



Using Box-Jenkins Models Forecasting Oil production in Saudi Arabia from 1962-2014

Abuzar Yousef Ali Ahmed

Department of Mathematics Faculty of Science, Jeddah University, Saudi Arabia,
abuzarjeha@gmail.com

Abstract:

Energy is one of the pillars of the economies of consuming and producing countries alike, making the oil market unaffected

By market laws, they are governed by a set of conflicting policies and strategies between the interests of industrialized countries

And between the oil-producing countries as well as between the bodies and organizations representing each party, especially on the one hand determining prices.

And what we are witnessing today in the changes in oil production, which in turn affect the movement of exports in general

And the revenues of countries in particular, know that the mechanisms of development of energy production levels help in making appropriate decisions

Our study aims at proposing a suitable statistical model for the production of petroleum in Saudi Arabia and forecasting it in Short-term.

1. Introduction:

Saudi Arabia is a major oil producer around the world, and the history of oil production in Saudi Arabia began in 1933 when King Abdulaziz Al Saud, the founder of modern Saudi Arabia, awarded Standard Oil of Sucall a concession to explore oil in the Kingdom.

After five years of exploration, the oil was discovered in commercial quantities in March 1938 in the well of No. 7, which was a turning point in the history of the Kingdom.

Saudi Arabia's oil production has grown from 1.64 million bpd in 1962 to 9.71 million bpd in 2014.



Saudi Arabia's largest daily oil production in 1980 was 9.9 million barrels and the lowest in 1962 was 1.64 million bpd

The objective of this paper is to forecast Oil production in Saudi Arabia from 1962-2014 by using box-Jenkins and ARIMA method. Data were obtained from According to official data from the Saudi Monetary Agency The rest of this research is organized as follows. Section (2) briefly reviews the literature, while section (3) discuss the data and methodology. The results and conclusion remarks are given in section (4) and (5) respectively.

2. Literature Review:

Introduction to Time Series Data

A great deal of information relevant to public health professionals takes the form of time series. Time series are simply defined as a sequence of observations measured at regular time intervals. For example, daily blood pressure measurements taken on a single individual are a time series, as are daily counts of emergency room visits for asthma.

Researchers might be interested in asking several different questions about time series data. These questions include:

- Can patterns identified in past observations of a single time series be used to predict its future values?
- How do the values of a single time series compare before and after an intervention?
- Are the values of one time series associated with the values of another time series (e.g., daily particulate matter measurements and daily mortality)?

In order to answer these questions, we must first note that the structure of time series data presents a unique challenge for researchers, such that traditional regression approaches do not yield valid results. Uncorrelated residuals are a key assumption of many regression methods. However, in a single time series, we find that observations that are close together in time tend to be more similar to each other than those that are farther away in time, leading to correlated residuals. This phenomenon is called "autocorrelation." Models that fail to account for autocorrelation will have correct parameter estimates, but incorrect standard errors.



Introduction to ARIMA Models

One type of model that does account for autocorrelation is the Autoregressive Integrated Moving Average (ARIMA) model, which is fit using a methodology developed by George Box and Gwilym Jenkins (1970). The application of ARIMA models in health sector is varied, however, it has been used extensively for (i) outbreak detection in the arena of infectious diseases and in (ii) the evaluation of population level health interventions in the format of interrupted time series analysis. Both of these methods require the formal characterization of the inherent pattern in a time series, and using this pattern to forecast future behavior of the time series. For outbreak detection, we forecast the 95% confidence interval for a time series, and deviation of the actual time series values from within 95% CI bounds would constitute a signal. In the interrupted time series, the time series is forecasted into the future, and deviations of actual values from the forecasted values is considered to be a causal effect of public health intervention.

Note:

- ARIMA models do NOT predict rare “black swan” events, as there is no pattern in the time series to suggest a future event of this type.
- The causal framework for ARIMA model differs slightly from Epidemiology frame, and is more consistent with the Granger definition of a cause from economics.

Data Requirements

The data requirements to fit an ARIMA model are:

- A univariate time series (count or continuous) with at least 50-100 observations
- If the time series consists of count data, the interval over which the count is taken must remain the same over time
- If the time series consists of continuous data, the interval between measurements must remain the same over time
- Data must be presented in a vertical vector (column of data)

Components and Fitting of ARIMA Models

Overview:

The ARIMA model divides the pattern of a time series into three components: the autoregressive component, p , which describes how observations are related to each other as the result of being close together in time; the differencing component, d ,



which is used to make a time series stationary (see below); and the moving average component, q , which describes outside “shocks” to the system.

Stationarity Assumption:

A key requirement of ARIMA models is that the data set of interest is stationary, meaning that it has a constant mean and variance over time. If a data set is not stationary to begin with, stationarity can be achieved by a process called “differencing,” which is represented by the “ d ” component of the model.

Identification:

The identification steps involve fitting the autoregressive component (variable “ p ”), the moving average component of the ARIMA model (variable “ q ”), as well any required differencing to make the time series stationary or to remove seasonal effects (variable “ d ”). Together, these user-specified parameters are called the order of ARIMA. The formal specification of the model will be ARIMA (p,d,q) when the model is reported.

The first step in model identification is to ensure the process is stationary. Stationarity can be checked with a Dickey-Fuller Test. Any non-significant value under model assumptions suggests the process is non-stationary. The process must be converted to a stationary process to proceed, and this is accomplished by the differencing the time series using a lag in the variable as well as removing any seasonality effects. The lagged values used to difference the time series will constitute the “ d ” order.

Ex. An additive difference of 1 and seasonal difference of 12 is reported as $d=(1,12)$

Once the process is stationary, we fit the autoregressive and moving average components. To fit the model we use the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) in addition to various model fitting tools provided by software. There are various sets of rules to guide p and q fitting in lower order processes, but generally we let the statistical software fit up to 12-14 orders for AR and MA, and suggest combinations that minimize an AIC or BIC criterion. This part is as much an artform as it is a structured process. The goal during this phase is to minimize the AIC/BIC criterion.

Estimation:

The estimation procedure involves using the model with p , d and q orders to fit the actual time series. We allow the software to fit the historical time series, while the user checks that there is no significant signal from the errors using an ACF for the error residuals, and that estimated parameters for the autoregressive or moving average components are significant.



Forecasting:

After a model is assured to be stationary, and fitted such that there is no information in the residuals, we can proceed to forecasting. Forecasting assesses the performance of the model against real data. There is an option to split the time series into two parts, using the first part to fit the model and the second half to check model performance. Usually the utility of a specific model or the utility of several classes of models to fit actual data can be assessed by minimizing a value such as root mean square.

ARIMA (p,d,q) modeling To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q

- Ensuring Stationarity
- Determine the appropriate values of d
- Identification: Determine the appropriate values of p & q using the ACF, PACF, and unit root tests • p is the AR order, d is the integration order, q is the MA order
- Estimation : Estimate an ARIMA model using values of p, d, & q you think are appropriate.
- Diagnostic checking: Check residuals of estimated ARIMA model(s) to see if they are white noise; pick best model with well behaved residuals.
- Forecasting: Produce out of sample forecasts or set aside last few data points for in-sample forecasting.

3. Data and Methodology:

The data of the study obtained from According to official data from the Saudi Monetary Agency, consists of annual data Oil production in Saudi Arabia from 1962-2014. We use ARIMA model for forecast one period a head of the series by applying Box-Jenkins approach. An ARIMA is a generalization of an ARIMA model. The model is generally referred to as ARIMA (p, d, q) model, where p, d and q are integers greater than or equal zero and refer to the order of autoregressive integrated and moving average aspects.

The Box-ARIMA model is a combination of the AR (Autoregressive) and MA (moving average) model as follows:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p} - \alpha_1 U_{t-1} - \alpha_2 U_{t-2} - \dots - \alpha_q U_{t-q} + U_t$$

The Box-Jenkins methodology is a five step process for identifying, selective and Assessing conditional means models.

4. Results and Discussion:



The ARIMA methodology is summarized in four stages in which the most appropriate model is selected for evaluation in the time series model, the stages are:

Phase I: The first stage: stability analysis: It checks the stability of the time series, and if it is unstable, conversions are applied. So it is necessary to make it stable.

Phase II: Appreciation ARIMA Model: C5 estimates at each iteration

Iteration	SSE	Parameters			
0	2.09464	0.100	0.100	0.134	
1	1.35592	-0.005	-0.050	0.105	
2	1.09764	-0.079	-0.200	0.074	
3	1.02840	-0.132	-0.315	0.041	
4	1.02702	-0.138	-0.317	0.034	
5	1.02701	-0.139	-0.317	0.034	
6	1.02701	-0.139	-0.317	0.034	

Relative change in each estimate less than 0.0010

During which the parameters of the standard model are estimated.

Phase III: Personal examination. Unable to reduce sum of squares any further

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	-0.1388	0.1357	-1.02	0.311
MA 2	-0.3168	0.1371	-2.31	0.025
Constant	0.03357	0.02922	1.15	0.256
Mean	0.03357	0.02922		

The appropriate ARIMA (1,1,2)

Differencing: 2 regular differences:

Number of observations: 52

Residuals: SS = 1.02694 (backforecasts excluded)

MS = 0.02096 DF = 49

Note that the total square errors are very small and that's what we want in the model

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6.8	13.2	20.1	21.2
DF	9	21	33	45
P-Value	0.654	0.901	0.962	0.999

Note that random errors are worth more than 5% of 0.962 indicating that the model is reliable



www.mecsaj.com

During which the form is checked to make sure it is relevant to the designated time series and when not appropriate, go back to the second stage, otherwise we will proceed to the fourth stage.

Phase IV: Prediction:

- If the hypothesis of the model is validated and the latter is statistically acceptable, the observable phenomenon can be predicted.
- Based on the proposed model, the calculated projections are short-term projections and are not valid for long periods.
- This study specifically examined the use of ARIMA prediction and
- tested its ability to establish accurate baseline. In particular, as retrospective.
- Analysis of secondary data using oil production data in the Kingdom. This study examined the use of the ARIMA model.
- Models and their ability to establish accurate basic lines for use in oil production data in the Kingdom.

To accomplish this task, this study presented one research question: Can ARIMA models expect oil production data in the Kingdom?

Conclusion Remarks:

Through the implementation of the oil production chain in Saudi Arabia, the following conclusions are drawn:

- The oil production in Saudi Arabia has an increasing general trend, which means that it is not fixed. We create the natural logarithm and convert it to a stable time series.
- Note that the time series is unstable. We create the first difference and turn it into a stable time series and note that the chain becomes stable.
- We find the function of automatic link and automatic partial correlation.
- From the automatic correlation function we note that the coefficient of MA = 1.
- From the automatic partial correlation function we observe that the coefficient of AR = 2 and the degree of difference = 1
- Check the properties of the residual material in the ARIMA model (1, 1, 2), P-Value greater than 0.05 and exceed the testing and diagnostic stage.
- The ideal model for predicting oil production in Saudi Arabia is ARIMA (1, 1,2).
- There is a convergence between predictive values and actual values during the period (1962-2014).
- There is an increase in oil production in Saudi Arabia (2017-2026).

Reference:



www.mecsaj.com

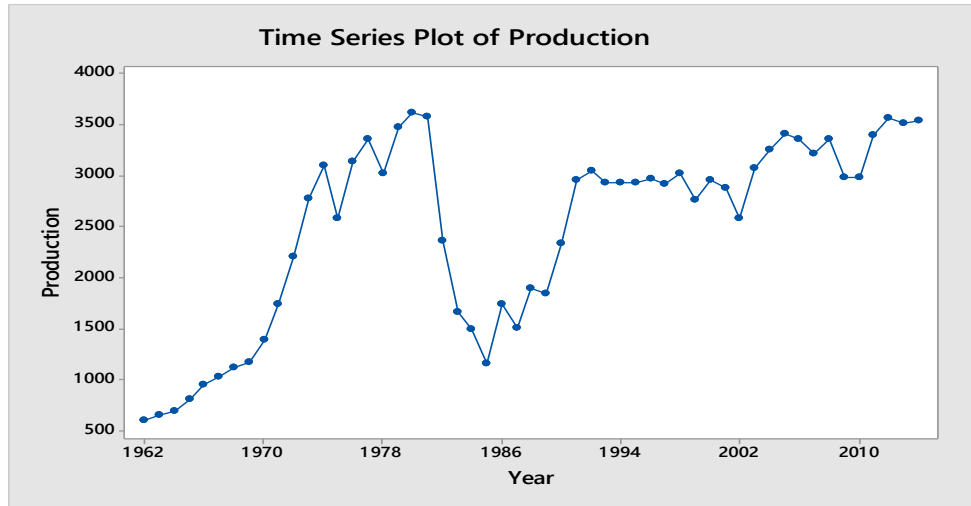
- 1- Abdalla, M. M. (2010). Inflation Determinants in Sudan 1970-2009. Central Bank of Sudan Publications, Khartoum, Sudan.
- 2- Abdulrahman, B. M. A. (2014). Inflation and Economic Performance in Sudan: An Analysis Study. Researchjournali's Journal of Economics, 2(3), 1-8.
- 3- Abdulrahman, B. M. A. *et al.* (2016). The Relationship between Unemployment and Inflation in Sudan: An Empirical Analysis, 1992-2015. Research in Economics and Management, 1(2), 113-122.
- 4- Ascari, G and Sbordone, A. M. (2014). The Macroeconomics of trend inflation. Journal of Economic Literature, 52(3), 679-739.
- 5- Caprio, G. et al. (2005). Financial Crises: Lesson from the Past, Preparation for the Future. Brooking Institute Press, Washington DC, 2005, P 20.
- 6- Mankiw, N. G. (2015). Principles of Economics. Cengage Learning, USA, Seventh Edition, 494-496.
- 7- Mellor, M. (2010). The Future of Money: From Financial Crisis to Public Resources. Pluto Press, London, 53-54.
- 8- Schofield, N. C and T. Bowler. (2011). Trading the Fixed Income, Inflation and Credit Markets. Willey, a John Willey& Sons, Ltd., Publications, 2-5.

Appendices:

First: We draw the time series to know the components and general direction

Figure (1)

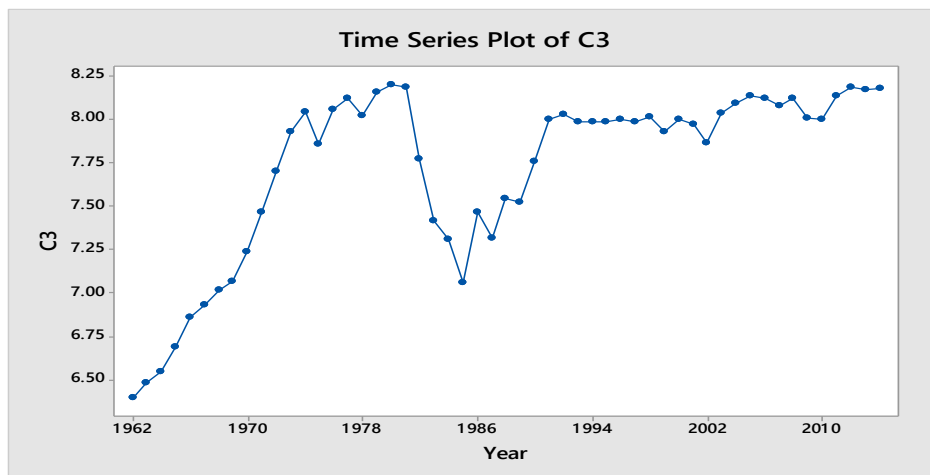
Draw the time series



From the graph we notice that the series is unstable in the medium and variance and has a general tendency to rise, and for its stability we take the natural logarithm.

Figure (2)

Time series after taking the natural logarithm

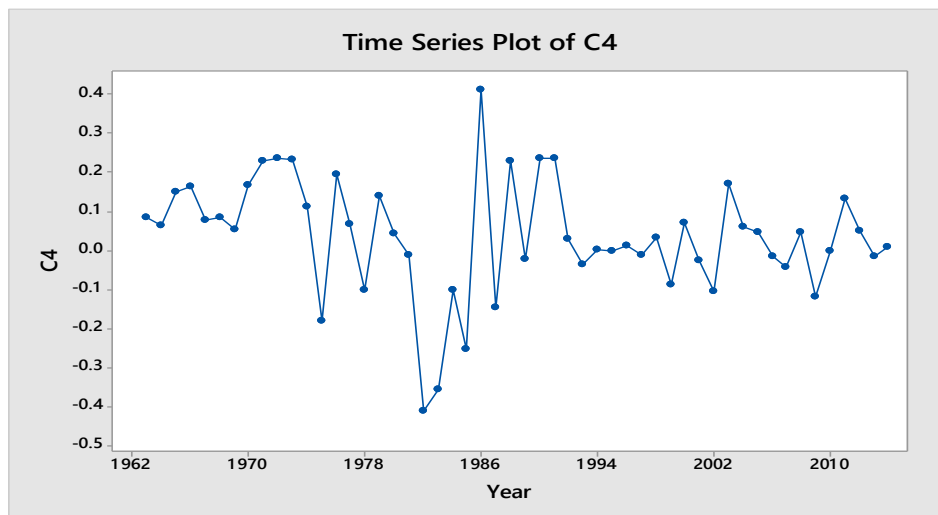




Note the discontinuity after taking the natural logarithm we turn to the use of differences.

Figure (3)

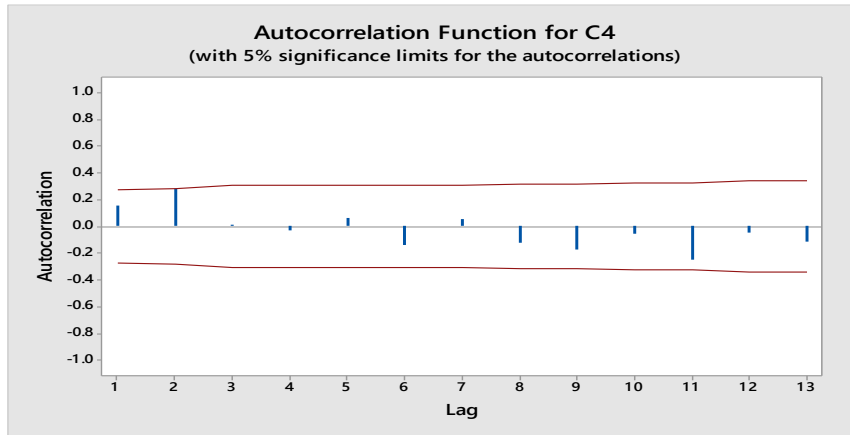
Series after taking the natural logarithm and taking the differences



Note that the series has become stable and rises and falls around the average and did not take a general direction and is valid for prediction and estimation, we draw the function of self-correlation and partial self-correlation.

Figure (4)

Self-correlation function



Note that errors and spikes have not passed the upper and lower limits of the data demonstrating the validity of the model and that it is predictable.

Figure (5)

Partial self-linking function

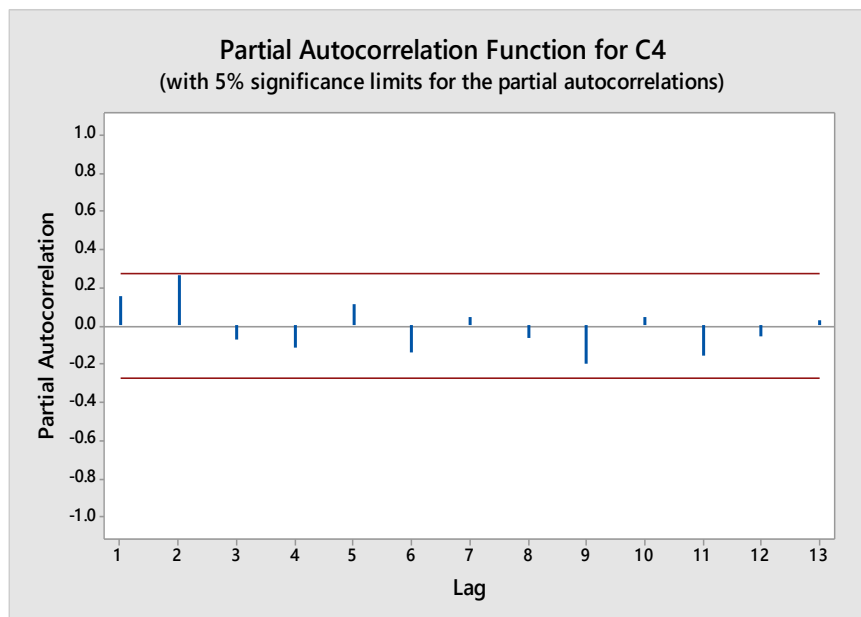
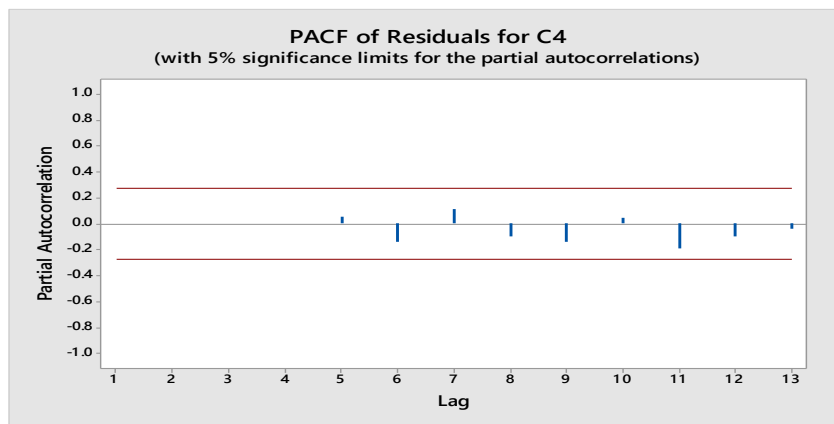




Figure (6)

Self-correlation function for errors and condoms



Note that errors and spikes have not passed the upper and lower limits of the data demonstrating the validity of the model

Figure (7)

Partial self-correlation function for errors and condoms

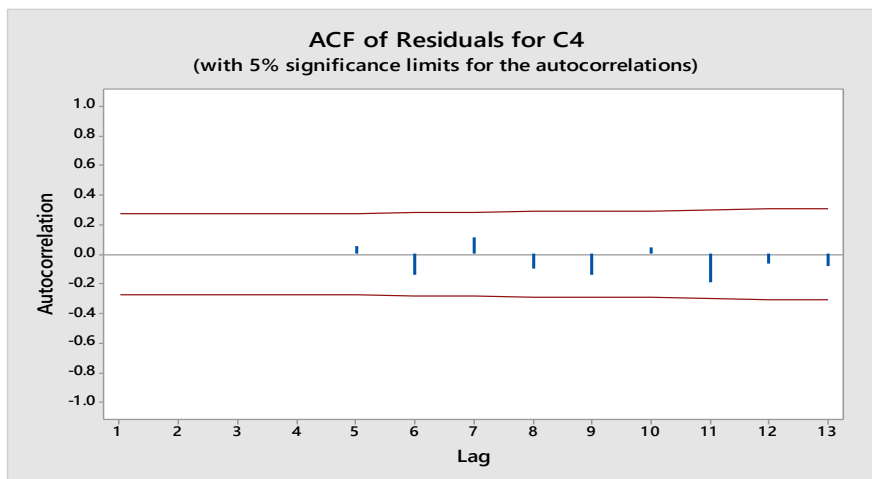
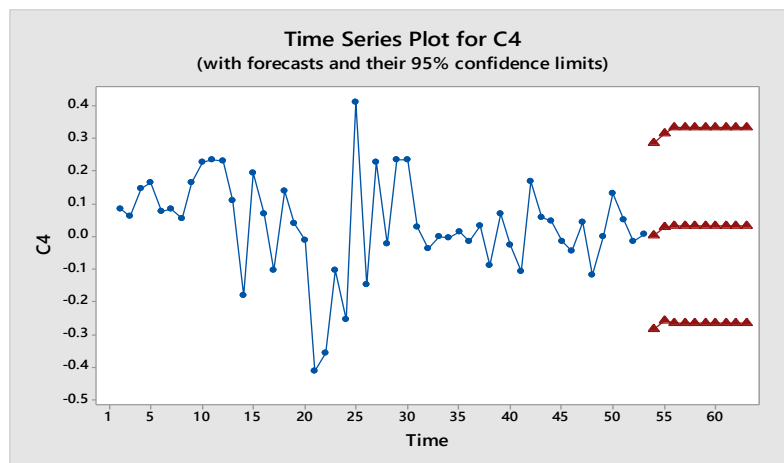




Figure (8)

Forecast for 10 years and minimum and upper limits for the period of confidence



Time Series Plot of Production

Autocorrelation Function

Lag	ACF	T	LBQ
1	0.154352	1.11	1.31
2	0.283887	2.00	5.84
3	0.009625	0.06	5.84
4	-0.031852	-0.21	5.90
5	0.065545	0.43	6.16
6	-0.141218	-0.92	7.38
7	0.053001	0.34	7.55



8	-0.126687	-0.81	8.58
9	-0.169247	-1.07	10.45
10	-0.052243	-0.32	10.63
11	-0.253315	-1.57	15.02
12	-0.049278	-0.29	15.19
13	-0.117579	-0.69	16.19

Autocorrelation Partial Autocorrelation Function

Lag	PACF	T
1	0.154352	1.11
2	0.266410	1.92
3	-0.070175	-0.51
4	-0.111008	-0.80
5	0.111593	0.80
6	-0.137668	-0.99
7	0.042813	0.31
8	-0.066346	-0.48
9	-0.199456	-1.44
10	0.043226	0.31
11	-0.158598	-1.14
12	-0.055441	-0.40
13	0.029607	0.21

Partial Autocorrelation

ARIMA Model

Estimates at each iteration

Iteration	SSE	Parameters		
0	2.09464	0.100	0.100	0.134
1	1.35592	-0.005	-0.050	0.105
2	1.09764	-0.079	-0.200	0.074
3	1.02840	-0.132	-0.315	0.041
4	1.02702	-0.138	-0.317	0.034
5	1.02701	-0.139	-0.317	0.034
6	1.02701	-0.139	-0.317	0.034

Relative change in each estimate less than 0.0010

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	-0.1388	0.1357	-1.02	0.311
MA 2	-0.3168	0.1371	-2.31	0.025
Constant	0.03357	0.02922	1.15	0.256



Mean 0.03357 0.02922

The appropriate ARIMA (1,1,2)

Number of observations: 52
Residuals: SS = 1.02694 (backforecasts excluded)
 MS = 0.02096 DF = 49

Note that the total square errors are very small and that's what we want in the model

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6.8	13.2	20.1	21.2
DF	9	21	33	45
P-Value	0.654	0.901	0.962	0.999

Note that random errors are worth more than 5% of 0.962 indicating that the model is reliable

Forecasts from period 53

Period	Forecast	95% Limits		Actual
		Lower	Upper	
54	0.001079	-0.282724	0.284883	
55	0.029115	-0.257409	0.315640	
56	0.033568	-0.266734	0.333870	
57	0.033568	-0.266734	0.333870	
58	0.033568	-0.266734	0.333870	
59	0.033568	-0.266734	0.333870	
60	0.033568	-0.266734	0.333870	
61	0.033568	-0.266734	0.333870	
62	0.033568	-0.266734	0.333870	
63	0.033568	-0.266734	0.333870	

إنتاج السعودية من النفط خلال الـ 53 سنة الماضية			
متوسط الإنتاج اليومي (مليون برميل)	التغير السنوي %	المجموع (مليون برميل)	السنة
1.64	--	599.76	1962
1.79	+ 9 %	651.71	1963
1.90	+ 7 %	694.13	1964
2.21	+ 16 %	804.94	1965
2.60	+ 18 %	948.57	1966
2.81	+ 8 %	1023.84	1967
3.04	+ 9 %	1113.71	1968
3.22	+ 5 %	1173.89	1969
3.80	+ 18 %	1386.67	1970
4.77	+ 26 %	1740.68	1971
6.02	+ 27 %	2201.96	1972
7.60	+ 26 %	2772.61	1973
8.48	+ 12 %	3095.09	1974
7.08	(% 17)	2582.53	1975
8.58	+ 22 %	3139.28	1976
9.20	+ 7 %	3357.96	1977
8.30	(% 10)	3029.90	1978
9.53	+ 15 %	3479.15	1979
9.90	+ 4 %	3623.80	1980
9.81	(% 1)	3579.89	1981
6.48	(% 34)	2366.41	1982
4.54	(% 30)	1656.88	1983
4.08	(% 10)	1492.90	1984
3.17	(% 22)	1158.80	1985
4.78	+ 51 %	1746.20	1986
4.12	(% 14)	1505.40	1987
5.16	+ 26 %	1890.10	1988



www.mecsjs.com

5.06	(% 2)	1848.50	1989
6.41	% 27 +	2340.50	1990
8.12	% 27 +	2963.00	1991
8.33	% 3 +	3049.40	1992
8.05	(% 4)	2937.40	1993
8.05	--	2937.90	1994
8.02	--	2928.54	1995
8.10	% 1 +	2965.45	1996
8.01	(% 1)	2924.28	1997
8.28	% 3 +	3022.27	1998
7.56	(% 9)	2761.10	1999
8.09	% 7 +	2962.60	2000
7.89	(% 3)	2879.46	2001
7.09	(% 10)	2588.98	2002
8.41	% 19 +	3069.74	2003
8.90	% 6 +	3256.30	2004
9.35	% 5 +	3413.94	2005
9.21	(% 2)	3360.90	2006
8.82	(% 4)	3217.77	2007
9.20	% 5 +	3366.34	2008
8.18	(% 11)	2987.27	2009
8.17	--	2980.43	2010
9.31	% 14 +	3398.52	2011
9.76	% 5 +	3573.40	2012
9.64	(% 2)	3517.62	2013
9.71	% 1 +	3545.14	2014



www.mecsaj.com

*وفقا للبيانات الرسمية الصادرة عن مؤسسة النقد السعودي