Far East Journal of Theoretical Statistics
Volume 41, Number 1, 2012, Pages 1-26

Published by Pushpa Publishing House, Allahabad, INDIA

MODELING AND FORECASTING FISH PRODUCTION USING UNIVARIATE AND MULTIVARIATE ARIMA MODELS

Medhat Mohamed Ahmed Abdelaal and Essam Fawzy Aziz

Mathematics and Statistics Department
Faculty of Commerce
Ein Shams University
Cairo, Egypt
e-mail: medhatal@hotmail.com
essamfawzy2002@hotmail.com

Abstract

This paper applies Auto-Regressive Integrated Moving Average model (ARIMA) with (multivariate model) and without (univariate model) explanatory variables to predict the fish catch in Lake Manzala, Egypt. Bootstrap technique has been applied to allow one to judge the uncertainty of estimators obtained from the suggested ARIMA model, without prior assumptions about the underlying probability distributions. This method is based on generating 1000 random samples with replacement from the original observations. Then refit the suggested ARIMA model to each sample to end with a probability distribution of the model parameters, which can be used to estimate the bias of the parameters. Also, Jackknife technique has been applied after bootstrapping the ARIMA model. Jackknife-after-bootstrap

---

2010 Mathematics Subject Classification: 62M10.

Keywords and phrases: auto-regressive integrated moving average (ARIMA), bootstrap, Jackknife-after-bootstrap, multivariate model, univariate model, periodogram, regressor variables.

Received August 10, 2012

(JAB) technique aimed to estimate the bias of the bootstrap estimates, by deleting each observation in turn to obtain n estimates based on $n-1$ observations; this procedure has been repeated for each bootstrap sample. The main aim of applying these techniques is to try to minimize the bias of the estimation of ARIMA model, so we can get accurate estimation of the parameters which consider the most accurate tool to help the decision maker to apply proper management policies to optimize the fishing effort and protect the fish stock. Evidence from the Egyptian fisheries revealed that using ARIMA coupled with both bootstrap and JAB leads to the most accurate prediction in the perspective especially with time series data.

1. Introduction

Time series models use past values of the dependent/independent variables to forecast the future values of the considered variable; they can be grouped in univariate and multivariate models depending on using only past values of the examined variable or also the past values of other explanatory variables. Auto-Regressive Integrated Moving Average (ARIMA) models are, in theory, the most general class of models for forecasting a time series which can be stationarized by transformations such as differencing and lagging. In fact, the easiest way to think of ARIMA models is as fine-tuned versions of random-walk and random-trend models: the fine-tuning consists of adding lags of the differenced series and/or lags of the forecast errors to the prediction equation, as needed to remove any last traces of autocorrelation from the forecast errors [5].

ARIMA lags of the differenced series appearing in the forecasting equation are called "*auto-regressive*" terms, lags of the forecast errors are called "*moving average*" terms, and a time series which needs to be differenced to be made stationary is said to be an "*integrated*" version of a stationary series. Random-walk and random-trend models, auto-regressive models, and exponential smoothing models (i.e., exponential weighted moving averages) are all special cases of ARIMA models [6].

In statistical analysis, the interesting in obtaining not only a point estimate of a statistic, but also an estimate of the variation in this point estimate, and a confidence interval for the true value of the parameter. Traditionally, we may apply the central limit theorem and normal distribution approximations to obtain standard errors and confidence intervals. These techniques are valid only if the statistic, or some known transformation of it, is asymptotically normally distributed [13]. Hence, if the normality assumption does not hold, then the traditional methods should not be used to obtain confidence intervals.

As a general term, bootstrapping describes any operation which allows a system to generate itself from its own small well-defined subsets. The bootstrap is a method allowing one to judge the uncertainty of estimators obtained from random samples, without prior assumptions about the underlying probability distributions. The method consists of forming many new samples of the same size as the observed sample, by drawing a random selection of the original observations, i.e., usually introducing some of the observations several times [12]. The estimator under study (ARIMA model parameters) is then estimated for every one of the samples thus generated, and will show a probability distribution of its own [15].

Jackknife is a statistical procedure in which estimates are formed of a parameter based on a set of n observations by deleting each observation in turn to obtain n estimates, each based on $n-1$ observations. This method has deposed distribution-based methods in many applications due to its simplicity, its applicability in complicated situations, and its lack of distributional assumptions, resulting in greater reliability in practice. Jackknife after bootstrap aimed to estimate the bias of the bootstrap estimates [15].

Building on the previous literature background, this paper aims at predicting the main species catch using ARIMA model and estimating the bias of the predictions using re-sampling technique. These can help to get an accurate catch prediction to help the decision maker to apply proper management policies to optimize the fishing effort and protect the fish stocks.

2. Methodology

This part of the study includes shedding light on the case study used, the collected data description and the applied statistical techniques.

2.1. Case study

The data used in this paper is extracted from the annual statistics books published by Central Agency of Public Mobilisation And Statistics (CAPMAS) and General Authority for Fisheries Resources Development (GAFRD). These statistics include monthly catch of each species, monthly number of fishing boats, and monthly number of fishermen. The available historical data consists of monthly catch of Tilapia species from 1986 to 2011.

Lake Manzala is considered to be one of the main sources of fish production in Egypt. Its fish production represents about 25% of all national fish production in Egypt. From the catch time series, it seems that the catch during 1986-2011 can be distinguished between three main periods; the 1st from 1986 to 1994, the 2nd from 1995 to 1998 and the 3rd period from 1999 to 2011. The annual catch average was 36 kilo tonnes (Kt) in first period; in the second period, it was 48 Kt, while it was 52 Kt in the third period. The average number of the registered vessels was 2982 during the 1st period, and 6181 during the 2nd period; while it was 4381 in the 3rd period. With respect of average number of fishermen, it was 8953 in the 1st period and 18635 in the 2nd period; while it was 12393 in the 3rd period.

Catch composition shows that Tilapia species represents about 76% of total catch of the lake during 1986 to 2011, this means that Tilapia species is considered the dominant species and it plays a vital role in management and development of the lake fisheries. So, any policies aiming at improving fish production should be directed toward sustainable development of this species.

2.2. Data description

A primary concern of fishery science has been dynamic systems modeling, specifically the construction of mathematical models that determine values of fishing dynamic system, this system is composed of an input, fishing effort, an output, catch, and a mathematical model which translates values of fishing effort to catch. The mathematical model can be constructed from two very different approaches: either driving the model from fundamental laws of nature; or using "the probabilistic or stochastic approach". Actually, the fundamental difference between those two modeling approaches is that the deterministic approach describes a model in advance which is subsequently used to fit the collected data, while the stochastic approach deduces a model solely from the actual data [22].

As shown in Figure 1, the time series plot of the monthly catches of Tilapia in kilo tonne (Kt) shows a slow upward trend from 1986 to 1994; from 1995 to 1998, there is much variation with increasing catch rate and more fluctuations and during 1999 to 2011, the catch rate decreased and has less variation than the second period but still higher than the catch rate in the first period. In order to show a clear history of catches and to estimate the growth or decay of Tilapia catches, a time series analysis for monthly data had been investigated.

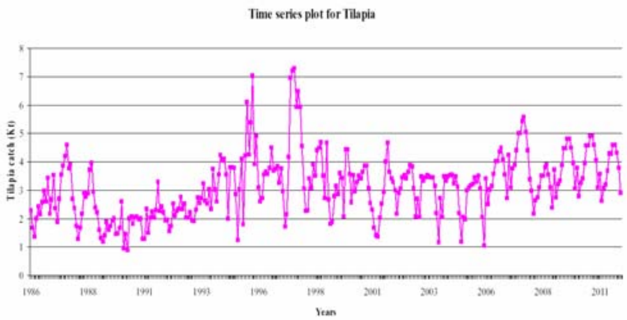


Figure 1. Monthly time series plot during 1986-2011 of Tilapia catch in Kt.

Estimated autocorrelation is very useful tool to test seasonal pattern or as a preliminary step in determining an appropriate time series model for the data [9]. Figure 2 shows the estimated autocorrelation between successive values of this time series. In Autocorrelation Function (ACF) plot, the vertical bars represent the coefficient of each lag and pair of lines at a distance from the base line which is multiples of the standard error at each lag to represent the confidence limits. Significant autocorrelation extends above or below the confidence limits. It is clear that there is a statistically significant correlation between the successive values which means that the time series during 1986-2012 are not stationary.

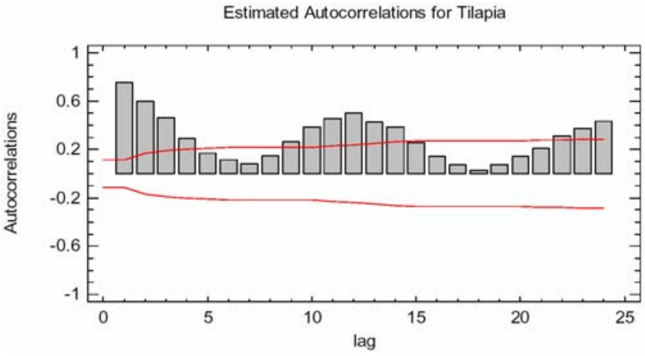


Figure 2. Estimated autocorrelation during 1986-2012 of Tilapia catch.

Periodogram

Periodogram is often used to identify cycles of fixed frequency in the data. The periodogram is constructed by fitting a series of sine functions at each of 312 frequencies. The ordinates are equal to the squared amplitudes of the sine functions. The periodogram can be thought of as an analysis of variance by frequency, the sum of the ordinates equals the total corrected sum of squares in an ANOVA table [21]. The periodogram plots the ordinates horizontally and frequency vertically, and plots the amplitudes of sinusoids (any curve derivable from the sine curve by multiplication by a constant or addition of a constant; a curve of the same shape as the sine curve but with possibly different amplitude, period or intercepts with the axis) for various frequencies against the frequencies [23]. Periodograms are computed using Fourier transformations. The value of the ordinate at each frequency F_i is given by the following formula if n (number of observation) is odd:

$$I(F_i) = \frac{n}{2}[a_i^2 + b_i^2], \quad i = 1, 2, 3, \ldots, \left(\frac{n-1}{2}\right), \tag{1}$$

where

$$a_i = \frac{2}{n}\sum_{t=1}^{n} Y_t \mathrm{Cos}(2\pi F_i t), \quad b_i = \frac{2}{n}\sum_{t=1}^{n} Y_t \mathrm{Sin}(2\pi F_i t), \quad F_i = i/n.$$

The time series data contain $n = 312$ observations which are even, so an additional term can be added:

$$I(0.5) = n\left(\frac{1}{n}\sum_{t=1}^{n} Y_t\right)^2. \tag{2}$$

If the periodogram plot shows one spike, that is, obviously larger than any other spike, then the largest magnitude is the length of seasonality ignoring the spike corresponds to frequency zero [6]. Figure 3 shows the periodogram plot of the time series of Tilapia catch; the largest ordinate value was 99.866 corresponding to 12 months seasonality length.

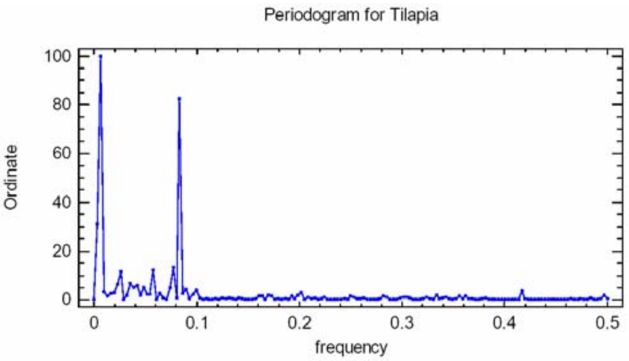


Figure 3. Periodogram plot during 1986-2011 of Tilapia catch.

To remove the trend, the first non-seasonal difference operator has been used. The estimated autocorrelation shows that there is no statistically significant correlation between the successive values of the first difference except around one or two years reflecting seasonality [5].

2.3. Study techniques

This part sheds light on the deployed analytical techniques: Box-Jenkins model and re-sampling technique (bootstrap and Jackknife after bootstrap).

2.3.1. Box-Jenkins model

It contains two main models:

(1) **A**uto-**R**egressive **M**oving **A**verage **(ARMA)** model that consists of five parameters $(p, q)(P, Q)S$. Statistically, ARMA model uses only the original data of the time series without any type of transformation.

(2) **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage model **(ARIMA)** that suggests data transformation (e.g., the differences).

It consists of seven parameters which can be expressed as $(p, d, q)(P, D, Q)S$. As the original data was non-stationary and the first difference operator has been used to convert it to stationary data, it has been decided to use the ARIMA model in this study. The definition of the model parameters can be expressed as follows [4]:

$p, P =$ non-seasonal and seasonal auto-regressive parameters, respectively,

$q, Q =$ non-seasonal and seasonal moving average parameters, respectively,

$d, D =$ non-seasonal and seasonal differences, respectively,

$S =$ seasonally length.

The general formula of (ARMA) can be as follows:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + e_i - \beta_1 e_{i-1} - \beta_2 e_{i-2} - \cdots - \beta_q e_{i-q}. \tag{3}$$

The general form of ARIMA model can be expressed as follows:

$$Z_t = \alpha_0 + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_p Z_{t-p} + e_i - \beta_1 e_{i-1} - \beta_2 e_{i-2} - \cdots - \beta_q e_{i-q}, \tag{4}$$

where Z_t is the operator, which can transform the non-stationary time series to stationary time series. So, if the first non-seasonal difference operator is applied, then $Z_t = Y_t - Y_{t-1}$.

Actually, to build ARIMA model, three main stages are required: (1) model identification, (2) parameters estimation and (3) model validation. This procedure is repeated until an adequate and valid time series model is specified. Model identification involves two stages: the model structure, and its corresponding model order. Using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) facilitates specifying model orders p and q. The ACF is used to estimate the autocorrelation between time series variable and the values for earlier time period. The PACF measures the linear relationship between time series observations separated by K time periods with the effects of intervening observations removed. A synopsis of ACF and PACF properties is listed below [4]:

- Mean non-stationary: The ACF decays slowly. The PACF possesses a large positive or negative value at lag 1.
- Strong seasonality: The ACF is zero except at the seasonal lags, i.e., S, $2S$, $3S$, … and decays very slowly.
- Auto-regressive behaviour: For a low order AR model, the PACF is nonzero for lags $K = 1, 2, 3, ..., p,$ and is zero thereafter. For seasonal AR model, the PACF is nonzero at lags $K = S, 2S, 3S, ..., pS,$ and zero elsewhere.
- Moving average behaviour: For low order MA model, the ACF is nonzero for lags $K = 1, 2, 3, ..., q,$ and zero thereafter. For seasonal MA model, the PACF is nonzero at lags $K = S, 2S, 3S, ..., qS,$ and is zero elsewhere.

When performing model identification, mean non-stationary and seasonality should be corrected by applying the difference operator Δ_s^d, where d is the level of differencing and S is the time period to which d is applied, before specifying the AR and MA model orders. Once a model has been tentatively specified, the model parameters are estimated. The most important validation procedure in ARIMA modeling approach is to inspect the ACF of residuals for any significant lags that still remain. An adequate ARIMA model is the one whose ACF is not statistically significantly different from zero. Also, diagnostic checking is usually done by inspecting the residuals, such as Portmanteau Test (sometimes called the *Box-Pierce-Ljung statistic*) and Goodness-of-Fit test determine if the residuals can be adequately modeled by normal distribution or not [16].

The tentative, parameterized model, at this point, is still a first-cut, informed guess. The most important validation procedure in ARIMA modeling approach is to inspect the ACF of the residuals for any significant lags that still remain. An adequate ARIMA model is one which the ACF contains no significant lags and thus resembles a white noise process. An important use of a time series model is to forecast future values of an output series. The forecasts are usually statistically valid only for next season of an output series; due to increasing prediction error as the forecast continues into the future [22].

2.3.2. Re-sampling techniques

2.3.2.1. Bootstrap

The bootstrap was first discussed by Efron [14] and is detailed further by Efron and Tibshirani [15]. Bootstrap was developed to provide standard errors and confidence intervals for a model coefficient and predicted values in situations in which the standard assumptions are not valid. The bootstrap method attempts to determine the probability distribution from the data itself, without recourse to Central Limit Theorem. The bootstrap algorithm, for a sample size of n, is as follows:

(i) Select B bootstrap samples of size n drawn with replacement.

(ii) Estimate parameters, $\hat{\theta}^*_{(b)}$, for each bootstrap sample $b = 1, 2, ..., B$.

The bootstrap estimates of the standard errors are given by equation (5), where $\hat{\theta}^*_{(.)}$ is the mean of the B bootstrap estimates. In order to obtain these estimates, Efron and Tibshirani [15] suggested that the number of bootstrap replicates, *B*, can normally be taken from the range 25 to 200:

$$\overset{\wedge}{SE} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*}_{(b)} - \hat{\theta}^{*}_{(\cdot)})^2}. \tag{5}$$

The bootstrap estimate of the bias, when the parameter estimates are $\hat{\theta}$, is given by equation (6). This estimate requires *B*, the number of bootstrap replicates, to be much larger than when estimating only the standard errors:

$$\overset{\wedge}{bias} = \hat{\theta}^{*}_{(\cdot)} - \hat{\theta}. \tag{6}$$

For each bootstrap sample, ARIMA model results are computed and stored. The bootstrap sampling process has provided *B* estimates of a parameter. The standard deviation of these *B* estimates of the bootstrap estimate is the standard error of this parameter. The bootstrap confidence interval is found the arranging the *B* values in sorted order and selecting the appropriate percentiles from the list [26]. A 90% bootstrap confidence interval for the slope is given by fifth and ninety-fifth percentiles of the bootstrap estimates values. The main assumption made when using the bootstrap method is that our sample approximates the population fairly well. Because of this assumption, bootstrapping does not work well for small samples in which there is little likelihood that the sample is representative of the population. Bootstrapping should only be used in medium to large samples [12].

2.3.2.2. Jackknife method

If the true parameter values are given by the vector θ and this is estimated by $\hat{\theta}$, then the bias of the estimate is given by equation (7):

$$bias = \theta - \hat{\theta}. \tag{7}$$

For a sample size *n*, let $\hat{\theta}$ be the estimated parameter values data including all the observations, and let $\hat{\theta}_i$ be the estimated parameter values on data excluding the *i*th observation, that is, $\hat{\theta}_i = \hat{\theta}(x_1, x_2, \ldots, x_{i+1}, \ldots, x_n)$.

Letting $\hat{\theta}_{(\cdot)}$ be the mean of all the fitted parameters on the reduced data, that is the mean of all $\hat{\theta}_i$, enables the Jackknife estimate of the bias to be given by equation (8):

$$\widehat{bias} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}). \tag{8}$$

The Jackknife corrected estimate of the bias, $\tilde{\theta}$, can then be given by equation (9):

$$\tilde{\theta} = \hat{\theta} - \widehat{bias}.$$

Therefore,

$$\tilde{\theta} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}. \tag{9}$$

The standard errors formula of the Jackknife is given by equation (10):

$$\widehat{SE} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}. \tag{10}$$

Efron and Tibshirani [15] stated that the Jackknife provides a simple procedure to estimate the bias and the standard errors. However, the Jackknife is only valid when the statistic $\hat{\theta}$ is smooth, that is, small changes in the data cause only small changes in the estimates of the parameters. Jackknife is a special kind of bootstrap. Each bootstrap sub-sample has all but one of the original elements of the list. For *n* observations, there are *n* Jackknife sub-samples. A simple approach that uses the information in the original bootstrap samples without further re-sampling was sought.

Efron [14] derived the Jackknife after Bootstrap (JAB) in an attempt to provide a solution to this problem. The JAB usually requires 100-1000 times less computation than JAB. Although the JAB is theoretically justified, situations arise in practice, where it performs poorly. These situations occur when a large number of bootstrap samples are not able to be drawn [5]. The JAB estimates in these situations have a tendency to be over-inflated and do not reflect the true variability in the bootstrap statistics but more the limitations of using a small number of bootstrap samples.

3. Study Results

3.1. Fitting univariate ARIMA time series model

Tilapia catch time series has been split to two data sets; the first data set from 1986 to 2010 is used to fit ARIMA model and last year 2011 has been used as a test data set to check the validity of the suggested model for forecasting. The best model has been found after many trials. This model is second order of non-seasonal auto-regressive and first order of seasonal auto-regressive model based on first non-seasonal difference. This can be expressed as ARIMA (2,1,0)(1,0,0)12, which means that the catch in a month depends on the catch in previous two months and previous twelfth months; that is, previous years levels. It is noted that the constant term is not significantly different from zero; we may assume that there is no specific level of the catch regardless of the effect of the previous levels so the model has been refitted without constant term. Figure 4 shows the time sequence plot of ARIMA model for Tilapia species catch.

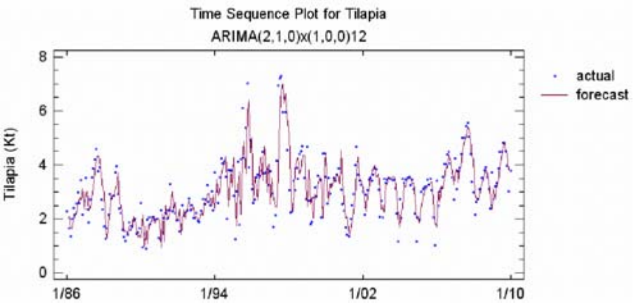


Figure 4. Time sequence plot of ARIMA (2,1,0)(1,0,0)12 during 1986-2010 of Tilapia catch.

The ARIMA model (2,1,0)(1,0,0)12 can be expressed as follows:

$$\begin{aligned}\Delta_1 C_t &= \beta_1 \Delta_1 C_{t-1} + \beta_2 \Delta_1 C_{t-2} + \beta_3 \Delta_1 C_{t-12} \\ &\quad - \beta_1 \beta_3 \Delta_1 C_{t-13} - \beta_2 \beta_3 \Delta_1 C_{t-14}, \qquad (11)\end{aligned}$$

where:

$\Delta_1 =$ The first difference operator; so $\Delta_1 C_t = C_t - C_{t-1}$,

$C_t =$ Monthly catch in time period t,

$\beta_1, \beta_2, \ldots, \beta_p =$ ARIMA model parameters and

$t =$ time in months 1, 2, 3, ….

This model can be simplified as follows:

$$C_t = (1 + \beta_1) C_{t-1} + (\beta_2 - \beta_1) C_{t-2} - \beta_2 C_{t-3} + \beta_3 C_{t-12}$$
$$- (\beta_1\beta_3 + \beta_3) C_{t-13} + (\beta_1\beta_3 - \beta_2\beta_3) C_{t-14} + \beta_2\beta_3 C_{t-15}. \qquad (12)$$

Table 1 displays the parameters in the ARIMA model without explanatory variables and the values of the estimated coefficients and their associated standard error, t-statistic, and p-values.

Table 1. Summary of univariate ARIMA model parameters

Parameter	AR1	AR2	SAR1	Portmanteau test	RUNM
Estimation	–0.246	–0.226	0.357	0.000	0.234
Standard error	0.057	0.058	0.056		
t-value	–4.317	–3.923	6.400		
p-value	0.000	0.000	0.000		

Table 1 summarizes the statistical significance of the terms in the forecasting model. Terms with p-values less than 0.05 are statistically significantly different from zero at the 95% confidence level. The p-value for the auto-regressive parameters is less than 0.05, so it is significantly different from zero. Also, the p-value for the SAR1 term is less than 0.05, so it is significantly different from zero. The estimated white noise variance is 0.535 with 308 degrees of freedom and the estimated white noise standard deviation is 0.720. It is noted that the auto-regressive parameters are negative which were expected because there is a downward trend in the time series.

The two tests Portmanteau Test and Runs above and below median (RUNM) have been run to determine whether or not the residuals come form a random sequence of numbers. A sequence of random numbers is often called *white noise*, since it contains equal contributions at many frequencies. The Portmanteau Test is used to determine if there is any pattern left in the residuals that may be modeled. The test is computed as follows:

$$Q(k) = N(N+2)\sum_{j=1}^{k} \frac{r_j^2}{N-j}, \tag{13}$$

$Q(k)$ is distributed as a Chi-square with $(K\text{-}p\text{-}q\text{-}P\text{-}Q)$ degrees of freedom. This is accomplished by testing the significance of the autocorrelations up to a certain lag. The Portmanteau Test is based on the sum of squares of the first 24-autocorrelation coefficients. Since the p-value for this test is less than 0.05, the hypothesis that the series is random at the 95% confidence level cannot be rejected. The second test RUNM counts the number of times, the sequence was above or below the median. Since the p-value for this test is greater than or equal to 0.05, the hypothesis that the residuals are random at the 95% or higher confidence level cannot be rejected. The ARIMA model for Tilapia species catch can be extracted as follows:

$$\begin{aligned} C_t &= 0.754C_{t-1} + 0.021C_{t-2} + 0.226C_{t-3} + 0.357C_{t-12} \\ &\quad - 0.269C_{t-13} - 0.007C_{t-14} - 0.080C_{t-15}. \end{aligned} \tag{14}$$

The most important validation procedure in ARIMA modeling approach is to inspect the ACF and PACF of the residuals for any significant lags that still remain. An adequate ARIMA model is one whose ACF and PACF are not significantly different from zero. Figure 5 shows the residuals autocorrelation and residuals partial autocorrelation for the Tilapia species catch, which give good evidence that the ARIMA model (2,1,0)(1,0,0)12 is the best model for all time series.

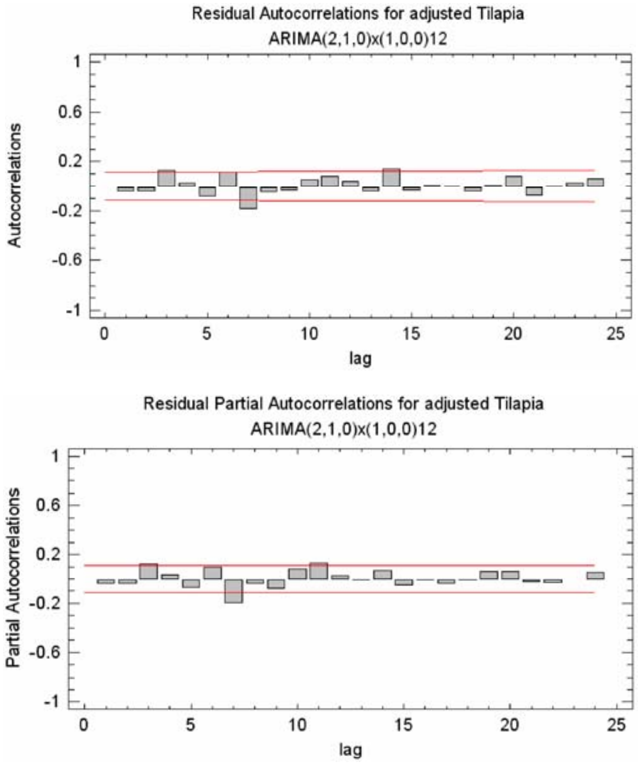


Figure 5. The residuals autocorrelation ACF and residuals partial autocorrelation PACF for Tilapia species catch.

3.2. Fitting multivariate ARIMA time series model

An analysis for the same time series was carried out using ARIMA model plus explanatory variables such as number of vessels, and number of fishermen; logically the effect of these variables cannot be ignored because the current catches may not only depend on the previous catches, but also on at least one of those explanatory variables. It was found that the only significant forms of the independent variables were the reciprocal of number of fishermen and number of vessels to power (–2). Table 2 shows the estimated parameters of ARIMA model with explanatory variables and the values of the estimated coefficients and their associated standard error, *t*-statistic, and *p*-values.

Table 2. Summary of multivariate ARIMA model parameters

Parameter	AR1	AR2	SAR1	Fleet^{-2}	Fishermen^{-1}	Portmanteau test	RUNM
Estimation	–0.247	–0.226	0.356	–383075	304.866	0.000	0.324
Standard error	0.057	0.0575	0.056	1E-6	0.001		
t-value	–4.322	–3.925	6.400	–3E-11	224789		
p-value	0.000	0.000	0.000	0.000	0.000		

It is noticeable that all terms are statistically significantly different from zero at the 95% confidence level. The estimated white noise variance is 0.53482 with 306 degrees of freedom and the estimated white noise standard deviation is 0.518. The parameter estimation of the explanatory variable Fleet^{-2} has a negative sign which means that the decreasing number of vessels will cause an increase in the catch, while the parameter estimation of the explanatory variable Fishermen^{-1} has a positive sign which means that the increase in number of fishermen size causes a decrease in the catch and vice versa which is quite logical finding, we conclude that the relationship between the explanatory variables and the catch is based on good reasons.

The ARIMA model (2,1,0)(1,0,0)12 with two explanatory variables can be expressed as follows:

$$C_t = (1+\beta_1)C_{t-1} + (\beta_2-\beta_1)C_{t-2} - \beta_2 C_{t-3} + \beta_3 C_{t-12} - (\beta_1\beta_3+\beta_3)C_{t-13}$$
$$+ (\beta_1\beta_3 - \beta_2\beta_3)C_{t-14} + \beta_2\beta_3 C_{t-15} + \gamma_1\,\text{Fleet}^{-2} + \gamma_2\,\text{Fishermen}^{-1}. \quad (15)$$

The ARIMA model with two explanatory variables for Tilapia species catch can be extracted as follows:

$$C_t = 0.753C_{t-1} + 0.021C_{t-2} + 0.226C_{t-3} + 0.356C_{t-12}$$
$$- 0.269C_{t-13} - 0.007C_{t-14} - 0.080C_{t-15}$$
$$- 383075\,\text{Fleet}^{-2} + 304.87\,\text{Fishermen}^{-1}. \quad (16)$$

By comparing the two models, it is noticeable that the ARIMA parameters are the same in both models. But the second model improves more accuracy of the prediction. Table 3 shows the performance of the two models.

Table 3. The performance of univariate and multivariate ARIMA models

Statistics	Univariate ARIMA model	Multivariate ARIMA model
RMSE	0.732	0.730
MAE	0.533	0.521
MAPE	19.333	18.180
ME	0.004	0.002
MPE	–4.154	–3.981

Table 3 summarizes the performance of the currently selected model in fitting the historical data, which are: the root mean squared error (RMSE); the mean absolute error (MAE); the mean absolute percentage error (MAPE); the mean error (ME); and the mean percentage error (MPE). Each of the statistics is based on the one-ahead forecast errors, which are the differences between the data value at time t and the forecast of that value made at time $t-1$. The first three statistics measure the magnitude of the errors. A better model will give a smaller value. The last two statistics measure bias. A better model will give a value close to zero. So, ARIMA with explanatory model can produce more accurate predictions than ARIMA without explanatory model.

The previous sections have introduced ARIMA model with and without explanatory variables, and have given fitted models for the Tilapia catch data as well as appropriate tests of significance and residuals analysis. These tests of significance for the fitted parameter values rely on the information matrix yielding good estimates of the standard errors. Although the true values of the standard errors are not known and so cannot be compared with the estimates, the validity of the estimates can be checked by means of simulation. The following section describes how to simulate data from ARIMA model and how this simulated data can be used to estimate the standard errors. The simulation study performed to validate the estimates of the standard errors highlighted a bias in the variance, and this requires further investigations. If the bias is too large, then the fitted models will be limited use.

The two methods introduced in the previous sections (bootstrap and Jackknife after bootstrap) have both been used to obtain estimates of the parameter bias for the ARIMA model without explanatory variables and ARIMA model with explanatory variables, equations (14) and (16). For each parameter, a histogram of the replicates is displayed with an overlaid smooth density estimate. The histograms in Figure 6 show that the distributions of replicated bias are normal (for example, the first and second parameters of the first ARIMA model and the parameters of the explanatory variables of the second ARIMA model), which mean that the bootstrap can produce unbiased estimates for the parameters, the same results are observed for the other parameters of the two models.

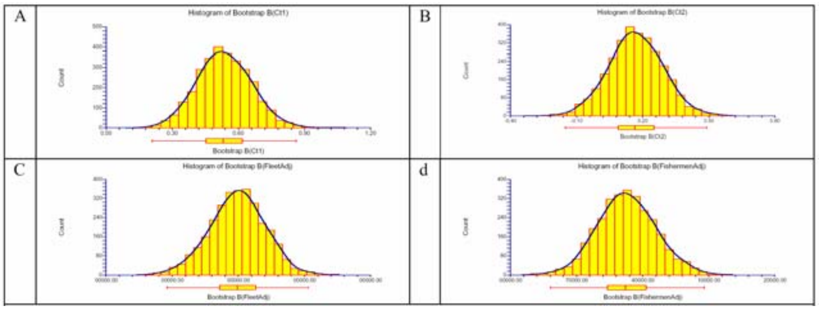
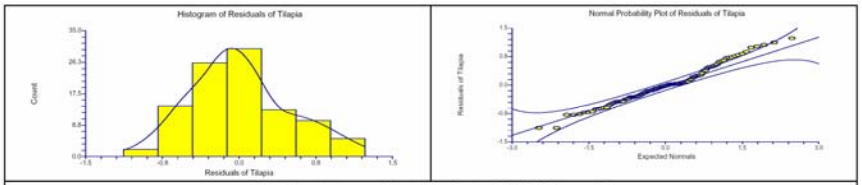


Figure 6. Histogram of replicated bias for the ARIMA model parameters (a) the C_{t-1} parameter of the first ARIMA model, (b) the C_{t-2} parameter of the first ARIMA model, (c) the Fleet parameter of the second ARIMA model and (d) the Fishermen parameter of the second ARIMA model.

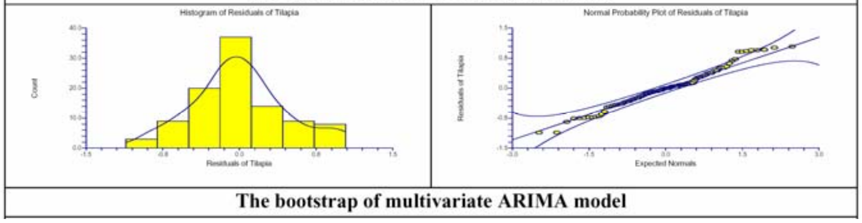
Figure 6 gives good evidence that bootstrap technique can produce accurate predictions. A normal probability plot has been used to further assess deviation from the normal distribution. Figure 7 gives an evidence that the residuals of the bootstrap are normally distributed for ARIMA model of Tilapia species with and without explanatory variables.

The bootstrap technique (for 1000 bootstrap replications) and JAB have been used to estimate the bias and the standard errors for each model. The estimates of the bias and the standard errors for both models are given in

Table 4. The results show that the Jackknife and bootstrap produce quite similar estimates of the bias.



The bootstrap of univariate *ARIMA* model



The bootstrap of multivariate ARIMA model

Figure 7. The normal probability plot and the histogram of the residuals of ARIMA models.

Table 4. Estimates of bias and standard errors

	Univariate ARIMA model				Multivariate ARIMA model			
	Bootstrap		JAB		Bootstrap		JAB	
Parameters	Bias	SE	Bias	SE	Bias	SE	Bias	SE
C_{t-1}	–0.001	0.099	0.001	0.105	–0.002	0.097	0.004	0.103
C_{t-2}	0.012	–0.463	0.013	–0.505	0.009	–0.458	0.010	–0.475
C_{t-3}	0.005	0.161	0.003	0.174	0.009	0.159	0.002	0.161
C_{t-12}	0.006	0.063	0.005	0.064	0.004	0.058	0.005	0.061
C_{t-13}	–0.002	0.142	–0.003	0.139	0.004	0.137	–0.002	0.139
C_{t-14}	0.000	–0.003	0.000	–0.003	0.002	–0.003	0.001	–0.003
C_{t-15}	–0.001	0.028	–0.001	0.029	–0.001	0.029	–0.008	0.031
Fleet^{-2}					8282.806	376214	1431.671	372935.901
Fishermen^{-1}					2.444	168.901	–0.393	175.994

The ratios of the bias to the standard errors for the Jackknife are given in Table 5.

Table 5. Ratios of bias to standard for Jackknife and bootstrap estimates

	Univariate ARIMA model		Multivariate ARIMA model	
	Bootstrap	JAB	Bootstrap	JAB
Parameters	Bias/SE	Bias/SE	Bias/SE	Bias/SE
C_{t-1}	2.64%	0.42%	0.75%	0.57%
C_{t-2}	1.98%	2.17%	2.69%	2.74%
C_{t-3}	6.22%	1.58%	3.17%	2.08%
C_{t-12}	7.62%	9.22%	8.82%	8.96%
C_{t-13}	3.45%	1.87%	1.52%	1.98%
C_{t-14}	6.04%	1.56%	0.90%	0.89%
C_{t-15}	3.85%	2.56%	4.75%	3.17%
Fleet^{-2}			2.20%	0.38%
Fishermen^{-1}			1.45%	0.22%

Two types of percentile estimates are calculated: empirical percentiles and bias-corrected and adjusted BCa percentiles. The empirical percentiles are easy to calculate, but may not be very accurate unless the sample size is very large. The BCa percentiles require more computation but are more accurate. For either type of percentile, using at least 1000 replications is recommended for accurate estimation. The empirical percentiles are simply the percentiles of the empirical distribution of the replicates. The BCa method transforms the specified probability values to determine which percentiles of the empirical distribution most accurately estimate the percentiles of the parameter. The percentiles of the empirical distribution corresponding to these values are then returned. To estimate the BCa percentiles, the bias correction (denoted Z_0) and the acceleration must be calculated. If these values are not specified (and they usually are not), then the bias correction will be obtained from the replicates, and the acceleration will be obtained using Jackknife. In Jackknife, the data are broken into groups and these groups are Jackknifined, the floor of the group size is

$(n/20)$, which will yield roughly 20 Jackknife replicates, depending on the magnitude of n.

Jackknife after bootstrap is a technique for obtaining estimates of the variation in functionals of the bootstrap distribution, such as the bias of the statistic, also it provides information on influence of each observation [15]. Table 6 shows the results of the bootstrap technique and JAB for ARIMA model of Tilapia species without explanatory variables and for ARIMA model of Tilapia species with explanatory variables using 95% confidence level.

Table 6. The empirical and BCa percentiles of ARIMA model without explanatory variables and ARIMA model with explanatory variables

	Univariate ARIMA model			Multivariate ARIMA model		
	Bootstrap		JAB	Bootstrap		JAB
Parameter	BCa	Empirical percentiles	Empirical percentiles	BCa	Empirical percentiles	Empirical percentiles
C_{t-1}	0.848	0.843	0.699	0.846	0.839	0.695
C_{t-2}	0.149	0.162	0.001	0.161	0.154	0.001
C_{t-3}	0.245	0.255	0.127	0.247	0.243	0.119
C_{t-12}	0.462	0.474	0.373	0.45	0.462	0.361
C_{t-13}	–0.051	–0.034	–0.171	–0.031	–0.034	–0.171
C_{t-14}	0.352	0.331	0.195	0.344	0.333	0.194
C_{t-15}	–0.076	–0.072	–0.187	–0.084	–0.089	–0.196
Fleet^{-2}				488820.201	597403.4	–926342.7
Fishermen^{-1}				3364.21	3359.33	1915.77

The corrected parameters with the estimated bias using bootstrap technique and JAB for both ARIMA models parameters with the ARIMA models parameters are shown in Table 7.

Table 7. ARIMA models parameters and the corrected parameters

	Univariate ARIMA model			Multivariate ARIMA model		
Parameter	ARIMA	Bootstrap	JAB	ARIMA	Bootstrap	JAB
C_{t-1}	0.754	0.751	0.754	0.753	0.753	0.754
C_{t-2}	0.021	0.02	0.031	0.021	0.033	0.035
C_{t-3}	0.226	0.236	0.228	0.226	0.231	0.229
C_{t-12}	0.357	0.361	0.362	0.357	0.362	0.362
C_{t-13}	–0.269	–0.264	–0.271	–0.269	–0.271	–0.271
C_{t-14}	–0.007	–0.007	–0.007	–0.007	–0.007	–0.007
C_{t-15}	–0.081	–0.082	–0.081	–0.081	–0.082	–0.081
Fleet^{-2}				–383075	–374792.19	–381643
Fishermen^{-1}				304.866	307.31	304.473

The corrected parameters and ARIMA models parameters have been used to predict the catch in 2011. Figure 8 shows the comparison between the actual data and the predicted values using the corrected parameters.

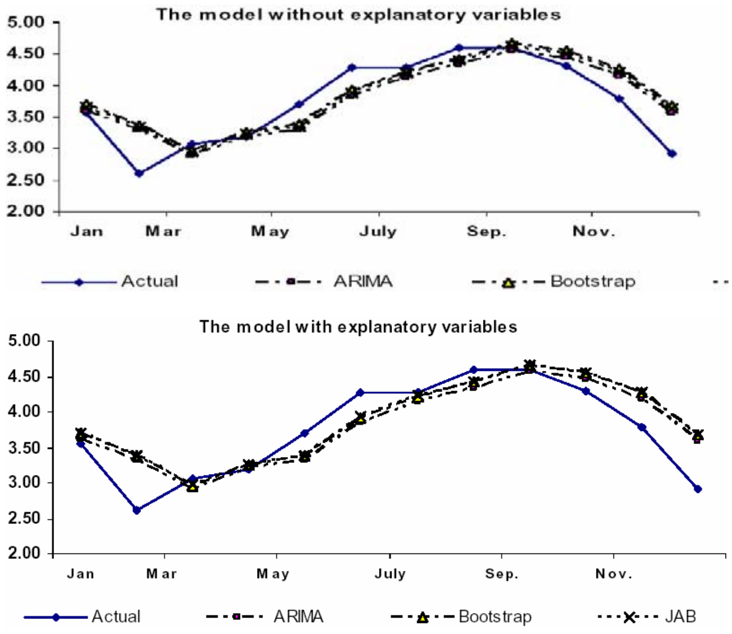


Figure 8. The actual data and the predicted values using the corrected parameters for both models.

From Figure 8, it is noticeable that ARIMA model parameters and the corrected parameters with the estimated bias produced by using bootstrap technique and the corrected parameters produced by using JAB can produce quite similar predictions. But the minimum difference between the predicted values and the actual values is observed in the use of the corrected parameters using the estimated bias produced by using JAB technique.

4. Conclusions

Evidence from the Egyptian fisheries revealed that using ARIMA coupled with both bootstrap and JAB leads to the most accurate prediction in the perspective especially with time series data. These two additional analytical techniques bootstrap and JAB have been used to estimate the bias of the estimated parameters. Consequently, it is strongly recommended adopting these two consecutive techniques with similar situations to minimize the bias of the estimated parameters.

References

[1] A. Aczel and N. Josephy, Using the bootstrap for improved ARIMA model identification, J. Forecast. 11 (1992), 71-80.

[2] T. W. Anderson, Goodness-of-fit for autoregressive process, J. Time Ser. Anal. 18 (1997), 321-339.

[3] O. D. Anderson and M. R. Perryman, Covariance Structure of Sample Correlations from ARIMA Models in Time Series Analysis, North-Holland, New York, 1981.

[4] B. L. Bowerman and R. T. O'Connel, Forecasting and Time Series: Applied Approach, 3rd ed., Duxbury Press, Belmont, California, 1993.

[5] G. E. P. Box and G. M. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, 1976.

[6] P. J. Brockwell and R. A. Davies, Time Series: Theory and Methods, 2nd ed., Springer-Verlag, New York, 1991.

[7] Central Agency of Public Mobilisation And Statistics (CAPMAS), Statistics Year Book from 1986 to 2011, Published Report, Cairo, Egypt, 2011.

[8] Charisios Christodoulos, Christos Michalakelis and Dimitris Varoutas, Forecasting with limited data: combining ARIMA and diffusion models, Technological Forecasting and Social Change, Elsevier, 2010, pp. 558-565.

[9] W. S. Cleveland and S. J. Devlin, Calendar effects in monthly time series: detection by spectrum analysis and graphical methods, J. Amer. Statist. Assoc. 75 (1980), 487-496.

[10] E. S. Dagum, Moving Averages in Encyclopaedia of Statistical Sciences, Vol. 5, John Wiley & Sons Inc., New York, 1985.

[11] N. Davies, C. M. Triggs and P. Newbold, Significance levels of the Box-Pierce Portmanteau statistic in finite samples, Biometrika 64 (1977), 517-522.

[12] A. C. Davison and D. V. Hinkley, Bootstrap Methods and their Applications, Cambridge University Press, New York, 1999.

[13] D. A. Dickey and W. A. Fuller, Distribution of the estimators for autoregressive time series with a unit root, J. Amer. Statist. Assoc. 74 (1979), 427-431.

[14] B. Efron, Jackknife-after-bootstrap standard errors and influence functions, J. Roy. Statist. Soc. Ser. B 54 (1992), 83-127.

[15] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[16] Yue Fang, Using least squares to generate forecasts in regressions with serial correlation, J. Time Ser. Anal. 15 (2008), 324-331.

[17] General Authority for Fisheries Resources Development (GAFRD), Statistics Year Book from 1986 to 2011, Published Report, Cairo, Egypt, 2011.

[18] P. Hall, The Bootstrap and Edgeworth Expansion, Springer-Verlag, New York, 1992.

[19] M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, 4th ed., Griffin, London, 1977.

[20] H. R. Künsch, The Jackknife and the bootstrap for general stationary observations, Ann. Statist. 17 (1989), 1217-1241.

[21] G. M. Ljung and G. E. P. Box, On a measure of lack of fit in time series models, Biometrika 65 (1978), 297-303.

[22] W. Makridakis, S. C. Wheelwright and U. E. McGee, Forecasting: Methods and Applications, 2nd ed., John Wiley, New York, 1983.

[23] D. A. Pierce, Data revisions with moving average seasonal adjustment procedures, J. Econometrics 14 (1980), 95-114.

[24] J. Shao and D. Tu, The Jackknife and Bootstrap, Springer-Verlag, New York, 1995.

[25] W. N. Venables and B. D. Ripley, Modern Applied Statistics Using *S*-plus, Springer-Verlag, New York, 1999.

[26] Nazuha Muda Yusof, Malaysia crude oil production estimation: an application of ARIMA model, Science and Social Research (CSSR), International Conference 2010, pp. 1255-1259.