



RESEARCH ARTICLE - MATHEMATICS

The Use of Time Series Analysis to Foretell the Number of Patients with Malignant Neoplasms in Ninewah Governorate

Zahraa Tariq Mohammed Taher

Medicine College- Family and Community Medicine/ Ninevah University- Mosul-Iraq

* Corresponding author E-mail: zahraa.mohammed@uoninevah.edu.iq

Article Info.	Abstract
<i>Article history:</i> Received 25 May 2024 Accepted 2 September 2024 Publishing 30 September 2025	There are limited published data regarding the recent incidence trends of patients with malignant neoplasms in ninewah governorate. This study aims to analyze the time series properties using the (BOX & Jenkins) approach in the analysis (estimation, identification, and selecting the appropriateness of model for prediction). In this study the capabilities of A time series are employed to determine the best and most efficient statistical model for the purpose of prediction the number of people with this disease. The findings of the data analysis indicated that ARIMA (2,1,0) is the most suitable model for predicting the quantity of people with malignant neoplasms in Ninewah Governorate depending on the data for the period of (2016 - 2021). The results explained the type of this function are an nonstationary series on average, and there is a clear general trend in the series. The stationarity of the time series was achieved after taking the first difference of the data, and after matching the auto and partial correlation coefficients of the time series with the theoretical behavior of the auto and partial correlation. The number of patients with malignant neoplasms was foretold using this model.

This is an open-access article under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>)
The official journal published by the College of Education at Mustansiriya University

Keywords: Time Series Analysis, Malignant Neoplasms, Moving Average Model, Integrated Mixed Sample, Estimation Process.

1. Introduction

On the extent of the past thirty years, our beloved country has suffered from the disaster of wars that affected its material and human resources, especially the American attack that destroyed its infrastructure and polluted its air and land, which requires a comprehensive renaissance in all fields and the economy[1] It is also a major role in the destruction of the infrastructure, so it is outdated to address the health aspect of its importance at the developmental level, because it is the human element that bears the responsibility for construction and reconstruction and keeping pace with progress and civilizational development [2][3]. The prevention of all illnesses, including malignancies, which account for a disproportionately high number of fatalities as compared to other diseases, is a crucial part of creating health. In view of the recent increase in the number of people that infected with this disease, this study came in order to reveal this phenomenon, which has worsened in Ninewah Governorate, which is one of the governorates that affected by bacterial and biological weapons, and the acute shortage of health and treatment care as a result of destroying most of its health centers [4]. The study depended on the monthly data of the number of patients with malignant neoplasms for the period of (2016 – 2021) as a time series for the purpose of analyzing it in order to reach the best model to predict this disease in later periods of taking measures that needed to minimize this phenomenon in the future [5].

2. Study Objectives

A time series function study are engaged to determine the best and most efficient statistical model for the purpose of predicting the number of people with malignant neoplasms in Ninewah Governorate for the period of (2016-2021).

3. Basic Concept for the Statistical Model

This study depended on descriptive statistics, which dealt with the (Box and Jenkins) map in time series analysis (identification, estimation, experiencing the suitability of the diagnosed model, and future foretelling), and supporting the study course for the theoretical aspect with the applied aspect, which depended on true data on the number of people with malignant neoplasms in Ninewah Governorate to attain the best mathematical sample for predicting the number of persons with this illness for later periods [6]. The last part of the research included the significant conclusions and recommendations, then the appendices, sources and tools are used for the statistical program (Minitab) and (SPSS V.10). To verify this statistic model, it is important to include the following parts [2] [7].

3.1 Theoretical Part

This item deals with some general concepts and presents the stages of building a time series model. The time series analysis depends on the algorithm drawn by the two researchers (B-J, 1976) that which begins with the first stage, which is the identification of the suitable sample for the data, and then the level of rating the parameters of the diagnosed sample. After that, the stage of examining the suitability of the diagnosed model comes, and if the model is appropriate, the last stage comes, which is the future foretelling [8].

3.2 Time Series

We can determine time series as a collection of observations of the appeared values that are taken at specific times (the intervals between observation and the next one may be equal or unequal, and in most cases they are equal). It is expressed if it is equal $(Z_{t1}, Z_{t2}, \dots, Z_{tn})$ at the time period $t_1, t_2, \dots, t_n$, where n represents the number of observed values, and the statistical time series can be represented as follows [9]:

$$Z_t = f(t) + a_t \quad t = 0, \pm 1, \pm 2, \dots \dots \dots (1)$$

Where:

$f(t)$ represents the regular that can be expressed by a mathematical function.

a_t represents the random part, also called noise. The time series may be of the specified type as shown in equation (2)[10]:

$$Z_t = Z_{t+s} \quad \forall t \quad t = 0, \pm 1, \pm 2, \dots \dots \dots (2)$$

So, (S) is the period of the series. To distinguish between two types of stationary and nonstationary time series, where there are two states of stationary, which are (Stationary in Mean) and (Stationary in Variance), where stationary in mean is the state of the series when it does not show a general direction, so it can be converted to stationary by using the differences [11].

3.3 Moving Average Model (MAM)

Moving average model with degree (q) and backscatter factor (B) can be formed as follows [12]:

$$Z_t = \Phi_0 + (I - \phi_1 B - \phi_2 B^2 - \dots - \phi_q B^q) a_t \quad \dots \dots \dots (3)$$

While the main form of the model are illustrated in equation (4):

$$Z_t = \Phi_0 + a_t - \phi_1 a_{t-1} - \phi_2 a_{t-2} - \dots - \phi_q a_{t-q} \quad \dots \dots \dots (4)$$

That is, ϕ_i is the parameter of the moving average model ($i=1,2,3,\dots,q$), and its value ($-1 < \phi < 1$) while (q) represent model degree. The autocorrelation indicator of the sample (MP) approaches zero after the displacement (q), and at the same time the partial autocorrelation indicator (PACI) diminishes exponentially [13].

3.4 Integrated Mixed Sample

Some time series samples can be stationary on their own, but become stable after many transitions or variations. Thus, the model that expresses this process differs from the original sample because it must contain the transformations or variations that have been made on these stationary models, which are called merged mixed samples. ARIMA models are extremely the most used time series samples because all samples can be derived from them, whether they are autoregressive, moving average, or mixed average. These forms are consisting of three parts [14]:

Part 1: Autoregressive sample commonly used in time series foretelling.

Part 2: It represents the moving average model.

Part 3: It shows the differences the chain needs to be stationary.

Therefore, it expresses the non-seasonal (ARIMAM) samples according to the ARIMA formula (p, d, q), that is:

P : is AR(P) rank

q : is MA(q) rank

d : is the number of differences through which the series becomes stationary.

By using the backscatter factor (B) for the following formula:

$$\Phi(B) = (1 - B)^d X_t = \Phi_0 + \phi(B)a_t \dots\dots\dots(5)$$

It will be:

$$\Phi(B) = (1 - \Phi_1 B - \dots - \Phi_p B^p) \dots\dots\dots(6)$$

$$\phi(B) = (1 - \phi_1 B - \dots - \phi_q B^q) \dots\dots\dots(7)$$

It becomes the basic form of the integrated mixed model as shown in equation (9).

$$Z_t = \Phi_0 + \Phi_1 Z_{t-1} + \dots + \Phi_p Z_{t-p-d} + a_t - \phi_1 a_{t-1} + \dots + \phi_q a_{t-q} \dots\dots\dots(9)$$

According to previous equations we can consider the ARIMA models to be the same stationary ARMA models with the difference in rank.

4. Building a Time Series Model

Building the time series model takes place in four stages, which are: “diagnosing the suitable sample for the data, estimating the information of the specific sample and testing the suitability of the specified model for future foretelling as explained below.

4.1 Defining the Model

Defining time series models is the most important step in building or diagnosing time series and the first stage of the algorithm for which a basis has been laid. The defining or diagnosing stage must cross the data preparation stage. If the data is stationary by observing the drawing of the original data and its partial and autocorrelations, then the data is prepared for identification or defining [15]. In case that the data is nonstationary in the mean and variance, then the nonstationary in the mean is treated, by taking the first difference ($d = 1$), but if it is nonstationary, we take the second difference ($d = 2$), and sometimes it settles in the first difference. As for the nonstationary in the variance, it is dealt with

by conducting the appropriate conversion of the data. When it is stationary, the time series begins the process of defining the sample. The goal here is to get an idea of the values q , d , p , so you need them in the general linear model ARIMA, whose formula is shown in the equation (1-7). After that, initial estimates of the model parameters are obtained. The two functions used to diagnose the model and determine its degree are the autocorrelation indicators (ACI) and the partial autocorrelation indicator (PACI), that is, they are drawn graphically, and then the autocorrelation and partial coefficients are matched with their theoretical behavior [16]. if it was:

- The autocorrelation indicator decreases gradually and exponentially or the behavior of the diminished sine function and stating the partial autocorrelation indicator is discontinued after the displacement (P), so the appropriate model for the data is AR (P).
- The autocorrelation indicator is stopped after the displacement (q) and the partial autocorrelation indicator is gradually decreasing exponentially or in the manner of the diminishing sine function. The appropriate model for the data is MA(q).
- If the autocorrelation and partial indicator decreases gradually and exponentially or the behavior of the diminishing sine function, then the appropriate model for the data is ARMA (p , q).

4.2 Estimation Process

The process of estimating the model is the second stage of studying and analyzing the time series and comes after the process of determining the suitable sample for the time series. In order for the sample to achieve the main objective of its construction, which is foretelling, we must ensure the quality of its estimation and its suitability for the time series. There are many ways to estimate model parameters, the most important are [17]:

1. The Ordinary Small Squares Method: This method works on the basis of reducing the sum of the squares of the estimation error, and making it at its smallest end.
2. The Method of Greatest Possibility: In which the matrix of the parameters of the model to be estimated is determined according to the principle of maximizing the possible function.

4.3 Testing the Suitability of the Model

After estimating the model, the validity of the model must be tested to represent the seasonal time series data. There are several methods for this, including:

- a. Suitability of the model: It must be statistically and significantly indicated, that is, it should not be close to zero, and for this, the t-test can be used. If it is non-significant, then one of the ranks AR or MA must be excluded.
 - b. Residual Analysis: The following tests may be used as in:
1. Testing Trust Limits: The values of the autocorrelation indicator for the estimated residuals are $a_t^\wedge$, which must be between $(\mp 1.96/\sqrt{n})$ with probability (0.95). If this is achieved, the residuals are distributed randomly, and the model provides an adequate representation of the data and can be used for foretelling, and the autocorrelations of the residuals are distributed normally with an arithmetic mean of zero and variance $(1/n)$. Also, the AIC information criterion (AIC) will be used to test the best sample whose variance is weak and decreases with the increase in the number of parameters and the sum of the few residuals. The AIC criterion is defined by the following mathematical formula [17]:

$$AIC_{(p)} = \ln(\alpha^2) + \frac{2(p+q)}{n} \quad \dots \dots \dots (10)$$

Where: α^2 represents the variance of the model and $(p+q)$ represents the number of features.

The above formula has been reformulated to give more weight to the models used for the largest number of observations:

$$MAIC = \frac{AIC}{n} \dots\dots\dots(11)$$

Schwartz criterion can be used according to the following mathematical formula:

$$SBC = \text{Log}(n) \left(\frac{p+q}{n} \right) \dots\dots\dots(12)$$

2. The method that depends on Q(Ljung & Box) in order to choose the null hypothesis that states:

$$H_0 = \rho_1 = \rho_2 = \dots = \rho_s = 0$$

based on the autocorrelation of the residuals. Where its mathematical formula is [18]:

$$Q_S = n(n+2) \sum_{k=1}^s \frac{1}{n-k} r_k^2(a) \dots\dots\dots(13)$$

The test measure Q is distributed in an X^2 form.

m = number of coefficients

K = number of displacements

S = the largest displacement that is taken.

Practical Part:

Data were collected, which consist of a time series of about (60) observations from January 2017 to December 2021. This data includes the number of patients with malignant neoplasms, which were obtained from Mosul Hospital for Oncology and Nuclear Medicine. As shown in Table (1):

Table 1.: Number of Patients with Malignant Neoplasms in Ninewah Governorate

2021	2020	2019	2018	2017	Year
					Month
2997	867	2052	2814	993	January
2588	1003	2154	1668	1010	February
2372	1757	1613	3160	646	March
2632	1258	2450	2866	1144	April
4242	2352	2302	2262	843	May
3100	1967	3502	1817	1092	June
2962	3187	5401	3761	131	July
3008	2500	2950	3283	493	August
6629	5837	5253	2064	972	September
5855	4124	2631	2555	1044	October
7130	4907	5724	4510	961	November
7392	6342	6002	3615	795	December
50907	38255	42035	34379	10124	Total

5. Data Initialization

From this data a scatterplot is drawn. The autocorrelation and partial coefficients are also extracted and the special confidence limits of the autocorrelation indicator are drawn for this data, using the statistical program (Minitab). Figure (1) shows the following [19].

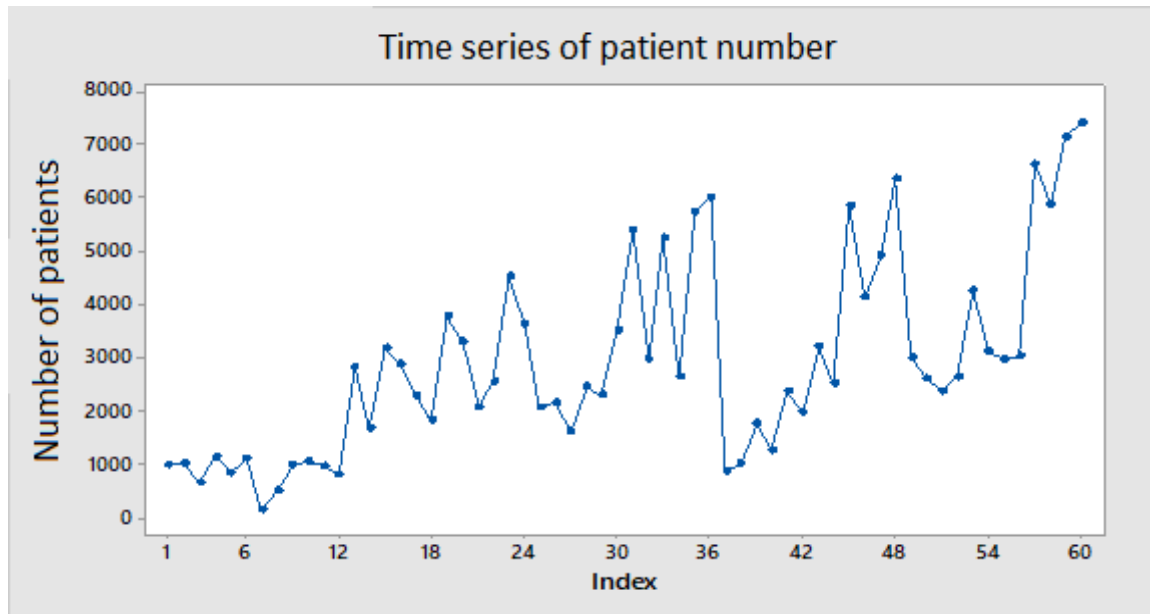


Fig.1. Curve of the number of patients with malignant neoplasms for the period of (2017-2021)

Through Figure (1), we notice that the variance tends to be stationary, but a general trend in a state of increase and decrease with time, which indicates the nonstationary of the series on average. This has been confirmed by the values of the autocorrelation and partial coefficients as shown in Figure (2) [22].

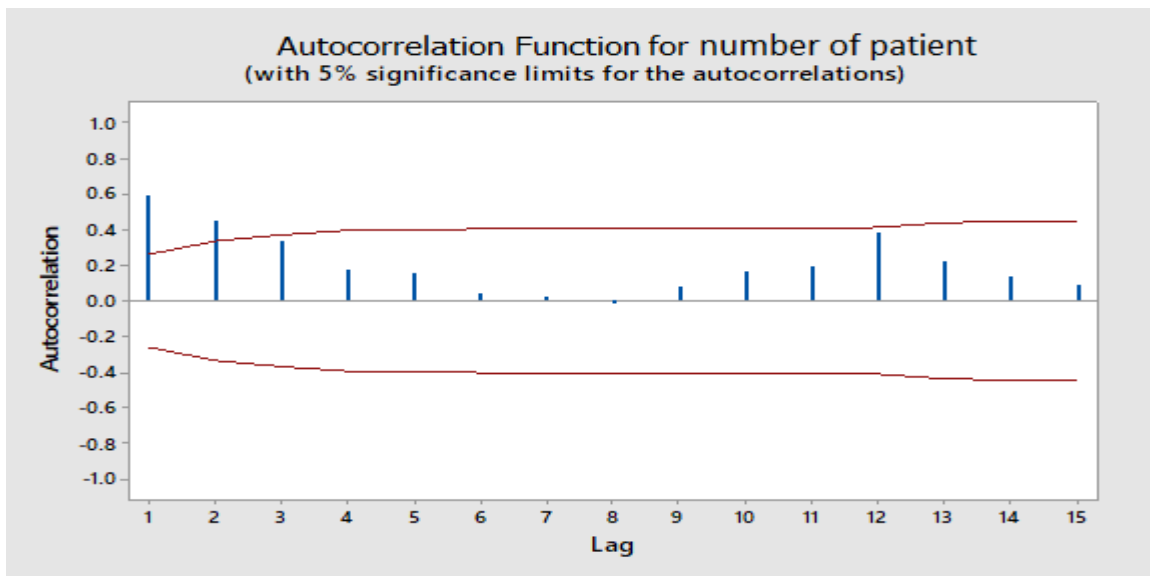


Fig.2. Autocorrelation and partial coefficients of the time series

In Figure (2), the values of the autocorrelation coefficients and partial up to lag (15) are significantly different from zero, and in order for the series to be stationary, all values of the autocorrelation coefficients of the sample must be entered within the trust limits, except for the first or second displacement, which may fall outside the trust limits for the data to be at a level of (95%), which are $(-0.41 < r_k < 0.41)$. This confirms the nonstationarity of the time series on average. Thus, we reject the null hypothesis, which indicates that the autocorrelation coefficients are equal to each other and equal to zero. We accept the alternative hypothesis, which means that the time series is nonstationary [20]. To address this case, we have to take the first difference of its data ($\nabla X_t = (X_t - X_{t-1})$) and Figure (3) shows the drawing of the curve (X_t^*):

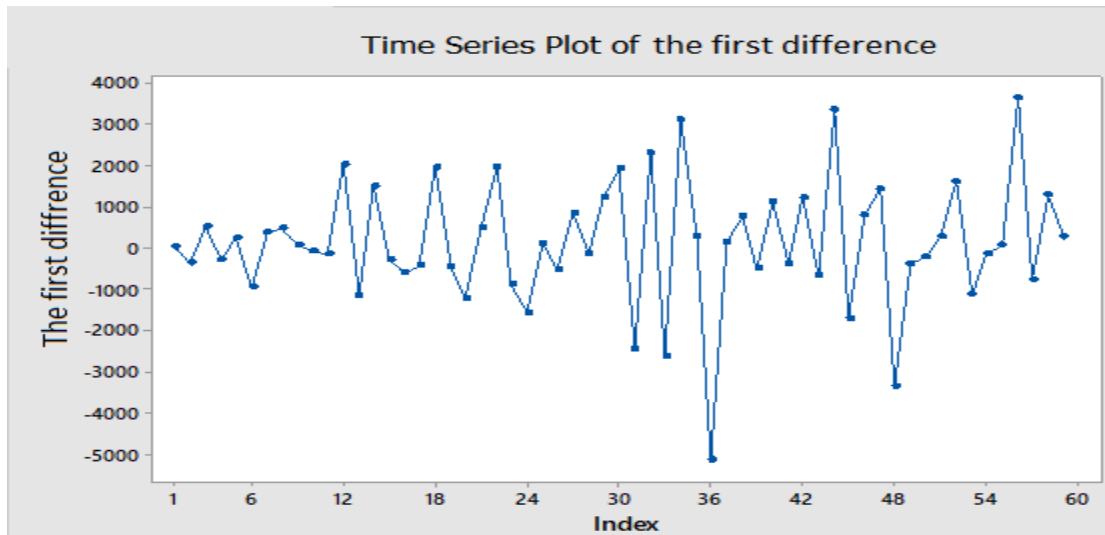


Fig.3. Time Series Curve After Taking the First Difference

In Figure (3) the loss of the general trend in its behavior, which indicates the stationarity of the series in the average (the displacement of general trend) and that the values of the autocorrelation and partial coefficients confirmed as shown in Figure (4).

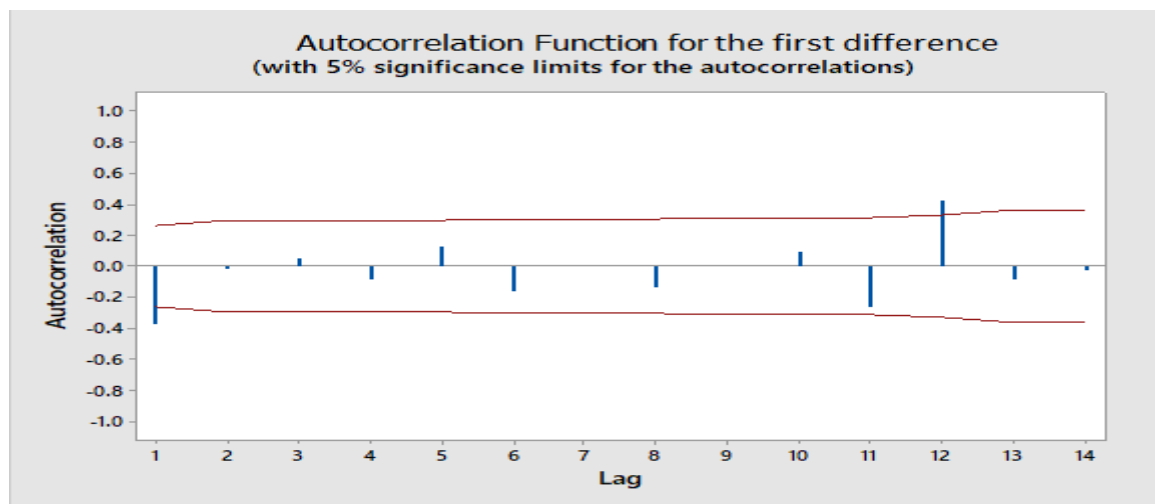


Fig.4. The Values of Autocorrelation and Partial Coefficients for the Time Series After the First Difference

Through figure (4), we notice that all the auto and partial correlations of the sample are within the limits of trust and that they are significant at the first and second limits, and this confirms the stationarity in the average and the absence of seasonal effects in the series, so the data became ready to be applied in the first stage of studying and analyzing time series models[10,21].

6. Predicting Future Values

In this research, the model (2,1,0) was used to foretell the number of patients with malignant neoplasms in Ninewah Governorate for the period of (2022-2023). As shown in Table (2) [21].

Table 2. The foretell of the patient's number with malignant neoplasms for the period of (2022-2023)

Month \ Year	2022	2023
January	7392	23199
February	8131	25519
March	8944	28071
April	9838	30878
May	10822	33966
June	11904	37362
July	13095	41098
August	14404	45208
September	15845	49729
October	17429	54702
November	19172	60172
December	21090	66190

Figure (5) shows the time series data representation of these predictions, which showed the same method as the original series.

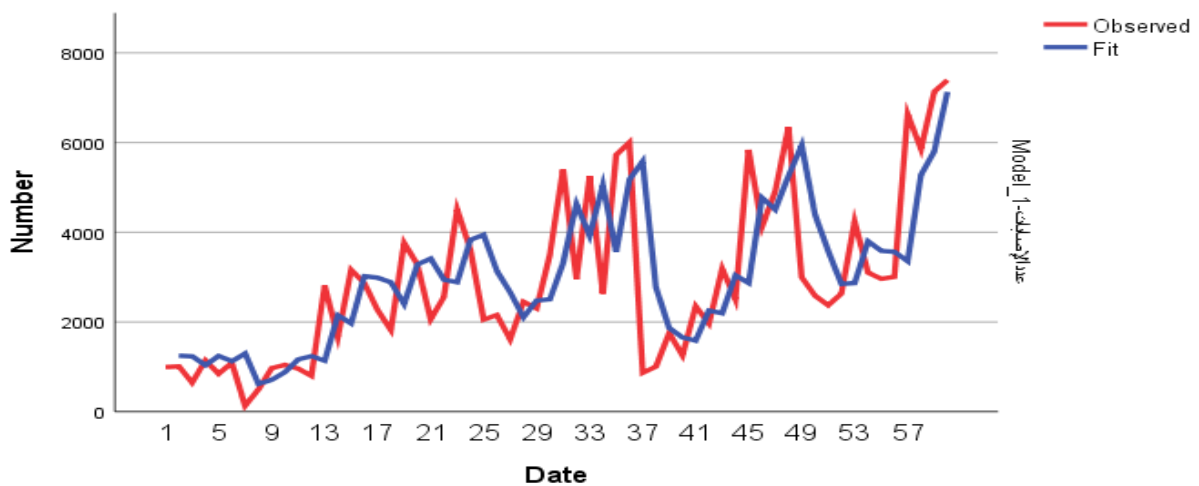


Fig.5. The Foretelling Curve of the Time Series

6.1 Predicting Accuracy Tests

To test the foretelling accuracy between the two models, we compare them according to the scale of foretelling accuracy. Table (3) shows this vision [22][12].

Table 3. Shows the difference between the two models depending on the foretelling accuracy criterion

Model (1,1,0) ARIMA	Model (2,1,0) ARIMA	Criterion
1434.459	1409.679	PMSE
999.252	1008.08	MAE
59.736	58.705	MAPE

In Table (3) it is shown that the model (1,1,0) records the highest values for squares average (999.252), absolute values average for error (59.736), and absolute values average for error

(1409.679). We conclude that the model gave the best fit for the series according to accuracy measures of the foretelling that used above is model (2,1,0)[12] [23].

7. The Results Analysis

1. The vibrant general results from this study can be represented by the series numbers of patients with malignant neoplasms in Ninewah Governorate, explained an nonstationary series on average. After taking the first difference of the data, and after matching the auto and partial correlation coefficients of the time series with the theoretical behavior of the auto and partial correlation. The stationarity of the time series was achieved. It has been shown that the autocorrelation indicator gradually decreases with the increase in the displacement periods and it is in the form of a sine wave when the partial and autocorrelation indication is observed after the second displacement. Furthermore, the results verified that such model, are suitable for the series data, which is the ARIMA (2,1,0) integrated autoregressive model. This model was used to foretell the number of people with malignant neoplasms in Ninewah Governorate for the period of (2022-2023), as the results showed that the foretelling values were consistent with the original values of the series.

8. Recommendations

1. Depending on the results of this research, which shows an increase in the number of people with malignant neoplasms over time, which must be taken by the competent authorities that are able to take a limit on this matter because the governorate hospitals lack devices for early detection of this disease.
2. Generalizing this study to identical studies in the governorates that were exposed to conditions that are similar to Ninewah Governorate, in order to present a comparison between them.

References

- [1] A. Wahaib Abdallah and S. Siham Dawood, "The Impact of Gross Domestic Product Response to the Money Supply Shock in the Iraqi Economy for the Period (2004-2021)," *J. Econ. Adm. Sci.*, vol. 29, no. 137, pp. 131–145, 2023, doi: 10.33095/jeas.v29i137.2758.
- [2] U. S. Srinivasan, V. Pavithra, K. Sutha, S. Ramachandiran, and N. Indumathi, "An Innovative Analysis of Time Series-Based Detection Models for Improved Cancer Detection in Modern Healthcare Environments †," *Eng. Proc.*, vol. 59, no. 1, pp. 1–9, 2023, doi: 10.3390/engproc2023059114.
- [3] Cancer Information Service, "Cancer Statistics in Japan 2013," p. 22, 2013, [Online]. Available: http://ganjoho.jp/en/professional/statistics/brochure/2013_en.html
- [4] W. Leneenadogo and S. P. U, "A Comparative Study of Fourier Series Models and Seasonal - Autoregressive Integrated Moving Average Model of Rainfall Data in Port Harcourt," *Asian J. Probab. Stat.*, no. January, pp. 36–46, 2021, doi: 10.9734/ajpas/2020/v10i330249.
- [5] M. Ekşi *et al.*, "Machine learning algorithms can more efficiently predict biochemical recurrence after robot-assisted radical prostatectomy," *Prostate*, vol. 81, no. 12, pp. 913–920, 2021, doi: 10.1002/pros.24188.
- [6] R. Ali, O. Sameer, and Q. Hassan, "Stability and Stabilization of Integro-Differential Perturbed Nonlinear System Lemma (2 . 1) : The integro-differential nonlinear problem (1) is equivalent to the double integral equation Proof: Let us assume that satisfies the integro-differential n," *Mustansiriyah J. Pure Appl. Sci.*, vol. 2, no. 2, pp. 8–27, 2024.
- [7] N. Islam *et al.*, "Excess deaths associated with covid-19 pandemic in 2020: Age and sex disaggregated time series analysis in 29 high income countries," *BMJ*, vol. 373, 2021, doi: 10.1136/bmj.n1137.
- [8] W. Zhu, L. Xie, J. Han, and X. Guo, "The application of deep learning in cancer prognosis prediction," *Cancers (Basel)*, vol. 12, no. 3, pp. 1–19, 2020, doi: 10.3390/cancers12030603.

- [9] P. Ström *et al.*, “Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study.,” *Lancet. Oncol.*, vol. 21, no. 2, pp. 222–232, Feb. 2020, doi: 10.1016/S1470-2045(19)30738-7.
- [10] M. Alahmadi, P. Atkinson, and D. Martin, “Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data,” *Comput. Environ. Urban Syst.*, vol. 41, pp. 167–176, 2013, doi: <https://doi.org/10.1016/j.compenvurbsys.2013.06.002>.
- [11] Z. S. Ahmed and S. S. Mahmood, “New Formula for Conjugate Gradient Method to Unconstrained Optimization,” *Mustansiriyah J. Pure Appl. Sci.*, vol. 1, no. 2, pp. 21–27, 2023.
- [12] N. Tomašev *et al.*, “A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury,” *U.S. Dep. Veterans Aff.*, vol. 572, no. 7767, pp. 116–119, 2020, doi: 10.1038/s41586-019-1390-1.A.
- [13] S. K. H. Bukhari, A. Jalil, and N. H. Rao, “Detection and Forecasting of Islamic Calendar Effects in Time Series Data: Revisited,” *State Bank Pakistan, Work. Pap. Ser.*, no. 39, 2011.
- [14] L. Xie, “Time Series Analysis and Prediction on Cancer Incidence Rates,” *J. Med. Discov.*, vol. 2, no. 3, 2017, doi: 10.24262/jmd.2.3.17030.
- [15] E. Y. Hamid and M. H. Osman Abushaba, “Use of Exponential Holt Model and the Box-Jenkins Methodology in Predicting the Time Series of Cement Production in Sudan,” *J. Al-Qadisiyah Comput. Sci. Math.*, vol. 12, no. 1, 2020, doi: 10.29304/jqcm.2020.12.1.666.
- [16] G. T. Wilson, “Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1,” *J. Time Ser. Anal.*, vol. 37, no. 5, pp. 709–711, 2016, doi: 10.1111/jtsa.12194.
- [17] S. Halabi *et al.*, “Prognostic Model for Predicting Survival in Men With Hormone-Refractory Metastatic Prostate Cancer,” *J. Clin. Oncol.*, vol. 21, no. 7, pp. 1232–1237, Apr. 2003, doi: 10.1200/JCO.2003.06.100.
- [18] A. G. E. P. Box, D. A. Pierce, and G. E. P. Box, “Distribution of Residual in Autoregressive-Autocorrelations Integrated Moving Average Time Series Models,” vol. 65, no. 332, pp. 1509–1526, 2011.
- [19] L.-M. Liu and G. Hudak, “Forecasting and Time series analysis usng the SCA statical system,” *Sci. Comput. Assoc.*, vol. 142, no. 3, p. 614, 1992, [Online]. Available: http://books.google.com/books?id=aY0QAQAIAAJ&q=inauthor:wei+1990&dq=inauthor:wei+1990&hl=&cd=3&source=gbs_api
- [20] S. C. Wheelwright and S. Makridakis, “Foregasting with adaptive filtering,” *Rev. française d’automatique, informatique, Rech. opérationnelle. Rech. opérationnelle*, vol. 7, no. V1, pp. 31–52, 1973, doi: 10.1051/ro/197307v100311.
- [21] C. Fitzmaurice *et al.*, “Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study.,” *JAMA Oncol.*, vol. 5, no. 12, pp. 1749–1768, Dec. 2019, doi: 10.1001/jamaoncol.2019.2996.
- [22] G. Gravis *et al.*, “Prognostic Factors for Survival in Noncastrate Metastatic Prostate Cancer: Validation of the Glass Model and Development of a Novel Simplified Prognostic Model,” *Eur. Urol.*, vol. 68, no. 2, pp. 196–204, 2015, doi: <https://doi.org/10.1016/j.eururo.2014.09.022>.
- [23] R. Ali, O. Sameer, and Q. Hassan, “Stability and Stabilization of Integro-Differential Perturbed Nonlinear System Lemma ” , *Mustansiriyah J. Pure Appl. Sci.*, vol. 2, no. 2, pp. 8–27, 2024.