International Journal of Advanced Statistics and Probability, 6 (1) (2018) 24-29

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v6i1.8751
Research paper

Using analysis of time series to forecast the number of patients with tuberculosis: a case study in Khartoum state from 2007 to 2016

Abu Elgasim Abbas Abow Mohammed [1, 2] *

[1]College of Business and Economics, Qassim University, Kingdom of Saudi Arabia
[2]College of Economic and Political Science Omdurman Islamic university, Sudan
**Corresponding author E-mail: gasimtas@gmail.com*

Abstract

This paper used time series analysis to predict the number of tuberculosis (TB) patients in Khartoum state. It is based on data obtained from TB patients the period from 2007 to 2016. The study was able to determine the best model of order (2) ARIMA (2, 1, 0) for data. The most important result of the study is to estimate the number of patients with TB the next four years in quartile basis. So, the forecasting value represented the source time series data and was observed to decrease.

Keywords: *Time Series; Tuberculosis; Estimation; Model; Forecasting.*

1. Introduction

Time series of important studies that deal with the behavior and interpretation of phenomena over time is the importance of the observation in obtaining a precise description of the time series of TB data and building an appropriate model for predicting the number of TB patients in the period (2007-2016) in Khartoum state and using the results in the future. That helps officials to develop plans and policies in the face of endemic illnesses, and a case study of Khartoum state is very representative of the rest of the country. Tuberculosis has an impact on human resources, thus affecting production, especially when the disease rates increase, and low standards of living also in turn lead to aggravating this rate .This study is an attempt to see the increase in the number of patients in the future.

2. Problem statement

This paper is trying to determine a suitable model that leads to estimates that can be used to know the number of TB cases in Khartoum state, as well as the possibility of applying the method of Box and Jenkins in data analysis.TB is endemic in wide areas in Sudan; it is one of the diseases linked to malnutrition unhygienic health condition. The researcher chose Khartoum State because it is the capital and stationed more people out as well as specialized hospital to treat the disease and to the many of injured comes with the disease the rest of other state in addition to existing issues in Khartoum .The researcher summarized the problem in the following question:
Is the number of TB patients increasing significantly?
To what extent this can increase in the future?

3. Objectives of paper

The study aims to forecast the number of TB patients in Khartoum State.

4. Importance's of paper

The impotence of this study stems from the use of (TB) data and hence the estimation of the number of patients in the future, benefiting patients and specialists like.

5. Limitations of study

The limitations of the study are the tuberculosis data are for the Khartoum State only in the period 2007 to 2016

6. Study hypotheses

The study looked in to the hypotheses that follow:

i). The series of patients with tuberculosis is non-stationary
ii). The number of patients with tuberculosis is increase with time

7. Methodology

The study was based on the theoretical approach that dealt with the method of Box-Jenkins (diagnosis, estimation, model suitability, perdition), and supported the practical side that depends on the data of TB from the state of Khartoum. The paper used SPSS and program of EVIEWS.

8. Literature review

The researchers study Naomi and El Sharout (2000) that relate to regarding the pest model to forecasting the number of people with malignant tumors in Gadisiyah Province using analysis intervention time series impact factor of economic blockade divided in two parts 1990-1993 and 1994-1997 .It was noted with tumors a series is non-stationary in the mean and variance and there is a trend after 1993.This shows that detrimental the impact of the economic blockade factor to increase the number of injured.

Jobbery (2009) in the time series binary variables predicts the rate of inflation related the rate of exchange of the dinar against US dollar starting in January (2004) to December (2008). The study concluded that the vector inflation direction of the exchange rate follows autoregressive binary variables model second order VAR1MA (2, 1, 0).

Tomah [1] using BOX and Jenkins Method (Identification, Diagnostics Checking of model, forecasting) to find the best forecasting model to the number of patient with malignant tumors in An bar Province by using the monthly data for the period (2006-2010). The result of data analysis shows that the proper and suitable model is integrated Auto regressive model of order (2) ARIMA (2, 1, 0).

9. Theoretical formulation

9.1. Time series

Time series is a collection of observation generated sequentially through time. The special features of a time series are that the data are ordered with respect to time and that successive observation is usually expected to be dependent. Indeed, it is this dependence from one time period to another which will be exploited in making reliable forecasts .The order of an observation is denoted by subscript t. Therefore, we denote by z_t the observation of tth time series .The preceding observation is denoted by z_{t-1} and the next observation as z_{t+1}, It also will be useful to distinguish between a time series process and a time series realization. The observed time series is an actual realization of underlying time series process. By a realization we mean a sequence of observation data point, and not just a single observation. The objective of time series analysis is to describe succinctly this theoretical process in the form of observation model that has similar properties to those of the process itself.

9.2. Stationary

We indicated that it is not advisable to estimate the mean for each time period on just one realization of general stochastic process. But if there is no trend in the series, we might be willing to assume that the mean is constant for each time period and the observed value at each time period is representative of that mean, we must restrict the mean of the series to be constant. Such an assumption could be quite plausible. This assumption is just one of the conditions for stationary. The Second condition for stationary is that the variance of the process be constant B-J [2].

9.3. Autocorrelation function of stationary series

Once it has been reasonably ascertained that the series is stationary, the next step is identifying a model to determine what type of autoregressive or moving average or mixed model adequately fits the data. The actual structure of the ARIMA (p, d, q) model is obtained by comparing the sample acf of stationary series with theoretical population acf' s .We should keep in mind that there is no rule that stipulates that after differencing a series consecutively or seasonally to induce stationary. We must now include no seasonal or seasonal parameters. Indeed, it is quite possible that after appropriate stationary transformations have been made, the new series is simply a white noise series and there is no need to include any parameter whatsoever. This is specifically the case for a random walk series. Only if autocorrelations are large at lags corresponding to the span and possibly multiplies thereof should we include seasonal parameters Anderson [3].

9.4. Partial autocorrelation function of ARIMA model

Partial autocorrelation faction (pacf) is one more characteristic of ARIMA model which will help us to distinguish one model from another. Any ARIMA (p, d, q) model, as we have seen can always be expressed as a pure autoregressive model. These autoregressive models process acf, which, although they die out quickly, could stretch out to infinity. The Partial autocorrelation constitute a device for summarizing all the information contained in the acf of an AR process in a small number of nonzero statistics. For an AR (p) process, only p such nonzero statistics are necessary, rather than an infinite number of nonzero autocorrelation Anderson [3].

9.5. Q Statistics

Rather than considering each autocorrelation individually, quite often we might want to see if, as a group of autocorrelation, show evidence of model inadequacies Box and Pierce [4] show that for a purely random process, that is, a model with all $P_k = 0$, the statistic.

$$Q_{(k)} = n(n+2)\sum_{k=1}^{n}\frac{r_k^2}{n-k} \quad (1)$$

Therefore $Q_{(k)}$ test statistic, n observation number, r_i Autocorrelation for residual and Q_k is distributed approximately as a χ^2 (chi-square) distribution with k degrees of freedom $\chi^2_{(k)}$. This test using Q Statistic is sometimes called the portmanteau test, if the computed value of Q is less than table value of the χ^2 statistic with degrees of freedom, given a prespecified significance level, the group of autocorrelations used to calculate the test can be assumed to be not different from 0. This indicates that the data generating the autocorrelations are. If the computed value of Q statistic is larger than the χ^2 value from a χ^2 table, the autocorrelations are significantly from 0, indicating the existence of some pattern.

9.6. Box and Jenkins models

9.6.1. Autoregressive model

Autoregressive process if a current value of time series can be expressed as a linear function of previous value of the series $z_{t-1}, z_{t-2}, ..., z_{t-p}$ and a random a_t shock formed AR (P) we can express this relationship model autoregressive with rank p as follows(jobbery 2010):

$$z_t = \phi_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} + a_t \quad (2)$$

9.6.2. Moving average model

Where is the error a_{t-1} at period $t-1$ and θ_q is called the moving average parameter which describes the effect of past error on z_t and which needs to be estimated. For reasons that will become clear, it is also customary to θ_q write a negative sign in front of the parameter .As with an autoregressive process, the random shocks in a moving average process are assumed to be normally and independently distributed with mean zero and constant variance σ_a^2 and moving average model expresses the current value of the series z_t as follow:

$$z_t = \theta_0 + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \quad (3)$$

Where, θ is called the moving average parameter.

9.6.3 Mixed Autoregressive and Moving Average Model

We represented a process with an autoregressive and moving Average as ARMA (p, q). The p refers to the number of autoregressive parameters and q to the number of moving average parameters. Such a model can also be obtained as follows B-J [2]:

$$z_t = \phi_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} + a_t$$

$$-\theta_1 a_{t-1} - ... - \theta_q a_{t-q} \quad (4)$$

9.6.4. Autoregressive process integrated moving average model

Note that, although the random walk model (2.1) is no stationary series, it can often be transformed in to stationary series by means of differencing. By working with changes in the realizations, one can often induce stationary so that some of AR, MA, or ARMA filters discussed above can then be used to model the differenced series. Now, suppose that the first differences of a series are stationary then by defining the difference between consecutive values of what as Kaiser [5]:

$$w_t = z_t - z_{t-1} \quad (5)$$

We could replace z_t and z_{t-1} in the mixed autoregressive moving average model, the ARMA (1, 1) model with w_t and w_{t-1} to obtain

$$w_t = \phi_1 w_{t-1} + a_t - \theta_t a_{t-1} \quad (6)$$

The process for z_t by (5) and (6) is called an autoregressive integrated moving average (ARIMA) model.

$$z_t = w_t + w_{t-1} + w_{t-2} + ... \quad (7)$$

Actual realization, z_t an infinite sum of differences

$$z_t = \phi_0 + \phi_1 z_{t-1} + ... + \phi_p z_{t-p} + dz_{t-p-d} + a_t$$

$$-\theta_1 a_{t-1} + ... + \theta_q a_{t-q} \quad (8)$$

10. Stages of building model

The approach starts with the assumption that the process that generated the time series can be approximated using an ARMA model if it is stationary or an ARIMA model if it non-stationary. The model building and that it an iterative approach that consists of the following steps:

10.1. Identification

Basic tools at this stage are auto correlation function (ACF) and partial autocorrelation function (PACF), where each of them shows the appropriate model for any data to be autoregressive model with rank p that decreases function exponentially towards zero or in granular waves if the function (PACF) interrupted after displacement of p. As is appropriate model moving average with rank q was PACF function exponential decreases to ward zero or granular waves and function ACF interrupted after displacement q. And the appropriate model is mixed model ARMA if trod each of ACF function PACF behavior of exponential decreasing towards zero after displacement q-p of function PACF, B-J [2] .

The identification step is further broken down into Assess whether the time series is stationary, and if not, how many differences are required to make it stationary.

Identify the parameters of an ARMA model for data Brownlee [6].

10.2. Estimation

At this stage we get precise estimate of the coefficients of the model chosen at the identification stage. For example, if we tentatively choose equation (1) as our model, we fit this model to a variable data series to get an estimate that involves using numerical methods to minimize or lessen error term.

10.3. Diagnostic checking

Box and Jenkins suggest some diagnostic checks to help determine if an estimated model is statistically adequate. A model that fails these diagnostic tests is rejected. This stage may also indicate how a model could be improved. This leads us back to the identification stage (B-J [2]). We repeat the cycle of identification estimation, and diagnostic checking until we find a good final model to look for evidence that the model is good fit for data Pankratz [7].

10.4. Forecasting

Once a fitted model has been judged as adequately representing the process governing the series, it can be used to generate forecasts for future periods. Let's the current period, the origin date, be period n, and suppose we want to forecast h time periods ahead to period $n+h,$, that is, we want to know the value of the yet unrealized observation z_{n+h}. the time interval $n+h$ is called the forecast horizon. The forecast for z_{n+h} made at time period n for h periods a head, is denoted by $z_n(h)$ what is, for $h=1, z_n(1)$ is the one step ahead forecast of for z_{n+1}, for $h=2, z_n(2)$ is the forecast made at time n for period $n+2,$, using only observations z_1 through z_n Bandael [8] it is according to Douglas equation: $\hat{z}_{t+1} = E\left[z_{t+1} | z_t, z_{t-1}, z_{t-2}, ...\right]$ for $l \geq 1$ Douglas [9].

Since the variable to be forecast, $z_{n+h},$, is am random variable, it can only be fully described in terms of its forecast distribution, a probability distribution which is conditional on past and present data as well as on the specification of the ARIMA model .We will denote the forecast distribution of z_{n+h} by $f_{n,h}(z)$.

11. Application

In this part we apply these concepts to time series data of the number of TB patients, using the data obtained from the Ministry of Health in Khartoum state during the period 2007-2016.And they are implemented through the models of Box and Jenkins as the following stages.

11.1. Identification

As a first step in analyzing time series of TB to laying the plot which the data consist quartiles observation covering the period from the first quarter 2007 to the fourth quarter 2016, then draw from examining this plots. We observe that for the entire to quarter year period the TB has been steadily growing in some quarter, and lowing in another, and because of this trend this series is therefore non-stationary as the figure (1) bellow

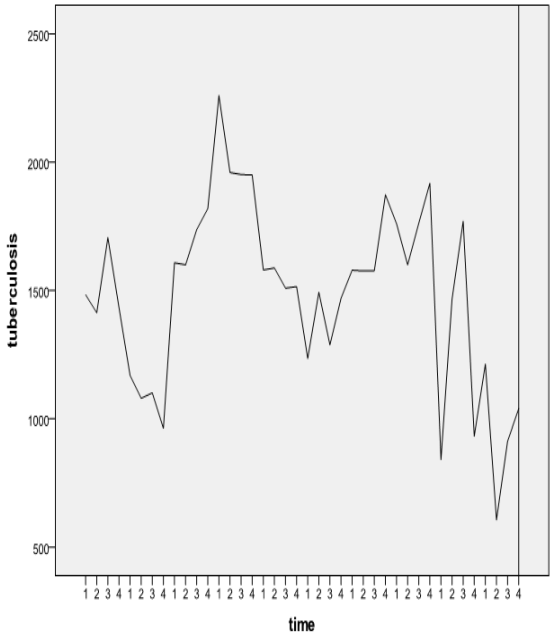

Fig. 1: Non-stationary TB.

From the above figure, it is clear that the time series is non-stationary, so we take the first difference. Then the series becomes stationary as figure (2) bellow shows:

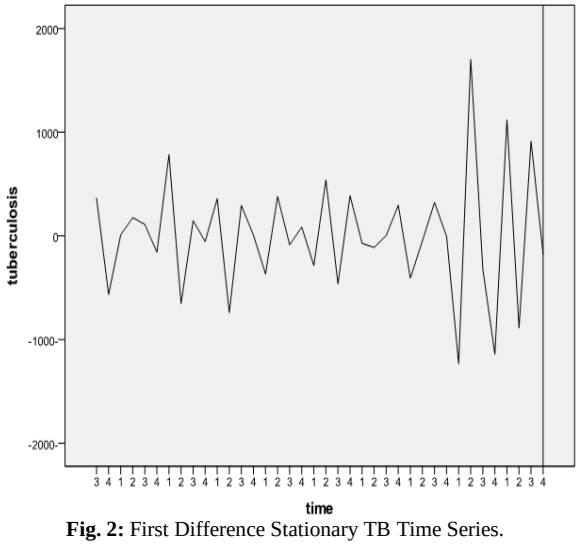

Fig. 2: First Difference Stationary TB Time Series.

The auto correlation function and the partial correlation function are two tools that identify the appropriate series model by comparing the figures, in case that the figure of the estimated functions of series data do not fully apply to the theoretical form but approximate them. When examining the autocorrelation function(acf), after the first lag falls to zero, whereas the partial autocorrelation function (pacf) contains two spikes the first at lag 1 and the other at lag 2. This shows that the model is ARIMA (2, 1, 0) as the following proposed model shows:

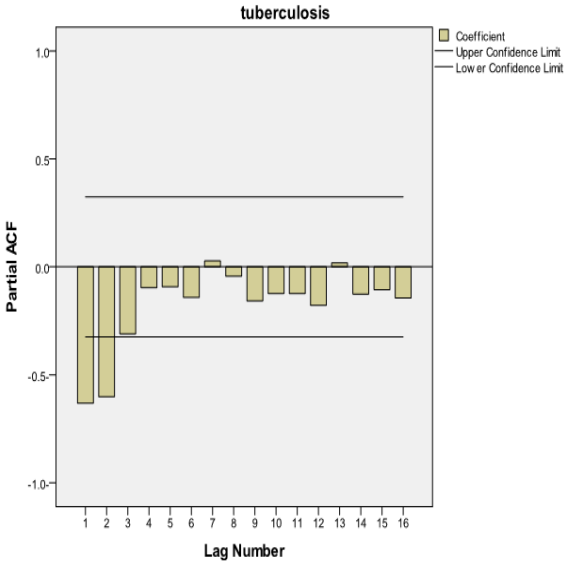

Fig. 3: Autocorrelation Coefficient.

Table 2: ARIMA Model Para meters

TB	Estimate	SE	t	sig
Constant	34.12	137.56	.248	0.81
AR-lag1	-1.015	.138	-7.37	0.0
AR-lag2	-.624	.142	-4.39	0.0
Time-Lag0	-13.6	54.04	-.252	0.80

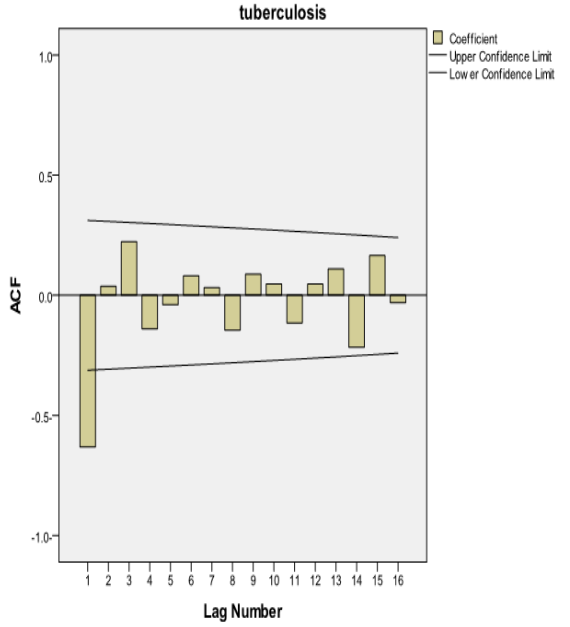

Fig. 4: Partial Correlation Coefficient.

From figure (4), we find that all partial autocorrelation coefficient within the limit of confidence $-0.5 \le r \le 0.5$.

11.2. Estimating the parameters of the model

The statistics and parameters of the model were estimated using statistical packages (SPSS) according to the following table (1) and table (2)

Table 1: Model Statistics

TB	Estimate
Stationary R square	0.624
R square	0.067
DF	16
L Jung- Box Q(18) statistic	28.399
sig	0.028
No outliers	0

Therefore, the proposed model is

$$z_t = -1.015 z_{t-1} - 0.624 z_{t-1} + 0.142$$

11.3. Checks the validity of the model

Box and Jenkins has proposed sets to test the model as follows:

1) Analysis of stationary.
2) Residual analysis.
3) Delete the parameters of the model.
4) Add new parameters if needed.

11.3.1. Analysis of stationary

After achieving stationary in the capabilities of the model, we take that as evidence of the adequacy of the model to represent the data since the stationary condition for (weak) stationary of the second – order regression model is as follows:

$$\Phi_2 + \Phi_1 < 1,$$

$$\Phi_2 - \Phi_1 < 1,$$

$$|\Phi_2| < 1,$$

All Φ_1, Φ_2 between -2, 2and it is residual investigated.

11.3.2. Analysis of residuals

The analysis of the errors of the model depends on the estimates of the residuals and not the real values of this error. The most important method in the analysis of the residuals is a test of the Box Pierce$Q_{(k)}$as follows:

The calculated value $Q_{(18)} = 28.399$ is greater than $\chi^2(16, 0.05) = 26.30$ we reject the null hypothesis that the errors are purely random.

11.3.3. Add new parameters if needed

Wheel and Wright used two types of random and nonrandom time series and found that the increase of parameters in the equation of time series reduces its mean sum square of error while the process repetition to obtain optimal parameter values is constant Wheel, Wright [10]

12. Forecasting

After verifying the validity of the model, it has been used to predict future observations of the phenomenon. For the model to be a good predictor, the prediction must have the least square mean of error.

12.1. Test of sample's ability to predict

Prediction is considered one of the significant aims in time series for, through it, the future path of phenomenon can be acquainted to assist in the process of planning, controlling and decision taking. Prediction studies phenomenon's development for a long time, as a factor that shows the impact of all factors on this phenomenon. Phenomenon changes along the time, from month to another, and from year to another. In itself, time is not considered to have an effective impact on the development of economic phenomena, as a future subjective indicator about human actions. However, time is adherers the development of economical phenomenon, and hence, it can be linked between phenomenon and instance that face this state, or between the phenomenon's development and the period of time elapsed or in which these developments shall occur resulting from elements other than time, impacting the phenomenon and leading to change it quantitatively and qualitatively. The estimated sample ability to forecast can be tested through the use of Theiler's coefficient criterion equality, as shown in the following table (3) and figure (5), (6).

Table 3: Forecast TB

year	quartile	Q_1	Q_1	Q_1	Q_1
2017		1259	1251	1243	1235
2018		1228	1220	1212	1204
2019		1197	1189	1181	1174
2020		1166	1159	1152	1144

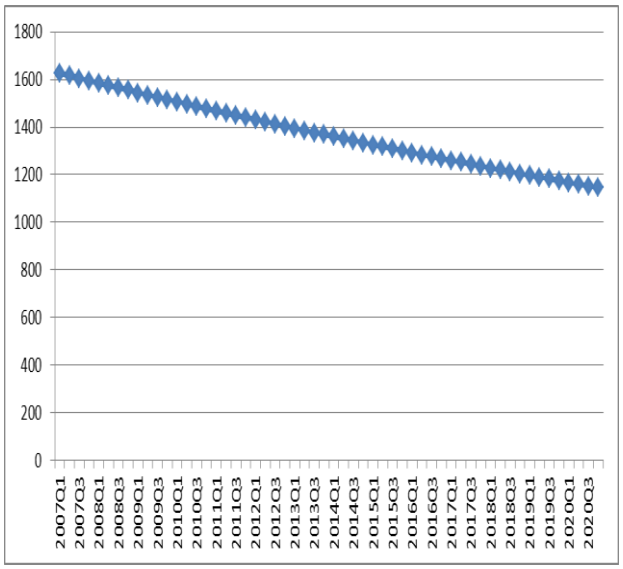

Fig. 5: Forecast.

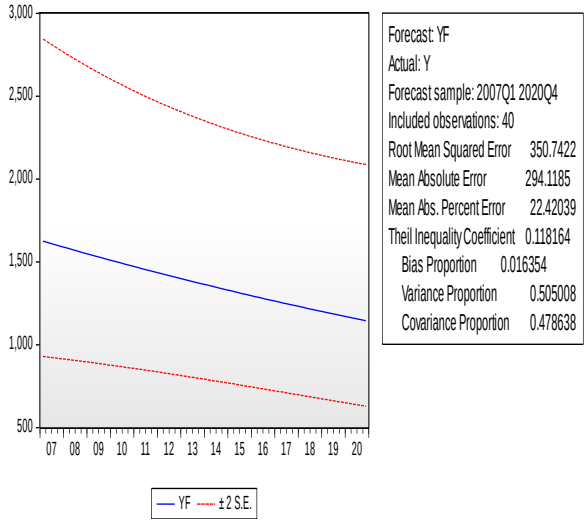

Fig. 6: Forecast.

13. Results

i). Time series of TB patients is not stationary.
ii). The number of TB patients decreases with time (series decreases)
iii). The model is order 2 of ARIMA (2,1, 0)

14. Recommendation

1) We recommend that the people in charge benefit from the result of the study.
2) The officials concerned should continue their efforts to reduce tuberculosis.

References

[1] S. A. Tomah, using analysis of time series to forecast number of patient's malignant tumors, Al- unbar university journal of economic, 8(4) ,1 (2012)371-393

[2] G. E. P. Box and Jenkins. M, time series Analysis Forecasting and control, Holden day, London (1976)

[3] R.I. Anderson, Distribution of theseseries analysis correlation coefficient, and, Mat statistic, vol, 13, 3(1942)113-129.

[4] G. E P., D.A Box and Piece, Distribution of residual Autocorrelation in Autoregressive –Integrated Moving Average time series model, JAsa, vol 65, 4(1970), 1520-1526.

[5] R. A. Maravall, Kasiser, Notes on time series Analysis ARIMA Models and signal Extraction, Ben co de esponaservicio studios, (2001)

[6] Brownlee, Jason, , "Gentle introduction to Box- Jenkins method for time series forecasting" elearningmastery.com/gentle-introduction - Box- Jenkins- method- time series- forecasting, (2017) 1-8 .

[7] Alan pankratz, Forecasting with univariate Box-Jenkins models concepts and cases, John whale and sons, Canada (1983)

[8] Walter. Vandael, Applied time seriesand Box-Jenkins models" Academic press, New York (1983)

[9] C M, J.G, Douglas and contreas, note on forecasting with Adaptive filtering, O.P. Q, Vol 24, No 4, (1976) 87 – 90

[10] Wheel Wright, S.C and Marked is, S (1973)," An Examination of the use Adaptive filtering in forecasting ", O.P.Q, Vol 24, No.1, 60-64.

Appendix

TB patient number of Khartoum state –period 2007 to 2016.

year	quartile	Q_1	Q_2	Q_3	Q_4
2007		1483	1413	1706	1433
2008		1169	1080	1101	963
2009		1608	1600	1737	1819
2010		2260	1970	1952	1950
2011		1580	1588	1509	1515
2012		1235	1493	1288	1470
2013		1579	1576	1577	1873
2014		1670	1600	1762	1918
2015		841	1466	1770	932
2016		1213	606	1129	1040