Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 2494-2500

Research Article

A Comparison of Some Information Criteria to Select a Weather Forecast Model

[1]Negar Nawzad Ali, [2]Anwaar Dhiaa Abdul Kareem

[1] College of Administration and Economics / University of Kirkuk
Anwaar Dhiaa Abdul Kareem
[2] College of Education for Pure Sciences / University of Kirkuk
[1]neekar.nawzad@uokirkuk.edu.iq
[2]anwaar71@uokirkuk.edu.iq

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: The purpose of using the criteria for selecting models is to determine an appropriate model that leads to estimates that we can use in making future predictions. In this study, we presented a number of information criteria that help to choose the best model in the time series, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), Akaike corrected information criterion (AIC_C) and Pham information criterion (PIC). We aim to obtain a appropriate model for a time series used to predict the lowest temperatures in Erbil governorate for the next three years depending on the Box-Jenkins methodology for the purpose of building time series models.

The best model was chosen from among the estimated models based on the value of the aforementioned statistical criteria. The model was used to predict the lowest temperatures for the next three years in Erbil governorate, as the predictive values were consistent with the real values of temperature ranges.

Keywords: Box-Jenkins models, model selection criteria, information criteria, AIC, BIC, AIC_C, PIC .

1. Introduction

Predicting future behavior is an important subjects in statistical sciences due to the need for it in different areas of life. Most countries depend in their development programs on advanced scientific foundations and methods in order to reach results that benefit us in this field. Time series analysis have the main role in building these programs through the analysis with knowledge of the past and forecasting of the future and its needs according to the available possibilities. It is certain that the analysis of time series at the global level has witnessed a significant development in the second half of the twentieth century, especially in the last three decades, It is also certain that this development is due to the modern methodology presented by the two scholars Box and Jenkins in the early seventies of the same century which has since become the most widely accepted and common performance in theoretical, and applied circles, as this methodology has proven to be highly efficient in modeling and predicting time data. The use of time series to predict the future behavior is of very huge importance to use it in different fields like forecasting weather such as (relative humidity, temperature, amount of rain, atmospheric pressure) and consumption of electrical energy, market conditions, prices, and others. Those interested in these studies have put in place a number of criteria to choose the best model to use in future predicting.

The aim of this research is to choose the best time series model for real data for the lowest temperatures for the period (January 1993 - December 2019) in Erbil governorate - Iraq. And choosing the appropriate order for the model by comparing several comparative models through criteria used in this field. Thus, determining the best model that is adopted to predict future periods.

2. Model building and predicting:[1][10]

Box and Jenkins suggested an iterative method for model building. This method includes three steps, as follows:

1- Model diagnosis
2- Estimating the model
3- Diagnostic examination

Each of these three steps can be illustrated. Diagnosis is the choice of an experimental model, as the diagnostic stage requires historical data to diagnose the appropriate model. We note that the experimentally diagnosed model contains unknown parameters usually and it is necessary to estimate, after the model is estimated, in this step the estimated model is verify. And making sure that it is the appropriate model devoid of the autocorrelation and moving average combination, This is done by examining the autocorrelation coefficients and the partial autocorrelation coefficients for the residuals in the model not the original series. If all the autocorrelation coefficients for a number of gaps fall within the confidence interval 95%, then the autocorrelation between the random error limits is not significant, in this case, the model is considered appropriateness of estimating and predicting. Except that iterative cycle of diagnosis, estimating and diagnostic examination are repeated until we reach a suitable model by analyzing the residuals of the model using the Ljung - Box Q statistic, it is symbolized by LBQ, which is used to test the following hypotheses:

H_0: $\rho_1 = \rho_2 = \ldots = \rho_k = 0$
H_1: $\rho_1 \neq \rho_2 \neq \ldots \neq \rho_k \neq 0$
Where ρ is autocorrelation coefficients for the residuals of the model.
The LBQ statistic is given in the following form:

LBQ = n(n+2) $\sum_{k=1}^{m}[\frac{\hat{\rho}_k^2}{n-k}]$

Where (m) is the number of previous time gaps entered into the test, (n) is the number of observations used in the estimation, the series is not stationary when the calculated value (LBQ) is greater than x^2 with a degree of freedom (m), Where the null hypothesis is rejected which states that all autocorrelation coefficients are equal to zero and vice versa. In the prediction stage, the final model is used to obtain forecasts, while the data becomes available.

3. Criteria for Selecting the Best Model : [3]

Model selection is one of the important issues in statistical analysis, as it is one of the main goals in statistical research and represents the final solution to many problems in practice. The process of determining the best model when there is a large number of explanatory variables and in the presence of a large sample is a difficult process, so the number of interpreted variables must be reduced because the process of introducing all the variables is an expensive process in terms of effort, time and money. On the other hand, the number of parameters should not be small, which may lead to unrealistic and biased predictions, meaning that the abbreviation mechanism does not affect the obtained information, as we obtain the same information as if we used all the variables. The difficulty lies in choosing the incoming variables and the excluded variables, meaning that we make a trade-off between what the independent variable adds in the interpretation of the dependent variable and between what it adds in terms of an increase in the variance in the event that its effect weakens. The following are the most important criteria used in selecting models.

3.1. Akaike Information Criterion (AIC): [6][8]

Akaike suggested a criterion for determining the best model, and he called it the Akaike Information Criterion, symbolized by (AIC), defined as follows :

AIC = -2 ln(l ($\hat{\theta}_{MLE}$ |y))+ 2k

Where: (l ($\hat{\theta}_{MLE}$ |y)) is maximum likelihood function, (k) is the number of estimated parameters.
The model with the lowest AIC value considered to be the best model.

3.2 Bayesian Information Criterion (BIC): [4][7]

Schwarz has presented a Bayesian method for estimating the model rank, since it is assumed that there is a set of models M_K with Previous probabilities P(M_K) with parameters θ_k. the Bayesian information criterion, symbolized by BIC, defines as the following:

BIC = -2 ln $\left(L(\hat{\Theta}|y)\right)$ + k ln (n)

Where: $L(\hat{\Theta}|y)$ is the maximum likelihood function, k is the number of estimated parameters.
For the purposes of model selection, BIC is calculated for each model and the model that produces the lowest value for this criterion is chosen.

3.3. Corrected Aakaiki Information Criterion (AICc): [9][2]]

Both of Davis and Brockwell suggested correcting the bias state in the AIC criterion by adding 2k (k + 1) / ((n-k-1)) to the AIC formula, so that the corrected criterion is as follows:

AIC_c= AIC + $\frac{2k(k+1)}{(n-k-1)}$

Where: k is the number of estimated parameters, (n) is the sample size.
AICc is used when k is large relative to the sample size n, and the model with the lowest AICc value is chosen.

3.4. Pham Information Criterion (PIC) : [5]

Pham suggested a criterion that takes into account a larger penalty term when adding many of the estimated parameters in the model, when there is a very small sample. This criterion symbolized by PIC, the value of the criterion is calculated as follows:

PIC = SSE + k ($\frac{n-1}{n-k}$)

Where: n is the number of observations in the model, k is the number of estimated parameters in the model, SSE is the sum squares of error.
The model with the lowest PIC value is determined.

4. Application for Minimum Temperature Data:

We will present the results of the application side of the study. The data is the average of the minimum temperatures in the Erbil Governorate for the period (January 1993 - December 2019), where we begin by presenting a simplified statistical description of the time series data through statistical measures and graphs in order to give a general idea of the nature of the data that will be modeled according to the Box-Jenkins methodology. The data were analyzed and the models were estimated using the programs Excel, Eviews and Gretl.

4.1 Data Description

To know the nature of the data whether the series it is stationary or not, a timeline of the minimum temperature averages for the period (January 1993 -December 2019) was drawn as in the figure (1).

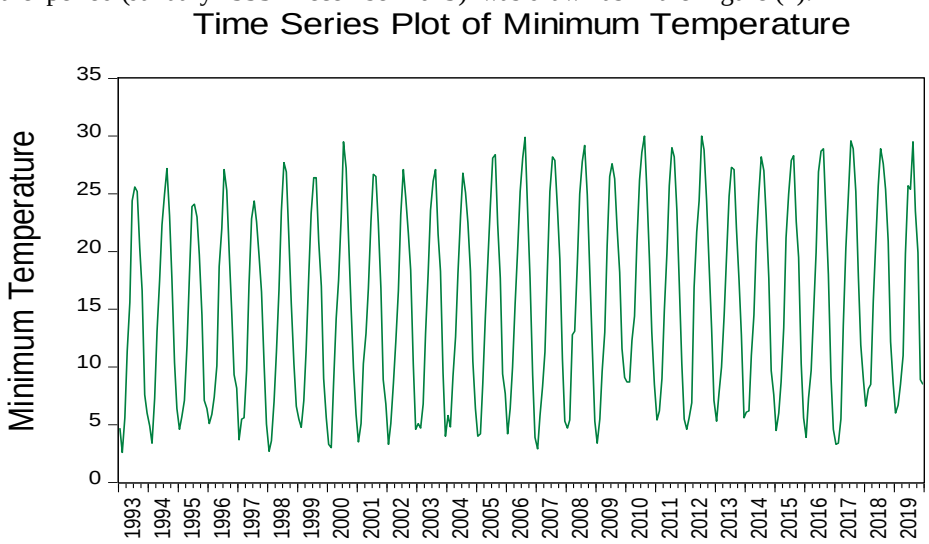

Figure (1) the time series for the minimum temperature data

We notice from the figure (1) that the time series suffers from the increasing Secular Trend and the seasonal, this indicates that the series is not stationary. To ensure that, we perform stationary tests.

4.2. Time Series Stationary Test

The Dicky-Fuller test was used to ascertain whether the series is stationary or not, Table (1) shows the result of the test. Then we can calculate and draw the autocorrelation function and the partial correlation function and confidence limits for the correlations for the purpose of testing the stationary of the original data**.**

Table (1) Dickey-Fuller test result (ADF)

		t-Statistic	P-value
ADF		-1.9989	0.2874
Test critical value	1 % Level	-3.4512	
	5 % Level	-2.8706	
	10 % Level	-2.5717	

We notice from the table (1) that the values of (p-value) of the test result are greater than the level of significance at all levels, and thus we accept the null hypothesis which states that the series is not stationary.

To get a stationary time series in the average, we resort to taking the first difference of the series($\nabla Z_t = Z_t - Z_{t-1}$) , As for the seasonal effects, they are removed by taking seasonal differences for the series and then performing the stationary tests again. Figure (2) shows the stationary of the series after taking the first difference.

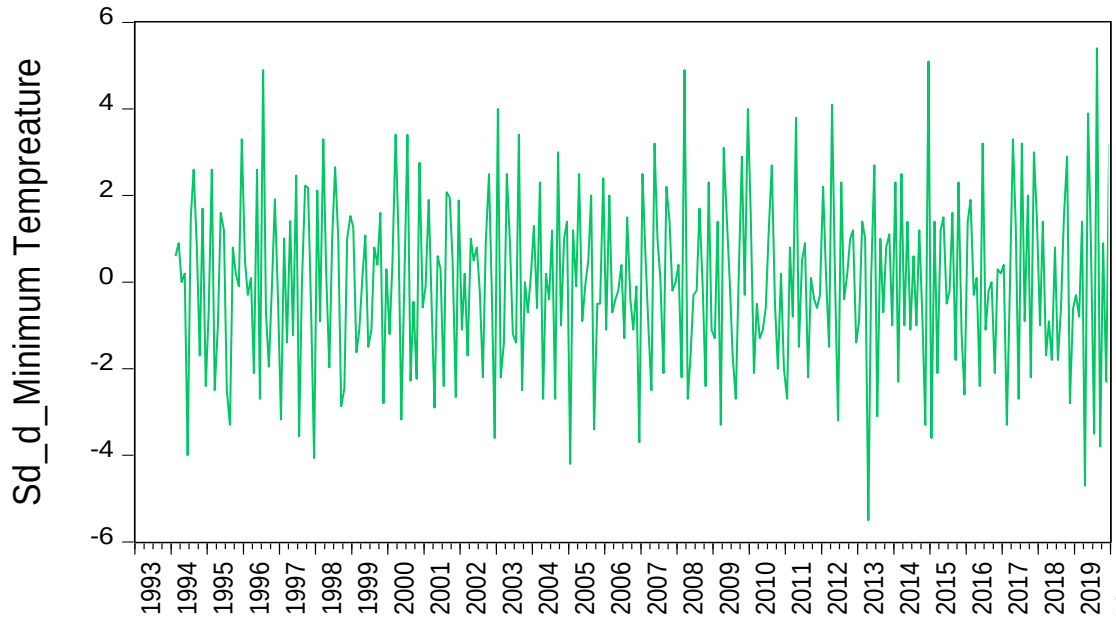

Figure (2) Stationary of the time series after taking the first difference and removing seasonal effects

To ensure the stationary of the modified series, we perform the Dickey - Fuller test again, as shown in Table (2).

Table (2) Dickey-Fuller test result (ADF) for the modified series

		t-Statistic	P-value
ADF		-7.5998	0.0000
Test critical value	1 % Level	-3.4521	
	5 % Level	-2.8710	
	10 % Level	-2.5719	

The aresults in the table (2) indicates that the value of (P-value) is smaller than the values of the level of significance at all levels, thus we reject the null hypothesis that the time series is not stationary and accept the alternative hypothesis that states the stationary of the time series.

4.3. Selecting the Best Model

After the series become stationary , we have to choose the model rank in such a step the model rank will be determined through the two autocorrelation and partial function diagrams to initially know the diagnosed model, then a number of models close to the diagnosed model are tested to choose the best ones based on some statistical criteria. It has been shown the initial diagnostic model is SARIMA(1,1,1)x$(0,1,1)^{12}$. Table (3) shows the comparison between these models according to information criteria (note the bold number indicates the lowest value of the criterion and thus the best model for the time series) .

Table (3): the values of information criteria for the proposed models for the minimum temperature series

No.	Model	AIC	AIC_C	BIC	PIC
1	SARIMA(1,1,1)x$(0,1,0)^{12}$	1198.5	311.2	1213.5	817.3
2	SARIMA(0,1,1)x$(0,1,1)^{12}$	1082.3	181	1097.2	539.1
3	SARIMA(1,1,0)x$(0,1,1)^{12}$	1130.5	229.7	1145.4	629.8
4	SARIMA(1,1,1)x$(1,1,0)^{12}$	1125.3	233.9	1144	633.1
5	SARIMA(1,1,1)x$(0,1,1)^{12}$	**1062.4**	**155.8**	**1077.4**	**497.5**
6	SARIMA(2,1,2)x$(0,1,1)^{12}$	1125.6	162.1	1151.8	499.6
7	SARIMA(2,1,2)x$(0,1,0)^{12}$	1199.7	312.9	1222.2	807.7
8	SARIMA(0,1,0)x$(0,1,1)^{12}$	1178.2	271.3	1189.4	725.4
9	SARIMA(0,1,0)x$(1,1,1)^{12}$	1177.2	258	1192.2	689.4
10	SARIMA(0,1,1)x$(1,1,1)^{12}$	1084.2	182.5	1102.9	537.5
11	SARIMA(0,1,2)x$(1,1,1)^{12}$	1069.6	161.7	1092.1	499
12	SARIMA(1,1,0)x$(1,1,1)^{12}$	1131	227.3	1149.7	620
13	SARIMA(2,1,0)x$(1,1,1)^{12}$	1110.1	207.4	1132.5	577.1
14	SARIMA(2,1,0)x$(0,1,1)^{12}$	1109.2	208.7	1127.9	584.2

Through the values of the criteria AIC, BIC, PIC and AIC_C in the table (3) we find that the model number (5) SARIMA(1,1,1)x$(0,1,1)^{12}$ is the best among the proposed models because it has the smallest value for the

criteria. Therefore, according to the model selection criteria, the appropriate model that has been reached for time series data for the minimum temperatures will be SARIMA(1,1,1)x(0,1,1)^{12} . Table (3) shows the parameters estimation of the selected model as follows.

Table (3): the estimated model coefficient SARIMA(1,1,1)x(0,1,1)^{12}

Models	Coefficient	Std. error	Z	P-value
AR(1)	0.2998	0.0588	5.102	0.0000
MA(1)	-0.9598	0.0209	-45.84	0.0000
SMA(1)	-0.9002	0.0510	-17.66	0.0000

Therefore, the proposed model SARIMA(1,1,1)x(0,1,1)^{12} will be the model according to which the minimum temperatures (series) are generated:

$= Z_{t-1} + Z_{t-12} - Z_{t-13} + \emptyset_1 Z_{t-1} - \emptyset_1 Z_{t-2} - \emptyset_1 Z_{t-13} + \emptyset_1 Z_{t-14} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \Theta_1 \varepsilon_{t-12} + \Theta_1 \theta_1 \varepsilon_{t-13} Z_t$

We substitute the values of the coefficients in the above formula as follows:

$= Z_{t-1} + Z_{t-12} - Z_{t-13} + 0.2998 Z_{t-1} - 0.2998 Z_{t-2} - 0.2998 Z_{t-13} + 0.2998 Z_{t-14} + \varepsilon_t + 0.9598 \varepsilon_{t-1} + Z_t$ $0.9002 \varepsilon_{t-12} + 0.86401196 \varepsilon_{t-13}$

In order to know the extent of the preference of the proposed model that has been identified, errors (residuals) are examined and diagnosed by knowing the residual distribution does it have a normal distribution that matches the assumptions that $\varepsilon_t \sim IID(0, \sigma_a^2)$, this can be known from the drawing of the residuals by using the histogram drawing of the residuals model closer to the normal distribution which indicates its randomness and this is confirmation of the quality of the model. As for the selected model, it is clear from figure (3) and through the shape of the drawing we notice that it is symmetrical and has the shape of a normal distribution, that is confirmation of the quality of the selected model.

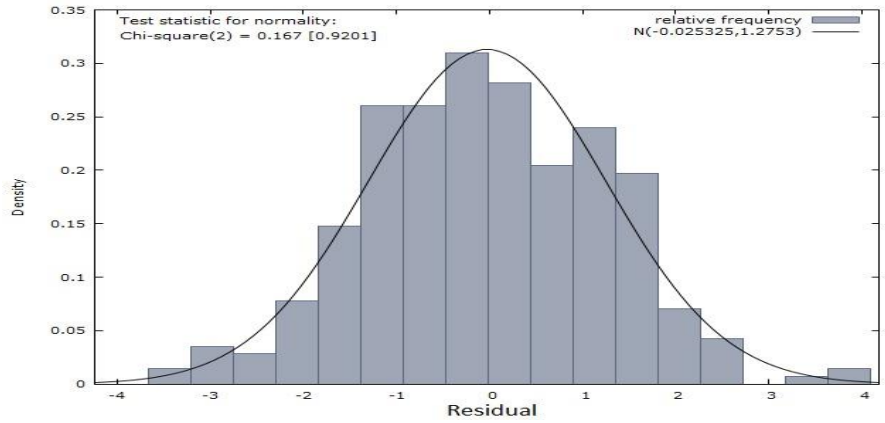

Figure (3): Test for the normal distribution of residuals

4.4. Forecasting the future Temperature

After going through the steps to identify the appropriate model for the minimum temperature data, estimate its parameters and examine the model, we use the model to predict future values of the minimum temperature averages for Erbil governorate for the next period from (January 2020 - December 2022), and as in table (4), it shows the results of the monthly averages of the minimum temperatures. Where the table includes new forecasts for three years and they were compared with the original values and build confidence limits 95% for these predictions. Figure (4) shows a drawing of the time series for the real data, the limits of confidence and the new predictions.

Table (4): The minimum temperature predicted with 95% confidence limits

Period	Forecast	Lower bound	Upper bound
Jan-2020	5.91864	3.43072	8.40656
Feb-2020	6.72545	4.09769	9.35320
Mar-2020	9.40486	6.75343	12.0563
Apr-2020	14.2467	11.5873	16.9061
May-2020	20.3129	17.6486	22.9772
Jun-2020	25.5936	22.9252	28.2621
Jul-2020	28.4598	25.7874	31.1321
Aug-2020	28.5059	25.8297	31.1821

Sep-2020	23.9172	21.2372	26.5972
Oct-2020	18.9987	16.3149	21.6825
Nov-2020	11.2603	8.57268	13.9479
Dec-2020	7.46030	4.76891	10.1517
Jan-2021	5.65935	2.93970	8.37900
Feb-2021	6.70011	3.97099	9.42923
Mar-2021	9.44966	6.71475	12.1846
Apr-2021	14.3125	11.5727	17.0523
May-2021	20.3850	17.6407	23.1294
Jun-2021	25.6677	22.9188	28.4166
Jul-2021	28.5343	25.7809	31.2877
Aug-2021	28.5807	25.8228	31.3385
Sep-2021	23.9920	21.2296	26.7543
Oct-2021	19.0736	16.3068	21.8404
Nov-2021	11.3351	8.56387	14.1064
Dec-2021	7.53513	4.75945	10.3108
Jan-2022	5.73418	2.92906	8.53930
Feb-2022	6.77494	3.95946	9.59043
Mar-2022	9.52450	6.70246	12.3465
Apr-2022	14.3873	11.5597	17.2150
May-2022	20.4599	17.6270	23.2928
Jun-2022	25.7425	22.9044	28.5806
Jul-2022	28.6092	25.7659	31.4525
Aug-2022	28.6555	25.8070	31.5039
Sep-2022	24.0668	21.2132	26.9204
Oct-2022	19.1484	16.2897	22.0071
Nov-2022	11.4099	8.54610	14.2738
Dec-2022	7.60996	4.74101	10.4789

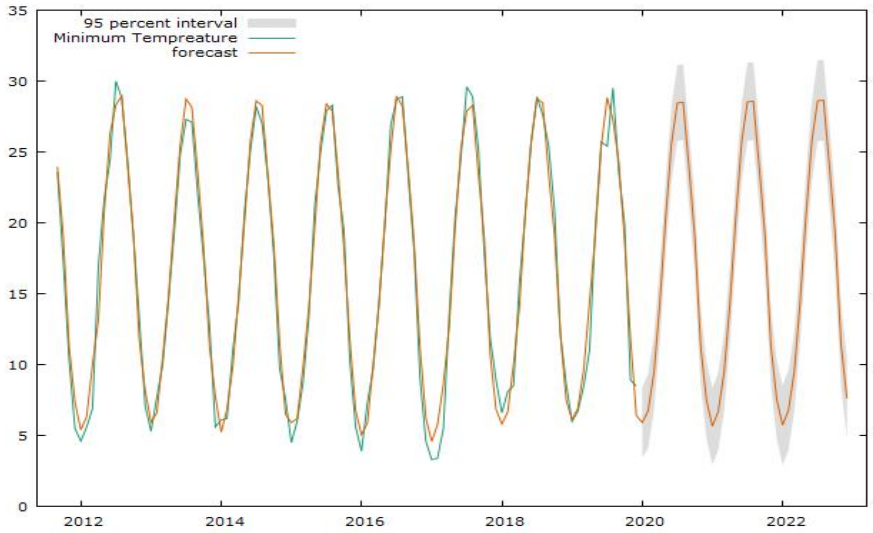

Figure (4) Real and predicted values for the minimum temperature data for the next three years

5. Conclusions

The study dealt with the use of time series models in modeling weather data and the most important information criteria available in selecting the best model in the time series. Box and Jenkins methodology was applied to analyze the minimum temperature data in Erbil governorate - Iraq, the results of the application showed that the information criteria used in the study all agreed in determining the same model to represent the time series data and that the predictive values that resulted from using the selected model were very close to the real values of minimum temperature .

References

1- Ahmad, T., 2018, Using the Box-Jenkins Methodology to Build a Standard Model for Predicting the Number of Syrian peoples, Tishreen University Journal for Research and Scientific Studies - Economic and Legal Sciences Series, 40(6), pp. 18-19.

2- Hurvich, C. M., & Tsai, C. L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of time series analysis*, *14*(3), 271-279.

3- Najm al-Din, A. K. & Salih, M. A., 2018, A comparison between linear and nonlinear regression models for studying the causes of premature infant mortality in Babel, Karbala University Scientific Journal, 16(2), pp. 152.

4- Neath, A. A. & Cavanaugh, J. E. ,2012, The Bayesian information criterion: background, derivation, and applications, Wiley Interdisciplinary Reviews: Computational Statistics, 4(2), pp.199-200

5- Pham, H. ,2019, A New Criterion for Model Selection. Mathematics, 7(12), pp. 1-12

6- Portet, S. ,2020, A primer on model selection using the Akaike Information Criterion. Infectious Disease Modeling, 5, pp. 123.

7- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*(2), 461-464.

8- Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*(1), 117-126.

9- Snipes, M., & Taylor, D. C. , 2014, Model selection and Akaike Information Criteria: An example from wine ratings and prices, Wine Economics and Policy, 3(1), pp.2-3 .

10- Tohma, Saadia Abdul-Karim (2012). "Using Time Series Analysis to Predict the Number of People with Malignant Tumors in Anbar Governorate", Anbar University Journal of Economic and Administrative Sciences, Issue (8), Volume (4).(pp.381)