum_{i=1}^{7} \alpha_i X_i \tag{3.1}$$

$$\text{Model (II)} \qquad \ln y = \alpha_0 + \sum_{i=1}^{7} \alpha_i (\ln X_i) \tag{3.2}$$

$$\text{Model (III)} \qquad y = \alpha_0 + \sum_{i=1}^{7} \alpha_i X_i + \sum_{i=1}^{7} \beta_i X_i^2 \tag{3.3}$$

$$\text{Model (IV)} \qquad \ln y = \alpha_0 + \sum_{i=1}^{7} \alpha_i (\ln X_i) + \sum_{i=1}^{7} \beta_i (\ln X_i^2) \tag{3.4}$$

$$\text{Model (V)} \qquad y = \alpha_0 + \sum_{i=1}^{7} \alpha_i X_i + \sum_{i=1}^{7} \beta_i X_i^2 + \sum_{i}^{7} \sum_{i \neq j}^{7} \gamma_{ij} X_i X_j \tag{3.5}$$

$$\text{Model (VI)} \qquad \ln y = \alpha_0 + \sum_{i=1}^{7} \alpha_i (\ln X_i) + \sum_{i=1}^{7} \beta_i (\ln X_i^2) + \sum_{i}^{7} \sum_{i \neq j}^{7} \gamma_{ij} (\ln X_i X_j). \tag{3.6}$$

Eviews package and MATLAB software are used for implementing ARIMA process and regression's models, respectively. Table 2 shows the best results acquired by applied methods.

As shown in Table 2, the ARIMA model is unable to forecast the considered time series for gasoline consumption. This observation could be resulted due to nonlinear behavior of the considered time series. Although the regression's models dominate ARIMA model, their capability to forecasting the considered time series is less

TABLE 2. The results of applied methods.

Regression's models	MAPE (%)
Hybrid ARIMA-ANN	5.73*
ARIMA	14.0
Regression Model (I)	7.31
Regression Model (II)	7.14
Regression Model (III)	7.13
Regression Model (IV)	7.19
Regression Model (V)	6.63
Regression Model (VI)	7.19

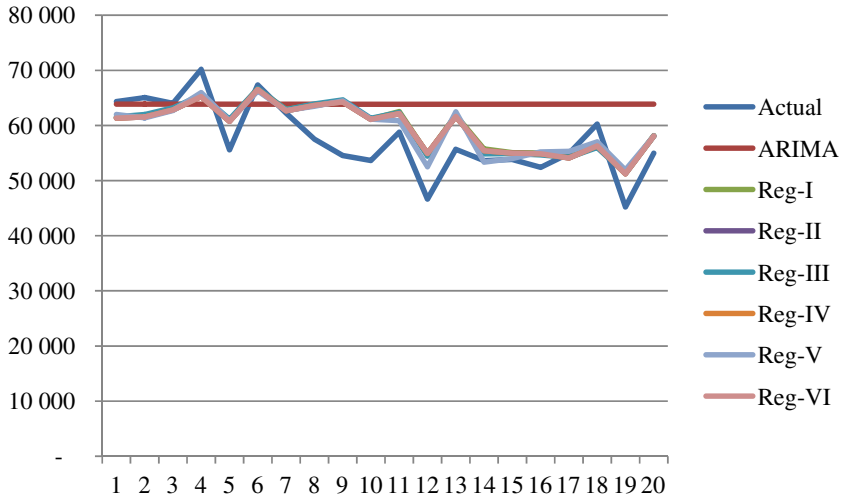

FIGURE 8. Forecasted values of different approaches.

than proposed hybrid ARIMA-ANN model. Also the regression's model (V) is superior to other developed regression's models. Figure 8 shows the forecasted values of ARIMA process and different regression's models against actual values.

4. Concluding remarks

In this paper we have proposed a hybrid ARIMA-ANN model for prediction of gasoline consumption. For this model a number of effective input data are extracted from the ARIMA process and applied in the structure of the ANN model. To show the applicability and superiority of the proposed framework actual data for robust period is used. MLP network is used and applied with the seven input variables. After tuning, the optimum number of neurons is determined in layers. The acquired results of the proposed hybrid ARIMA-ANN model are compared with different regression's models and ARIMA process. Although eliminating subside from gasoline price has led to appearing noisy data in gasoline consumption in Iran, the acquired results show high accuracy of about 94.27%. Also the results show that the ARIMA process is unable to model the nonlinear behaviour of time series even by using nonlinear ARCH method. Moreover, the proposed approach dominates well-known regression's models which are commonly used in forecasting areas. It can be concluded from the acquired results that the proposed hybrid ARIMA-ANN model can be effectively used in gasoline consumption forecasting. Although the proposed approach is applied on a real case of gasoline consumption, it can be successfully used for other types of time series that have the similar structure.

Acknowledgements. The author appreciates for the Iran National Science Foundation (INSF) for the financial support of this study. Also, the author thanks the anonymous referees for their valuable comments.

References

[1] M. Khashei and M. Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Appl. Soft Comput.* **11** (2011) 2664–2675.

[2] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50** (2003) 159–175.

[3] A. Azadeh, S.M. Asadzadeh and A. Ghanbari, An adaptive network-based fuzzy inference system for short-term natural gas demand estimation: Uncertain and complex environments. *Energy Policy* **38** (2010) 1529–1536.

[4] M. Tarafdar Hagh and N. Ghadimi, Radial basis neural network based islanding detection in distributed generation. *Int. J. Eng. Trans. A: Basic* **27** (2014) 1061–1070.

[5] A. Azadeh, R. Babazadeh and S.M. Asadzadeh, Optimum estimation and forecasting of renewable energy consumption by artificial neural networks. *Renew. Sustain. Energy Rev.* **27** (2013) 605–612.

[6] A.Bagheri, N. Narimanzadeh, A.S. Siavash and A.R.Khoobkar, Gmdh type neural networks and their application to the identification of the inverse kinematics equations of robotic manipulators (Research note). *Int. J. Eng.* **18** (2005) 135–143.

[7] L.A. Diaz-Robles, J.C. Ortega, J.S. Fu, G.D. Reed, J.C. Chow, J.G. Watson and J.A. Moncada-Herrera, A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmos. Environ.* **42** (2008) 8331–8340.

[8] J.D. Hamilton, Time series analysis. Princeton University Press, New Jersey (1994).

[9] P. Box and G.M. Jenkins, Time Series Analysis: Forecasting and Control. Holden-day Inc., San Francisco, CA (1976).

[10] H.B. Hwang, Insights into neural-network forecasting time series corresponding to ARMA (p; q) structures. *Omega* **29** (2001) 273–289.

[11] C.S. Ong, J.J. Huang and G.H. Tzeng, Model identification of ARIMA family using genetic algorithms. *Appl. Math. Comput.* **164** (2005) 885–912.

[12] M. Haseyama and H. Kitajima, An ARMA order selection method with fuzzy reasoning. *Signal Process.* **81** (2001) 1331–1335.

[13] M. Turhan, Neural networks and computation of neural network weights and biases by the generalized delta rule and back-propagation of errors. Rock solid images (1995).

[14] N. Austina, P.S. Kumarb and N. Kanthavelkumaranc, Artificial neural network involved in the action of optimum mixed refrigerant (domestic refrigerator). *Int. J. Eng.* **26** (2013) 1025–2495.

[15] P.I. Werbos, *Beyond regression: new tools for prediction and analysis in the behavior sciences*. Ph.D. thesis, Harvard University, Cambridge, MA (1974).

[16] D.E. Rumelhart and J.L. McClelland, Parallel distributed processing: explorations in the microstructure of cognition, Vol. 1. Cambridge, MA: Foundations, MIT Press (1986).

[17] J.H. Park, Y.M. Park and K.Y. Lee, Composite modeling for adaptive short term load forecasting. *IEEE Trans. Power Syst.* **6** (1991) 450–457.

[18] M. Khashei and M. Bijari, An artificial neural network (p, d, q) model for time series forecasting. *Expert Syst. Appl.* **37** (2010) 479–489.

[19] L. Wang, H. Zou, J. Su, L. Li and S. Chaudhry, An ARIMA-ANN Hybrid Model for Time Series Forecasting. *Systems Res. Behavior. Sci.* **30** (2013) 244–259.