ISSN 1022 - 2502, E-ISSN 2408- 8528
Dhaka Univ. J. Sci. 72(1): 71-76, 2024 (January) DOI: https://doi.org/10.3329/dujs.v72i1.71192

Comparing ARIMA, Neural Network and Hybrid Models for Forecasting Fish Production in Bangladesh

Maisha Binte Saif and Murshida Khanam

Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh

(Received: 3 October 2023 ; Accepted: 13 December 2023)

Abstract

Time series forecasting is a commonly applied method for scientific predictions. There are several econometric methods for forecasting time series observations and predicting the systematic pattern of underlying data. ARIMA model is most renowned in this aspect for its linearity. Nowadays a machine learning model namely artificial neural network (ANN) is gaining popularity for its nonlinear characteristics. Inconsistent conclusions may frequently be drawn when evaluating whether ARIMA models or neural networks are better at forecasting future events. For this reason, a hybrid methodology has been established in this study to get advantage from both linear and nonlinear modeling. The annual dataset of fish production in Bangladesh from 1990 to 2020 has been evaluated in this case. Formulating three measurement errors, RMSE, MAE and MAPE it has been demonstrated that the hybrid approach has high level of forecasting accuracy than the other two models in forecasting fish production data.

Keywords: ARIMA, ANN, Hidden Layer, Learning Rate, Hybrid, RMSE, MAE, MAPE.

I. Introduction

Forecasting is an essential task of any time series dataset, it is also essential for using an appropriate model so that the forecasting accuracy can be made maximum. There are various techniques for forecasting and predicting time series and economic variables[1].

Two of the them are Autoregressive Integrated Moving Average (ARIMA) model and Artificial Neural Network (ANN) model. ARIMA model is appropriate for forecasting univariate dataset whereas ANN is a very popular nonlinear approach.Another newly developed model which is merged using both ARIMA and ANN models has also gathered much attention in some research fields.

A method to fit time series data into a member of the type of ARIMA models was first described by Box and Jenkins in 1970. The technique of time series analysis and forecasting was greatly impacted by this[2]. Model selection, parameter estimations, model checking, and forecasting are iteratively cycled through in the Box-Jenkins method for creating ARIMA forecasting models[3].

ANNs are computer based programs created to replicate how the human brain processes information. They are digitalized models of the human brain. Unlike people, ANNs do not learn through programming but rather through experience with suitable learning exemplars. The patterns and relationships in data are what neural networks use to build their expertise[4]. There are several kinds of neural network models that can be used in different circumstances like image identification, speech recognition and so on[5].

Some studies had been conducted for comparing the forecasting performance of ARIMA and ANN models.A study had been done for comparing the performance of these two models to forecast jute production in Bangladesh. In this case, a supplementary data set of yearly jute production in Bangladesh from 1972 to 2013 collected from FAO (Food and Agricultural Organization) website had been analyzed. The results of this study showed that ARIMA model performed better than ANN model[6]. A similar study had been performed for comparing the forecasting accuracy of ARIMA and ANN models by analyzing annual rice production data in Bangladesh. Here, the yearly data of rice production from 1972 to 2013 that was collected from FAO had been used. The results of this study suggested that, ARIMA model showed better forecasting performance than ANN model in the case of rice production[7].Moreover, there are a few study related to Hybrid ARIMA and Neural Network Model. A study had been done by utilizing this model to enhance the forecasting behavior of ARIMA and ANN models to obtain high precision. Eight input variables were used by the model to forecast a time series data called PM-10. The outcome of this study exhibited that hybrid model performed better than ARIMA and ANN models[8].

In practical situation, it is difficult to decide whether one method is superior to the other for prediction purposes. As a result, forecasters struggle to decide which strategy would work best in their particular situation[9]. Hybrid methodology, which merges ARIMA and ANN models, has been employed in this study. This methodology which is the combination of both ARIMA and ANN models is less widespread. This method helps to overcome the circumstances of using linear and nonlinear modeling in a particular way. It is hard to find a perfect linear or nonlinear time series in actual situation and so it is hard to use an ARIMA or ANN model separately. Therefore, more accurate models can be created for data with complex autocorrelation structures. It is also possible to get better forecasting results using this method[10].

*Author for correspondence. e-mail: murshidakhanam@yahoo.com

The fishing industry is one of the most productive and innovative businesses in Bangladesh, with enormous potential for future growth in the agricultural economy.According to the Bangladesh Statistical Year Book report, which was released by Bangladesh Bureau of Statistics (BBS) in 2017, fish is the main source of protein in human diet. The fishing industry provides approximately 60% of animal protein. Fisheries have long been a vital part of the economy of Bangladesh[11]. Increasing fish production is crucial to satisfy the requirements of people. So forecasting the future production of fisheries of this country could be useful.

In this paper, forecasting performance of fish production dataset of Bangladesh has been illustrated using an appropriate model among ARIMA, ANN and Hybrid models. This study will greatly enhance the present literature as there hardly exists any study comparing ARIMA, ANN and Hybrid models using the dataset of fish production in Bangladesh. In section II the models that have been used in this study are shortly discussed. Section III includes the data of this study. In section IV the overall results of the analysis along with some tables and graphs have been showed. Section V comes with the findings of this study. Finally, in section VI some recommendation and conclusion of this study have been indicated.

II. Models and Theories

In this paper, three models have been illustrated and among these three the best fitted model for forecasting fish production in Bangladesh has been chosen.

For any time series dataset three basic characteristics have to be determined. They are stationarity, trend and seasonality[12]. A time series that is stationary has properties that are constant regardless of the time point at which it is seen. Therefore, time series with trends are non-stationary. The behavior of time series will fluctuate depending on trend and seasonality. Stationarity of a time series data can be tested by incorporating two procedures: Graphical Analysis and Augmented Dickey-Fuller (ADF) test.

The building of an appropriate ARIMA model corresponds to Box-Jenkins methodology (1970).These models are used to represent "non-seasonal" time series for showing patterns that is not just random noise. In $ARIMA(p,d,q)$ model, autoregressive or AR process has order p, moving average or MA process has order q and the time series is made stationary using d amount of differencing. Now, if a time series becomes stationary after first difference then it is known as integrated of order 1 or $I(1)$. If a time series is $I(2)$, it becomes stationary after second difference. As a whole, a time series being stationary after d^{th} difference is called integrated of order d. It is crucial to understand which process a time series corresponds to. Whether it follows an *AR* process, *MA* process, *ARMA* process, or *ARIMA* process. It is also important to be aware of the values for *p*, *q*, and *d*. This is where the B-J technique works. Here, a specific model from the broad ARIMA class is initially selected for investigation based on the examination of statistics derived from the data. The decision is made after analyzing the sample autocorrelations and partial autocorrelations of the original data. After choosing a particular model from the general class, the parameters of the model are subsequently calculated using effective statistical methods. The fitted model must next be evaluated to see if it accurately captures the behavior of the data[2]. This can be done by observing the AIC, BIC, RMSE, MAE, and MAPE values. After a suitable model has been fitted to the data, it can be forecasted forward to get predictions of future values for an underlying dataset.

ANN models are used in supervised learning to approximate discriminative functions for classification. It consists of neurons that store and process values inside the network.There are a number of inputs, weights, and bias values included within each node. An input is multiplied by a weight value when it enters a node after that output is seen or sent to next layer of neural network. An activation function which is employed in the neural network by linked information processing units to convert input into output. Among the activation functions, logistic activation function is a popular one.This function accepts any real value as an input and output values are between 0 and 1.It is frequently applied to models where predicting the probability as an output is crucial. There are numerous learning rules, but the most common is the back-propagation rule. Back-propagation of the error during the training or learning phase is used to optimize the weights[4].In order to reduce the mean squared error (MSE) this machine-learning technique, makes a backward pass to modify the parameters of the model[13]. It is crucial to assess how well the neural network model operates after training and it is easy to assess a neural network using a cost function. In order to estimate parameters for process systems, neural networks are used. The aim of estimating the parameter is to get a response from the process model that closely matches the real process data[14]. There are various kinds of neural networks and one of the most popular one is Feed-Forward Neural Network. This neural network is most basic type where input data only flows in one way, going via artificial neural nodes and out through output nodes.

Purely linear or nonlinear time series are infrequent in real-world data. Linear and nonlinear forms can be found frequently. If so, neither ARIMA model nor ANN model would be sufficient for modeling and predicting time series as it is difficult to handle for ARIMA model the non-linear relationships and neural network itself cannot manage linearity and nonlinearity in similar effective manner. To simulate complicated autocorrelation structures in the data properly, ARIMA and ANN models are combined to a hybrid model[15]. The suggested hybrid technique comprises of two parts. The linear part is assessed using an ARIMA model in the first stage. The residuals from the ARIMA model are modeled using a neural network model in the second stage. In this procedure, the forecasted values of

ARIMA and ANN are combined to enhance the entire modeling[9].In order to classify various patterns, the hybrid method makes use of both the distinctive quality and power of ANN model as well as ARIMA model. Hybrid methodologies are capable of combining linearity and non-linearity patterns which is a suitable technique for real-world application. This model can also provide better forecasting performance. The hybrid framework can be more resistant to any changes in the way the data is organized[16.]

A comparison among these three models can be done using three measurement error formulas[17].

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{E((f_t - Y_t)^2)} \quad (1)$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|f_t - Y_t| = \frac{1}{n}\sum_{t=1}^{n}|e_t| \quad (2)$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{\frac{1}{n}\sum_{t=1}^{n}|f_t - Y_t|}{Y_t} \times 100 \quad (3)$$

The best model can be chosen after evaluating the measurement errors.

III. Data and Variables

In this paper, data on Fish production in Bangladesh have been obtainedthrough the Bangladesh Statistical Year Book published by the Bangladesh Bureau of Statistics (BBS) in 2021. Annual time series data on fish production in Bangladesh (Metric tons) have been used here from 1990 to 2020. The major variable of interest under this study is the fish production dataset which is a univariate time series dataset.

IV. Results

The production value for 1992-1993 is missing from the dataset. This could be due to unavoidable circumstances. To fill up this gap imputation method has been done. Again, in the actual dataset there exists other subsections like fish production in river, lake, ponds and so on. Total production values have been used in the final data after trimming and normalizing these sections. Further testing and analysis havebeen carried out after imputation, trimming and normalization.

Stationarity Test

At first the overall dataset has been divided into two parts: Training and Testing datasets. Training data is from 1990 to 2015 and testing data is from 2016 to 2020. For checking the stationarity pattern of the fish production dataset both graphical analysis and ADF test have been done. These two techniques have the same conclusion that, after taking second difference fish production data can be made stationary.

ARIMA Model

The overall procedure for selecting ARIMA model has been done using the Box-Jenkins methodology. By observing the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plots and for different values of *p*, *d* and *q* five ARIMA models have been chosen. Here, *p* is the order of autoregressive (AR) process, *q* is the order of moving average (MA) process and the underlying time series data has been made stationary using *d* amount of differencing[18] (here, *d=2*). The five selected models are, $MA(1), AR(1), ARMA(1,1), AR(2)$, and $ARMA(2,1)$. Among these five models $MA(1)$ or $ARIMA(0,2,1)$ model has been chosen by observing the lowest AIC and BIC values.

Table 1. AIC and BIC Values of Five Selected ARIMA Models

Model	AIC	BIC
arima(0,2,1)	***587.44***	***589.7921***
arima(1,2,0)	589.45	591.8061
arima(1,2,1)	587.84	591.3765
arima(2,2,0)	588.49	592.0205
arima(2,2,1)	589.79	594.5005

It has been clearly shown in Table 1 that $ARIMA(0,2,1)$ model has the lowest AIC and BIC values. So this model has been selected for forecasting fish production in Bangladesh and the parameters are then estimated.

Table 2. Estimated Values of *ARIMA(0,2,1)*

	ARIMA(0,2,1)
MA(1)	-0.7148
σ^2 (Estimated)	2.052e+09
Log-likelihood	-291.72
AIC	587.44

Table 2 shows the estimated values of the parameters of the selected $ARIMA(0,2,1)$ model.

Now the predicted values from the year 2016 to 2020 as well as the actual values has been shown in Table 3.

Table 3. Forecasting Performance of ARIMA(0,2,1) Model for Fish Production Data

Year	Actual Value	Predicted Value	Error	MAPE
2016	3878324	3832539	45785	
2017	4134434	3980832	153602	
2018	4276641	4129126	147515	2.50
2019	4384221	4277420	106801	
2020	4503371	4425714	77657	

From Table 3 it is observed that the MAPE score is less than 10% which indicates a good MAPE score[19]. As a result, the $ARIMA(0,2,1)$ model has produced pretty accurate forecasts.

The actual values of 2016 to 2020 are very close to the predicted values of 2016 to 2020. Hence, the selected ARIMA model provides good forecasting.

ANN Model

To form the appropriate ANN model training dataset has been used. In constructing the neural network model back-propagation algorithm has been utilized. As for activation function logistic function has been employed to make the calculation simple. Learning rate and threshold value both have been set as 0.01 along with two hidden layers and three nodes or neurons. Selection of these values have been done using trial and error method. After forming the neural network model, testing dataset has been applied to evaluate this model[7]. Here, testing dataset is considered as 2016 to 2020 in the fish production data.

Using the appropriate ANN model forecasted values from the year 2016 to 2020 along with the actual values has been shown in Table 4.

Table 4. Forecasting Performance of ANN Model for Fish Production Data

Year	Actual Value	Predicted Value	Error	MAPE
2016	3878324	2086131	1792193	
2017	4134434	2086131	2048303	
2018	4276641	2086131	2190510	50.61
2019	4384221	2086131	2298090	
2020	4503371	2086131	2417240	

Here, the predicted values are all same over the five years. Also, the MAPE score is greater than 50 which is an extremely high value[16].

Thus, the predicting capability of ANN model is poor for forecasting fish production Bangladesh.

Hybrid Model

For the dataset of fish production in Bangladesh, it is important to merge the linear autocorrelation structure with the nonlinear component to fit the hybrid model. In other words, building up a hybrid model using ARIMA and ANN models. Hybrid model can be estimated applying the package "ARIMAANN" in RStudio.

The forecasted values from the year 2016 to 2020 accompanying the actual values has been shown in Table 5.

Table 5. Forecasting Performance of Hybrid Model for Fish Production Data

Year	Actual Value	Predicted Value	Error	MAPE
2016	3878324	3837922	40402	
2017	4134434	3986184	148250	
2018	4276641	4134478	142163	1.07
2019	4384221	4282772	101449	
2020	4503371	4431065	72306	

The value of MAPE is less than 10% so it represents a very good forecasting approach[16].

As the predicted values are very close with the actual values with a very small MAPE score, hybrid model has been considered a better choice to predict the dataset of fish production in Bangladesh.

Comparison

Three models have been compared to observe which will provide a better forecasting performance for the data of fish production. Three measurement errors are used for checking the accuracy of the models.

Table 6. Comparison among ARIMA, ANN and Hybrid Models

Model	MAE	MAPE	RMSE
ARIMA	106272.0	2.501099	113918.9
ANN	2149267	50.61338	2160097
Hybrid	***100910.5***	***1.066996***	***108938.5***

By observing Table 6, it can be clearly stated that the three measurement errors (MAE, MAPE and RMSE) are lowest for the hybrid model. Thus by using the hybrid approach a good forecasting result is possible to achieve.

Forecasting

As hybrid model has the lowest measurement errors this model has been further used to forecast the future production of fish in Bangladesh.The available data on fish production covers time period from 1990 to 2020. In this paper, hybrid model has been established to forecast the fish production data for the next five years.

Table 7. Forecasting Upcoming Five Years Using Hybrid Model

Year	Forecasted Value
2021	4658293
2022	4804671
2023	4951049
2024	5097428
2025	5243806

The predicted values from 2021 to 2025 for the dataset of fish production in Bangladesh has been shown in Table 7.

Therefore, hybrid model shows a superior forecasting performance than the other two models.

V. Discussion

The aim of this paper is to evaluate the forecasting performance of three models (ARIMA, ANN, and Hybrid) that are relevant to a univariate time series dataset of fish production in Bangladesh. A model with a superior pattern

of prediction has been thus picked for future forecasting of production.

Initially it has been determined whether or not the underlying data set is stationary. The underlying data of fish production exhibits non-stationarity pattern for both the level form and the first difference, according to graphical analysis and the ADF test. When there is a second difference, the data becomes stationary for both circumstances.

The ACF and PACF plots have been observed in order to discover the best ARIMA model for modeling and predicting fish production. Five potential ARIMA models have been suggested based on these plots. Among them $ARIMA(0,2,1)$ model has been selected according to the lowest AIC and BIC values (Table 1).

The fish production dataset has been split into training data and test data for finding best neural network model.Three nodes or neurons, two hidden layer, 0.01 for learning rate and threshold value, and the backpropagation technique have all been employed to build the correct ANN model.

Both ARIMA and ANN models have been combined to create hybrid model. The forecasted values of this model are then produced for comparing them to those of ARIMA and ANN, the other two models which has been shown in Table 5.

Among three models the hybrid model performed better on the basis of three measurement errors (MAE, MAPE and RMSE) which has been shown in Table 6.

Lastly, the fish production for several recent and upcoming years has been provided in Table 7. using the best model chosen, which is the hybrid model.

Finally, among ARIMA, ANN and Hybrid models, hybrid method has been chosen. There are two reasons for doing so. Firstly, forecasting performance of ANN model is not good as it showed similar production values in future forecasting. Also, the MAPE value is much greater than 10. This could be the reason of small dataset. Secondly, between ARIMA and Hybrid model the MAPE score of hybrid model is lower. Again, the forecasting performance of hybrid model is much better as the predicted values are closer to the actual value.

Therefore, it can be concluded that hybrid model can be a better fit for forecasting fish production in Bangladesh with an upward production pattern which can fulfill the needs of mass people of Bangladesh.

VI. Conclusion

The results of this paper indicate that the Hybrid ARIMA and Neural Network model outperforms the ARIMA and ANN models. The hybrid strategy has been superior to the other two models in terms of prediction accuracy. So, a hybrid model for predicting fish production in Bangladesh can be constructed.

This study has been done using the fish production data which is a univariate time series dataset. So further research can be conducted using more than one variable or another time series data. This study is a combination of econometric and machine learning models. Hence there is a scope of enhancing the current literature by organizing the present study in diverse aspects.

References

1. Cryer, J. D., & Chan, K. S. 2008. Time series regression models. *Time series analysis: with applications in R*, 249-276.
2. Newbold, P. 1983. ARIMA model building and the time series analysis approach to forecasting. *Journal of forecasting*, **2(1)**, 23-35.
3. Gujarati, D. N. (2022). *Basic econometrics*. Prentice Hall.
4. Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, ***22*(5)**, 717-727.
5. Aversano, L., Bernardi, M. L., Cimitile, M., Iammarino, M., & Verdone, C. 2023. A data-aware explainable deep learning approach for next activity prediction. *Engineering Applications of Artificial Intelligence*, **126**, 106758.
6. Hossain, M. M., Abdulla, F., & Hossain, Z. 2017. Comparison of ARIMA and Neural Network Model to Forecast the Jute Production in Bangladesh. *Jahangirnagar University Journal of Science*, **40 (1)**, *11*, *18*.
7. Sultana, A., & Khanam, M. 2020. Forecasting rice production of Bangladesh using ARIMA and artificial neural network models. *The Dhaka University Journal of Science*, **68(2)**, 143-147.
8. Wongsathan, R., &Seedadan, I. 2016. A hybrid ARIMA and neural networks model for PM-10 pollution estimation: The case of Chiang Mai city moat area. *Procedia Computer Science*, **86**, 273-276.
9. Zhang, G. P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**, 159-175.
10. Luxhøj, J. T., Riis, J. O., &Stensballe, B. 1996. A hybrid econometric—neural network modeling approach for sales forecasting. *International Journal of Production Economics*, **43(2-3)**, 175-192.
11. Shamsuzzaman, M. M., Mozumder, M. M. H., Mitu, S. J., Ahamad, A. F., & Bhyuian, M. S. 2020. The economic contribution of fish and fish trade in Bangladesh. *Aquaculture and Fisheries*, ***5*(4)**, 174-181.
12. Granger, C. W. 1981. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, **16(1)**, 121-130.
13. Graves, A., Wayne, G., & Danihelka, I. 2014. Neural turing machines. Ar Xiv preprint arXiv: 1410.5401.

14. Samad, T., & Mathur, A. 1992. Parameter estimation for process control with neural networks. International Journal of Approximate Reasoning, **7(3-4)**, 149-164.

15. Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, **5(4)**, 559-583.

16. Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., ... & Winkler, R. 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of forecasting, **1(2)**, 111-153.

17. Hyndman, R. J., & Koehler, A. B. 2006. Another look at measures of forecast accuracy. *International journal of forecasting*, **22(4)**, 679-688.

18. Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. 2000. *Time series analysis and its applications* (**3**). New York: springer.
https://stephenallwright.com/good-mape-score/