

Forecasting the Number of Muslim Pilgrims Using NARX Neural Networks with a Comparison Study with other Modern Methods

ABSTRACT

Pilgrimage (Hajj) of Muslims is considered the largest human gathering all over the world in which more than three millions move together through a very limited space in a short time period. The yearly number of pilgrims coming from outside Saudi Arabia, denoted by NPO for short, is more than two third of the total number of Pilgrims. Therefore forecasting the NPO is considered by Saudi Arabia as the most important indicator in determining the planning mechanism for future secure and comfortable hajj seasons. The main objective of this article is to employ the NARX neural networks to forecast the yearly series of NPO and to show that it gives better forecasts than Box–Jenkins and Bayesian Procedures. In order to achieve our objective, the NARX is used to forecast the future five observations and the results are compared with the results given by Shaarawy et. al(2014).

Keywords: *Forecast, NARX, Bayesian Analysis, Box and Jenkins Methodology, Mean Average Percentage Error(MAPE).*

1. Introduction

Pilgrimage (Hajj) of Muslims is considered the largest human gathering in which more than three millions Pilgrims move together through a very limited space in a short time period. This important event is repeated annually at the same time period and location, and the number of Pilgrims is increasing year after year. Moreover, Hajj is one of the main sources of gross national product in Saudi Arabia and very essential source of livelihood

for many citizens and foreigners living in the Kingdom. In Hajj season, The Saudi government should provide many services to pilgrims such as security, food, housing, electricity, transportation and tents.

Of special interest for Saudi government is forecasting the yearly series of pilgrims coming from outside Saudi Arabia, denoted by NPO for short, since it is one of the most important indicators in determining the planning mechanism regarding these services for future Hajj seasons.

From statistical viewpoint, it is almost impossible to use cross section approach to forecast such data because it is very difficult to confine all independent variables that might affect the NPO and it is almost impossible to succumb some variables to quantification such as religious consciousness. In addition, it is almost impossible to obtain the data on some independent variables such as the average annual income in Islamic countries and countries that have Muslim communities. For these reasons and others, one should use time series analysis to the forecast such data. However, one may trace three important different procedures to forecast time series. The first and well-known one is the methodology of Box and Jenkins (1970). Their analysis has grown in popularity and is today the prevailing methodology of time series analysis. They assume that a parsimonious stationary and invertible ARMA process could present the time series at hand (or a transformation of the series) such that one can perform the four phases of time series analysis (Identification (order determination), estimation, diagnostic checking, and forecasting). Their methodology have been widely used and explained by many others such as Newbold (1981), Priestly (1981), Harvey (1993), Wei (2005), Box et al (2008) and liu (2009). However, their identification technique is highly nonobjective and requires very good experience and skills.

The second approach of time series analysis is the Bayesian one. This approach is being developed and most of the Bayesian contributions have been occurred within the last three decades. For the theory and analysis of Bayesian procedure, the reader is referred to Newbold (1973), Chow (1975), Monahan (1983), Shaarawy and Broemeling (1984), Broemeling (1985), Broemeling and Shaarawy (1988), Shaarawy and Ali (2003) and Shaarawy et al (2007).

Although ARIMA models are quite flexible and can be used for a large number of time series, the main limitation is given by the assumption of a linear form for the model. Therefore, an ARIMA model can not capture nonlinear behavior that is quite common in non-linear time series. The approach of nonlinear modeling the time series is suitable for most real-world problems. The real world is nonlinear and complex rather than linear because there are so many possible nonlinear structures and relationships. Most nonlinear models developed during the last two decades are parametric.

In order to use the parametric model, it must be specified before forecasting. Therefore, they cannot be used if the data characteristics do not match the model assumptions. Parametric approach is very suitable for nonlinear problems with a complex structure but there is a lack of theories that suggest a specific form for the structure.

Artificial neural networks are algorithms and techniques that can be used for statistical modeling and is an alternative to linear ARIMA models. In addition to the nonlinear modeling capability, neural networks have several other features that make them valuable for time series forecasting. First, neural networks are data-driven nonparametric methods that do not require many restrictive assumptions on the underlying stochastic process that might generate the data. As such, they are less susceptible to the model misspecification problem than parametric methods. This “learning from data” feature is highly desirable in various forecasting situations where time series data are usually easy to collect but the underlying data-generating mechanism is not known. Second, neural networks have been shown to have universal functional approximating capability in that they can accurately approximate many types of complex functional relationships. This is an important and powerful characteristic as a time series model aims to accurately capture the functional relationship between the variable to be forecast and its historical observations. The combination of the above-mentioned characteristics makes neural networks a quite general and flexible tool for forecasting (Zhang, P. (2012)).

It is shown that the recurrent neural networks (RNN) with a sufficiently large number of neurons is a realization of the nonlinear ARMA (NARMA) process (Haykin, S.(1999), Gao, Y.(2005), and Mandic, D.P.(2001)). Many research papers are published in scientific literature and some commercial companies claim or market the so-called advanced statistical programs, using neural networks, for modeling and prediction.

However, some difficulties and limitations remain nevertheless of current actuality and cause researches for new NN models and learning techniques to be conducted(Lin T.(1996), Gao Y.(2005), Lin T.(1997), and Siegelmann H.(1997)).

The classical techniques for time series prediction require stationarity before performing the four phases of time series analysis, while most real time series are non-stationary. After NNs have been introduced one can forecaste the original time series without the need to transform it to a stationary one.

The problem of long time dependencies is related to the problem of vanishing gradient or forgetting behavior. The main idea of RNN is providing a weighted feedback connection between layers of neurons and adding time significance to entire artificial NN. For all the previous reasons RNNs are considered the most suitable approach for time series analysis. The main objective of this article is to use the NARX approach to forecast the NPO and show that it gives better forecasts than Bayesian and Box-Jenkins approaches.

The remainder of this paper is structured as follows: Section 2 introduces and explains the proposed NARX network structure. Section 3 is devoted to forecast the NPO the proposed NARX network. Section 4 is dedicated to evaluate the forecast performance of the proposed NARX network and compare the achieved numerical results with Bayesian and the traditional Box and Jenkins approaches. Finally, the paper is concluded in Section 5.

2. NARX Networks

The “Nonlinear Autoregressive models with eXogenous input (NARX model)”, which are therefore called NARX recurrent neural networks (Haykin S.(1999), Lin T.(1996), and Gao Y.(2005)). The NARX network uses the past values of the actual time series to be predicted and past values of other inputs to make predictions about the future value of the target series. This is a powerful class of models which has been demonstrated that they are well suited for modeling nonlinear systems and specially time series. One principal application of NARX dynamic neural networks is in control systems. Also, is a class computationally equivalent to Turing Machines (Haykin S.(1999)). Some important

qualities about NARX networks with gradient-descending learning gradient algorithm have been reported: (1) learning is more effective in NARX networks than in other neural network (the gradient descent is better in NARX) and (2) these networks converge much faster and generalize better than other networks(Lin T.(1996), and Gao Y.(2005))).

The simulated results show that NARX networks are often much better at discovering long time dependences than conventional recurrent neural networks. An explanation why output delays can help long-term dependences can be found by considering how gradients are calculated using the back-propagation-through-time (BPTT) algorithm. Recently, several empirical studies have shown that when using gradient-descent learning algorithms, it might be difficult to learn simple temporal behavior with long time dependencies(Lin T.(1997) and Yao J.(2000)), in other words those problems for which the output of a system at time instant k depends on network inputs presented at times $r \ll k$. The researchers have analyzed learning algorithms for systems with long time dependencies and showed that for gradient-based training algorithms, the information about the gradient contribution m steps in the past vanishes for large m . This effect is referred to as the problem of vanishing gradients, which partially explains why gradient descent algorithms are not very suitable to estimate systems and signals with long time dependencies. For instance, common recurrent neural networks encounter problems when learning information with long time dependencies, a problem in the prediction of nonlinear and non stationary signals. The vanishing gradients problem makes the learning of long-term dependencies in gradient-based training algorithms difficult if not virtually impossible in certain cases (Haykin S.(1999)). A state space representation of recurrent NARX neural networks can be expressed as[7]:

$$z_k(k+1) = \begin{cases} \Phi(u(k), z_i(k)), & i = 1, \\ z_i(k), & i = 2, 3, \dots, N. \end{cases}$$

where the output $y(k) = z_i(k)$ and $z_i, i = 1, 2, \dots, N$, are state variables of recurrent neural network. The recurrent network exhibits forgetting behavior, if:

$$\lim_{m \rightarrow \infty} \frac{\partial z_i(k)}{\partial z_j(k-m)} = 0, \quad \forall k, m \in \mathbf{K}, i \in \mathbf{O}, j \in \mathbf{I}.$$

where z are state variables, $\mathbf{I}$ denotes the set of input neurons, $\mathbf{O}$ denotes the set of output neurons and $\mathbf{K}$ denotes the time index set. Several approaches have been suggested

to get around the problem of vanishing gradient in training RNNs. Most of them rest on including embedding memory in neural networks, whereas several others propose improved learning algorithms, such as the extended Kalman filter algorithm, Newton type algorithm, annealing algorithm, etc.

2.1 NARX Archticture

The NARX model can be implemented by using feed forward neural network with embedded memory (first tapped delay line) plus a delayed connection from the output of the second layer to input (second taped delay line) see figure (2.1).

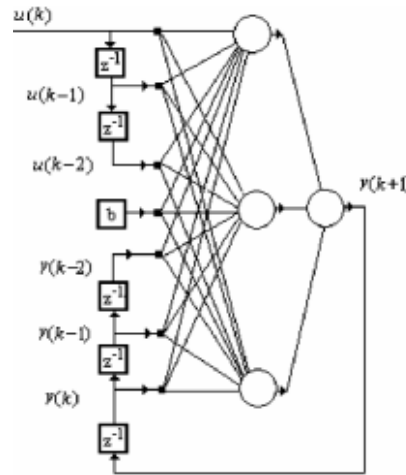


Figure2.1: NARX model with tapped delay line at input.

For the model shown in figure (2.1) , let $RNN(d_u, d_y; N)$ denote the neural net with $d_u \equiv$ input delays, $d_y \equiv$ output delays and $N \equiv$ the number of neurons in the input layer. The general prediction equations for computing the next value of time series (output) $y(k+1)$ using the model in figure (2.1) with:
past observations: $u(k), u(k-1), \dots, u(k-d_u)$ and past outputs: $y(k), y(k-1), \dots, y(k-d_y)$,
as inputs may be written as (DIACONESCU E.(2008)),

$$y(k+1) = \Phi_0 \left[w_{b0} + \sum_{h=1}^N w_{ho} \cdot \Phi_h \left(w_{ho} + \sum_{i=0}^{d_u} w_{ih} u(k-i) + \sum_{j=0}^{d_y} w_{jh} y(k-j) \right) \right]$$

Moreover, let

$u_1(k), u_1(k-1), \dots, u_1(k-d_u) \equiv$ the past observations of the first time series (I),
 $u_2(k), u_2(k-1), \dots, u_2(k-d_u) \equiv$ the past observations of the second time series (II) and
 $y(k), y(k-1), \dots, y(k-d_y) \equiv$ the past outputs (III).

The model in figure (2.2) takes (I), (II), and (III) as inputs and may be written as

$$y(k+1) = \Phi_0 \left[w_{b0} + \sum_{h=1}^N w_{ho} \cdot \Phi_h \left(w_{ho} + \sum_{i1=0}^{d_{u1}} w_{i1h} u_1(k-i1) + \sum_{i2=0}^{d_{u2}} w_{i2h} u_2(k-i2) + \sum_{j=0}^{d_y} w_{jh} y(k-j) \right) \right]$$

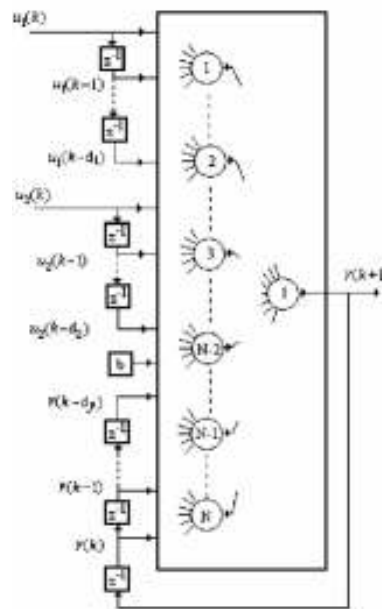


Figure2.2: NARX with two tapped delay lines for two time series applied at NARX input.

3. NARX Analysis of NPO data

The time series of number of pilgrims coming from outside the kingdom of Saudi Arabia (NPO) (as shown in table (3.1)) consists of 44 observations (from year 1390AH* to year 1433 AH).

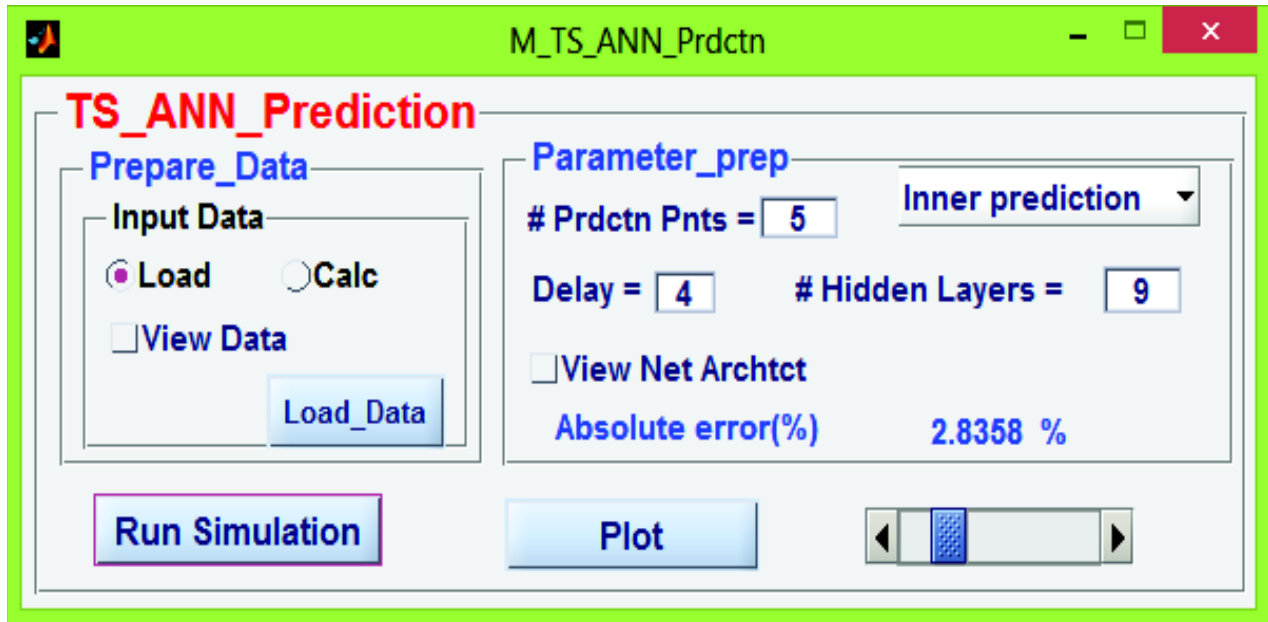
Table(3.1)

Number of Pilgrims coming from outside the Kingdom (NPO)

Year	Data	Year	Data	Year	Data	Year	Data
1390	431270	1401	879368	1412	1012917	1423	1431012
1391	479339	1402	853555	1413	992813	1424	1419706
1392	645182	1403	1003911	1414	995611	1425	1534759
1393	607755	1404	919671	1415	1043274	1426	1557447
1394	918777	1405	851761	1416	1080465	1427	1654407
1395	894573	1406	856718	1417	1168591	1428	1707814
1396	719040	1407	960386	1418	1132344	1429	1729841
1397	739319	1408	762755	1419	1056730	1430	1613965
1398	830236	1409	774560	1420	1267555	1431	1799601
1399	862520	1410	828993	1421	1363992	1432	1828195
1400	812892	1411	720102	1422	1354184	1433	1752932

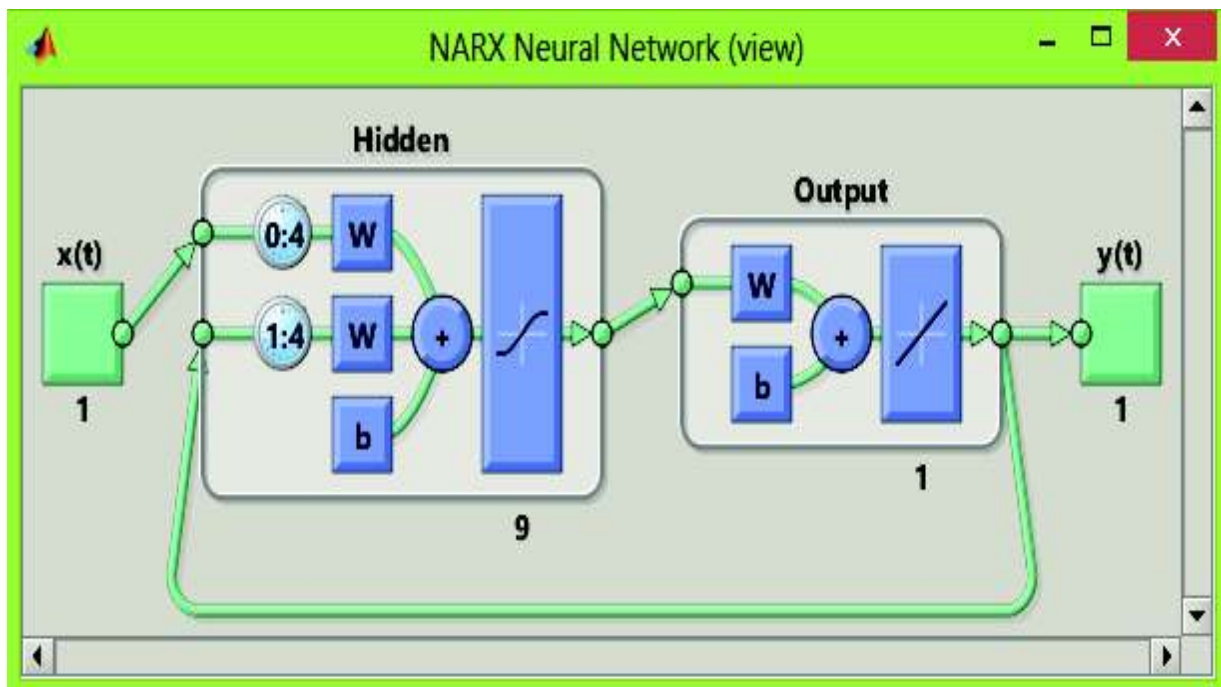
In this experiment, NARX-net of *Matlab Neural Network Toolbox* was deployed to develop GUI of the prediction program for the future number of pilgrims (see table (3.1)), shown in figure (3.1) below.

Figure (3.1): program demo used for NARX prediction.



The net architecture shown in figure (3.2).

Figure (3.2): NARX-net with nine hidden layers and four input delay.



Before forecasting the future observations, the NARX-model has been used to check its ability to forecast the future observations. In order to evaluate the forecast performance of our proposed approach, a small portion of the NPO data at the end of the data are reserved solely for forecasting evaluation. In statistical literature, these data are referred to as *hold-out sample*, or *post-sample*, and in principle are not used in model or forecasting when evaluating forecast performance. The last 5 observations (about 12% of the whole data) are reserved as the *hold-out sample (post-sample)*. The first 39 observations were used to forecast the next five observations using NARX approach; then the five forecasts were compared with the five real observations and the percentages of the absolute errors were calculated. The results are reported in tables (3.2).

Table (3.2)

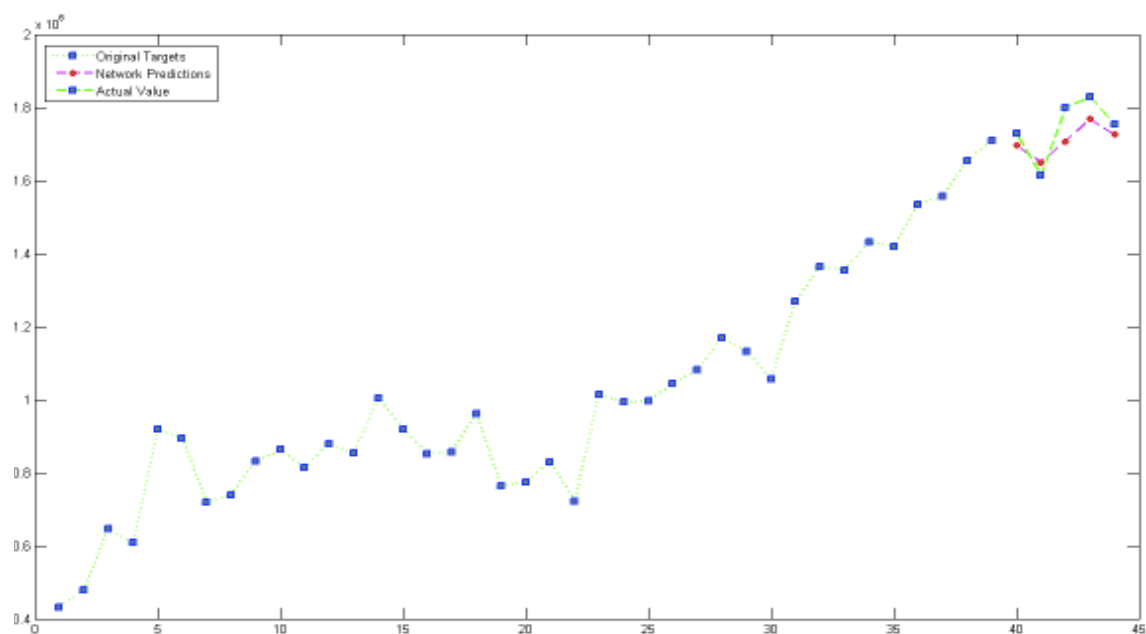
The last five observations and their forecasts

Year	Real observations	Forecasts	Percentage of the absolute error%
1429	1707814	1695484	1.986
1430	1729841	1648932	2.167
1431	1613965	1706119	5.195
1432	1799601	1767791	3.304
1433	1828195	1725265	1.578
Mean			2.846

Inspecting the results given table (3.2), one may conclude that NARX- model gives very good forecasts since the mean of the percentages of the absolute errors is 2.846%. Figure (3.3) shows the NPO forecasts of the last five observations, and figure (3.4) shows the NPO forecasts of the five future observations. It might be important to mention that the percentages of the absolute errors are calculated using the formula

$$P.A.E = \left| \frac{\text{real observation} - \text{forecast}}{\text{real observation}} \right| \cdot 100$$

Figure (3.3): The True NPO and forecasts of the last five observations



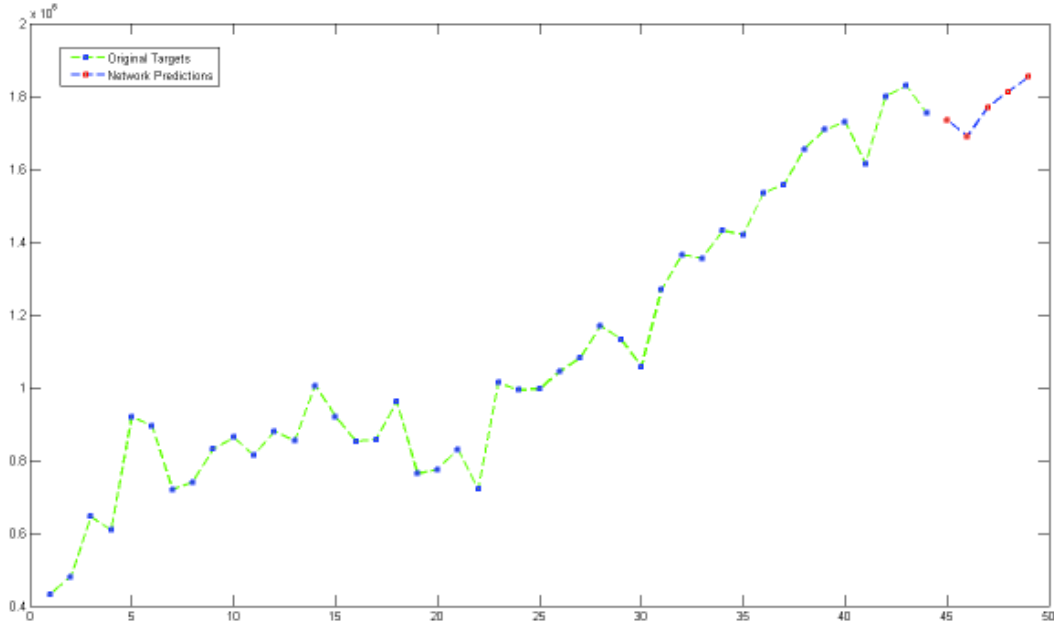
Finally, the NARX- model has been used to forecast the next five future observations. The point forecasts for these observations are given by table (3.3).

Table (3.3)

The future five forecasts

Year	1434	1435	1436	1437	1438
Point forecast	1852816	1811862	1769373	1689825	1734789

Figure (3.4):
The True NPO and future five observations



4. A Comparative Study

As we have mentioned before, Shaarawy et.al (2014) analyzed the same NPO data using Bayesian and the traditional Box-Jenkins approaches. In the previous section, we used their proposed evaluation technique to evaluate our proposed NARX approach in forecasting the *hold-out sample (the last five observations)*. The main objective of this section is to compare the numerical results achieved by the proposed NARX approach, given in the previous section, with the numerical results achieved by Shaarawy et.al (2014). In order to achieve the main objective, the numerical results achieved by Shaarawy et.al (2014) are reported in tables (4.1) and (4.2).

Table (4.1)**The forecast performance of Bayesian approach**

Year	Real observations	Forecasts	absolute percentage error(APE) %	absolute deviation (AD)	squared error (SE)
1429	1707814	1756113	1.52	48299	2332793401
1430	1729841	1805778	11.88	75937	5766427969
1431	1613965	1856848	3.18	242883	58992151689
1432	1799601	1909363	4.44	109762	12047696644
1433	1828195	1963362	12.00	135167	18270117889
Mean			5.7000	122409.6	19481837518.4

Table (4.2)**The forecast performance of Box-Jenkins approach**

Year	Real observations	Forecasts	absolute percentage error(APE)%	absolute deviation (AD)	squared error (SE)
1429	1707814	1748546	1.08	40732	1659095824
1430	1729841	1852292	14.77	122451	14994247401
1431	1613965	1904772	5.84.	290807	84568711249
1432	1799601	1985503	8.61	185902	34559553604
1433	1828195	2066444	17.89	238249	56762586001
Mean			9.600	175628.2	38508838815.8

However, there are several criteria to evaluate forecast performance of a model, including mean absolute percentage error (MAPE), mean absolute deviation (MAD), and root mean squared error (RMSE) as defined below

$$MAPE = \frac{1}{m} \sum_{t=1}^m \left| \frac{real\ value - forecast}{real\ value} \right| \cdot 100$$

$$MAD = \frac{1}{m} \sum_{t=1}^n |true\ value - forecast|$$

$$RMSE = \left[\frac{1}{m} \sum_{t=1}^n (true\ value - forecast)^2 \right]^{1/2}$$

Where, m is the total number of observations in the *hold-out sample (post-sample)*.

These criteria are also computed using our NARX and the results are given by table(4.3)

Table (4.3)
The forecast performance of NARX

Year	Real observations	Forecasts	Percentage of the absolute error%	absolute deviation (AD)	squared error (SE)
1429	1707814	1695484	1.986	12330	152028900
1430	1729841	1648932	2.167	80909	6546266281
1431	1613965	1706119	5.195	92154	8492359716
1432	1799601	1767791	3.304	31810	1011876100
1433	1828195	1725265	1.578	102930	10594584900
Mean			2.846	64026.6	3830798419.4

Inspecting the results given by tables (4.1), (4.2) and (4.3) one may conclude the flowing:

1. The mean absolute percentage error (MAPE) achieved by Bayesian approach is higher than the corresponding value achieved by our proposed NARX approach with more than 100%. Our MAPE is almost one-half of Bayesian MAPE.

2. The mean absolute percentage error (MAPE) achieved by Box and Jenkins approach is higher than the corresponding value achieved by our proposed NARX approach with more

than 237.3%. Our MAPE is less one-third of their MAPE.

3. The mean absolute percentage error (MAPE) achieved by Box and Jenkins approach is higher than the corresponding value achieved by Bayesian approach with more than 68%.

4. The mean absolute deviation (MAD) achieved by Bayesian approach is higher than the corresponding value achieved by our proposed NARX approach with more than 91%.

5. The mean absolute deviation (MAD) achieved by Box and Jenkins approach is higher than the corresponding value achieved by our proposed NARX approach with more than 174%. Our MAD is almost one-third of their MAD.

6. The root mean squared error (RMSE) achieved by Box and Jenkins approach is higher than the corresponding value achieved by our proposed NARX approach with more than 217%. Our RMSE is less than one-third of their RMSE.

7. The root mean squared error (RMSE) achieved by Bayesian approach is higher than the corresponding value achieved by our proposed NARX approach with more than 125%. Our RMSE is less than one-half of the Bayesian RMSE.

From the foregoing numerical results, we conclude that our proposed NARX approach is much more accurate than Bayesian and Box-Jenkins approaches in forecasting the NPO.

6. SUMMERY AND COCLUSION

The authors have proposed to use the NARX approach to forecast the series of number of Pilgrims coming from outside the Kingdom of Saudi Arabia from year 1390 AH¹ to year 1433 AH. Point forecasts for the next five future years are provided by the authors using the proposed NARX approach. The forecasts of the last five observations achieved by our proposed NARX approach are compared with the results of Bayesian and Box-Jenkins approaches given Shaarawy et.al (2014). It has been shown that the proposed NARX approach gives much better forecasts than those achieved by Bayesian and the traditional

¹ The years are written using Islamic (Lunar) Calendar (AH). For instance, the year 1433 corresponds to year 2012 A.D. For more details see http://en.wikipedia.org/wiki/Islamic_Calendar

References

1. Box, G. and Jenkins, G.(1970). Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.
2. Broemling, L. and Shaarawy, S. (1988). Time Series Analysis: A Bayesian Analysis in the Time Domain. Bayesian Analysis of Time Series and Dynamic Model, edited by Spall, J.
3. Chatfield, C. (1980). The Analysis of Time Series: Theory and Practice. Chapman and Hill Ltd, London.
4. Chatfield, C. (2004). The Analysis of Time Series: An Introduction. 6th ed. Boca Raton, Fla. : Chapman & Hall/CRC, - Texts in statistical science.
5. Chow, G.(1975).Multi-period Prediction from Stochastic Difference Equations by Bayesian Methods. Chapter8 of Studies in Bayesian Econometrics and Statistics, edited by S.E. Fienberg and A. Zellner. North-Holland, Amsterdam.
6. DIACONESCU E.(2008), The use of NARX Neural Networks to predict Chaotic Time Series, 9th WSEAS International Conference on AUTOMATION and INFORMATION (ICAI'08), Bucharest, Romania.
7. Gao Y., Er M. J.(2005), NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches, Fuzzy Sets and Systems, Vol. 150, No. 2, pp.331-350.
8. Harvey, A. (1993). Time Series Models, 2nd edition. The MIT Press.
9. Haykin S.(1999), Neural Networks, Second Edition, Pearson Education.
10. Lin T., Bill G. H., Tino P., Giles C(1996), Learning long-term dependencies in NARX recurrent neural networks, IEEE Transactions on Neural Networks, Vol. 7, No. 6, pp. 1329-1351 .
11. Lin T., Giles C. L., Horne B. G., Kung S.Y.(1997), A Delay Damage Model Selection Algorithm for NARX Neural Networks, IEEE Transactions on Signal Processing, "Special Issue on Neural Networks", Vol. 45, No. 11, pp. 2719-2730.
12. Lin T., Horne B.G., Tino P., Giles C.L.(1997), A delay damage model selection algorithm for NARX neural networks, IEEE Trans. Signal Process. 45(11)2719-2730.

13. Liu, L. (2009), Time Series Analysis and Forecasting. 2nd edition. Scientific Computing Association Corp, USA.
14. Mandic D.P., and Chambers J.A.(2001), Recurrent Neural Networks for Prediction, JohnWiley&Sons.
15. Maria J. P. and Menezes Jr.(2008), Long-term time series prediction with the NARX net- work: An empirical evaluation Advances in Neural Information Processing (ICONIP 2006), Volume 71, Issues 16-18, Pages 3335-3343.
16. Monahan, J. (1983). Fully Bayesian Analysis of ARIMA Time Series Models, Journal of Econometrics, Vol. 21, pp. 307-331.
17. Newbold, P.(1973). Bayesian Estimation of Box and Jenkins Transfer Function Model for Noise Models. Journal of the Royal Statistical Society, Series B, vol. 35. No. 2, pp. 323-336.
18. Priestley, M. (1981). Spectral Analysis of Time Series, Academic Press, London.
19. Shaarawy, S. and Ali, S. (2003). Bayesian Identification of Seasonal Autoregressive Models. Communications in Statistics-Theory and Methods, Vol. 32, Issue 5, pp.1067-1084.
20. Shaarawy, S. and Ali, S. (2008). Bayesian Identification of Multivariate Autoregressive Processes. Communications in Statistics-Theory and Methods, Vol. 37, Issue 5, pp.791-802.
21. Shaarawy, S. and Ali, S. (2008). Bayesian Identification of Multivariate Autoregressive Processes. Communications in Statistics-Theory and Methods, Vol. 37, Issue 5, pp.791-802.
22. Siegelmann H. T., Horne B. G. and Giles C. L.(1997), Computational capabilities of recurrent NARX neural networks, IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol. 27, No.2, 208-215.
23. Tong, H. (1990). Non-Linear Time Series: A Dynamical System Approach. New York: Oxford University Press.
24. Wei, W.W.S. (2005). Time Series Analysis: Univariate and Multivariate Methods. Addison Wesley, Reading, MA.
25. Yao J., Tan C. L.(2000), A case study on using neural networks to perform technical forecasting of forex, Neurocomputing, 34, pp. 79-9.
26. Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons. Inc, New York.

27. Zellner, A. and Reynolds, R. (1978). Bayesian Analysis of ARMA Models. Presented at the Sixteenth Seminar on Bayesian Inference in Econometrics, June 2-3.
28. Zhang G. P.(2012), “Neural networks for time-series forecasting,” in Handbook of Natural Computing, G. Rozenberg, J. N. Kok, and T. Back, Eds., pp. 461477, Springer, Berlin, Germany.