International Journal of Applied Mathematics
Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 7, Issue 4, Jun - Jul 2018; 9-18
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

APPLICATION SARIMA MODELS ON TIME SERIES TO FORECAST THE NUMBER OF DEATH IN HOSPITAL

Hanaa Elgohari[1], Mohammed AbdulMajeed[2] & Ahmed Elrefaey[3]

[1]Department of Applied Statistics, Faculty of Commerce Mansoura University, Egypt

[2]Faculty of Administration, Faculty of and Economics Salahaddin University, Iraq

[3]Professor, Department of pediatric, Mansoura University, Egypt

ABSTRACT

This paper aims to predict the number of deaths at Mansoura University Children's Hospital by using SARIMA models. It is necessary to use death data to determine the health requirement for hospital and measure medical efficiency within the hospital. We take the death data in hospital from Jan. 2011 to Dec 2017. We concluded that the model SARIMA (1,1,1) (0,1,1) is the best model which gives us the lowest value for each of RMSE and BIC, approximately lowest value for MAE and the largest value for R2.

KEYWORDS: *Time Series, SARIMA Models, BIC, RMSE, MAE, MAPE, R2, ACF, PACF*

Article History

Received: *19 Apr 2018* | ***Revised:*** *10 May 2018* | ***Accepted:*** *22 May 2018*

INTRODUCTION

The records that are gathered over time refer to Time Series analysis, because of the importance of the time order of data. One differentiating characteristic is that the applications of time series applications are very various and the records are dependent in time series. In addition, data may be gathered hourly, daily, weekly, and monthly and yearly, this depends on various applications. Moreover, notation can be used to symbolize "T" for a time series of length and the unit of the time scale implied in these notations such as {Xt} or {Yt} (t =1,⋯,T). We start to introducea number of real data that are used to indicate the modeling and forecasting of time series.

The term of seasonally refers to a regular model of changes which repeat for S time period, in which S refers to the numbers of timer periods till the pattern repeats again. Surly, seasonality causes the time series to be no stationary, a difference between a value and a value with lag and it refers to a multiple of S is called seasonal distinguishing.

The term of time series defined as data series that indexed (listed or graphed) in time order. Generally, a time series refers to the word " sequence" that is taken at equally, successive, and spaced points in time. Therefore, it is the sequence of separated time data. In addition, to, time series are frequently plotted through line charts. Also, time series applied in signal processing, statistics, the forecasting of weather, econometrics, the finance of mathematics, transport, earthquake prediction, the forecasting of trajectory, astronomy electroencephalography, communications, and control engineering, and broadly in any field of applied science and engineering that includes temporal measurements.

On the other hand, time series analysis can analyze time series data by involving some methods that elicit significant statistics and features of the data. The forecasting of Time series defined as using the model to make the prediction of future values that focused on early values, "time series analysis" does not refer to this type of time series analysis. It compares the values of a single or multiple time series at various points in time. Also, the data of time series have a temporal ordering in which natural ordering of the observations are not in it. Time series analysis is extracted from the analysis where the observations are related to geographical areas.

The stochastic model shows that observations are close together in time. In addition, to, time series models employed the natural one-way of time ordering, because the values will be illustrated over a specific period as it elicits some way from past values, rather than future values.

In addition to, the techniques of time series analysis can be classified as "parametric" and "non-parametric" approaches. In detail, the parametric approaches suppose the basic stationary stochastic process has a specific structure that could be characterized as using a few numbers of parameters.

On the other hand, time series analysis approaches or methods may be classified into" linear" and "non-linear", and "univariate" and "multivariate".

Theoretical Aspect

Autoregressive Integrated Moving Average (ARIMA)

$\{Y_t ; t \in Z\}$ process is an *autoregressive moving average (ARMA)* process of order *(p, q),* denoted with *$Y_t \sim$ ARMA(p, q),* if :

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + u_t - \theta_1 u_{t-1} - \cdots - \theta_q u_{t-q} \ldots \quad (1)$$

Where $u_t \sim$ *WN (0, σ^2),* and $\phi_0, \phi_1, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$ are *(p+q+1)* constants and the polynomials $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ and

$\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ Have no common factors.

ARIMA models are used in the data that indicates non-stationary evidence, which is first distinguishing step (identical to integrate part of the model) that can be used more times to ignore non-stationary.

The part of "AR' 'in ARIMA shows the regression of interests' developing variable on its lagged such as prior values. The part of "MA" shows a linear collection of error terms and its values happened at different times in the past as the regression error. The part of "integrated" model shows the values that are exchanged with the distinction between the prior values and their values (this process may have been implemented more than once). Moreover, the aim of each feature has specific aim that is making the model be suitable for the data.

Non-seasonal models are symbolized ARIMA (p, d, q) in which parameters "p, d, and q" are non-negative integers. "P" refers to the number of time lags of the autoregressive model, "d" refers to the degree of variation such as the number of times in which the data have past values subtracted, and" q" refers to the order of the "moving-average model".

Seasonal ARIMA Model

Both "non-seasonal" and "seasonal" factors in a multiplicative model are integrated by the seasonal ARIMA model. For the model, one shorthand notation is: SARIMA (p, d, q) × (P, D,Q)S

"p" refers to non-seasonal AR order "d" refers to non-seasonal differencing

"q" refers to non-seasonal MA order "P " refers to seasonal AR order

"D" refers to differencing" Q" refers to seasonal MA order

S = time span of repeating seasonal pattern.

This model may be stated without differencing operations as:

$$\Phi(B^s)\emptyset(B)(X_t - \mu) = \Theta(B^s)\theta(B)u_t \ldots \quad (2)$$

The non-seasonal components are:

$$\text{AR: } \emptyset(B) = 1 - \emptyset_1 B - \ldots - \emptyset_p B^p \ldots \quad (3)$$

$$\text{MA: } \theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q \quad (4)$$

The seasonal components are:

$$\text{Seasonal AR: } \Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps} \quad (7)$$

$$\text{Seasonal MA: } \Theta(B^s) = 1 + \Theta_1 B^s + \ldots + \Theta_Q \quad (6)$$

Where B is operating on Yt, has the effect of shifting the data back one period.

$$BY_t = Y_{t-1} \ldots \quad (7)$$

Two applications of B to Yt shifts the data back two periods:

$$B(BY_t) = B^2 Y_t = Y_{t-2} \ldots \quad (8)$$

and so on

By the sample autocorrelation coefficients that are the series of quantities, significance guide to the persistence in a time series are used to measure the correlation at different times between observations. A group of autocorrelation coefficients sorted as a separation function in time that is the sample of autocorrelation function (rk), or the ACF.

$$r_k = \frac{C_k}{C_0} = \frac{\sum_{t=1}^{N-k}(Y_t - \overline{Y})(Y_{t+k} - \overline{Y})}{\sum_{t=1}^{N}(Y_t - \overline{Y})^2} \ldots \quad (9)$$

Where $C_k = \frac{1}{N}\sum_{i=1}^{N-k}(Y_t - \overline{Y})(Y_{t+k} - \overline{Y})$; $k = 0,1,2, \ldots, K \leq \frac{N}{4}$ is the auto covariance?

The symbol of $\overline{Y}$ refers to the mean of the time series and N refers to the number of the observations.

The partial autocorrelation coefficients $\widehat{\varphi}_k$ are calculated as follows:

$$\widehat{\varphi}_1 = r_1 \; ; \; \widehat{\varphi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2} \; ; \; \ldots \ldots \text{etc.}$$

$$\widehat{\varphi}_k = \frac{\begin{vmatrix} 1 & r_1 & r_2 & \ldots & r_{k-2} r_1 \\ r_1 & 1 & r_1 & \ldots & r_{k-3} r_2 \\ r_2 & r_1 & 1 & \ldots & r_{k-4} r_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & r_{k-3} & \ldots & r_1 r_k \end{vmatrix}}{\begin{vmatrix} 1 & r_1 & r_2 & \ldots & r_{k-2} r_{k-1} \\ r_1 & 1 & r_1 & \ldots & r_{k-3} r_{k-2} \\ r_2 & r_1 & 1 & \ldots & r_{k-4} r_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & r_{k-3} & \ldots & r_1 1 \end{vmatrix}} \ldots \quad (10)$$

The collection $\{\hat{\varphi}_k\}$ is called the sample partial autocorrelation function (SPACF).

Stationary

If it has first and second moment time-invariant, yt is called stationary.

- E(yt)= μy for all $t \epsilon T$
- $E[(yt - \mu y)(yt-h - \mu y)] = \gamma h$ for all $t \epsilon T$ and all integers h such that $t - h \epsilon T$.

In the equation, one, a stationary stochastic process should fluctuate around a constant mean and does not have direction because all members of a stationary stochastic process have the same constant mean.

Fitting Model

It is very important that the selection of the model "Under-fitting a model" probably not express the true nature of the variability in the outcome variable. On the other hand, an "over-fitting model" loses generality. Akaike Information Criteria (AIC) is a way of choosing the model which balances the drawbacks. When a best model is chosen, the traditional method of null-hypothesis testing can be used on the best model to determine the correlation between particular variables and the interest outcome:

$$AIC = 2K - 2\log\left(L\left(\hat{\theta}/y\right)\right)\ldots \quad (11)$$

The denotation $\log\left(L\left(\hat{\theta}/y\right)\right)$ refers to the log at the maximum point in the model estimated but "K" refers to the number of estimable parameters such as degrees of freedom. Further refined this estimate for correcting for small data samples:

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}\ldots \quad (12)$$

n refers to the sample size and K and AIC are defined above. The correction is negligible and AIC is sufficient if n is large with respect to K,. AIC c is more general, however, and is generally used in place of AIC. The best model is with the lowest of "AIC c "(or AIC) score. It is essential to concentrate on the AIC and AIC c scores that are ordinal.

Moreover, Bayesian Information Criteria (BIC) is an estimate of the posterior probability function of a model as being true, under specific Bayesian setup, so that a lower BIC is a model to be the true model:

$$BIC = 2\log n - 2\log\left(L(\hat{\theta}/y)\right)\ldots \quad (13)$$

The Box-Ljung test is considered as a diagnostic tool that is used to test the lack of fit of a time series model. In addition, It is used to apply the residuals of a time series after fitting an ARMA (p, q) model to the data.

The test investigates autocorrelations of the residuals. If the autocorrelations are so small, we deduce that the model does not exhibit lack of fit.

Forecasting values, (where n is the number of forecasted errors):

$$\text{Mean Square Error MSE} = \frac{\sum e_i^2}{n} \ldots (14), e_i = y_i - \hat{y}_i .$$

$$\text{Root Mean Square Error RMSE} = \sqrt{MSE} \quad (15)$$

$$\text{Mean Absolute Percentage Error MAPE} = \frac{1}{n}\sum\frac{|e_i|}{y} * 100\% \quad (16)$$

$$\text{Mean Absolute Deviation MAD} = \sum\frac{|e_i|}{n} \quad (17)$$

Practical Aspect

We take the data of deaths in hospital for the period from Jan. 2011 in Dec. 2017 (the table below), to forecasting the daily rate of deaths periods for the future months, using the seasonal time series model (SARIMA)for the period from Jan. 2011 to Dec. 2017in the Figure (1) for the original data below, we notice the increasing and decreasing in the following of every all month's (2011-2017), a spatially increasing in the end months (Sep.in Dec.) of the years 2011- 2017 and decreasing after 2017 to 2011 in the first months .

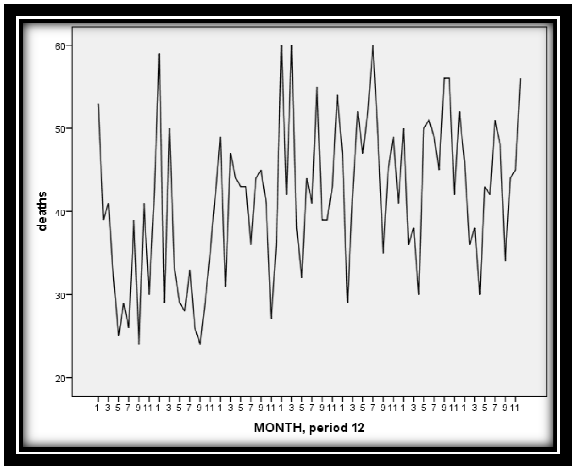

Figure 1: Seasonal Time Series of Deaths at Months in Hospital (2011-2017)

And you have tested the stationary of the series to know the stationary and Equality of mean, but the variance not stationary by t-test with Levenes Test on table(1) and table(2) :

Table 1: Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means	
		F	Sig.	T	DF
visitors	Equal variances assumed	.916	.341	-6.625	82
	Equal variances not assumed			-6.625	80.060

Table 2: Independent Samples Test for Equality of Means

		T-Test for Equality of Means				
		Sig. (2-Tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
visitors	Equal variances assumed	.000	-2031.90476	306.68507	-2641.99906	-1421.81046
	Equal variances not assumed	.000	-2031.90476	306.68507	-2642.22048	-1421.58905

After the Analyses the Seasonal Factors of Months in Table (3):

Table 3: Seasonal Factors % in Each Months from Years (2011-2017)

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
Seasonal factor %	15.4	30.8	46.2	61.5	76.9	92.3	107.7	123.1	183.5	153.8	169.2	184.6

From the above table, we are noticing the seasonal fluctuations increasing monthly, which are very high during the end four months of the year, decreasing in the following first months, and so on.

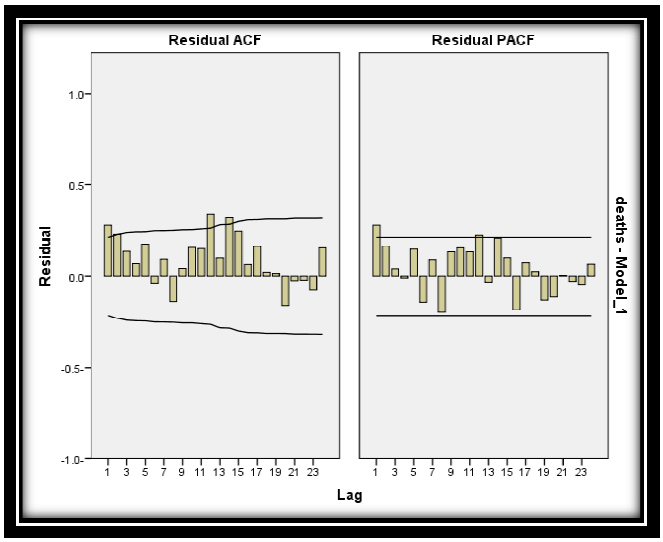

Figure 2: Autocorrelation Function (ACF) & Partial Autocorrelation Function (PACF) in Original Data

We suggest that we should use a seasonal difference after using transformation natural logarithm of the data of deaths in the series, It is also apparent from Figure (1) after non-stationary variable is difference, it becomes stationary. by first-differencing, it not necessary to show that the number of times a variable require to be distinguished to deduce stationary that depends on the number of unit natural Log to become equals for a variation of errors and used difference one(d=D=1) in models.

In the analysis that follows, we will try to improve these models through the addition of seasonal SARIMA terms:

Table 4: Statistics of Seasonal SARIMA Models

No.	SARIMA model	R2	RMSE	MAE	MAPE	BIC
1	(0,1,0)(0,1,1)12	0.320	11.822	9.112	22.418	5.120
2	(0,1,0)(1,1,0)12	0.256	12.487	9.895	24.159	5.229
3	(0,1,0)(1,1,1)12	0.329	11.699	9.061	22.319	5.159
4	(0,1,1)(0,1,0)12	0.323	11.088	8.844	22.040	4.992
5	(0,1,1)(0,1,1)12	0.473	9.610	7.654	19.025	4.706
6	(0,1,1)(1,1,0)12	0.472	9.608	7.646	18.989	4.765
7	(0,1,1)(1,1,1)12	0.485	9.598	7.513	18.682	4.823
8	(1,1,0)(0,1,1)12	0.396	11.126	8.542	21.128	5.059
9	(1,1,0)(1,1,0)12	0.341	11.652	9.095	22.418	5.151
10	(1,1,0)(1,1,1)12	0.410	10.985	8.436	20.877	5.093
11	(1,1,1)(0,1,0)12	0.353	11.051	8.754	21.894	5.045
12	(1,1,1)(0,1,1)12	0.508	9.603	7.496	18.681	4.824
13	(1,1,1)(1,1,0)12	0.501	9.526	7.614	18.920	4.808

From the table above, we conclude that the model (SARIMA (1,1,1) (0,1,1)) is the best, which gives us the lowest values for each of RMSE, and BIC, and approximately lowest value for MAE and the largest value for R2. So, we will rely on this model to estimate the predictions of the next months of the years 2018 and 2019 .The test of the parameters of the model is:

Table 5: Test of the Parameters of Predictions Electricity Interruption for Years 2017-2018 by the Model SARIMA (1,1,1) (0,1,1)

Parameters	Estimate	S.E.	t	Sig.
Constant	-0.029	0.029	1.006	0.318
AR – Lag 1	0.355	0.150	2.372	0.021
Difference	1			
MA – Lag 1	1.000	34.309	0.029	0.971
Seasonal Difference	1			
MA, Seasonal Lag 1	0.629	0.164	3.840	0.000
Natural Log Lag 0	0.017	0.017	0.951	0.345

The autocorrelation functions (ACF) and partial autocorrelation functions (PACF) can present useful information on particular properties of than stationary in the residual for the model of the figure (3):

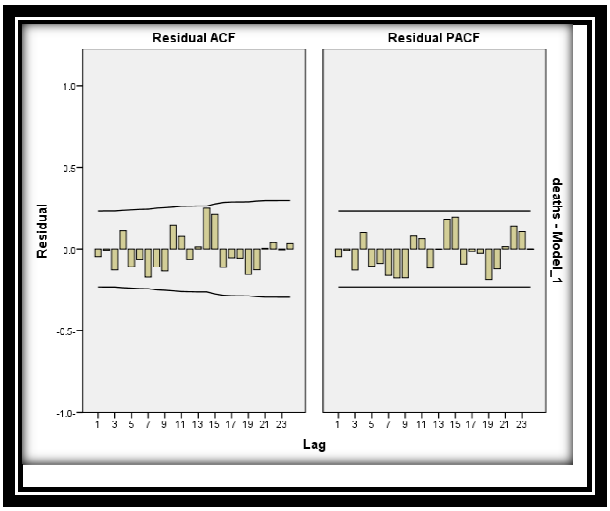

Figure 3: Residual of (ACF) & (PACF) for SARIMA (1,1,1) (0,1,1)12

Therefore, the forecasting values of the daily average of deaths per month in hospital during the years 2018 and 2019, using the above model SARIMA (1,1,1) (0,1,1)12, will be as follows:

Table 6: Forecasting of Deaths Per Month of SARIMA (1,1,1) (0,1,1)12 Since 2018-2019

Year	Months	Forecast	LCL	UCL
2018	Jan.	54	35	81
	Feb.	36	22	55
	Mar.	42	26	64
	Apr.	34	21	52
	May	41	25	63
	Jun.	43	27	66
	Jul.	47	29	72
	Aug.	46	28	71
	Sep.	39	24	60
	Oct.	47	69	72
	Nov.	45	28	69
	Dec.	54	34	84
2019	Jan.	52	31	82
	Feb.	34	20	54
	Mar.	39	23	62
	Apr.	32	19	50
	May	38	23	61
	June	40	24	64

	July	44	26	69
	Aug.	43	25	69
	Sep.	37	22	59
	Oct.	44	26	71
	Nov.	43	25	69
	Dec.	53	31	85

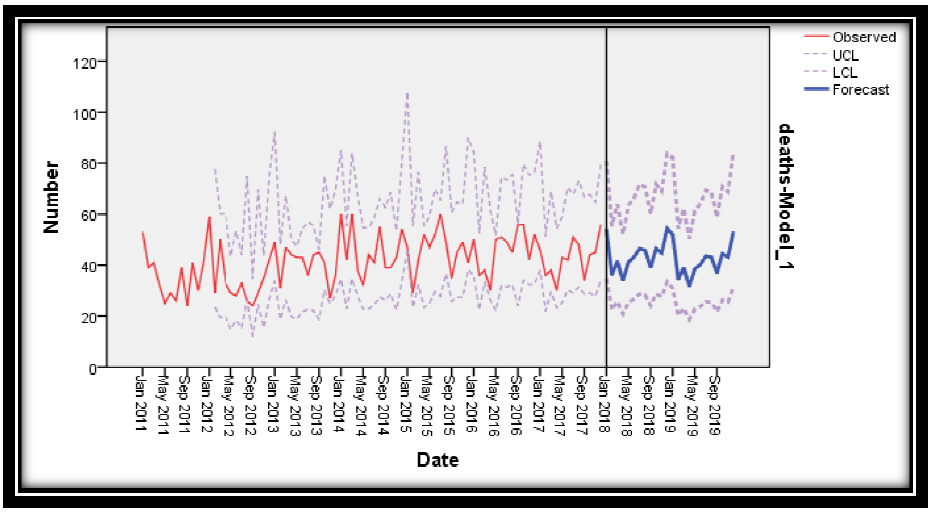

Figure 4: Fitting Model for Predictions of Deaths of SARIMA (1,1,1) (0,1.1) 12

Therefore, in table (6) appears that the forecasting values will increasing for death per month during the year 2018 and 2019 spatial in months Jan. and Dec., and by using the model SARIMA (1,1,1)(0, 1.1)12, the large forecast of deaths will be (54) in Jan. and Dec. in year 2018 as it's shown in table (6) and Figure (4) fitting the model by original data and forecasting for years 2018,2019

CONCLUSIONS AND NOTES

This study aims to predict the number of deaths at Mansoura University Children's Hospital by using SARIMA models and we are finding that SARIMA is the best model for forecasting from the other models by using the model SARIMA (1,1,1) (0,1.1)12 in this research. Also, the model SARIMA (1,1,1) (0,1.1)12 and SARIMA (1,1,1) (1,1.0)12 shown best results from the other models.1In general forecasting of deaths will increase for the next months in years 2018 and 2019. So, It should attention and study for health sector to know why increasing of deaths in the last years.

REFERENCES

1. *Adhistya, E.P., Indriana, H., Isna, A.(2013),"SARIMA (Seasonal ARIMA) Implementation on time series to forecast the number of Malaria incidence", Information Technology and Electrical Engineering, conference on Yogyakarta, Indonesia .*
2. *Akapanta, A. C., Okorie, I. E., Okoye, N. N.(2015)"SARIMA Modeling of frequency of Monthly Rainfall in Umuahia, Abia State of Nigeria, American Journal of Mathematics and Statistics, 5(2):82-87.*
3. *Brock well, P., Davis, R.(2002),"Introduction to time series and forecasting ",New york: springer .*
4. *Chan, N. H. (2002),” Time Series Applications to Finance”, John Wiley & sons, INC., publication, USA; ISBN 0-471-41117-5.*

5. *Gikungu, S. W., Waititu, A.G. (2015),"Forecasting Inflation Rate in Kenya Using SARIMA Model American Journal of The oretical and Applied Statistics 4,15-18 .*
6. *Cl, Chayalakshmi, Ds Jangamshetti, And Savita Sonoli. "Time Series Analysis And Forecasting Of Boiler Efficiency."Of Boiler Efficiency*
7. *Luo, C. S., Zhou, L., Qingfeng, W.(2013)," Application of SARIMA Model in Cucumber Price Forecast", Applied Mechanics and Materials,Vols.373-375,pp.1686-1690 .*
8. *BIRĂU, FELICIA RAMONA. "Forecasting Financial Time Series Based On Artificial Neural Networks." IMPACT: International Journal of Research in Business Management (IMPACT: IJRBM), ISSN (E) (2014).*
9. *Mira, S.K., Ahmad M.R.(2015),"Time Series Models for Average monthly Solar radiation in Malaysia", Research and Education in Mathematics, International Conference Kuala Lumpur, Malaysia .*