Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 1391–1415
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5745

A statistical approach and analysis computing based on autoregressive integrated moving averages models to predict COVID-19 outbreak in Iraq

Ali Abdul Karim Kazem Naji[a,*], Asmaa Shaker Ashoor[b]

[a] *College of Education for Pure Science, University of Babylon, Iraq*
[b] *College of Basic Education, University of Babylon, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

A time series has been adopted for the numbers of people infected with the Covid-19 pandemic in Iraq for a whole year, starting from the first infection recorded on February 18, 2020 until the end of February 2021, which was collected in the form of weekly observations and at a size of 53 observations. The study found the quality and suitability of the autoregressive moving average model from order (1,3) among a group of autoregressive moving average models. This model was built according to the diagnostic criteria. These criteria are the Akaike information criterion, Bayesian Information Criterion, and Hannan & Quinn Criterion models. The study concluded that this model from order (1,3) is good and appropriate, and its predictions can be adopted in making decisions.

Keywords: Autoregressive Models, ACF, PACF, COVID-19, Unit Root Test.

1. Introduction

The spread of diseases threatens human societies and limits their development and growth. The Covid-19 pandemic is considered a global characteristic in terms of its impact, which has left many negative effects on various fields of life due to the growing number of infected people, as well as the emergence of advanced strains of the virus. Hence the importance of studying the phenomenon from a statistical point of view using time series by analyzing the behavior of the numbers of people infected with it, as well as predicting the future levels of the numbers of people infected with the

*Corresponding author
Email address: ali.babylon1982@gmail.com (Ali Abdul Karim Kazem Naji)

Received: May 2021 *Accepted:* October 2021

Covid-19 pandemic in Iraq, which leads to develop strategic plans that would limit the spread of the epidemic.

There has been an increasing interest in analyzing time series and methods for predicting their future values. It was considered one of the most prominent statistical methods used in forecasting for many applications and scientific fields. Therefore, forecasting is an important issue for a long time, and this topic has remained the focus of researchers in all fields. There found the process of forecasting the future values of time series, which represents a particular phenomenon, is useful for planning and decision-making, whether at the regional or international level.

The prediction method for the time series can be divided into many directions predictions based on proposed mathematical models such as the Box-Jenkins model (Autoregressive Integrated Moving Averages Models) which needs to achieve a set of hypotheses about the distribution of errors and achieve stability and the size of the sample is relatively large.

The use of Box-Jenkins models in different fields were developed to represent stationary and non-stationary time series. So, these models are useful to predict the future values of these series. Moreover, they give good predictions when the time series has a linear component, which can be represented by a linear model. However, when the series has a non-linear component, these predictions are accompanied by an increase in the prediction error. It can be said that autoregressive integrated moving averages models give predictions with high accuracy when there are linear relationships for the system of parameters, but the prediction fails if the time series has a non-linear component.

2. Related Work

Hernandez-Matamoros, and et. al. in [2] presented a predictability of virus behavior that can be used to plan future response. This work presented an analysis of COVID-19 by 6 geographic regions (continents) which are in the same geographic region to predict the progression of the virus. Countries in the same geographic area have quantifiable and non-quantifiable variables which influence the spread of the virus. ARIMA model template is built for 145 countries that are distributed in 6 regions. Next, the researcher builds a model for these regions using ARIMA parameters.
The main result of this research has shown that the opportunity to create more models to predict COVID-19 behavior using variables such as humidity, climate, culture, etc.

Panda in [**?**] used ARIMA and Holt-Winters models to predict the number of people infected with the covid-19 pandemic in the short term for India and its six states such as Odisha, Delhi, Maharashtra, Karnataka, Andhra Pradesh and West Bengal. The model is evaluated using correlation chart and ADF, AIC and MAPE test to understand the accuracy of the proposed prediction model.

Ribeiro, M. H. D., and et. al in [ ] used autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR), and stacking-ensemble learning are evaluated in the task of time series forecasting with one, three, and six-days ahead the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily inci- dence. In the stacking-ensemble learning approach, the CUBIST regression, RF, RIDGE, and SVR models are adopted as base-learners and Gaussian process (GP) as meta-learner. The models' effectiveness is evalu- ated based on the improvement index, mean absolute error, and symmetric mean absolute percentage error criteria. In most of the cases, the SVR and stacking-ensemble learning reach a better performance regarding adopted criteria than compared models. In general, the developed models can generate accu- rate forecasting, achieving errors in a range of 0.87%–3.51%, 1.02%–5.63%, and 0.95%–6.90% in one, three, and six-days-ahead, respectively. The ranking of models, from the best to the worst regarding accuracy, in all scenarios is SVR, stacking-ensemble learning, ARIMA, CUBIST, RIDGE, and RF models. The use of eval- uated models is recommended to forecasting and monitor the ongoing growth of COVID-19 cases, once these models can assist the managers in the decision-making support systems.

Theoretical Background:

3. Box-Jenkins Time Series Models [12, 9, 14]

The statistical phenomenon that arises during time according to the probabilistic law is called a random process, including the time series. The time series are a group of observations linked to each other that is observed consecutively with time t, or the time series is defined as a group of observations in equal intervals of time mostly and for a period of time, which arises from its relationship to time according to the probabilistic law. The time series can be classified according to the properties of the random variable into:

- Continuous time series as the series observations recorded continuously from its relationship with time , $\{x_t,\ t \in T\}$. This case appears when the series is continuous, but the observations for this series were obtained and recorded at fixed time periods such as the time series of the human voice where the time period is divided into small periods and after this takes the observation at a certain fixed time to measure the sound characteristic.

- Discrete time series in which the observations of the time series were recorded at fixed intervals of time $\{x_t\ , t = \ \mp 1, \mp 2, \ldots\ \}$. Also from a statistical point of view, the time series may be stationary, and this means that its data fluctuates around a constant mean of the series, that is, there is no change in its arithmetic mean and in its variance. Therefore the stationary series has a constant mean and variance that does not depend on time t. The time series is either strictly stationary jointly distribution, or that the time series is stable of the second degree, meaning that it has weakly stationary with the following characteristics:

- $E\left(X_t\right) = \ \mu \ for\ all\ t$
- $\gamma_k = cov\left(X_t, X_{t+k}\right) = E\left(X_t - \mu\right)\left(X_{t+k} - \mu\right) \qquad k = 0, \pm 1, \pm 2...$

Where k represents the lags, which is the time interval between observations.

3.1. Stationary in the Time Series [10, 5, 7]

Stationary is achieved in the time series before starting the analysis process and building appropriate models for it, and this is done by make stationary in variance and mean. There are two main types of power and logarithmic transformation to achieve stationary in variance transformation that are used for original time series. Also, the stationary of the mean is achieved through several methods, like estimating models, (linear, second-order polynomial and exponential model) and dealing with their errors as a stationary time series but the most important of which is the difference method.

The nonstationary in mean removed from the time series using the method of difference, where d is taken from the appropriate differences to the original data, the difference factor is symbolized by ∇, and when applied to the time series x_t , we get the first difference in relation:

$$\nabla\ x_t = x_t - \ x_{t-1}$$

The non stationary time series often turns into a Stationary one after the second difference, in general the d difference defines:

$$\nabla^d \; x_t = (1-B)^d x_t$$

Stationary in mean is detected by examining the autocorrelation function coefficients (in short, acf) acf if the majority of the autocorrelation coefficients fall within the limits:

$$-2 \;\; se_e \leq \; r_k \leq \;\; 2 \; se_e$$

where se_e Standard error of the autocorrelation coefficients.

So we conclude that the series is stationary in mean then that the autocorrelation coefficients enter within the above limits after the second or third lags, provided that they do not show a general trend. But if these coefficients do not behave this behavior, then this is evidence that the series is unstationary and this should be treated in order to convert it into a stationary series.

3.2. Box-Jenkines Methodology [10, 1, 2, 13]

Identification

Draw the time series data to note the basic series components such as the general trend, seasonal effect, or cyclical effect, then calculate the autocorrelation coefficients, partial auto-correlation and any other information with the drawing and choose items from the ARIMA model in order to estimate an appropriate value to the order of the model p, d, q.

Estimation

The model parameters are estimated, the parameters of the autoregressive model and the parameters of the moving average model and for the selected model according to the order diagnosed for the model in the previous stage, by any appropriate method.

Diagnostic Checking

The successful model is examined in terms of its acceptability through statistical tests and its relationship to errors, considering that the errors of the model are a random series where the autocorrelation factors are calculated and tested statistically in order.

Forecasting

Use the estimated model parameters to describe the model and then calculate the predictive values for the time series, whether point or interval prediction with efficiency measures of predictive values. The following figure explain the stages of building Box-Jenkins models.

3.3. Autocorrelation Function ACF [10, 5, 7, 12]

Box and Jenkins models are diagnosed by testing the behavior of the autocorrelation function, and partial auto-correlation function, to the values of the stationary time series observations. As these values represent the original time series, they may represent the transformed time series through logarithmic transform formulas, or they may represent the time series after taking appropriate differences into the series in order to convert it to the stationary state and based on the values of the treated time series (Working Series). The autocorrelation coefficients of the sample (in short sac) at the lags k are defined r_k and estimated by the formula:

$$r_k = \frac{\sum_{t=1}^{n-k} ((x_t - \overline{x})(x_{t+k} - \overline{x}))}{\sum_{t=1}^{n} (x_t - \overline{x})^2}, \qquad k = 1, 2, 3, \ldots, \left(\frac{n}{4}\right)$$

If the value of the autocorrelation coefficient is close to 1, then this means that the observations separated by the lag k have a strong tendency to move together in a linear and with a positive slope, but if the value of the autocorrelation coefficient is close to -1, this means that the observations separated by the time lag k have a strong direction to move with us, linearly as well, but with a negative slope.
The autocorrelation coefficients are often calculated based on autocovariance, which we denote by γ_k and are calculated by the formula:

$$\gamma_k = (n-k)^{-1} \sum_{t=1}^{n-k} ((x_t - \overline{x})(x_{t+k} - \overline{x}))$$

At that, Autocorrelation is estimated by the relationship:

$$r_k = \gamma_k / \gamma_0 \qquad , \quad k = 1, 2, 3, \ldots \ldots , (\frac{n}{4})$$

Where k represents the lags, which is the time interval between observations, γ_0 represents the variance of the time series. Autocorrelation is widely used in time series analysis and when it is mapped to it is called (Correlagram) or (autocorrelation plot) , which is used during the identification diagnosis stage to determine the order for the Box-Jenkins models, and that the standard error of a sample of the autocorrelation coefficients is equal $se_{r_k} = n^{-0.5}$and thus the approximate confidence limits equal to $\mp 2n^{-0.5}$ the auto-correlation coefficients r_k are statistically significant if they fall outside the confidence limits.

3.4. Partial Autocorrelation PACF [13, 7, 9]

Partial autocorrelation (in short pacf) is a measure of the degree of relationship between observations of the time series,x_t and ,x_{t-l} when the effect of other time periods is fixed. Partial autocorrelation coefficients are estimated depending on the autocorrelation coefficients r_k and are used in analyzing the time series and helping to determine the type and degree of the model. Through the synchronization relationship between the autocorrelation coefficients and the partial autocorrelation, the group of partial autocorrelations at specific periods denote $k = 1, 2, 3, \ldots,$ its estimates with symbol $\emptyset_{kk}$ and these coefficients are calculated by the formula:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}}$$

Where replaced ρ_i by the estimates r_i and these coefficients for partial autocorrelations are calculated by the following formulas:
Or, the Partial Autocorrelation function PACF may be computed with the following formula and with the lag k:

$$\emptyset_{kk} = r_1 \qquad\qquad if \quad k = 1$$

$$\emptyset_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j}\ r_j}{1 - \sum_{j=1}^{k-1} r_{k-1,j}\ r_j} \qquad if \quad k = 2,3,4 \ldots\ldots\ldots\ldots\ldots (1.8)$$

Where as

$$r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j} \qquad\qquad for\ j = 1,2,\ldots,k-1$$

Whereas, these parameters can be in the form of a list in addition to figures in a graph that effectively illustrates the behavior of the partial autocorrelation function. As for its test, it is through the t-test according to the relationship, when this relationship is achieved, it indicates that there are no significant differences in partial autocorrelation when the k displacement is, and that the best suggestion to number of k is what was presented by Box - Jenkins (1976), who suggested that n be at least equal to 50 and
$\#\ k = [0.25\ n]$.

3.5. Unit Root Test[13, 5, 12, 14]

The unit root test aims to examine the properties of the time series through the time period of observations, to ensure their stationary, and to determine the order of integration if the original time series is stationary at its original values it is said to be an integral of order zero I (0). In general, the time series x_t is integral of degree d if it stationary after taking the difference d is said to be an integral of order I(d).

3.5.1. Simple Dickey - Fuller Test

The test equation is described by the first difference of the time series:

$$\nabla x_t = \mu + (\emptyset - 1)x_{t-1} +$$

The parameters model can be estimated by using Ordinary Least Square (in short OLS), also the equation called the simple Dickey-Fuller model. And the null hypothesis is tested that the parameter is *($\phi_1 = 1$).* That is mean, the presence of a unit root in the time series meaning that it is not stationary versus alternative hypothesis *($\phi_1 < 1$)* meaning that it is stationary.

3.5.2. Augmented Dickey-Fuller Test

One of the techniques used to check the stationary of the time series is the unit roots test. The stationary condition is achieved when the unit roots of the series lie within the unit circuit. One of the most important methods used to detect the stationary of the time series is the augmented dickey-fuller test, which relies on three mathematical equations that assume the existence of a random process x_t.

- test without constant

$$\nabla x_t = \emptyset_1 x_{t-1} + \sum_{j=1}^{p} B_j\ \nabla x_{t-j} + e_t$$

- test with constant

$$\nabla x_t = \emptyset_0 + \emptyset_1 x_{t-1} + \sum_{j=1}^{p} B_j\ \nabla x_{t-j} + e_t$$

- with constant and trend

$$\nabla x_t = \emptyset_0 + \emptyset_1 x_{t-1} + \sum_{j=1}^{p} B_j\ \nabla x_{t-j} + \delta t + e_t$$

The hypothesis of the test is expressed by:

H_0:$\emptyset_1$=0 , Unstationary (having a unit root)

H_1:$\emptyset_1 \neq 0$, stationary (no root unit)

The test statistic $DF_t = \emptyset_1/SE(\emptyset 1)$ compares with the tabular values (Dickey-Fuller tables). If the computed t value is greater than the tabular value, the null hypothesis is rejected and the alternative hypothesis accepted. Therefore, the time series is stationary or through a p-value if the (p-value) is less than 0.05, the alternative hypothesis is accepted, i.e. the series is stationary.

4. Autoregressive Model AR (p)[10, 1, 13, 9]

The model is described for stationary time series

$$x_t = \mu + \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + ... + \emptyset_p x_{t-p} + \varepsilon_t.$$

where μ is the mean of the series, $\emptyset_i$, $i = 1, 2, 3 \ldots .p$ the parameters of the model also these are weights associated with the series values, and $\varepsilon_t$random errors with zero mean and variance σ_ε^2, p represents the degree of the model, where the autoregressive model of the p-degree is symbolized by AR (p), and by using the backshift operator B, the model written by:

$$\left(1 + \emptyset_1 \mathrm{B} + \emptyset_2 \mathrm{B}^2 + \cdots + \emptyset_p \mathrm{B}^p \right) x_t = \mu + \varepsilon_t$$

or

$$\emptyset_p (B) \, x_t = \mu + \varepsilon_t$$

$\emptyset_p (B)$ Polynomial of order p, for AR(p) model

- The autocorrelation function coefficients r_k decrease exponentially
- The coefficients of the partial autocorrelation function $\emptyset_{kk}$ cut off after the p lags

From that the autocorrelation function of AR(p) model define:

$$r_k = \emptyset_1 r_{k-1} + \emptyset_2 r_{k-2} \cdots + \emptyset_p r_{k-p}$$

with variance to time series

$$var\left(x_t\right) = \frac{\sigma_e^2}{1 - \emptyset_1 \rho_1 - \emptyset_2 \rho_2 - \cdots - \emptyset_p \rho_p}$$

And we can rely on graphs to describe the relationship between the behavior of the autocorrelation function and the partial autocorrelation function for model AR(p).
Roots of polynomials can be found by solve the equation of the roots if the values of the roots $\emptyset_p (B) = 0$, whose number p in absolute value is greater than one, then the series is stationary.

- When p = 1 it is said to the time series that it is autoregressive process of the first order or a Markov model and it is denoted by the symbol AR (1) and that its equation is in the form

$$x_t = \emptyset_1 x_{t-1} + e_t$$

The autocoveriance function of AR (1)

$$\gamma_k{=}\sigma_e^2\left(\frac{\emptyset^k}{1-\emptyset^2}\right)\ , \qquad k=1,2,\ldots$$

Autocorrelation function for AR (1)

$$\rho_k{=}\emptyset_1\ \rho_{k-1}$$

$$\widehat{\emptyset}_{11}=\widehat{\rho}_1=r_1 \quad , \quad k=1$$

- When p = 2 it is said to the time series that it is autoregressive process of the second order denoted by the symbol AR (2) and that its equation is in the form:

$$x_t{=}\emptyset_1 x_{t-1}+\emptyset_2 x_{t-2}+e_t$$

The necessary and sufficient conditions for stationary are as follows

$$\phi_2-\phi_1<1\ ,\ \phi_1+\phi_2<1\ ,\ \ |\phi_2|<1$$

5. Moving Average Model MA(q)[10, 2, 6, 3, 7]

The model is written in terms of errors for the time series the previous values of errors and the general formula for this model of degree q, symbolized by MA (q):

$$x_t=\mu+\ a_t-\theta_1 a_{t-1}-\theta_2 a_{t-2}-\cdots-\theta_q a_{t-q}$$

or

$$x_t=\ \mu+\theta_q\left(B\right)a_t$$

$$\theta_q\left(B\right)=1{-}\theta_1 B{-}\theta_2 B^2{-}\cdots-\theta_q B^q$$

Where: $a_1,a_2,\ldots,\ a_{t-q}$ white noise , $\theta_1,\theta_2,\ldots,\theta_q$ are parameters
Whereas, the polynomials $\theta_q\left(B\right)$ are of the q order, of the moving average model and that q determines the order of the model. The errors are independently distributed with mean zero, variance equal to σ_a^2 , and specification of the model.

1. the autocorrelation function declines to zero after the q rank, and is described

$$\rho_k=\left(\begin{cases}\dfrac{-\theta_k+\theta_1\theta_{k+1}+\ldots+\theta_{q-k}\theta_q}{1+\theta_1^2+\theta_2^2+\ldots+\theta_q^2} & k=1,2,...,q\\ 0 & k>q\end{cases}\right.$$

2. The partial autocorrelation function $\emptyset_{kk}$ is decreasing exponentially.

6. Autoregressive Moving Average Model ARMA (p, q)[10, 1, 5, 9]

This model contains the characteristics of autoregressive and moving average models, is written in the following form

$$x_t{=}\mu + \emptyset_1 x_{t-1} + \cdots + \emptyset_p x_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

$$\emptyset_p\left(B\right) x_t = \ \mu + \theta_q\left(B\right) a_t$$

It is denoted by ARMA (p, q) as they are $\emptyset_p(B)$ $\theta_q\left(B\right)$ p-order and q-order polynomials for autoregressive and moving average to achieve the stationary condition, the roots of $\emptyset_p(B)\ = \mathit{0}$ must lie outside the unit circle.

The condition of inevitability of the model is that the roots of the equation $\theta_q\left(B\right){=}\ \mathit{0}$ must be outside the unit circle.

But if the differences are used to convert the time series into a stable series, then the model is described

$$\emptyset_p\left(B\right)\left(1-B\right)^d x_t = \ \mu + \theta_q\left(B\right)\ a_t$$

where d represents the non-seasonal difference, which converts the time series to the stationary attribute at the first and second difference in most cases, then it denotes the ARIMA(p,d,q) model and it is called autoregressive integrated moving average model.

And the model has the following characteristics:

1. The autocorrelation coefficients r_k are decreasing exponentially

   so the autocorrelation function define as:

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ... + \phi_p\rho_{k-p} + \frac{\sum_{j=1}^{q}\theta_j\psi_{j-k}}{\left(\sum_{j=0}^{\infty}\psi_j^2\right)} \qquad for\ \ 0 \le k \le q$$

   As for the autocorrelation function for k> q define by:

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ... + \phi_p\rho_{k-p}$$

2. The coefficients of the partial autocorrelation function $\emptyset_{kk}$ are also decreasing exponentially.

   It is possible to rely on the graphs of the autocorrelation function and the partial autocorrelation through which the rank of the ARMA(p,q) model is determined.

7. Methods for Estimating [10, 7, 12, 9]

7.1. Methods for Estimating the Parameters of the Autoregressive Model: AR (p)

There are several methods for estimating the model, where by the maximum likelihood function of the time series $\mathrm{x_t}$ consisting of n observations, as defined by Box-Jenkins, is as:

$$L\left(\phi,\theta,\sigma_a \mid \mathrm{x}\right){=}(2\pi\sigma_a^2)^{\frac{-n}{2}}\left|M_n^{(p+q)}\right|^{\frac{1}{2}} exp\left[\frac{-S(\phi,\theta)}{2\sigma_a^2}\right]$$

and by taking the logarithm to the function, then:

$$linL\left(\phi,\theta,\sigma_a \mid \mathrm{x}\right) \ = -\left(\frac{n}{2}\right) ln\left(2\pi\right) \ -\left(\frac{n}{2}\right) ln\sigma_a^2 + \left(\frac{1}{2}\right) ln\left|M_n^{(p+q)}\right| - \left[\frac{S\left(\phi,\theta\right)}{2\sigma_a^2}\right]$$

This formula is derived with respect to$(\phi\,,\theta\,,\sigma_a)$ and then equal to zero, as follows:

$$\frac{\partial \ell in\ L}{\partial \sigma_a} = -\frac{n}{\sigma_a} + \frac{S(\emptyset,\theta)}{\sigma_a^3}$$

$$\widehat{\sigma}^2 = \frac{S\ (\widehat{\emptyset},\widehat{\theta})}{n} \ldots.$$

This formula is difficulty of deriving in relation $(\phi\,,\theta\,)$ to obtaining the estimates It is possible to describe a modified formula, and in the case of medium and large samples, which characterize the time series, where the sum of the squares of errors is defined for any model in the formula:

$$S\left(\phi\,,\theta\,\right) = \sum_{t=1}^{n} \left(a_t/\phi,\theta,X_t\right)^2$$

For autoregressive model and when q=0 then:

- $M_n^{(p)}$ The inverse of the covariance matrix is written as follows:

$$M_n^{(p)} = \sigma_a^2 \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix}^{-1}$$

- S Sum of squares of the estimated model error of the time series from p + 1 to n.

Thus, by minimizing the sum of the squares of errors of the model, we obtain the approximate maximum likelihood estimates, which are equal to the least squares estimates. The linear regression model of the order p-order and in terms of deviations from the meanμ is described.

$$X_t - \mu = \phi_1\left(X_{t-1} - \mu\right) + \phi_2\left(X_{t-2} - \mu\right) + ... + \phi_p\left(X_{t-p} - \mu\right) + a_t$$
$$\ldots\ldots\ldots\ldots.(1.27)$$

Model parameters are estimated by the least squares method, by minimizing the sum of squares of errors, which is described:

$$S = \sum_{t=p+1}^{n-p} \left[\left(X_t - \mu\right) - \sum_{i=1}^{p} \phi_i\left(X_{t-i} - \mu\right)\right]^2 \ ...$$

The model parameters are estimated by the formula:

$$\underline{\phi} = \left(X'X\right)^{-1} X'Y$$

Where $\underline{\phi}$ *a* px1 dimension vector of the autoregressive model parameters X is a matrix with dimensions *n-p* $\times$ *p* for the lags independent variables, Y is a vector with dimensions of nx1 for values of deviations of the time series X_t .

It is also possible to estimate the parameters of the autoregressive model AR(p) in other ways, the estimation of the parameters based on the autocorrelation function, and since the autocorrelation function of the AR(p) is described:

$$\rho_k = \emptyset_1\rho_{k-1} + \emptyset_2\rho_{k-2} + \cdots + \emptyset_p\rho_{k-p} \qquad , k = 1, 2 \ldots , p$$

A set of equations as much as the number of parameters in order to obtain the estimations, as the set of these equations are called Yule - Walker equations. Thus, the estimations are expressed in the form of matrices

$$\widehat{\emptyset} = R_1^{-1}\ R_2$$

R_1 It is a matrix of dimensions p×p that contains the autocorrelation estimates, R_2 is a vector with dimensions of p×1 for autocorrelations. $\widehat{\emptyset}$ Vector of estimated paramters with dimensions of p×1.

$$R_2{}^t = (\ r_1\ , r_2, \ldots , r_p\ )$$

$$R_1 = \begin{pmatrix} 1 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & 1 & r_1 & \ldots & r_{p-2} \\ \vdots & \vdots & & & \vdots \\ r_{p-1} & r_{p-1} & & \cdots & 1 \end{pmatrix}$$

7.2. Estimating the Parameters of the Moving Averages Model MA(q)

The method of least squares cannot be applied to estimate the parameters of the moving averages model because the sum of squares of errors does not constitute a linear function in the features. Therefore, Box-Jenkins proposed a method for estimation based on the autocorrelation function, so the MA(1) model describes:

$$x_t = \mu +\ a_t - \theta_1 a_{t-1}$$

And the estimation is given values that are bound by the autocorrelation function of the model, where we choose$|\theta_1| < 1$ then it is possible to calculate the sum of the squares of the errors of the model:

$$a_t = x_t - \mu + \theta_1 a_{t-1} \qquad t\text{=}1, 2, 3 \ldots .n$$

And choose the estimated parameters that reduce the sum of squares of error:

$$s = \sum_{t=1}^{n} a_t^2$$

If the moving averages model is of the second degree MA (2), then we can impose initial values to θ_2 , θ_1 achieve the characteristic that describes the conditions of reflectivity of the MA(2) model.

$$x_t = \mu +\ a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

And the initial values that meet the following conditions:

$$\theta_1 + \theta_2 < 1 \quad , \quad \theta_1 - \theta_2\ < 1 \quad , \quad -1 < \theta_2\ < 1$$

The residuals are then computed sequentially, depending on the model with $0 = a_1 = a_0$

$$a_t = x_t - \mu + \theta_1 a_{t-1} + \theta_2 a_{t-2} \qquad t{=}2, 3, 4 \ldots . n$$

Then sum of error squares calculated and then the process is repeated with other default values until we choose the parameters that achieve the lowest value for the sum of error squares.

8. Model Selection Criterions [1, 5, 7]

A range of measures can be used to find rank for the Box-Jenkins model:

1. **Akaike information criterion**

   This criterion was proposed by the Japanese mathematical scientist Heitro Akaike and was known by his name and is considered more advanced compared to the FPE standard test to determine the rank of the optimal model Box-Jenkins, symbolized by an AIC, this test of an AR (K) time series model is described as:

$$aic\,(k) = lin\ \left(\widehat{\sigma}_k^2\right) + \frac{2k}{n}$$

   And that the $aic\,(k)$ scale should be smaller, where k is chosen, which represents an estimator of the model's degree for the lowest value of this magnitude, and this test is often estimated and choosing a large value to the rank of the model p.

2. **Schwarz criterion**

   The test format takes the following form

$$sc\,(k) = lin\ \left(\widehat{\sigma}_k^2\right) + \frac{k\ lin(n)}{n}$$

   Also, the sample test scale is with degree k for the lowest value of the Schwarz test, the estimated scale score approaches the model degree by increasing the sample size but if it is a series white noise then both scales are equal.

$$aic\,(0) = sc\,(0) = lin\ \left(\widehat{\sigma}_k^2\right)$$

3. **Hannan & Quinn Criterion**

   Researchers Hannan and Quinn proposed a new criterion for determining the rank of the box-Jenkins model, symbolized by an abbreviation H-Q and is defined as follows

$$H - Q\,(k) = lin\ \left(\widehat{\sigma}_k^2\right) + \frac{2\ c\ k\ lin\,(n)}{n}$$

   C represents a constant magnitude greater than *2, i.e (c > 2).*

4. **Bayesian Information Criterion**

   The Bayesian Information Criterion, which is derived from the Bayes modification of the standard AIC, takes the following form:

$$BIC\,(k) = n\ lin\ \left(\widehat{\sigma}_k^2\right) - (n-k)\,lin\left(1 - kn^{-1}\right) k\ lin\,(n) + k\ lin\left\{\left(\ k^{-1}\right)\left(\ \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_e^2} - 1\right)\right\}$$

9. Diagnostic Checking[4, 13, 5, 14]

The series of errors is an independent and identically distributed random variable IID If this is achieved through these tests, then the model can be approved and accepted, and thus the level of the resulting errors is accepted. If this is not achieved, then the stationary theory must be used in order to build a better model.

9.1. The Sample Autocorrelation Function Test

The autocorrelation is calculated to the time series of the estimated model errors, and since these autocorrelations are approximately a normal distributed, then if *95%* of the autocorrelation coefficients are within the boundaries $\pm 1.96 n^{-0.5}$, then the series of observations error is independent and with identical distribution.

$$-1.96 \;\; n^{-0.5} = \hat{r}_k\left(\hat{a}\right) \leq 1.96 \;\; n^{-0.5}$$

9.2. Ljung Box Test

It is a developed test and takes the following formula:

$$Q_{LB} = n\left(n+2\right)\sum_{k=1}^{f}\left(n-k\right)^{-1} r_k^2\left(\hat{a}\right) \;\sim\; x^2_{(f)}$$

Whereas, Q_{LB} has a better asymptotic distribution than Q and Q_{LB} has a chi-square distribution with a degree of freedom f .
And the null hypothesis is rejected: $Q_{LB} > \; x^2_{1-\alpha}(f)$

9.3. Fitting an Autoregressive Model

Autoregressive model AR(p) is fitted for the error of the time series and compute p-model rank, which makes the AIC statistic as less as possible, so if the suggested rank is equal to zero, p=0 this indicates that the series data is White Noise and there is no remaining model in the errors.

10. The Practical Application

10.1. Box-Jenkins Models

10.1.1. Box-Jenkins Models for Registered COVID-19 Pandemic Patients in Iraq

The data represent observations of a time series for the number of people with Covid 19 Pandemic for the years (2020-2021), and the number of weekly observations reached 53, and the unit of measure was the number of patients registered with the Pandemic, and the data were described in Table 1.

Table 1: The Number of Patients Registered with the Covied-19 Pandemic in Iraq for the Period (2020-2021)

Year	**2020**					**2021**				
Mon	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Jan						**5841**	**5370**	**5290**	**6044**	**5547**
Feb			**17**	**25**		**9643**	**16387**	**21646**		
Mar	**37**	**53**	**147**	**364**						
Apr	**402**	**347**	**196**	**273**						
May	**499**	**472**	**736**	**1078**	**1807**					
June	**6613**	**7834**	**11361**	**14475**						
July	**15124**	**17460**	**15058**	**17892**	**19301**					
Aug	**21713**	**26534**	**27852**	**26949**						
Sept	**29752**	**29794**	**28378**	**30710**						
Oct	**29383**	**22488**	**25241**	**24720**	**23303**					
Nov	**23032**	**19809**	**15915**	**15092**						
Dec	**13466**	**9957**	**9373**	**7183**						

10.1.2. Stationary in Time Series

The first step of analyzing the time series is through drawing the time series to identify some of its components such as the general trend, seasonal changes, cyclical changes, and random component, and based on the data table 1 the time series was drawn, and Figure 1 illustrates the absence of a general trend uniform of increase Positive in the time series, and the data does not fluctuate around a constant average, and from this we infer the unstatinary time series of the average, i.e. achieve weak stability to the time series of numbers of people with Covid-19 disease, but the series contains high fluctuation, so it is unstatinary in variance.

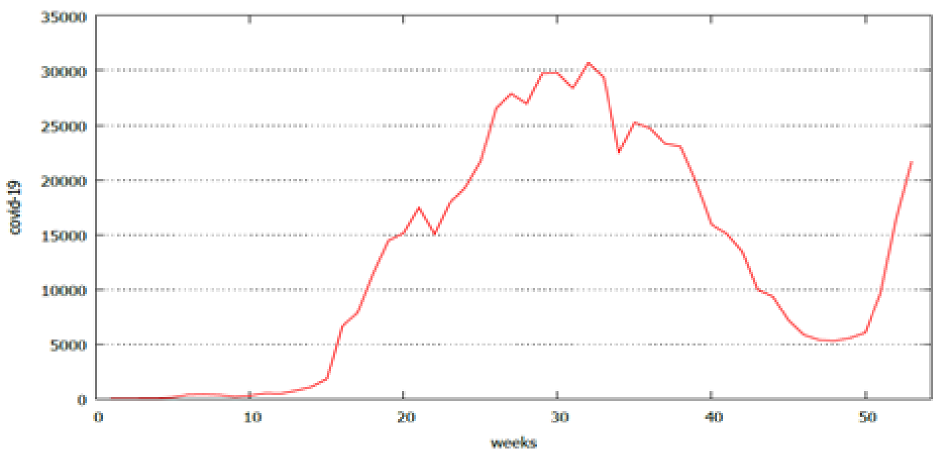

Figure 1: The Time Series for the Number of Patients Registered with the Covied-19 Pandemic in Iraq for the Period (2020-2021)

The stationary of the time series is verified whether it is strong or weak, and by that we mean stationary in the average and stationary in the variance, and for more accuracy and to confirm doubts, the parameters of the autocorrelation function (AC) and the coefficients of the partial autocorrelation function) PAC) were calculated and drawn for the original series and the results were as follows:

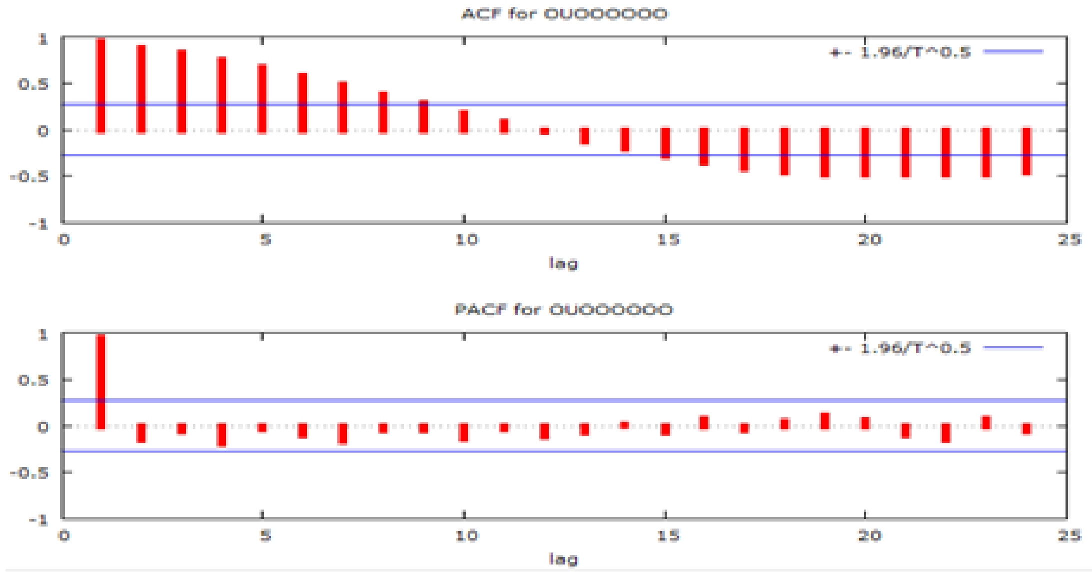

Figure 2: A graph of autocorrelation (AC) and partial autocorrelation (PAC)coefficients for time series for number of people with Covid-19 pandemic

The autocorrelation coefficients are decreasing exponentially and outside the boundaries of confidence and at a significant level (95%) in relation to the lag shift, and the partial autocorrelation coefficients are cut off after the first lag displacement and fall within the acceptance limits for the values of the transactions, which is evidence of the instability of the time series, which necessitates the initial taking of the necessary differences to achieve stationary.

11. Unit Root Tests

For more accuracy and to detect unstable chains, it is through the use of statistical tests, including the extended dickey - Fuller test based on the auto-regression model of the first degree. The Phillips - Peron test is a development of the dickey - Fuller test. The two tests were calculated for the study series variable, and the results are shown in Table 2 and 3:

Table 2: The Augmented Dickey - Fuller Test results for the Unit Root and the Original Time Series Data, as well as the test results after the first and second difference

	LEVEL			1 ST DIFFERENCE			2 nd DIFFERENCE		
test	**intercept**	**trend And intercept**	**none**	**intercept**	**trend And intercept**	**none**	**intercept**	**trend And intercept**	**none**
t-statistic	**-2.0888**	**-3.0511**	**-0.5431**	**-0.8949**	**-0.5051**	**-0.7916**	**-9.8659**	**-9.9887**	**-9.8790**
p-value	**0.2499**	**0.1294**	**0.4766**	**0.7817**	**0.9800**	**0.3680**	**0.0000**	**0.0000**	**0.0000**

Table 3: The Phillips - Perron Test results for the Unit Root and the Original Time Series Data, as well as the test results after the first difference

	LEVEL			1 ST DIFFERENCE		
test	**intercept**	**trend And intercept**	**none**	**intercept**	**trend And intercept**	**none**
t-statistic	**-1.4115**	**-1.4590**	**-0.3158**	**-4.7461**	**-4.6850**	**-4.6629**
p-value	**0.5697**	**0.8390**	**0.5670**	**0.0003**	**0.0022**	**0.0000**

The unit root test results in Table 2 indicate that the original time series data contain the unit root because the results of the tests indicated acceptance of the null hypothesis of the existence of the unit root for the intercept term, trend and intercept, and none (none trend and intercept) term because all calculated p-values for the tests is greater than the level of significance 0.05. In the case of the extended Dickey-Fuller test, stationary was not achieved until after the second difference, where we find that the p-value values are less than 0.05, which means rejecting the null hypothesis that indicates the existence of the unit root and accepting the alternative hypothesis that indicates to the absence of a unit root, thus achieving stationary in the time series after second difference.

And the Phillips - Perron test results in Table 3 showed that the p-value calculated for the test is greater than the level of significance 0.05, and from it we infer that the original time series is not stationary and it is necessary to take the differences to achieve the mean stationary and when re-tests on the data of the difference series (after the difference First) we find that the p-value for test are less than 0.05, which means rejecting the null hypothesis that indicates the existence of the unit root and accepting the alternative hypothesis that indicates the absence of the unit root and thus achieving stationary in the time series after the first difference.

Among the results of the two tests, we find a wide difference in the test results. The augmented dickey - Fuller test indicated the need to take the second difference in order to achieve stability in the time series, while the Phillips - Perron test indicated that the first difference was sufficient to achieve stationary in mean for time series.

The coefficients of the autocorrelation function and the partial autocorrelation of the transformed time series (the series of first differences) have also been calculated, as we find that most of the coefficients fall within the limits of confidence, which indicates the achievement of the stationary of the time series after the first difference, which is weak stationary in the mean.

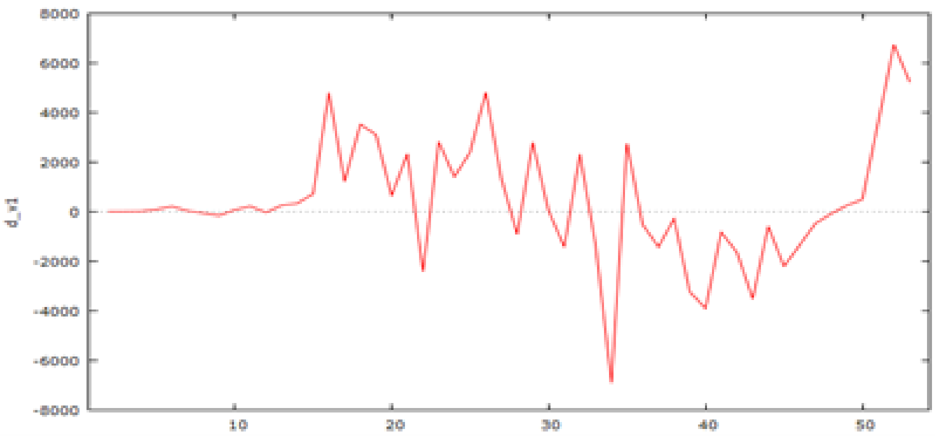

Figure 3: The time series for the number of patients registered with the covied-19 Pandemic in Iraq after the first difference for the period (2020-2021)

Where the autocorrelation coefficients and partial autocorrelation coefficients were calculated for the series of first differences, the results as follows:

Table 4: The values and the test of the Autocorrelation (AC) and Partial Autocorrelation (PAC) Coefficients for the Time Series for the number of people with the Covid 19 Pandemic in Iraq after the first difference

LAG	ACF		PACF		Q-STAT	P-VALUE
1	0.3539	**	0.3539	**	6.8964	[0.009]
2	0.2513	*	0.1441		10.444	[0.005]
3	0.3876	***	0.304	**	19.0512	[0.000]
4	0.0471		-0.2238		19.181	[0.001]
5	0.0949		0.0583		19.7195	[0.001]
6	0.23	*	0.1322		22.9497	[0.001]
7	-0.0213		-0.1014		22.978	[0.002]
8	-0.0914		-0.185		23.5116	[0.003]
9	0.0651		0.0821		23.7883	[0.005]
10	-0.1129		-0.048		24.6398	[0.006]
11	-0.1154		-0.0379		25.5515	[0.008]
12	-0.0758		-0.1476		25.9548	[0.011]
13	-0.2842	**	-0.1583		31.7696	[0.003]
14	-0.2122		0.0106		35.096	[0.001]
15	-0.2161		-0.1386		38.6407	[0.001]
16	-0.1912		0.0776		41.4925	[0.000]
17	-0.1755		-0.087		43.9642	[0.000]
18	-0.321	**	-0.2554	*	52.4721	[0.000]
19	-0.2426	*	-0.0099		57.4811	[0.000]
20	-0.1426		0.079		59.2666	[0.000]
21	-0.1434		0.0247		61.1297	[0.000]
22	-0.1475		-0.1873		63.1653	[0.000]
23	-0.0798		-0.0502		63.7819	[0.000]
24	-0.1115		0.0839		65.0277	[0.000]

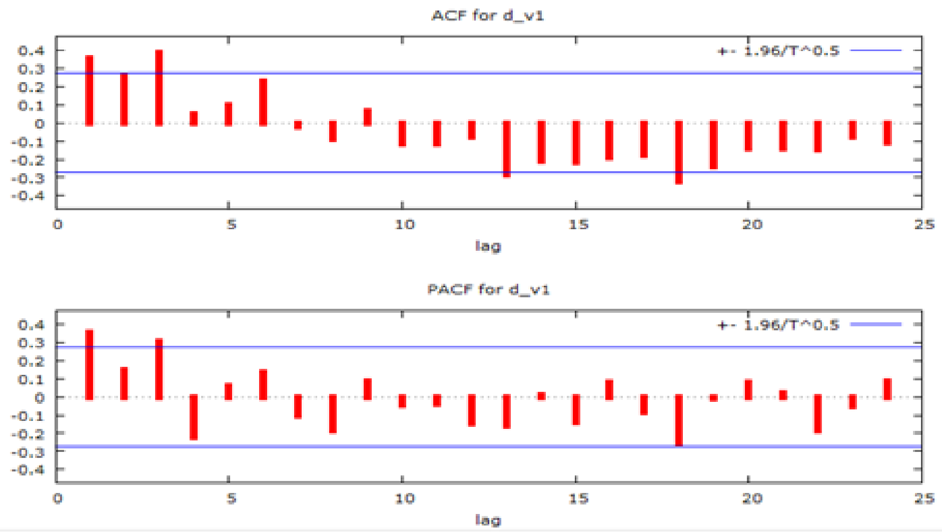

Figure 4: The Graph of the Autocorrelation (AC) and Partial Autocorrelation (PAC) Coefficients for the Time Series for the Number of People with the Covid 19 Pandemic in Iraq After the First Difference

12. Model Order Diagnosis

In order to diagnose the model through the Box-Jenkins method, it is done after achieving stationary in the time series in the mean and variance, and by looking at the relationship between the autocorrelation coefficients and the partial autocorrelation.

Autocorrelation (AR) determines its order through the number of partial autocorrelations that differ significantly from zero, when the decrease is exponential in partial autocorrelations, the model is of the type (MA) and its order is determined by the number of autocorrelations with statistical significance, also when the autocorrelations behave and partial autocorrelations are an exponential behavior in their decline and approach to zero, the model is of the mixed model type (ARMA).

But in the practical side these specifications are often not achieved in the time series data, so we resort to reconciling a set of models and choosing the best model based on some statistical measures and from these tests are the Akaike Information Criterion (AIC), the Hannan-Quinn scale (HQ), and the Schwarz Information Criterion or Bayesian Information Criterion (BIC), so different models have been built up to the level third, with differences and no differences until the second difference and choosing the lowest values for errors of the model to determine the best model. Below are some of the selected models as in Table 5:

Table 5: Some proposed Box-Jenkins Models up to the third Order for the Time Series Data for the number of people infected with the Covid 19 Pandemic in Iraq

Model Selection Criteria Table				
Dependent Variable: COVID19				
Date: 02/23/21 Time: 14:30				
Sample: 1 53				
Included observations: 53				
Model	**Log L**	**AIC***	**BIC**	**HQ**
ARMA(1,3)	-480.054034	18.341662	18.564714	18.427437
ARMA(2,1)	-481.955573	18.375682	18.561559	18.447161
ARMA(2,2)	-481.755483	18.405867	18.628919	18.491642
ARMA(3,1)	-481.854793	18.409615	18.632667	18.495390
ARMA(2,3)	-480.937475	18.412735	18.672962	18.512806
ARMA(3,3)	-480.019372	18.415825	18.713228	18.530192
ARMA(3,2)	-481.132104	18.420079	18.680307	18.520150
ARMA(3,0)	-484.027726	18.453876	18.639753	18.525356
ARMA(2,0)	-485.371335	18.466843	18.615544	18.524026
ARMA(1,1)	-486.272696	18.500856	18.649558	18.558040
ARMA(1,2)	-486.263603	18.538249	18.724126	18.609728
ARMA(1,0)	-489.767626	18.595005	18.706531	18.637892
ARMA(0,3)	-502.230489	19.140773	19.326650	19.212252
ARMA(0,2)	-516.607214	19.645555	19.794257	19.702739
ARMA(0,1)	-534.213555	20.272210	20.383736	20.315097
ARMA(0,0)	-565.069862	21.398863	21.473213	21.427454

By comparing the models of different orders up to the third order and according to the Akaike Information Criterion (AIC) test, and by choosing the lowest values for the best model, we find that the results are represented in Table 4, which shows the extent of the decrease in the Akaike Information Criterion (AIC) scale from the change in the order of the comparative models, and from it the best model is ARMA (1,3) and the model according to this case is plausible.

13. Model Estimate

Through the use of the approximate maximum Likelihood method with the mathematical formula of the diagnosed time series model, which is the ARMA (1,3) model, and after dealing with the data, the model was estimated according to the approximate maximum likelihood method, and the results are shown in Table 6 with the accompanying statistical tests to the quality of the model and the parameters of the model were according to the following.

Table 6: The Estimated Model and Time Series Data for the Number of People Infected with the Covid 19 Pandemic in Iraq

Dependent Variable: COVID19				
Method: ARMA Maximum Likelihood (BFGS)				
Date: 02/23/21 Time: 14:30				
Sample: 1 53				
Included observations: 53				
Convergence achieved after 13 iterations				
Coefficient covariance computed using outer product of gradients				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13172.11	6738.141	1.954858	0.0566
AR(1)	0.902188	0.090434	9.976206	0.0000
MA(1)	0.476748	0.119934	3.975073	0.0002
MA(2)	0.532683	0.160578	3.317285	0.0018
MA(3)	0.577818	0.113730	5.080593	0.0000
SIGMASQ	3932061.	764261.6	5.144914	0.0000

Statistical tests and statistical measures were also calculated to judge the efficiency of the estimated model:

Table 7: The tests and the Estimated Statistical. Measures of the Chosen Model ARMA (1,3)

R-squared	0.963148	Mean dependent var	12603.45
Adjusted R-squared	0.959227	S.D. dependent var	10428.33
S.E. of Regression	2105.713	Akaike info criterion	18.34166
Sum squared resid	2.08E+08	Schwarz criterion	18.56471
Log likelihood	-480.0540	Hannan-Quinn criter.	18.42744
F-Statistic	245.6729	Durbin-Watson stat	2.038292
Prob(F-statistic)	0.000000		

The model is generally accepted because the probability of **F**-testing for the model as a whole Prob(F-statistic) which is equal to 0.0000 are less than 0.05, and coefficient of determination of the model (R-squared) equal to **0.963148.** It means that the explanatory variables have interpreted the changes in the time series variable by a percentage 96%. The estimated model can be written in the following simplified form:

$$y_t = 13172.11 + 0.902188\ y_{t-1} + 0.476784\ a_{t-1}\ + 0.532683\ a_{t-2} +\ 0.577818\ a_{t-3}$$

The results in general indicate the significance of the model parameters estimated according to the t-statistic test and the probability values Prob where its results are less than **0.05**, in addition to the significance of the model as a whole according to a value. Prob for (F-statistic), which amounted to **0.0000** and less than 0.05 also, and coefficient of determination of the model (R-squared) equal to **0.963148** and it indicates the value interpreted in the time series in terms of the explanatory variables of the model. It means that the explanatory variables have interpreted the changes in the time series variable by a percentage 96%.
that confidence limits for estimated model parameters:

Table 8: Confidence Limits for Estimated Model Parameters t(49, 0.025) = 1.960

95 confidence interval	Coefficient	Variable
(1205.25, 25139.0)	13172.1	const
(0.779073, 1.02530)	0.902188	phi_1
(0.265034, 0.688463)	0.476749	theta_1
(0.234141, 0.831226)	0.532683	theta_2
(0.267426, 0.888208)	0.577817	theta_3

14. Check the Fit of the Model

Autocorrelation Coefficents Test for Residues

The time series residual test was performed by calculating the autocorrelation coefficients and partial autocorrelation, and we find that most of the autocorrelation coefficients fall between the interval $\pm 1.96/\sqrt{n}$., In other words, it is noted that all results for the values of the autocorrelation coefficients fall within the following confidence limits:

$$-0.26923 \leq r_k\,(\varepsilon^\wedge) \leq +0.26923$$

which means that the error series is random and the model used is good in terms of estimation:

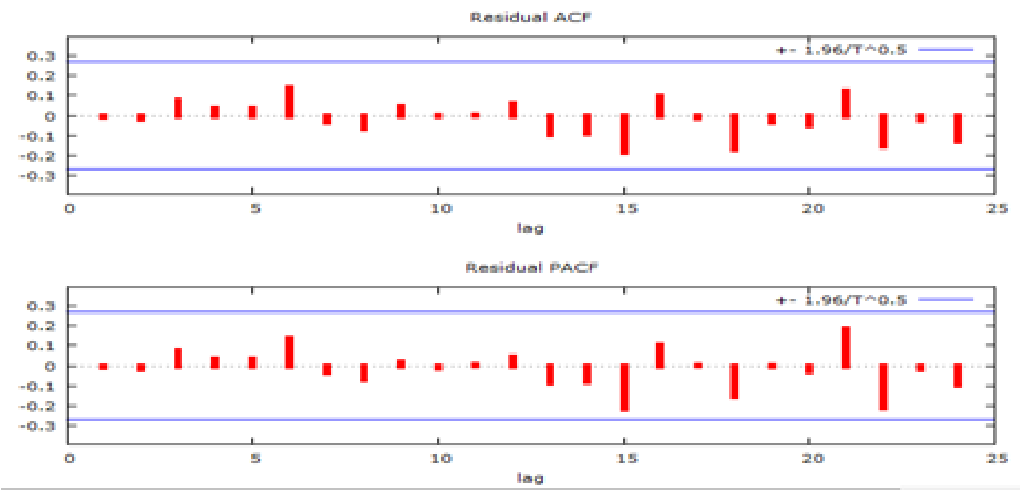

Figure 5: The Autocorrelation and Partial. Autocorrelation for the Residual of the Estimated Model

Whereas most of the autocorrelation and partial autocorrelation coefficients fall within the confidence limits, which means that the residual series is random, and the model is good and significant statistically, with a Q-stat test performed at each lag from lag 1 to 24, which showed its insignificance and refers to No remaining model remains in the time series and thus the approved model is accepted.

Table 9: Autocorrelation Coefficients and Partial Autocorrelation For Residual of the Estimated Model

Residual autocorrelation function

LAG	ACF	PACF	Q-stat.	[p-value]
1	-0.0092	-0.0092	0.0048	[0.945]
2	-0.0166	-0.0167	0.0206	[0.990]
3	0.0753	0.0750	0.3513	[0.950]
4	0.0338	0.0351	0.4193	[0.981]
5	0.0323	0.0357	0.4828	[0.993]
6	0.1401	0.1374	1.6996	[0.945]
7	-0.0342	-0.0352	1.7736	[0.971]
8	-0.0649	-0.0685	2.0464	[0.980]
9	0.0428	0.0182	2.1680	[0.989]
10	-0.0039	-0.0122	2.1690	[0.995]
11	0.0020	0.0049	2.1693	[0.998]
12	0.0589	0.0449	2.4160	[0.998]
13	-0.0964	-0.0849	3.0934	[0.998]
14	-0.0910	-0.0802	3.7119	[0.997]
15	-0.1818	-0.2159	6.2462	[0.975]
16	0.0966	0.1012	6.9817	[0.974]
17	-0.0144	0.0013	6.9984	[0.984]
18	-0.1691	-0.1540	9.3799	[0.950]
19	-0.0362	-0.0038	9.4923	[0.964]
20	-0.0488	-0.0286	9.7027	[0.973]
21	0.1201	0.1853	11.0175	[0.962]
22	-0.1525	-0.2101	13.2048	[0.927]
23	-0.0255	-0.0165	13.2681	[0.946]
24	-0.1297	-0.0964	14.9588	[0.922]

The autocorrelation coefficients and the partial autocorrelation of the residual squares variable were also tested to notice the extent of a model in the residual squares and the results were described in Table 9, where we find that the coefficients fall within the confidence limits, which means that the series of residual squares is random and that the matching model is good and significant statistically. With a Q-stat test performed at each displacement, which showed its insignificance at all displacements, thus the successful model is accepted.

Table 10: Autocorrelation Coefficients and Partial Autocorrelation For Residual Squares of the Estimated Model

Correlogram of Residuals Squared

Date: 03/06/21 Time: 22:49
Sample: 1 53
Included observations: 53

Lag	AC	PAC	Q-Stat	Prob
1	0.243	0.243	3.3137	0.069
2	-0.124	-0.195	4.1921	0.123
3	-0.030	0.059	4.2443	0.236
4	-0.060	-0.102	4.4597	0.347
5	0.036	0.093	4.5373	0.475
6	0.168	0.121	6.2955	0.391
7	-0.013	-0.087	6.3057	0.505
8	-0.031	0.049	6.3676	0.606
9	0.165	0.165	8.1791	0.516
10	-0.137	-0.255	9.4489	0.490
11	-0.045	0.145	9.5878	0.568
12	0.074	-0.049	9.9817	0.618
13	0.022	0.064	10.016	0.693
14	-0.122	-0.205	11.123	0.676
15	-0.179	-0.143	13.579	0.558
16	-0.089	0.068	14.209	0.583
17	0.134	0.116	15.658	0.548
18	0.267	0.129	21.582	0.251
19	0.002	-0.025	21.582	0.306
20	-0.085	-0.032	22.224	0.328
21	-0.108	-0.034	23.295	0.328
22	-0.057	-0.046	23.598	0.369
23	-0.009	-0.001	23.606	0.426
24	-0.111	-0.213	24.853	0.414

Q-Q Plot Test

A Q-Q plot test was performed for the residuals of the estimated model, drawing the data and noting the its spread, as in the figure 6, where we note that the spread of errors of the estimated model was around the direction line of the drawing and not far from it indicating that there are no major deviations between the line of the estimated model and that the errors are concentrated In the middle, with no outliers, errors in the lower or upper limit all indicate the quality of the Box-Jenkins estimator.

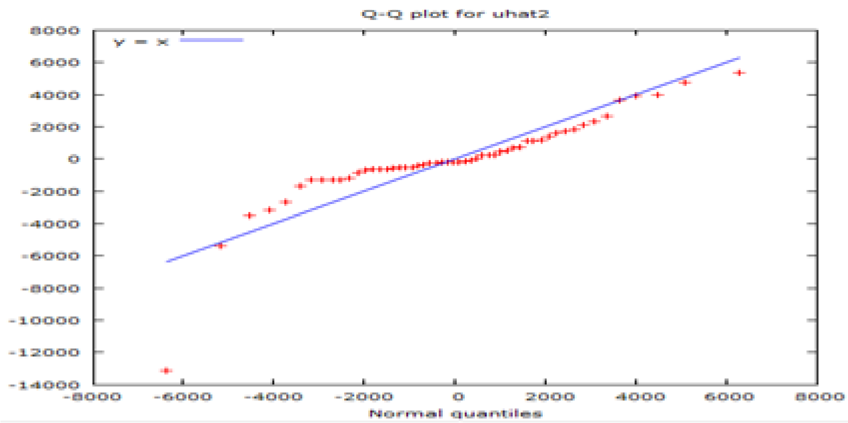

Figure 6: Q-Q Plot Test of the Residuals of the Estimated Model

Test the Statistical Distribution for Model Residuals

The statistical distribution of the estimated model residuals was also tested, if it is follow the normal distribution using jarque-bera, as it showed that the residuals follow the normal distribution, and the figure 7 illustrates that.

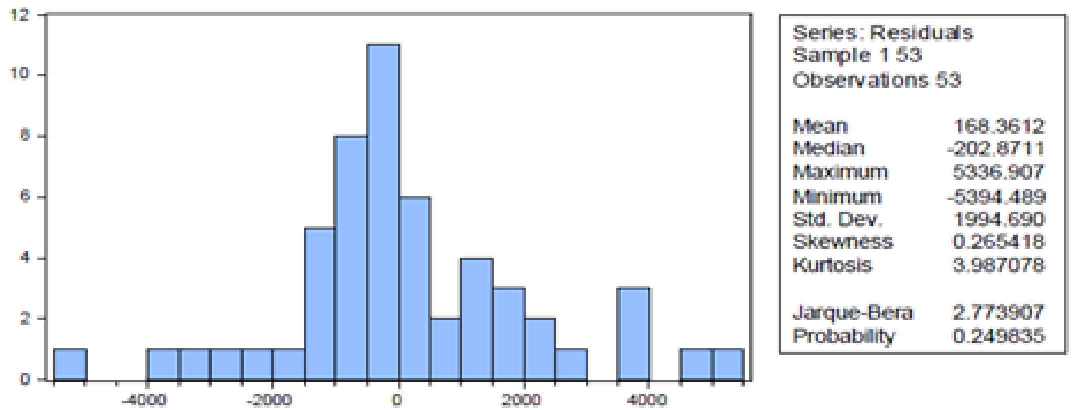

Figure 7: TShows that the Residuals Test if it Follow a Normal Distribution Using Jarque -Bera

Since the null hypothesis indicates that the residuals of the estimated model follows the normal distribution, where the chi-square value was **2.773907** with a probability value of **0.249836**, which is greater than the level of significance 0.05, then it is accepted that the residuals follow the normal distribution.

15. Estimated and Predictive Values

Estimated and predictive values were calculated according to the chosen Box-Jenkins model and reviewed for the study period , The following is the diagram of the form : 8

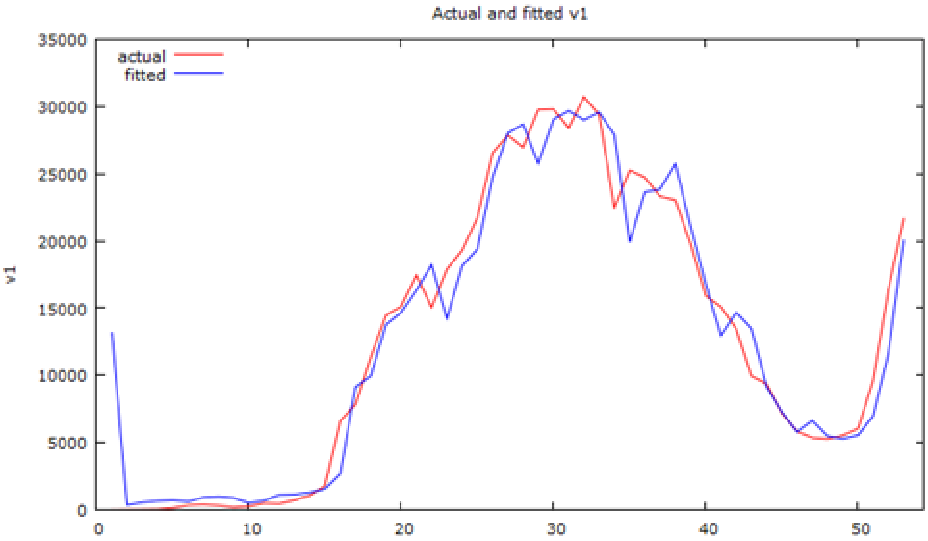

Figure 8: The Original Series and the Estimated Values from the Appropriate Model ARMA (1,3)

With the calculation of the confidence limits for the estimated values from the estimated model, with two standard deviations as show in figure 9.

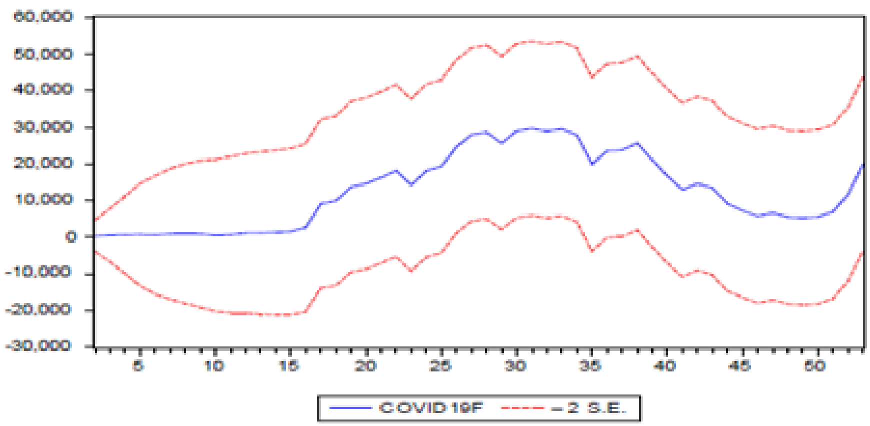

Figure 9: The Confidence Limits for the Estimated Values from the Estimated Model ARMA (1,3)

The measures of the efficiency of the estimated model were as follows:

Forecast: COVID19F	
Actual: COVID19	
Forecast sample: 1 53	
Adjusted sample: 2 53	
Included observations: 52	
Root Mean Squared Error	1977.520
Mean Absolute Error	1398.442
Mean Abs. Percent Error	114.2396
Theil Inequality Coefficient	0.060643
Bias Proportion	0.011359
Variance Proportion	0.010941
Covariance Proportion	0.977700

Then the predictive value series was calculated, and the results were as follows:

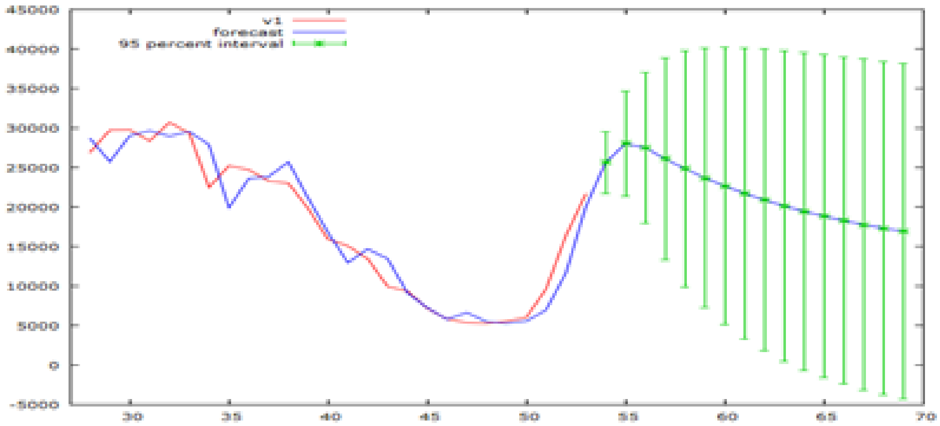

Figure 10: The Original Series with the Estimated Series and Predictive Values of the Time Series Data for the Number of People Infected with the Covid-19 Pandemic in Iraq.

And the predictive values with the following confidence limits:

Table 11: The Model's Predictive Values for the Time Series Data for the Number of People Infected with a Pandemic Covid-19 in Iraq (95% Confidence Intervals, z(0.025) = 1.96)

95% interval	std. error	prediction	Covid-19	Obs
(21760.1, 29533.1)	1982.94	25646.6	undefined	54
(21404.0, 34644.3)	3377.68	28024.2		55
(17942.8, 37074.9)	4880.73	27508.8		56
(13325.9, 38887.2)	6520.86	26106.5		57
(9947.95, 39734.8)	7598.84	24841.4		58
(7286.42, 40113.6)	8374.44	23700.0		59
(5116.27, 40224.3)	8956.28	22670.3		60
(3311.04, 40171.4)	9403.33	21741.2		61
(1789.33, 40016.8)	9752.09	20903.1		62
(494.321, 39799.5)	10027.0	20146.9		63
(-615.798, 39545.2)	10245.3	19464.7		64
(-1572.96, 39271.3)	10419.7	18849.2		65
(-2402.18, 38990.0)	10559.4	18293.9		66
(-3123.47, 38709.4)	10671.8	17792.9		67
(-3753.07, 38435.0)	10762.5	17341.0		68
(-4304.32, 38170.7)	10835.7	16933.2	undefined	69

16. Conclusions

1. The best estimated model for autoregressive integrated moving averages models is the model ARMA (1,3):

$$y_t = 13172.11 + 0.902188\ y_{t-1} + 0.476784\ a_{t-1}\ + 0.532683\ a_{t-2} + \ 0.577818\ a_{t-3}$$

This is because the model has the lowest scale value AIC, BIC, HQ.

2. The use of the autoregressive model integrated moving averages was efficient in building a model to predict the numbers of patients registered with the Covid-19 pandemic in Iraq, due to its accuracy in the results.

References

[1] S. Bisgaard, M. Kulahci ,*Time Series Analysis and Forecasting By Example* , John Wiley & Sons, Inc., Hoboken, New Jersey, (2011).

[2] G. E. P. Box and G. M. Jenkins, *Time series analysis forecasting and control*, Helden-Day, ISBN 0-8162-11043, 2 revised, 1976.

[3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, *Time Series Analysis: Forecasting and Control*, John Wiley and Sons Inc., Hoboken, New Jersey, 2016.

[4] A. Hernandez-Matamoros, H. Fujita, T. Hayashi and H. Perez-Meana, *Forecasting of COVID19 per regions using ARIMA models and polynomial functions*, Appl. Soft Comput., 96 (2020) 106610.

[5] E. Ghysels, M. Marcellino -, *Applied economic forecasting using time series methods*, Oxford University Press, 2018.

[6] E. Kočenda, A. Černý, *Elements of Time Series Econometrics: An Applied Approach*, Karolinum Press, Charles University, 2017.

[7] M. S. Paolella ,*Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH.* John Wiley & Sons, Series in Probability and Statistics, LCCN 2018032640 (ebook) | ISBN 9781119431855, 2018.

[8] M. Panda, *Application of ARIMA and Holt-Winters forecasting model to predict the spreading of COVID-19 for India and its states*, medRxiv,(2020), `https://doi.org/10.1101/2020.07.14.20153908`.

[9] T. C. Mills, *Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting*, 1st Edition, Academic Press, 2019 .

[10] A. Nielsen, *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*, O'Reilly Media,(2019).

[11] M. H. D. Ribeiro, , and others, *Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil* , Chaos, Solitons and Fractals, 135(2020) 109853 .

[12] R. H. Shumway and D. S. Stoffer ,*Time Series Analysis and Its Applications: With R Examples*, Springer Int. Publ.,2017 .

[13] E. T. Vieira , *Introduction to Real World Statistics with Step-by-Step SPSS Instructions*, Routledge, Taylor and Francis Group New York and London, 2017 .

[14] W. A. Woodward, H. L. Gray, A. C. Elliott, *Applied Time Series Analysis with R*, CRC Press, Taylor & Francis Group, LLC, ISBN 9781498734226 (pbk.) 2017.