IRAQI Academic Scientific Journals

Samarra Journal of Pure and Applied Science

SJPAS

www.sjpas.com

ISSN:2663-7405

Modeling Time Series for Prediction of Thalassemia in Nineveh Governorate

Shaymaa Riyadh Thanoon*
College of Nursing, University of Mosul (shaymaaalnuaimi@yahoo.com)

Article Information

Received: 01/04/2020
Accepted: 12/05/2020

Keywords:

Tim Seriese, Prediction,(ARIMA)

Abstract

The aim of this research is to analyze the time series of Thalassemia cancer cases by making assumptions on the number of cases to formulate the problem to find the best model for predicting the number of patients in Nineveh governorate using (Box and Jenkins) method of analysis based on the monthly data provided by Al Salam Hospital in Nineveh for the period (2014-2018). The results of the analysis showed that the appropriate model of analysis is the Auto-Regressive Integrated Moving Average (ARIMA) (2,1,0) and based on this model the number of people with this disease was predicted for the next two years where the results showed values consistent with the original values which indicates the good quality of the model.

Introduction

This study deals with the health aspect of the governorate, which went through the past period of suffering and negligence as a result of wars and control.is concerned with the human element, whose responsibility is the construction and reconstruction and development . The aim of the study is to uncover this disease, which has increased in recent times in Nineveh governorate. The city is affected by microbial weapons and severe shortage of health care due to the destruction of most of its health centers. relied on the theoretical aspect of the box and Jenkins map in time series analysis (diagnosis, estimation, future prediction،,testing model appropriateness. Practically ,it depends on real numbers of infections in tumor of stomach cancer to reach the best model of predication of numbers of infections. [1] used two types of random and nonrandom time series. They noted that increasing the number of parameters in the equation of random series minimizes mse , whereas repeating the process to get the value of parameters will be stable. Whereas in case of nonrandom series whose supposed values are (1,2,1,2,....) ,the number of parameters increased by increasing the frequency of the operation in order to obtain the lowest value of the mean error square.

The researcher [2] studied a set of Hybrid models (-Regressive - moving media) by simulation to the form [3]. The aim of this research is to predict and determine the best model to study the number of people with thalassemia in Nineveh governorate for the period (2020-2019) in order to take the necessary measures to reduce the disease in the future.

The Theoretical Side

Autocorrelation (AC)

It is an indicator that shows the degree of relationship between the values of the same variable at a different displacement period (K) and its value ranges between (-1,1) . That is $-1 \leq \rho_k \leq 1$ to be evaluated according

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}(m_t-\bar{m})(m_{t+k}-\bar{m})}{\sum_{t=1}^{n}(m_t-\bar{m})^2} \qquad (1)$$

The ACF function is a means of knowing the stability of the time series as it tends to either decline rapidly to zero as the displacement intervals increase k or break after a number of displacement periods. That is , K=q

$\rho_k = 0 \ \forall \ k > q$ Since the sample correlation function is only an estimate of the correlations, the values are likely to be small $r_k \neq 0 \ \forall \ k > q$ q. Whereas if the time series is not stable because of the existence of a rising or descending direction in average , the function (ACF) does not break slowly to zero .The residual Auto-correlation function and an important basket to examine the suitability of the model by randomizing residual errors are: $\rho = \begin{bmatrix} 1 & k = 0 \\ 0 & k \neq 0 \end{bmatrix}$

Partial correlation (PACF)

It is an index that measures the relationship between m_t و $m_{t-K}$for the same series assuming constant values of the time series. It is known as the last limit of the Auto-Regressive model of AR(P) degree .

The partial Auto-correlation function (PACF) is used for a stable time series that tends to decline rapidly to zero as the displacement intervals increase[4].

Stationary Time Series

The time series may be unstable in variability as well as unstable on average, which makes it have multiple fluctuations around data even when the series is homogeneous as these models are described as having unstable and homogeneous behavior and become stable when taking appropriate number of practical differences.

It is using the back difference factor and is denoted by the symbol ∇ then $\nabla m_t = m_t - m_{t-1} = (1-b)m_t$ then the time series becomes stable after taking (d) of the differences, but when the stability of the contrast is processed by taking the natural logarithm of series data or inverted data [5].

Box and Jenkens (B-J) models of time series

1- Autoregressive Model (AR)

The general formula of the Auto-regression model is as follows:

$$m_t = \emptyset_0 + \emptyset_1 m_{t-1} + \emptyset_2 m_{t-2} + \cdots \emptyset_p m_{t-p} + a_t \quad (2)$$

Therefore, the PACF is interrupted after the first displacement. If P = 2 is in equation (2), we get the second order autistic regression model AR (2) whose form is:

$$m_t = \emptyset_0 + \emptyset_{1\ 1} m_{t-1} + a_t \quad (3)$$

Which represents the first order Auto regression model AR (1) and the ACF of the model is $p_k = \emptyset_{p_{k-1}}$. This equation can be solved using ($\rho_0 = 0$) and getting ($\rho_k = \emptyset_1^k \quad k = 0,1,2, \dots$) The AR correlation function of the AR model (1) ,Therefore, the PACF is interrupted after the first displacement. If P = 2 is in equation (2), we get the second order autistic regression model AR (2) whose form is:

$$m_t = \emptyset_0 + \emptyset_{1\ 1} m_{t-1} + \emptyset_{1\ 2} m_{t-2} + \cdots a_t \quad (4)$$

showed that the Auto-correlation function of AR (2) is exponentially diminished if $\emptyset_1^2 + 4\emptyset_2 \geq 0$but if $\emptyset_1^2 + 4\emptyset_2 < 0$the correlation function (ACF) are dwindling sine waves, and the subjective ρ_{kk} partial Auto-correlations of the AR model (2) are as follows: [6], $\rho_{11} = \frac{\emptyset_1}{1-\emptyset_1}$, $\rho_{22} = \emptyset_2$, ρ_{kk}

So the partial correlation function (PACF) of the AR (2) model is interrupted after the second offset k> 2.

2- Moving average Model (MR)

The model of the moving media of class (q) can be represented using the backward recoil factor (B) as follows:

$$m_t = \emptyset_0 + \left(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_q B^q\right) a_t$$

And the general equation of this model will be

$$m_t = \emptyset_0 + a_t - \emptyset_1 m_{t-1} - \emptyset_{\ 2} m_{t-2} + \cdots \quad (5)$$

Where $\emptyset_t$ represents parameters of moving average model

$$-1 < \emptyset < 1 \ , i = 1,2,3, \dots, q$$

and q is the degree of the model and the Auto correlation function of the model (AM) breaks or approaches zero after displacement qwhereas (PACF) decreases exponentially.

3 -Autoregressive Moving Average Model (ARMA)

The form can be written in the general formula of grade (p, q)

$$m_t = \emptyset_0 + \emptyset_1 m_{t-1} + \cdots + \emptyset_{\ p} m_{t-p} + a_t - \emptyset_{\ 1} a_{t-1} - \cdots - \emptyset_{\ q} a_{t-q} \quad (6)$$

Autoregressive Integrated Moving Average Models (ARIMA)

Time series models may be unstable by themselves but become stable after many conversions. Therefore; the model that expresses this process will differ from the original model as it must include those conversions or differences made on the model called integrated hybrid models. ARIMA models are the most commonly used time series models since all models can be derived from either autistic regression, moving or mixed media .It comprises three parts . The first part is AR (p), which is used in prediction of time series. The second is the MA(q) model and the third part I(d) represents the differences required by the series in order to be stable, therefore, express the non-seasonal regressive moving average models according to the formula (p, d, q) ARIMA where:

P: is the rank of the Auto-regression model

q: rank of the moving media model

d: the number of differences that make the chain stable

And using the Bounce Factor (B) in the following formula:

$$\emptyset(B)(1-B)^{d}x_t = \emptyset_0 + \theta(B)a_t$$

[7], Where

$$\emptyset(B) = (1 - \emptyset_1 B - \cdots - \emptyset_P B^P)$$
$$\theta(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$$
$$(1-B)^d = \nabla^d$$

Assuming $\nabla^d x_t = M_t$the general formula of the integrated hybrid model is as follows:

$$m_t = \emptyset_0 + \emptyset_1 m_{t-1} + \cdots + dm_{t-p-d} + a_t - \emptyset_1 a_{t-1} - \cdots - \emptyset_q a_{t-q} \qquad (7)$$

ARIMA models can therefore be considered stable ARMA models with different ranks [8].

Building Time Series Models

1.Identification

2.Estimation

3.Diagnostic checking of model

1 - The method that depends on the test (ljung and box) Q in order to test the purpose of nothingness

$$H_0 = \rho_1 = \rho_2 = \cdots \rho_S = 0$$

The test(Q) equals is:

$$Q_s = n(n+2)\sum_{k=1}^{s}\frac{1}{n-k}r_k^2(a)$$

2 -It is the method by which we rely on the confidence limits of the Auto-correlations of the estimated residues. The akaikes information criterion (AIC), will be chosen to select the best model and defines the AIC standard: [9].

$$AIC(p) = IN(\sigma^2) + \frac{2(p+q)}{n}$$

Where σ^2 represents the variation of the model and p + q is the number of estimated features. The above formula has been modified to give more weight to the models used for the largest number of views :

$$MIAIC = \frac{AIC}{n}$$

4– Forecasting

It is the last step in the study and analysis of time series models and is the main objective of the study. After determining the appropriate data model is used to determine the values of the future phenomenon and for periods (L). [10] and forecasting can be calculated after steps of (L) from $\widehat{m}_{t+L} = E[m_t, m_{t-1}, m_{t-2}, \dots]$ FOR ≥ 1

Application

The data was collected from the visitors of Al-Salam Teaching Hospital in Nineveh Governorate, where the data consists of a time series consisting of 60 samples from 2014 to 2018 as these samples represent the number of Thalassemia sufferers as shown in Table 1.

Table 1: Numbers of thalassemia patients

year month	2014	2015	2016	2017	2018
January	40	65	100	150	190
February	55	59	65	70	88
March	43	43	45	55	70
April	59	65	70	75	99
May	67	70	88	90	108
June	81	88	98	99	150
July	80	89	100	105	145
August	90	100	105	160	190
Septembr	95	102	160	170	199
October	95	107	120	125	140
November	122	129	128	133	156
December	130	133	140	150	18

The data was prepared by drawing the propagation shape and extracting self-and partial correlation coefficients as showing in figure 1.

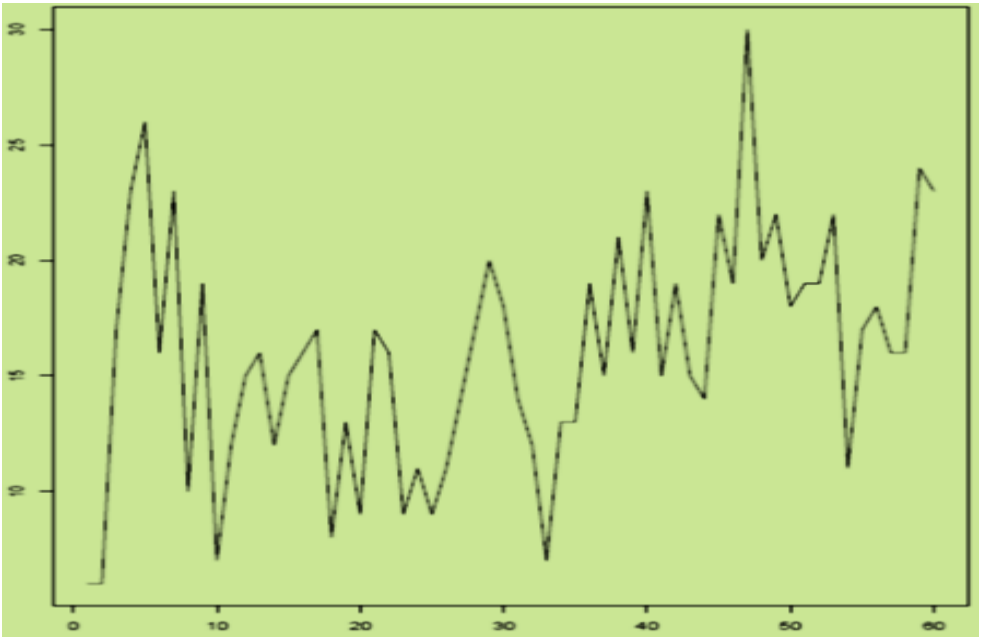

Fig.1: Represents the number of people with thalassemia

It is observed that the data tends to remain stable, but a general trend increasing with time is noted .

This has been confirmed by the correlation coefficients where they were different from zero.

In order for the chain to be stable, all values of the self-correlation coefficients of the sample must be entered within the confidence limits except the first displacement as the confidence limits are at an accuracy level of 95%.$-0.25 \leq r_k \leq +0.25$

To test the significance of the coefficients of a auto-correlation function (Q.stat)LjungandBox , the values were

$$(Q.\,\text{stat} = 221.44 > x^2_{(12,0.05)} = 18.55)$$

This confirms the instability of the time series on the average, therefore rejects the null hypothesis, which indicates that the coefficients of Auto-correlation are equal and zero, and accept the alternative hypothesis. This means that the series is unstable. No. (2) .The loss of the general trend is noted in his behavior which indicates the stability of the chain in the average, as showing in figures 2 and 3.

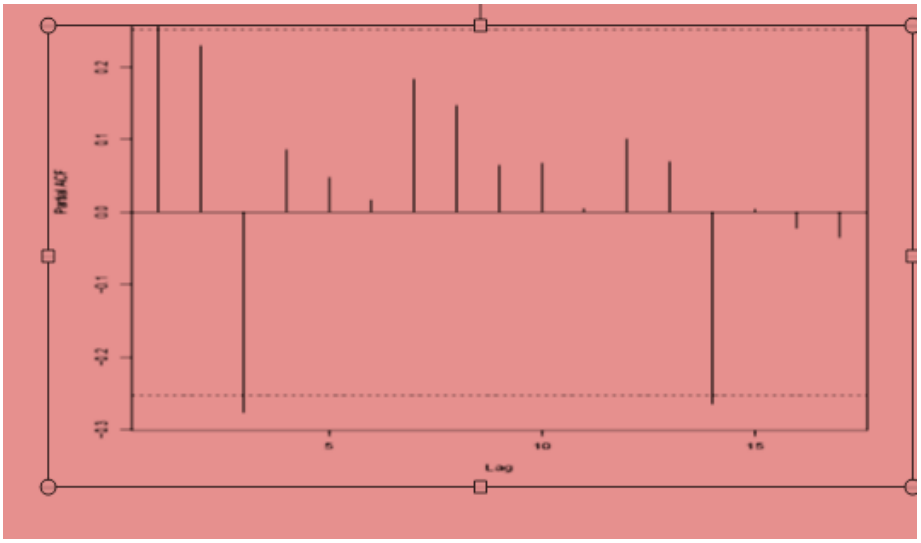

Fig.2: Self-Correlation and partial time series coefficients

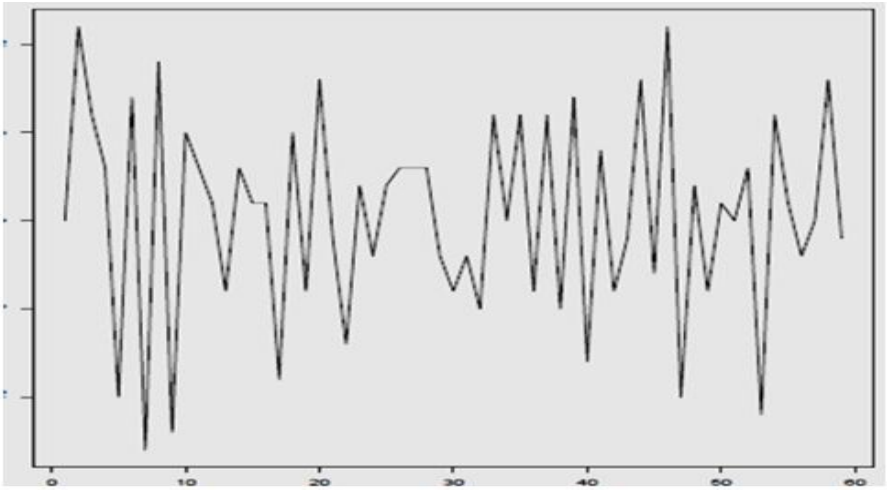

Fig.3 : The time series after taking the first difference

The confidence limits of the self-correlation function and the partial correlation of the sample were also drawn as shown in figure 4.

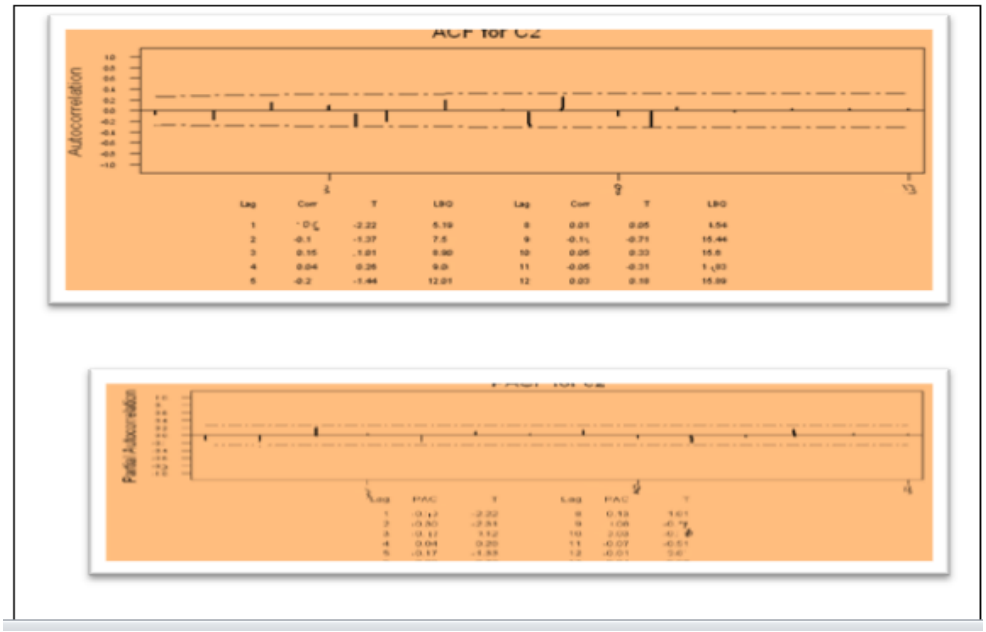

Fig.4: Self correlation and parttial correlation coefficients after taking the first difference

Diagnosis.

The diagnostic criteria have been applied based on the shape of the curve of the correlation function of the samples (ACF) and the shape of the partial correlation function curve (PACF). At the comparison of partial and auto values and taking the first difference ,it is noted that the behavior of the diminished sinusoidal function decreases gradually with increasing displacement intervals (K). A cut after the third displacement of the PACF function is also noted .Therefore , it is concluded that auto regression model is the appropriate one .ARIMA (2,1,0) was also compared with ARIMA (1,1,0) for the purpose of determining the exact gradient rank. Table 2 shows its significant significance.

Table.2: The sum of the residual squares with the variation of the model, the akek standard and the Schwarz standard.

Model order(p)	Adi.sse	Risiduals.v	AIC	MAIC	SBC
ARIMA(1,1,0)	534.8888	9.7651	365.1051	6.1297	350.6519
ARIMA(2,1,0)	487.4723	6.1209	355.1053	6.1109	344.6410

Differentiation criteria were used between the two models.

Thus, it is concluded that the second order model has the lowest value and therefore that the integrated model ARIMA (2,1,0) is the appropriate model of data, whose mathematical form is as follows:

$$m_t = \emptyset_0 + m_{t-1} + \emptyset_1(m_{t-1} - m_{t-2}) + \emptyset_2(m_{t-1} - m_{t-2}) + a_t$$

Evaluation

After verifying the suitability of the model and selecting the morale of its features, the model is estimated and the normal least squares method is applied to the data of the series which proved the variation of significance of parameters from zero as shown in table 3

Table 3: The results of the analysis that proved the difference in the moral parameters

T	STDEF	CEOFF	TYPE
-1.8532	0.1194	-0.1359	AR 1
-1.1362	0.1194	-0.1102	AR 2
0.2532	5.4932	1.460	constant
Analysis of variance:			
Df	Adj.sum of squares	residuals	
Residuals 49	350.4920	variance	
		5.8	
		Standard error=2.15420	
		Log likehood=-129	
		250.1 AIC	
		4.2 MAIC=	
		285.1 SBC=	

Testing the Accuracy Model

The residual and partial correlation coefficients of the residue were extracted in order to test the randomness of the residual sequence as shown in figure (5) since all the correlation coefficients have $r_k(\hat{a})$ the confidence limits $(-0.1 \leq r_k(\hat{a}) \leq +0.1)$and to ensure that Adaptability of the test The test ([Q.stat) (Ljung and Box)] has been applied.

It is noticed that the calculated value of $((5.5)q_{12}$ is less than the tabular value of $(x^2_{10.005})$ of (18.307) which leads to acceptance of the null hypothesis. Therefore, ARIMA (2,1.0) is the appropriate model for the data and by representing the sequence of residues of the estimated model. It is confirmed that the errors are distributed naturally, which shows the characteristics of the normal distribution which leads to use it in forecasting and accepting it.

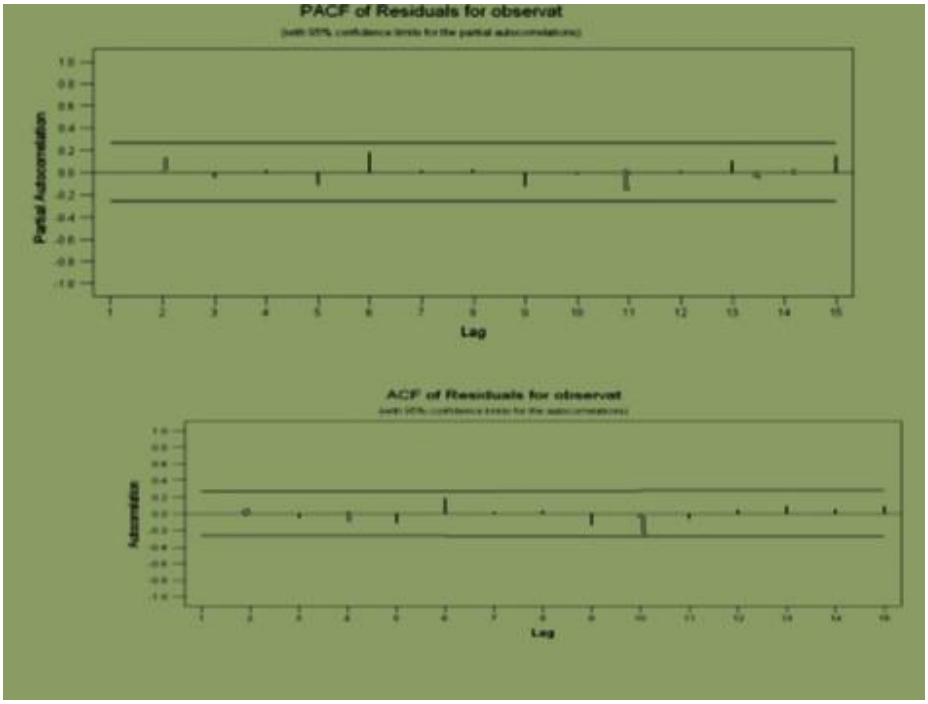

Fig.5: Self and partial correlation coefficient for estimated model protectors.

Forecasting

Form (222) was used to predict the number of patients with thalassemia for the period (2019-2021) as shown in Table 4

Table 4: The time series of these predictions are also represented as

year month	2019	2020
January	104	120
February	160	165
March	100	170
April	177	188
May	190	200
June	190	200
July	200	225
August	205	240
September	255	270
October	270	280
November	288	302
December	300	344

The data for the time series of these predictions were also represented as shown in figure 6 where they were similar to the behavior of the original series

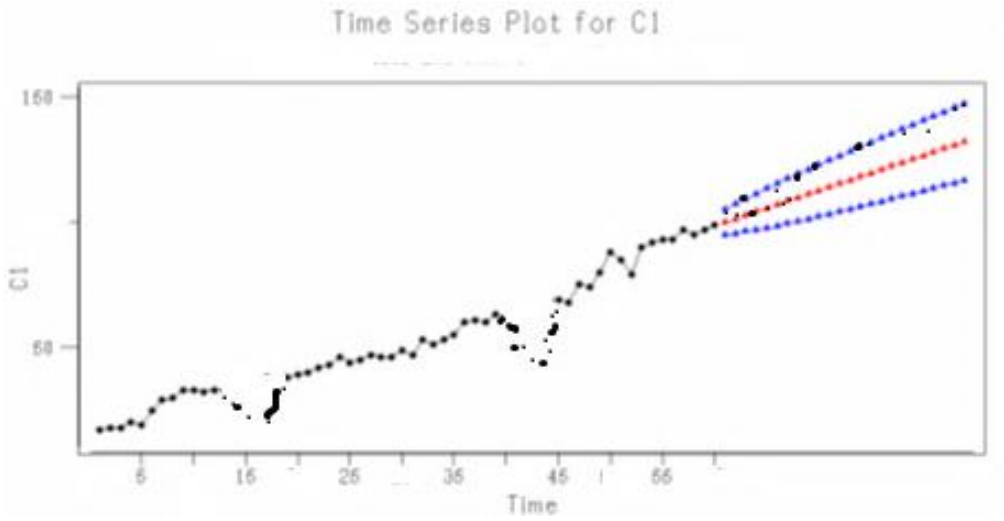

Fig.6: Predictive values for the thalassemia population chain

Discussion

In this paper, the stability of the time series was achieved after taking the first difference of the data and taking partial Auto-correlation of the sample after the second displacement. It is shown that auto correlation function gradually decreases with the increase of displacement intervals in a form of decreasing sine waves .Whereas, it is noted that there is a break in the partial auto correlation of the sample after second displacement .the study also observed through the study of the number of patients with thalassemia in Nineveh governorate that it is unstable on average and that there is a very clear general trend in the series after the events experienced by the governorate especially after the use of biological and bacteriological weapons which increased the number of infected and acute shortage of drugs, and after taking the appropriate efficient model of data series which is the integrated Auto-regression model, where this model was successfully used to predict the number of patients with the disease for the period (2019-2020). It is recommended that the results of this research, which show an increase in the number of patients with stomach cancer over time, which requires taking the necessary measures by competent authorities.

References

1. Wheelwright, S. C., & Makridakis, S. (1973). An examination of the use of adaptive filtering in forecasting. *Journal of the Operational Research Society*, *24*(1), 55-64.
2. عبد الغفور جاسم سالم. (2006). دراسة السلسلة الزمنية لعدد حالات الإصابة بمرض سرطان الرئة في مدينة الموصل (1980-1990)، *مجلة تكريت للعلوم الصرفة*، 11(2)، 38-41.
3. Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, *65*(332), 1509-1526.
4. Anderson, R.I, (1942),"distribution of the series analysis correlation coefficient", Ann,Mat.Statistic,.13: 113-129
5. Crato, N. (1996). Some results on the spectral analysis of nonstationary time series. *Portugaliae Mathematica*, *53*(2), 179-186.

6. Wegman, E. J. (1996). Time Series Analysis: Theory, Data analysis and computation. *Lecture Notes, George Mason University*.

7. Wojbor,A.Woyczyan,ski(2006).A First course in statistics for signal analysis.birkhauser boston.
8. Kaiser, R., & Maravall Herrero, A. (2000). *Notes on time series analysis, ARIMA models and signal extraction*. Banco de España. Servicio de Estudios.
9. Makridakis, S. (1998). Forecasting: Methods and applications, edited by SC Wheelwright and RJ Hyndman.
10. Douglas, C. Mandcontreas,J.G.,(1976). Note on forecasting with adaptive filtering,O.P.Q,24,(4) :87-90.

IRAQI Academic Scientific Journals
Samarra Journal of Pure and Applied Science
SJPAS
مجلة سامراء للعلوم البحتة والتطبيقية
www.sjpas.com
ISSN:2663-7405

نمذجة السلاسل الزمنية للتنبؤ بأعداد المصابين بمرض الثلاسيميا في محافظة نينوى

شيماء رياض ذنون
كلية التمريض، جامعة الموصل، العراق (shaymaaalnuaimi@yahoo.com)

معلومات البحث:

تأريخ الاستلام: 2020/04/01
تأريخ القبـــول: 2020/05/12

الكلمات المفتاحية:

السلسة الزمنية ،التنبؤ، ARIMA

الخلاصة:

يهدف البحث الى تحليل السلاسل الزمنية لعدد الاصابات بمرض سرطان الثلاسيميا بوضع افتراضات على عدد الاصابات لصياغة المسألة لايجاد افضل نموذج للتنبؤ بأعداد المصابين في محافظة نينوى باستخدام طريقة (box and Jenkins) في التحليل وذلك بالاعتماد على البيانات الشهرية التي تم تزويدنا بها من مستشفى السلام في محافظة نينوى للفترة (2014 -2018) وقد اظهرت نتائج التحليل ان النموذج الملائم للتحليل هو نموذج الاحدار الذاتي المتكامل من الدرجة الثانية arima (2,1,0) وبالاعتماد على هذا النموذج تم التنبؤ بأعداد المصابين بهذا المرض لسنتين قادمتين حيث اظهرت النتائج قيما متناسقة مع القيم الاصلية مما يدل على جودة النموذج.