Modeling CPUE series for the fishery along northeast coast of India: A comparison between the Holt-Winters, ARIMA and NNAR models

K. G. Mini*, Somy Kuriakose and T. V. Sathianandan
ICAR - Central Marine Fisheries Research Institute, Kochi - 682 018, Kerala, India.
*Correspondence e-mail: minikg02@gmail.com

Received: 05 Dec 2015, Accepted: 25 Dec 2015, Published: 30 Dec 2015

Abstract
Mathematical as well as statistical models not only help in understanding the dynamics of fish populations but also enables in short-term predictions on abundance. In the present study, three univariate forecasting techniques viz., Holt-Winters, Autoregressive Integrated Moving Average and Neural Network Autoregression were used to model the CPUE data series along northeast coast of India. Quarterly landings data which spans from January 1985 to December 2014 was used for building the model and forecasting. The accuracy of the forecast was measured using Mean Absolute Error, Root Mean Square Error and Mean Absolute Percent Error. Based on the comparison of the model, performance of Holt-Winter’s model was found to provide more accurate forecasts than the Autoregressive Integrated Moving Average and Neural Network Autoregression model. A Holt-Winters model with smoothing factors $\alpha = 0.172$, $\beta = 0$, $\gamma = 0.529$ was found as the suitable model. The presence of seasonality in the series is evident from gamma value. An ARIMA model with one non-seasonal moving average term combined with two seasonal moving average terms was found to be suitable to model the CPUE series based on the Akaike Information Criteria. Among the Neural Network Autoregression models used to fit the CPUE series, a configuration of 13 lagged inputs and one hidden layer with 7 neurons provided the best fit.

Keywords: Catch per unit effort, Holt-Winter’s model, Autoregressive Integrated Moving Average model, Neural Network Autoregression model, forecasting

Introduction
The northeast coast of India comprising of two maritime states Odisha and West Bengal has 638 km coast line with an average annual marine fish production of 4.68 lakh tonnes during the last five year period. As per the information generated in 2010 marine fisheries census, the states Odisha and West Bengal has marine fish population of 6.06 lakhs and 3.80 lakh respectively who depend on fisheries for their livelihood (CMFRI, 2012 a, b). According to census, there are 2,248 mechanized fishing crafts, 3,922 motorized crafts and 4,656 non-mechanized crafts operating in Odisha where as in West Bengal there are 14,282 mechanized fishing crafts and 3,066 non-motorized crafts. In Odisha, marine fish landings take place in 73 landing centres and in West Bengal there
are only 59 landing centres. In 2014, the contribution from Odisha towards total marine fish landings in the country was 1.39 lakh tonnes and that from West Bengal was only 0.77 lakh tonnes. Together, these two maritime states accounted for nearly 6% of the all India landings in 2014. The maximum landings observed so far in Odisha was 3.23 lakh tonnes and that in West Bengal was 3.65 lakh tonnes both in the year 2011. The fishery in this region is mainly by trawl nets, gill nets and hooks and lines. Important marine fishery resources in the region are croakers, penaeid prawns, ribbon fishes, lesser sardines, catfishes, hilsha shad, Bombay duck and non-penaeid prawns. This region is prone to cyclone and is badly affected during the southwest monsoon. The fishing season in this region normally starts in July and extends up to February, October to December being the most productive period.

Modelling of the abundance indices is a useful tool for understanding the dynamics of fish populations and to make short-term quantitative recommendations for fisheries management (Czerwinski et al., 2007). A number of efficient time series approaches with different levels of complexity have been developed and tested to obtain accurate and reliable forecasts of fish harvest/biomass. The well-known methods and models such as Exponential Smoothing, Autoregressive Integrated Moving Average (ARIMA), Vector Auto Regression (VAR), Neural Network, Wavelet etc. are commonly used. These methods were evaluated based on their efficiency to describe and forecast the dynamics of the series. As there is considerable variation in the season-wise landings in northeast region, the quarter-wise CPUE time series is used for modeling and forecasting the dynamics of fish abundance in the northeast region of India. Three modeling and forecasting methods viz., Holt-Winter’s model, seasonal ARIMA model and Neural Networks Autoregression (NNAR) were used. These methods were evaluated based on their efficiency to describe and forecast the dynamics of the series.

Material and methods

The quarterly CPUE time series data pertaining to the northeast region of India was taken from the National Marine Fisheries Data Centre (NMFDC) of Central Marine Fisheries Research Institute (CMFRI), Kochi for the study. The time series covers a total of 120 data points, spanning from first quarter of 1985 to fourth quarter of 2014. To set the platform for model estimation, testing and model comparison, the CPUE time series was split into two sets, one up to 2010 with 104 data points for the estimation part and the remaining 16 data points spanning the years 2011 to 2014 for testing and comparison. The data was used to investigate the most appropriate model to describe the dynamics in the series, using Holt-Winter’s model, seasonal ARIMA model and neural network forecasting approaches. The time series were initially tested for stationarity using Dickey Fuller test (Dickey and Fuller, 1979). The statistical procedures for the three approaches were implemented by developing necessary codes using the statistical computing environment R (R Core Team 2015).

Holt-Winters models

The Holt-Winters method (Makridakis and Wheelwright, 1978) captures patterns of increasing or decreasing trend with presence of seasonality and it uses simple exponential smoothing in order to forecast. The forecast is obtained as a weighted average of past observed values where the weights decline exponentially so that the values of recent observations contribute more to the forecast than the values of earlier observations. The basic model has three smoothing equations, each designed to capture either the presence of level, trend or seasonality in the series. This method can be used for short, medium and long-term forecasts.

The equations involved in the additive Holt-Winters model are as follows;

1. Level component: \( L_t = \alpha(y_t - s_{t,p}) + (1 - \alpha)(L_{t-1} + b_{t-1}) \)
2. Trend component: \( b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \)
3. Seasonal component: \( s_t = \gamma(y_t - L_t) + (1 - \gamma)s_{t,p} \)
4. Forecasting system: \( \hat{y}_{t+m} = (L_t + b_t m)s_{t+m} \)

where \( y_t \) is the observed series, \( \alpha, \beta, \) and \( \gamma \) are the smoothing parameters \( (0 \leq \alpha, \beta, \gamma \leq 1) \), \( L_t \) is the smoothed level at time \( t \), \( b_t \) is the change in the trend at time \( t \), \( s_t \) is the seasonal smooth at time \( t \), \( m \) is the number of seasons per year, and \( \hat{y}_{t+m} \) presents \( m \)-periods ahead forecast.

ARIMA model

ARIMA models are a class of models that can be used for short term forecasts. The development of these models was based on the methodology described in the classic
work of Box and Jenkins (1976). The procedure for setting up the ARIMA forecasting model starts by identifying the model, estimating the model parameters, performing the diagnostic check and considering alternative models based on diagnostic check. Once the model is found appropriate, it is then used for forecasting.

The mathematical expression for a general class of seasonal ARIMA model, denoted by ARIMA \((p, d, q) (P, D, Q)_s\), is explained below using the back shift operator \(B\) which operates on time series \(y_t\) such that \(B^s y_t = y_{t-s}\).

\[
\phi(B) \phi(B^s) y_t - \mu = \varepsilon_t,
\]

where \(y_t\) represents the time series. Different functions in the above model are polynomials in the back shift operator \(B\) defined as

\[
\phi(B) = 1 - \Phi_1 B - \ldots - \Phi_p B^p, \\
\phi(B^s) = 1 - \Phi_{1s} B^s - \ldots - \Phi_{ps} B^{ps}, \\
\theta(B) = 1 + \Theta_1 B + \ldots + \Theta_q B^q, \\
\theta(B^s) = 1 + \Theta_{1s} B^s + \ldots + \Theta_{qs} B^{qs}.
\]

The other quantities in the model are \(p\) and \(q\) the orders of non-seasonal autoregressive and moving average terms, \(P\) and \(Q\) the orders of seasonal autoregressive and moving average terms, \(d\) the order of non-seasonal difference, \(D\) the order of seasonal differences, \(\mu\) the drift, \(s\) the seasonality and \(\varepsilon_t\)'s are random error components assumed to be distributed independently and identically with zero mean and constant variance \(\sigma^2\).

The appropriate form of ARIMA for a data series is determined by means of an iterative series of steps utilizing the distribution of the autocorrelation (ACF) and partial autocorrelation (PACF) functions to establish among other considerations, model parsimony, the statistical significance of the coefficients, model stationarity, statistically independent residuals and satisfactory error for catch forecasts. An in-depth discussion on ARIMA models is given in Pankratz (1983).

**Neural Network Auto-Regression**

NNAR model allows the modeling of complex nonlinear relationships among input variables and output variables. In the case of a NNAR model, lagged values of the time series are used as input for the model and outputs are predicted values of the time series. Since there is seasonal component in the CPUE series, the model NNAR \((p, P, k)_m\) proposed by Hyndman and Athanasopoulos (2013) is used for the study. The structure of the NNAR \((p, P, k)_m\) is represented in Fig. 1.

![Diagrammatic representation of the NNAR (p, P, k)_m model](image)

The terms \(p\) and \(P\) denote the non-seasonal and seasonal lagged inputs, \(k\) is the number of nodes in the hidden layer and \(m\) is the seasonality. In the absence of a hidden layer, the NNAR \((p, P, k)_m\) is analogous to the Seasonal ARIMA model denoted as an ARIMA \((p, 0, 0) (P, 0, 0)_m\). An NNAR model with a feed forward neural network which involves a linear combination function and an activation function was used to model the series. The linear combination function is formulated as

\[
y_t = a_0 + \sum_{i=1}^{k} w_i x_i,
\]

The parameters \(a_0\)'s and \(w_i\)'s are “learned” from the data and \(x_i\)'s are lagged values of the time series. The values of the weights are often restricted to prevent them becoming too large. The parameter that restricts the weights is known as the “decay parameter” and is often set to be equal to 0.1. In the beginning, the weights take random values and later they are updated using the observed data.

In the hidden layer, this is then modified using a nonlinear function such as a sigmoid, \(f(x) = \frac{1}{1+e^{-x}}\) to give the input for the next layer. This tendency to reduce the effect of extreme input values, thus making the network robust to outliers.

**Comparison of forecast accuracy**

Ljung and Box (1978) \(\chi^2\) test is applied to see whether there is any autocorrelation among the residuals of fitted models. The comparative performance of the three types of models for forecasting CPUE is judged based on three criteria, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).
The MAE, RMSE and MAPE are defined as follows:

\[
\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} |\hat{y}_t - y_t|
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (\hat{y}_t - y_t)^2}
\]

\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\hat{y}_t - y_t}{y_t} \right|
\]

where \( \hat{y}_t \) is the difference between observed and fitted values for the time series for time \( t \).

The ability of each modeling approach to forecast catch was assessed by the relative magnitudes of these comparative statistics. The best model results in minimum values of MAE, RMSE and MAPE.

Results and discussion

The marine fish landings in this region has gone up from 0.7 lakh tonnes in 1985 to 6.8 lakh tonnes in 2011. There is a quantum jump in the landings over the past decade which instigated from 2000 onwards and reached its peak in 2011. This increase can be attributed to the highly efficient fishing craft and gear together with the increase in the number and size of crafts, introduction of multi-day fishing and extension of fishing grounds (Mini et al., 2012, 2013). But later, the catch has started declining and came down to 2.2 lakh tonnes in 2014 as majority of the small trawlers were not venturing into sea for fishing as they were incurring huge losses due to very poor catch of high value fish such as hilsa.

The data used were examined for the presence of seasonality and decomposition type. The time series decomposition plots (Fig. 2) separate the series into its constituent components i.e., the estimated trend component and the estimated seasonal component. The plots show that the estimated trend component follows an increasing trend since 2011. The present findings also indicated the existence of a seasonal pattern in the CPUE series, which can be related to the vulnerability of the area to tropical cyclones and the closure of fishery during the ban period.

Until 1990, landings in the northeast region were contributed by mechanized and non-mechanized sectors only and there after motorized sector also started contributing to the fishery. The fishing intensity by the non-mechanized sector reduced over years from 10.5 lakh fishing trips by non-mechanized crafts in 1985 to 2.6 lakh trips in 1997, nearly 75% reduction in 13 years. On the other hand the operation of mechanized crafts in this region increased from 1.6 lakh fishing trips in 1985 to 4.5 lakh trips in 1996, nearly 1.8 fold increase. There after it reduced and remained steady around 2.0 lakh trips with little fluctuations from 2000 to 2011 and reduced further in the last few years. In the mechanized sector, multi-day voyage fishing started from 2001 onwards in the northeast region and over years it intensified causing increase in catch per fishing trip (Fig. 3). The CPUE from mechanized crafts was 310 kg/unit (per boat trip) in 1985 which gradually increased over years to reach 1149 kg/unit in 2004 and thereafter it shoot up and reached 2822 kg/unit in 2010. Afterwards it reduced and fluctuated between 2000 and 2500 kg/unit. The non-mechanized sector showed a steady increasing trend in CPUE and reached 78 kg/unit in 2013 from 19 kg/unit in 1985. In the case of motorized sector though initially the CPUE was high to the tune of 201 kg/unit in 1985 it immediately came down and fluctuated between 50 and 100 kg/unit until 1999 and increased thereafter to reach the maximum of 175 kg/unit in 2003. Thereafter it is fluctuating between 100 and 150 kg/unit. The overall CPUE from this region combining the three sectors was 58 kg/unit in 1985 which gradually increased and reached 294 kg/unit in 2005 and had a sudden spike there after to reach 682 kg/unit in 2010 and fell down to reach the level of 306 kg/unit in 2014.

The Holt-Winters model is used when the data exhibits both trend and seasonality. Forecasted values are dependent on the level, slope and seasonal components of the series being forecast. An additive model is applied to the series. Model

![Decomposition of additive time series](image-url)

Fig. 2. Time series decomposition plots separating trend, seasonal and random components from the quarterly CPUE series along northeast region of India
parameters $\alpha$, $\beta$, $\gamma$ are initialized using the test data and the coefficients for level, trend and seasonal components of the best fitting model is presented in Table 1. The presence of seasonality in the series is evident from gamma value.

The same data set used for building Holt-Winters model was considered for ARIMA model and NNAR model. Selection of the most appropriate ARIMA model is crucial for the purposes of accurate forecasting. If the selected model fits the current and previous observations well then it is hoped that it will be capable of predicting future observations successfully. The orders $p$, $q$, $P$ and $Q$ were identified and estimated for the series following Box and Jenkins methodology. The model which minimizes the Akaike information criteria ($\text{AIC} = 1261$) is selected as the best model for CPUE series, that is ARIMA$(0, 1, 1)$ $(0, 0, 2)$. Summary results for the fitted seasonal ARIMA model are given in Table 2. The forecasted series using the developed models are presented in Fig. 4.

The third model, a feed-forward neural network with three layers; one input layer, one hidden layer and one output layer were also fitted to the data. The optimum numbers of neurons were chosen by trial and error method in the input layer and hidden layer. For the input and the output layers a...
linear activation function was used and sigmoid function was used for the hidden layer. A total of 20 networks were fitted, each with random starting weights. The network is trained for one-step forecasting. The final model is a feed-forward network NNAR (13,7,1) with 13 lagged inputs and one hidden layer with 7 neurons. Multi-step forecasts were computed recursively. For a good forecasting model, the residuals left over after fitting model should be white noise. The Ljung-Box test is applied to the residuals of the fitted models and results showed that all p-values for Holt-Winters and NNAR model exceeded 0.01 which indicates acceptance of model at 1% significant level (Table 3).

For the purpose of comparing the two methods it was necessary to calculate the error of the forecasts, meaning the forecasted value of each series was compared against the actual observed value in the relevant period. Both forecasted values were then compared against the actual observed values that were excluded from the analysis.

NNAR is the preferred forecasting method according to the three evaluative measures described above (Table 4). Based on this, NNAR out performs Holt-Winters and ARIMA in the case of training data set. Further, it can found that Holt-Winters performed better compared to ARIMA.

For the purpose of comparing the two methods it was necessary to calculate the error of the forecasts, meaning the forecasted value of each series was compared against the actual observed value in the relevant period. Both forecasted values were then compared against the actual observed values that were excluded from the analysis.

NNAR is the preferred forecasting method according to the three evaluative measures described above (Table 4). Based on this, NNAR out performs Holt-Winters and ARIMA in the case of training data set. Further, it can found that Holt-Winters performed better compared to ARIMA.

Table 2. Summary results for the fitted seasonal ARIMA model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drift</td>
<td>4.1439</td>
<td>1.5947</td>
</tr>
<tr>
<td>θ₁</td>
<td>-0.9080</td>
<td>0.0417</td>
</tr>
<tr>
<td>θ₂</td>
<td>0.2772</td>
<td>0.1082</td>
</tr>
</tbody>
</table>

Table 3. Ljung-Box test statistic and its significance level for different models

<table>
<thead>
<tr>
<th>Model</th>
<th>Box-Ljung z²</th>
<th>Significance probability (p &lt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holt-Winters</td>
<td>38.38</td>
<td>0.032</td>
</tr>
<tr>
<td>ARIMA</td>
<td>79.40</td>
<td>0.001</td>
</tr>
<tr>
<td>NNAR</td>
<td>25.50</td>
<td>0.379</td>
</tr>
</tbody>
</table>

Table 4. Comparison of forecasting accuracy measurement statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAPE (%)</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holt-Winters</td>
<td>45</td>
<td>68</td>
<td>29</td>
<td>176</td>
<td>229</td>
<td>58</td>
</tr>
<tr>
<td>ARIMA</td>
<td>55</td>
<td>83</td>
<td>45</td>
<td>178</td>
<td>205</td>
<td>67</td>
</tr>
<tr>
<td>NNAR</td>
<td>39</td>
<td>58</td>
<td>26</td>
<td>436</td>
<td>531</td>
<td>164</td>
</tr>
</tbody>
</table>

On the other hand, Holt-Winter’s model performed well in the case of test data set. (Table 3, Fig. 4). ARIMA models also performed in a similar way as evidenced by the resemblance in the frequency and amplitude of their residuals. NNAR model gave poor forecast among three models resulting in higher mean absolute forecast errors over the 16-time period. This is indicated by larger residuals. The larger residuals for all models occur in years when there were large changes (either positive or negative) in CPUE from the previous year, namely, third quarter of 1997 and 2009 and fourth quarter of 2010. (Fig. 4). The random fluctuations in the observed catch since 2007 are reflected in the NNAR and ARIMA model. Thus, their generalization capability, i.e. the ability to predict well if a strong deviation from the previous pattern occurs, was limited compared with the Holt-Winter’s model. From the study, it can be concluded that Holt-Winter’s model has better forecasting ability than ARIMA and NNAR based on the two-year ahead forecast (first quarter of 2011 to fourth quarter of 2014) for CPUE. Based on the developed Holt-Winters model, forecasts were obtained for next three years (12 periods ahead) and are given in Table 5.

Considering the importance of fisheries sector, it is essential to know about fish abundance, total biomass,
Modeling CPUE series for fishery: a comparison between three models

quarter is expected to have marginal reduction in CPUE to the tune of 21% and 24% respectively.

Acknowledgements

The authors are grateful to the Director, CMFRI for the support and facilities provided to carry out the study.

References


Table 5. Forecasted CPUE using Holt-Winters model

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>Holt-Winters Forecast</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>1</td>
<td>351</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>266</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>460</td>
<td>98</td>
</tr>
<tr>
<td>2016</td>
<td>1</td>
<td>363</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>106</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>277</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>472</td>
<td>116</td>
</tr>
<tr>
<td>2017</td>
<td>1</td>
<td>375</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>118</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>289</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>484</td>
<td>131</td>
</tr>
</tbody>
</table>


