Modeling and Forecasting by using Time Series ARIMA Models

Mustafa M. Ali Alfaki
Research Scholar,School of Basic Science
Sam Higginbottom Institute of Agriculture, Technology and Sciences
Allahabad, India

Dr. (Ms.) Shalini Bhawana Masih
Assistant Professor,School of Basic Science
Sam Higginbottom Institute of Agriculture, Technology and Sciences
Allahabad, India

Abstract:-- In this paper we introduce a brief review about Box-Jenkins models. "Auto-Regressive Integrated Moving Average", Known as ARIMA models. It is a good method to forecast for stationary and non stationary time series. According to the data which obtained from the monthly sales for Naphtha product in Azzawiya Oil Refining Company – Libya, then we determine an optimal model. The results of this study showed that the suitable and efficient model to represent the data of the time series according to AIC, SIC, and MSE criteria with the smallest values as well as the Box-ljung test is the ARIMA(1,1,1), whose equation is :

$$\hat{Z}_t = 0.6010\, Z_{t-1} + 1.1713\, \varepsilon_{t-1} + \varepsilon_t$$

According to the results of ARIMA (1,1,1), the amounts of the monthly sales for Naphtha product Azzawiya Oil Refining Company – Libya, have been forecasted for the period from Jan: 2015 to Dec: 2020. Those values showed harmony with their counterparts in the original time series. It provided us with a future image of the reality of the monthly sales for Naphtha product in Azzawiya Oil Refining Company – Libya.

Key Words : Forecasting, Box-Jenkins, Auto-Regressive Integrated Moving Average (ARIMA), Auto-Regressive (AR), Moving Average (MA), Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF).

INTRODUCTION

The planning process is the most important step in the develop plan of any country, it becomes adequate if the statistical methods that has been applied are based on powerful scientific basis.

The time series is a set of well-defined data items collected at successive points at uniform time intervals. Time series analysis is an important part in statistics, which analyzes data set to study the characteristics of the data and helps in predicting future values of the series based on the characteristics. Forecasting is important in fields like finance, industry, etc. The Autoregressive Integrated Moving Average (ARIMA) is based on ARMA Model. The difference is that ARIMA Modelconverts a non-stationary data to a stationary data before working on it. This analysis uses Box-Jenkins ARIMA modeling techniques to find an appropriate model for this time series. This model is the assessed to determine how well it's the data.

Non-seasonal Box-Jenkins Models for a Stationary Series:

That is known as (ARIMA) Models, it is an Autoregressive Model AR(p) as the first part, and Moving Average Model MA(q) a second part, and the third part I(d) represents the differences required by the time-series in order to be stationary. Some models of time series may be non-stationary of the same kind but they would become Stationary after a lot of differences and changes, so that the model which represents this process will vary from the original model because it will include those differences and changes made by the model by taking appropriate number of operational differences process.

Autoregressive Models AR(p):

Which has the general form:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + \varepsilon_t \quad , \quad t > p \qquad (1)$$

Where :

Z_t = dependent variable at time t.

$Z_{t-1}, Z_{t-2}, \dots, Z_{t-p}$ = dependent variable at time lags t − 1, t − 2, . . . , t − p respectively.

ϕ_i : Coefficients to be estimated, as:

$i = 1, 2, 3, \dots p \,, \quad -1 < \Phi_P < 1$

ε_t : Series of random errors, with mean zero and constant variance. $i.e. \quad \varepsilon_t \approx N(0, \sigma_\varepsilon^2)$.

Moving Average Model MA (q):

Which has the general form:

$$Z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - .... - \theta_q \varepsilon_{t-q} \qquad (2)$$

Where:

Z_t = dependent variable at time t.

ε_t : Series of random errors, with mean zero and constant variance. $i.e.\quad \varepsilon_t \approx N(0,\sigma_\varepsilon^2)$.

$\varepsilon_{t-1}\ ,\ \varepsilon_{t-1}\ ,\ \ldots\ \varepsilon_{t-q}$ = Errors in previous time periods that are incorporated in the response Z_t .

θ_i : Parameters of moving average model.

q : Degree Model.

Mixed Models ARMA (p , q):

Which has the general form:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

Autoregressive Integrated Moving Average Model ARIMA (p,d,q):

The general formula of Autoregressive Integrated Moving Average Models (ARIMA) That is written as :

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \ldots + d\, Z_{t-p-d} + \varepsilon_t - \qquad (4)$$

Building ARIMA Model and Forecasting:

The data used in this study consist of the monthly sales for Naphtha production in Azzawiya Oil Refining Company – Libya, for the period (2008 – 2014).

Stationary

In order to apply certain techniques for identifying and ARIMA model for the data, we must determine the form of stationary of the original time series and we will draw the autocorrelation function (ACF), and the partial autocorrelation function (PACF) of data and draw confidence interval of (ACF) and (PACF) to detect the stationary or non-stationary of time series, as well as the use of the Q test, (Box- Ljung test) where:

$$Q_m = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k} \sim \chi^2(m) \qquad (5)$$

Where:

$\hat{\rho}_k$ = the residual autocorrelation at lag k.

n = the number of residuals.

m = the number of time lags includes in the test.

And we will test of unit root by Augmented Dickey-Fuller (ADF) test on the original data provided evidence of the presence of a unit root we difference the data to counteract the effect of this unit root. We depended on Eviews.5 program to determine tables and figures we can obtain the following results:

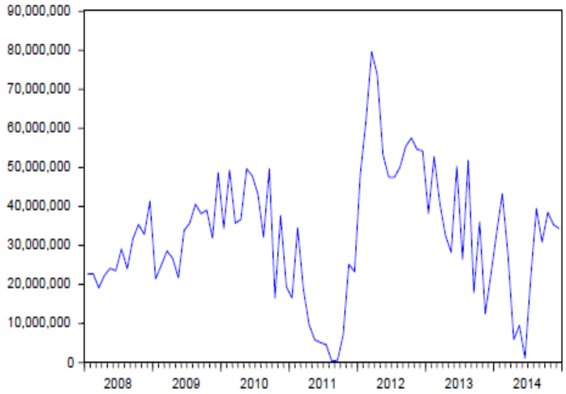

Figure (1): Graphical representation of the series Naphtha

Table (1): ACF and PACF of the series Naphtha

Date: 01/31/15 Time: 07:46
Sample: 2008M01 2014M12
Included observations: 84

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.694	0.694	41.961	0.000
		2	0.644	0.312	78.482	0.000
		3	0.411	-0.243	93.518	0.000
		4	0.317	-0.039	102.59	0.000
		5	0.217	0.073	106.92	0.000
		6	0.160	-0.001	109.29	0.000
		7	0.064	-0.132	109.68	0.000
		8	-0.022	-0.115	109.73	0.000
		9	-0.095	-0.015	110.59	0.000
		10	-0.220	-0.189	115.33	0.000
		11	-0.247	-0.026	121.38	0.000
		12	-0.290	0.029	129.82	0.000
		13	-0.367	-0.239	143.50	0.000
		14	-0.351	-0.001	156.22	0.000
		15	-0.405	-0.050	173.41	0.000
		16	-0.392	-0.093	189.75	0.000
		17	-0.355	0.044	203.32	0.000
		18	-0.357	-0.132	217.30	0.000
		19	-0.307	-0.025	227.79	0.000
		20	-0.287	-0.054	237.09	0.000
		21	-0.191	0.082	241.29	0.000
		22	-0.152	-0.007	243.99	0.000
		23	-0.072	-0.123	244.61	0.000
		24	-0.079	-0.083	245.36	0.000
		25	-0.071	-0.127	245.99	0.000

Through the Figure (1) and table (1) of the original series of Correlation Coefficients and figures of ACF and PACF, we note that there is non- stationary in the data of original series as there are some values outside the confidence interval, and that the significant value of the Coefficients autocorrelation function using the Q test :

$$Q = 245.99 > \chi^2_{25,0.05} = 37.652$$

And to make the series stationary we make differences from First level while taking the natural logarithm of the data and then we can obtain the following results:

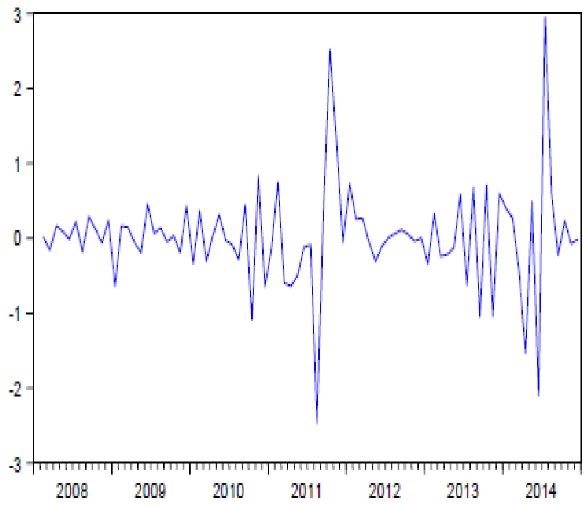

Figure (2): Graphical representation of the series D(Log(Naphtha))

Table (2): ACF and PACF of the series D (log(Naphtha))

Date: 01/31/15 Time: 08:16
Sample: 2008M01 2014M12
Included observations: 83

	AC	PAC	Q-Stat	Prob
1	-0.198	-0.198	3.3594	0.067
2	0.032	-0.008	3.4467	0.178
3	-0.125	-0.125	4.8192	0.186
4	-0.019	-0.071	4.8510	0.303
5	-0.009	-0.028	4.8579	0.433
6	-0.135	-0.169	6.5372	0.366
7	0.056	-0.021	6.8246	0.447
8	-0.074	-0.086	7.3358	0.501
9	0.088	0.014	8.0784	0.526
10	-0.103	-0.105	9.1062	0.522
11	0.080	0.013	9.7387	0.554
12	-0.066	-0.074	10.166	0.601
13	0.001	-0.052	10.166	0.680
14	0.014	-0.018	10.187	0.748
15	-0.062	-0.079	10.590	0.781
16	-0.043	-0.132	10.785	0.823
17	0.093	0.071	11.707	0.818
18	-0.096	-0.147	12.713	0.808
19	0.035	-0.038	12.847	0.846
20	-0.009	-0.041	12.856	0.883
21	0.017	-0.044	12.888	0.912
22	-0.046	-0.118	13.128	0.930
23	0.052	0.024	13.444	0.942
24	-0.034	-0.101	13.585	0.955
25	0.003	-0.053	13.586	0.968

Through the figure (2) and table (2) of modified series of correlation coefficients and figures of ACF and PACF we note that there is stationary in the data of series and most of the values within the confidence interval, and that the Significant value of autocorrelation coefficients by using the Q test was:

$$Q = 13.586 < \chi^2_{25,0.05} = 37.652$$

We then test by Advanced Dickey Fuller and the estimation of models is as follows:

Table (3): ADF test results for series D(Log(Naphtha))

ADF – Test	T -student	Test Critical Values 5%	Prob.
With constant	-10.9271	-2.8972	0.0001
With constant and trend	-10.8590	-3.4655	0.0000
Without constant and trend	-10.9945	-1.6142	0.0000

Through the data of the table (3) we note that all calculated statistics of Dickey Fuller in all models are less than the corresponding tabular values, then there is no a unit root in the series, and the series D(Log(Naphtha)) is stationary.

Model Identification:

Through the table (2), we observe the presence of moving averages models MA (1), MA (3) and MA (6), through the autocorrelation function. And the presence of autoregressive models AR (1), AR (3) and AR (6), through partial autocorrelation function, and for more accuracy in reconciling the best model among Box-Jenkins models, possible models of ARIMA (p, d, q) have been applied, and calculation of each of them SC, AIC and MSE are as shown in the following table:

Table (4): Compared to a set of valuesof AIC , SIC , MSE

Models	AR	MA	MSE	AIC	SIC
ARIMA (1,1,1)	1	1	0.3506	1.8629	1.9510
ARIMA (3,1,3)	3	3	0.4899	2.1993	2.2886
ARIMA (6,1,6)	6	6	0.4788	2.1793	2.2706

From the table (4), we find that the model ARIMA (1,1,1) is the one who gives the lower values of the previous standards, these values were as follows:

MSE = 0.3506 , AIC=1.8629 , SC=1.9510

So this model was relied on to be an appropriate model for this series.

Model Diagnostics:

Residual diagnostic tests and the over fitting process are used here to determine the goodness of fit of the ARIMA(1,1,1) model to the original time series.

➢ *Residual Diagnostics:*

We have been used diagnostic the ACF for the residuals. Here we see that the ACF values are all within the 95%, indicating that there is no correlation amongst the residuals. This plot is used as an indicator of the independence of the residual.

Table (5): ACF and PACF for Residual ARIMA (1,1,1)

Date: 01/31/15 Time: 11:20
Sample: 2008M03 2014M12
Included observations: 82
Q-statistic probabilities adjusted for 2 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.002	-0.002	0.0002	
		2	0.125	0.125	1.3371	
		3	-0.046	-0.047	1.5253	0.217
		4	0.006	-0.009	1.5289	0.466
		5	-0.010	0.002	1.5372	0.674
		6	-0.129	-0.133	3.0451	0.550
		7	0.019	0.022	3.0795	0.688
		8	-0.088	-0.058	3.7976	0.704
		9	0.040	0.024	3.9475	0.786
		10	-0.118	-0.102	5.2907	0.726
		11	0.025	0.012	5.3535	0.802
		12	-0.097	-0.088	6.2706	0.792
		13	-0.045	-0.057	6.4772	0.840
		14	-0.037	-0.033	6.6135	0.882
		15	-0.100	-0.095	7.6504	0.866
		16	-0.080	-0.118	8.3185	0.872
		17	0.035	0.066	8.4515	0.904
		18	-0.114	-0.161	9.8615	0.874
		19	-0.004	-0.030	9.8632	0.909
		20	-0.034	-0.046	9.9895	0.932
		21	-0.010	-0.058	10.001	0.953
		22	-0.056	-0.117	10.366	0.961
		23	0.026	0.017	10.444	0.973
		24	-0.041	-0.119	10.640	0.980
		25	-0.008	-0.059	10.648	0.987

From the table (5) we note that the most of the coefficients fall within the confidence interval, as well as the statistical (Q):

$$Q = 10 < \chi^2_{25,0.05} = 37.652$$

Therefore Residuals represent white noise.

Normal distribution test:

The residuals follow the normal distribution in the following diagram:

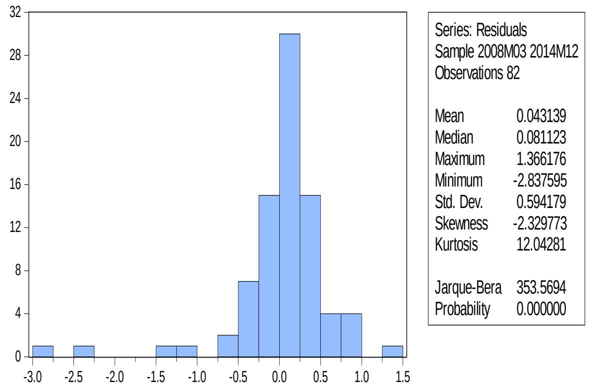

Figure (3): Testing of normal distribution of residuals series

The distribution of residuals are almost symmetrical.

Autocorrelation test for errors (Durbin-Watson's statistic):

The value of D-W = 1.9924, it is lies within the confidence interval, hence it has no Auto correlation of errors. *Then all the results of the residuals tests confirm the validity of the estimated model ARIMA (1,1,1) to represent the time series.*

Final Model

Therefore conclude that the ARIMA(1,1,1) model is the best ARIMA model for the original time series being analyzed (Naphtha product). The final model is of the following form:

Table (6): Estimated model parameters of Naphtha sales model

Type	Coef.	SE Coef.	T – Stat.	Prob.
AR (1)	0.6010	0.0882	6.8173	0.0000
MA (1)	– 1.1713	0.0718	– 16.3106	0.0000

We obtained the model in the form:

$$\hat{Z}_t = 0.6010\, Z_{t-1} + 1.1713\, \varepsilon_{t-1} + \varepsilon_t \quad (6)$$

Forecasting

After the identification of the model and its adequacy check, it is used to forecast the sales for Naphtha product in Azzawiya Oil Refining Company – Libya in the period (January 2015 to December 2020). The forecasting results are presented in figure (4)

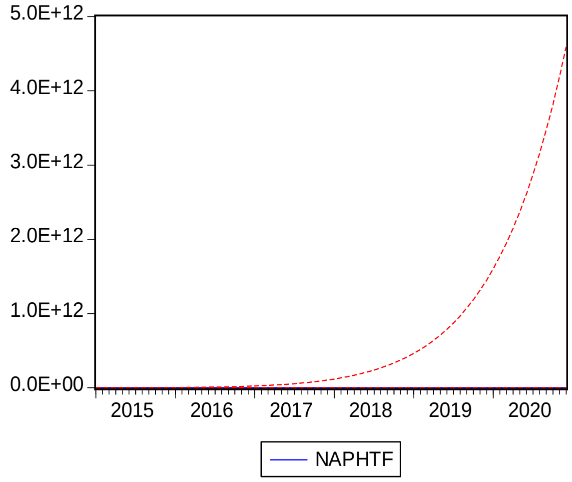

Figure (4): Forecasted Naphtha Sales

CONCLUSION

The aim of this analysis was to determine an appropriate ARIMA model for the sales Monthly data for Naphtha product. In particular we were interested in forecasting future for sales of company using this model. It is concluded that the ARIMA (1,1,1), where were the value of MSE, AIC and SC for the model are very small. Hence the model is appropriate to forecast the sales of Naphtha for the next six years.

REFERENCES

1. **David A. Bessler. (1982).** *Adaptive Expectations. The Exponentially Weighted Forecast and Optimal Statistical Predictors: A Revist.* Agricultural Economics Research., 34 (2), pp, 16-23.
2. **Abraham, B. and Ledoter. J, (1983).** "Statistcal Methods for Forecasting" , John Wiley, New York.

3. **Wiliam N. Busard, (1985).** *Comparison of Alternative Price Forecasting Models for Slaughter Hogs,* Ph.D. Thesis, University of Tennessee, Cited in Dissertation Abstracts International, 46(3): 7.
4. **Peter J. Brockwell & Richar A. Davis, (1987).** *Time Series: Theory and methods.* Berlin Heidelberg New York.
5. **Aidan Meyler, Geoff Kenny, and Terry Quinn (1998).** *Forecasting Irish Inflation using ARIMA models.*
6. **Chris Chatfield, (2000).** *Time – series forecasting,* [e-book] Wallingford: CRC Publishers. Available at : Google Books < www.crcpress.com > [Chapman and& hall 2000].
7. **Hamad Al-Ghannam (2003).** Time Series Analysis of Saudi Arabian Stock Market Index: *(Box-Jenkins Method),* King Saud University, Riyadh, 17(2), pp, 3-26.
8. **Rangsan Nochai and Titida Nochai (2006).** *ARIMA Model for forecasting Oil Palm Price Price. 2nd* IMT-GT Regioal Confrence on University Sains Malaysia Penang June 13-15, 2006.
9. **Azhar Salman, and Narjis Hadi. (2010).** *Prohibition of brick production in Iraq.* Al-Mansoor Journal, (14), Special Issue, (Part One).
10. **Sadia A. Tama. (2012).** Using Analysis of Time Series to Forecast numbers of the patients with Malignant Tumors in Anbar Province. Anbar University, 4: (8), pp, 371-393.
11. **Hala F. Hussain and Hosiba K. Muthnna (2013).** Annahereen University, 16(3), pp 48-61.
12. **Nor Hamizah Miswan, Pung Yean Ping, and Maizah Hura Ahmad (2013).** *On Parameter Estimation for Malaysian Gold Prices Modeling and Forecasting.,* University Technology Malaysia, Int. Journal of Math. Analysis, 7(22), pp, 1059 – 1068.