# COVID-19 Cases in Iraq; Forecasting Incidents Using Box - Jenkins ARIMA Model

Hadeel I. Mustafa
*Computer Information System Department*
*University of Basra*
Basra, Iraq
hadeelismu@gmail.com

Noor Y. Fareed
Pharmaceutics Department
*University of Basra*
Basra , Iraq
lanayo.ly@gmail.com

**Abstract: The pandemic outbreak of COVID-19 created panic all over the world. The mathematical principle in developing forecasting models aims to predict the number of future infections is considered crucial at this stage. The present investigation aims to analyze the time series using the Box-Jenkins method (Diagnostic, The Estimate, and selection, Forecasting) to find the best ARIMA model (Autoregressive Integrated Moving Average) for predicting the numbers of people infected with Covid-19 disease in Iraq. The data used were collected in the period between 1 –March and 31- July. The results showed that the appropriate forecasting model is ARIMA (2,1,5). Depending on this model, they predict the numbers of those infected with COVID-19 daily and for thirty days. Predictive values are consistent with original series values, indicating the efficiency of the model.**

***Keywords: ARIMA, Box-Jenkins, Minitab programming, Iraq, COVID-19.***

## I. Introduction

The quick spread and the highly contagious nature of COVID -19 created a serious crisis worldwide [1]. The absence of specific treatment for this decrease further raises the concerns of the public [2]. Therefore, world governments utilize all the possible measures to prevent the infection and decrease the disease's devastating outcomes [3].

A forecast is a quantitative, probabilistic statement about an unobserved event, outcome, or trend. Its surrounding uncertainty, conditional on previously observed data [4]. Time-series analysis is a powerful tool of forecasting, in which a mathematical model is established according to the regularity and trend of the observed historical values with time [8]. Box–Jenkins model is an autoregressive integrated moving average (ARIMA) model and is the most common time series prediction model [14][15].

The application of the mathematical principle of forecasting to predict the number of future infections relying on existing numbers has been used on many occasions [5],[7]. Forecasting would provide decision-makers with the necessary information required to prepare health care.

In COVID-19, several studies were conducted to predict the future burden utilizing the time series approach. For instance, the ARIMA model was used to forecast COVID-19 future infection in Nigeria[18]. Similarly, Researchers in China created a time series ARIMA model for new COVI-19 cases incidence and death [19]. Also, curve estimation models, Box-Jenkins and Brown/Holt linear exponential Smoothing methods were used to forecast COVID-19 cases in eight different countries[20].

The present investigation aims to postulate the suitable forecasting model for the COVID-19 outbreak in Iraq by modeling actual data using (Box- Jenkin ) models by the Minitab program.

## II. The Methodology Of Research

Time series is defined as a series of recorded values (observations) for a specific phenomenon in limited periods [9]. It represents a historical record over time, and under the influence of economic, social, and environmental factors [10]. The time series is considered stable if the variable values' deviation from the mean values is zero or converts to decay. The stable time series has fixed arithmetic mean, and its variance and co-variations are constant over time, [11] i.e.

$$E(\mathrm{Y}_{\mathrm{t}}) = E(Y_{\mathrm{t+K}}) = \mu \qquad (1)$$

$$\mathrm{Var}\,(\mathrm{Y}_{\mathrm{t}}) = \mathrm{E}[\mathrm{Y}_{\mathrm{t}} - \mathrm{E}(\mathrm{Y}_{\mathrm{t}})]^2 = \mathrm{Var}(\mathrm{Y}_{\mathrm{t+k}}) = \mathrm{E}[\mathrm{Y}_{\mathrm{t+k}} - \mathrm{E}(\mathrm{Y}_{\mathrm{t+k}})]^2 = \gamma(0) = \sigma^2 < \infty \qquad (2)$$

$$\mathrm{Cov}(\mathrm{Y}_{\mathrm{t}},\mathrm{Y}_{\mathrm{t+k}}) = \mathrm{E}[(\mathrm{Y}_{\mathrm{t}} - \mu)\,(\mathrm{Y}_{\mathrm{t-k}} - \mu)\,] = \mathrm{Cov}\,(\mathrm{Y}_{\mathrm{t-k}}, \mathrm{Y}_{\mathrm{t-k+s}}) = \gamma(k) \qquad (3)$$

The Autocorrelation Function(ACF) means that some variables or observations are related to each other during a specific time series period, which are important in clarifying the time series's characteristics. The mathematical formula for the Autocorrelation function as in "(4)":

$$\rho_{\tilde{K}} = \frac{\mathrm{COV}\,(\mathrm{Y}_{\mathrm{t}}, Y_{t+k})}{\sigma_{\tilde{Y}}^2} \quad , \mathrm{t} = 1,2,\dots,\mathrm{N} \ , \ \mathrm{k} = 0,1,2, \qquad (4)$$

The series is stable if it has an Autocorrelation function equal to or close to zero. Meaning, the lower the auto-correlations, the higher k, while for the unstable series, the differences are taken to them and to different degrees to convert them to stable [12]. Model stability is one of the important steps to implement the Box- Jenkins package [13]. To obtain a stable series, we use $\mathrm{W}_{\mathrm{t}}$ as a separate series as follows:

$$\mathrm{W}_{\mathrm{t}} = \nabla^1 Y_t = \mathrm{Y}_{\mathrm{t}} - \mathrm{Y}_{\mathrm{t-1}} \quad , \mathrm{t} = 2,3,\dots,\mathrm{N} \qquad (5)$$

978-1-6654-0432-7/20/$31.00 ©2020IEEE

In general, the temporal chain is stabilized after the difference d, according to the following formula [12]:

$$W_t = \nabla^d Y_t \qquad ,t=d+1,d+2,..,N \qquad (6)$$

The series $\nabla^d Y_t$ is stable, and the model in it is called ARIMA (p, d, q), which is written according to formula “(7)”.

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \ldots + \emptyset_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots, \theta_q e_{t-q} \qquad (7)$$

Where:

P: Degree of Autocorrelation.

d: Number of differences.

q: Degree of moving averages.

$\emptyset$ : Polynomial of degree p for Autocorrelation.

$\theta$ : Polynomial of degree q for parameters of the moving averages.

## III. Box-Jenkins Forecasting Methodology

Three steps must be followed before starting to use the Box-Jenkins models in forecasting, as follows [16] [17] :

A. . *The diagnostic stage includes*:

a. Preparing data by representing the data to stabilize the variance, and taking the differences to obtain a staging sequence.

b. Choosing the appropriate model by examining the data and using the ACF (Autocorrelation function) and PACF functions (Partial Autocorrelation function).

B. . *The Estimate and selection stage, which includes:*

a. Estimate appropriate model parameters and choose the best model using the MAE (mean Absolute Error) and BIC (Bayesian Information Criterion).

b. Diagnostic tests to examine the ACF (Autocorrelation function).

C. . *Forecasting stage This stage involves the use of the prediction model.*

IV. DATA COLLECTIONThe data used in this research compose a daily time series of 151observations representing the total number of positive cases with COVID-19 in Iraq.

The data used in this study are illustrated in Table I. These numbers were provided by the statements of the Iraqi Ministry of Health. The period through which data collected extended from (1 –March- 2020) to (31- July-2020). The average capacity was (813) infected cases while the minimum and maximum value were (3) and (3346) recorded on 5-March and 31-July, respectively. The chain was dispersed from the mean by an Std. Deviation equals (999).

TABLE I: NUMBER OF PEOPLE WITH COVID-19 DURING THE STUDY PERIOD

| | March | April | May | June | July |
|---|---|---|---|---|---|
| Number of people with COVID-19 | 681 | 1391 | 4454 | 42670 | 73500 |

## V. BOX-JENKINS TIME SERIES ANALYSIS MODEL

### A. *The model diagnostic stage*

The data used in this study were represented as a time series shown by Fig.1. It can be seen that the chain proceeds with an increasing trend with the presence of fluctuations. The Autocorrelation function for the series observations was illustrated in Fig. 2. Accordingly, it can be said that the chain is unstable.

To impart stability to the time series, the data were processed by using logarithmic transformation, then taking the first difference for logarithmic transformation. The stable series is shown in Fig. 3. the Autocorrelation coefficients, and the partial Autocorrelation coefficients of the stable series are shown in Fig. 4 and Fig.5, respectively. The time series is ready to identify the appropriate model for the data.

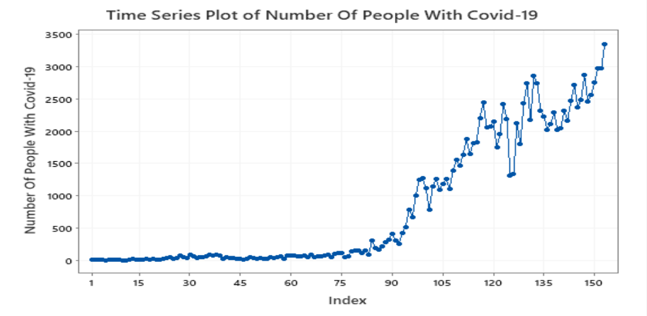

Fig.1: Number of people infected with COVID-19 in Iraq.

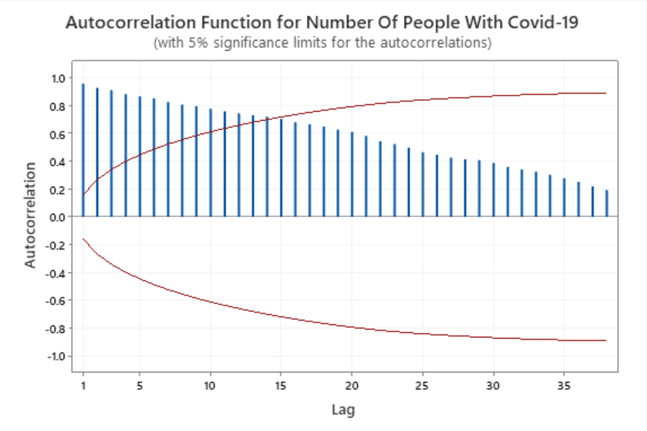

Fig.2. Autocorrelation coefficient of time series data.

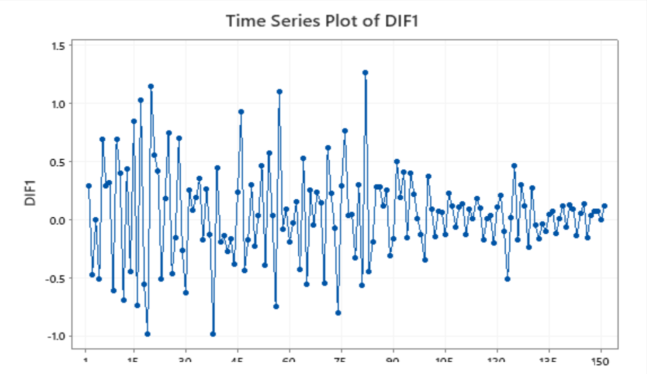

Fig.3. The stable series data.

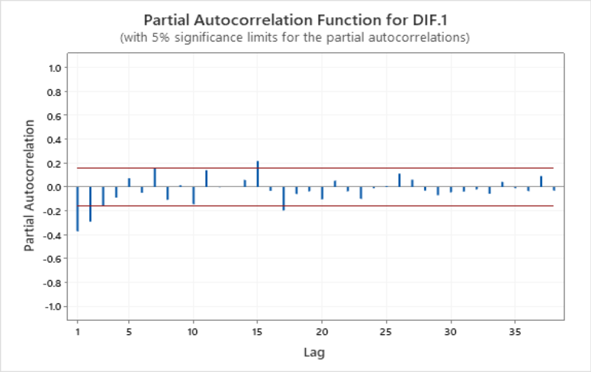

Fig.4. Autocorrelation coefficient of the stable time series.

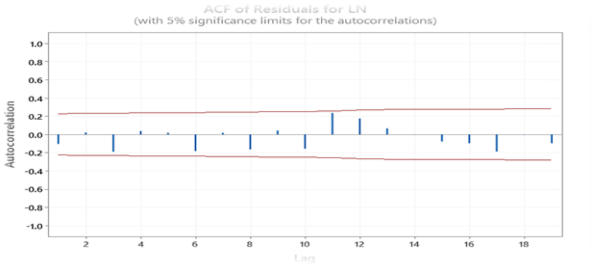

Fig. 5. Partial Autocorrelation coefficient of the stable time series.

*B. The Estimate and selection stage*

The stable series was analyzed by the ARIMA (p, d, q) model assuming d = 1 and (p, q = 0, ..., 5). Minitab statistical program was used for these estimations. All the models were determined by taking all possibilities to estimate the best predictive model. The best models' estimation relies on the principle of the lowest mean Absolute Error (MAE). The possible models are shown in Table II. The chosen model is ARIMA (2,1,5) since it has the lowest MAE with a value equals (0.253), BIC = -1.819. It is necessary to ensure the validity and efficiency of the model. This is done by testing the Autocorrelation coefficients of the residues (errors). Autocorrelation is shown in Fig. 6. It can be observed that all of the residues of the autocorrelation coefficients fall within the confidence limit [-0.2, 0.2]. The normal distribution of errors can be tested by applying Ljung & Box. The following null hypothesis $H_{0,}$ was tested. It states that:

$H_0$: The residues in the logarithmic time series do not follow the normal distribution

It was found that the value of p-value = 0 is (<0.05). Accordingly, we reject the null hypothesis since testing the hypotheses of symmetry. The residual chain's normal flatness indicates that the residual series has the properties of the normal distribution, as shown in Fig. 7.

In this way, the alternative hypothesis is accepted. It states that the errors are distributed in a normal distribution, and we conclude from this that the model is statistically accepted and, therefore, can be used for prediction.

TABLE II: SUGGESTED BOX-JENKINS MODELS FOR COVID-19 SERIES

| P,d,q | MSE | MAE | P,d,q | MSE | MAE |
|---|---|---|---|---|---|
| (0,1,1) | 0.125482 | 0.266 | (0,1,4) | 0.125748 | 0.264 |
| (5,1,4) | 0.115489 | 0.261 | (0,1,5) | 0.126422 | 0.264 |
| (1,1,2) | 0.121513 | 0.253 | (1,1,0) | 0.139021 | 0.274 |
| (1,1,3) | 0.121965 | 0.264 | (2,1,1) | 0.130502 | 0.264 |
| (1,1,4) | 0.127188 | 0.264 | (2,1,2) | 0.125060 | 0.264 |
| (1,1,5) | 0.122667 | 0.256 | (2,1,3) | 0.116865 | 0.257 |
| (2,1,0) | 0.128145 | 0.265 | (5,1,4) | 0.115489 | 0.261 |
| (0,1,2) | 0.126074 | 0.266 | (2,1,5) | 0.120222 | 0.252 |
| P,d,q | MSE | MAE | P,d,q | MSE | MAE |
| (3,1,2) | 0.124016 | 0.256 | (4,1,5) | 0.117305 | 0.261 |
| (3,1,4) | 0.114856 | 0.259 | (5,1,1) | 0.119262 | 0.253 |
| (3,1,5) | 0.124549 | 0.253 | (5,1,2) | 0.120222 | 0.262 |
| (5,1,4) | 0.115489 | 0.261 | (5,1,3) | 0.116629 | 0.253 |
| (4,1,1) | 0.128551 | 0.253 | (5,1,0) | 0.125466 | 0.263 |
| (4,1,2) | 0.118561 | 0.254 | (5,1,4) | 0.115489 | 0.261 |
| (4,1,3) | 0.127430 | 0.254 | | | |

Where *P: Degree of Autocorrelation.

*d: Number of differences.

*q: Degree of moving averages.

*MSE: Mean Squares error.

*MAE: Mean Absolute Error.

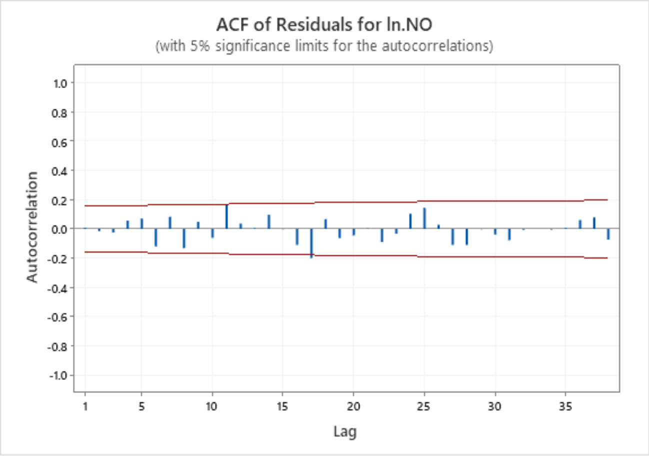

Fig. 6. Autocorrelation coefficient of the residual stable time series.

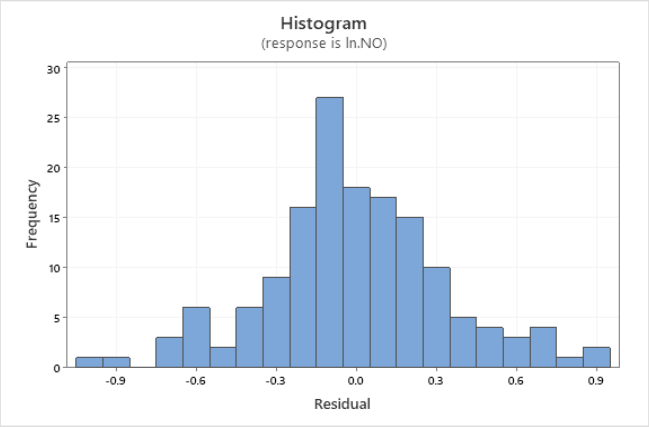

Fig. 7. Normal distribution of residual stable time series.

### C. Forecasting stage

The ARIMA (2,1,5) model is used to predict the number of people expected to be infected with COVID-19 for the next thirty days. The results are summarized in Table III after converting the logarithmic values to the original values. The predicted time series is shown in Fig.8. It follows the same behavior as the original series. Fig.9 illustrates the steps for forecasting.

TABLE .III: PREPARING PEOPLE WITH COVID-19 FOR A PERIOD OF THIRTY DAYS

| 1\8- 10\8 | 11\8 – 20\8 | 21\8 – 30\8 |
|---|---|---|
| 3163 | 4722 | 7105 |
| 3356 | 4845 | 7467 |
| 3394 | 4994 | 7887 |
| 3510 | 5294 | 8198 |
| 3768 | 5601 | 8471 |
| 3980 | 5778 | 8877 |
| 4058 | 5958 | 9363 |
| 4187 | 6285 | 9757 |
| 4464 | 6646 | 10097 |
| 3163 | 6885 | 10558 |

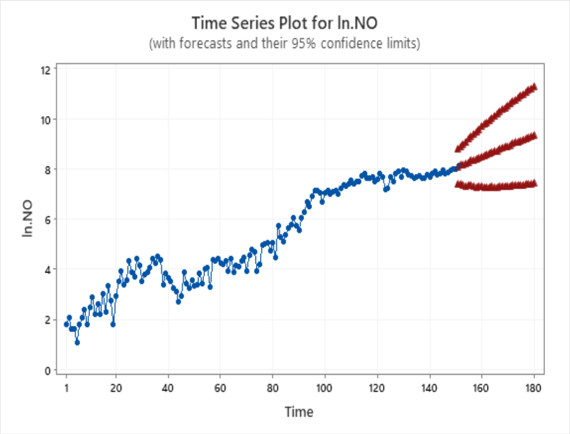

Fig. 8. Predicted time series representation

## VI. Conclusions

By Studying the number of positively confirmed cases with COVID-19 in Iraq, it can be noticed that there is increasing numbers and the series is unstable. Making logarithmic conversions and taking the first difference of the series imparted stability to the series. The ARIMA (2,1,5) model was found to be an efficient and appropriate model for string data as the predicted chain possesses the same characteristics as the original chain. It is also noticed an increasing trend in the number of infected individuals due to a lack of the people's commitment to preventive measures against the disease.

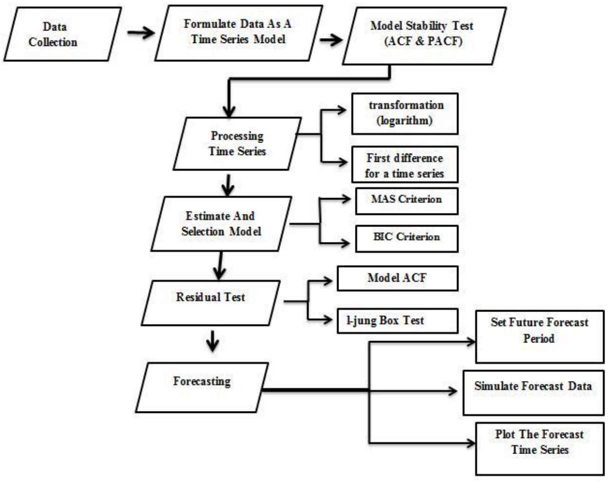
Data Collection
Formulate Data As A Time Series Model
Model Stability Test (ACF & PACF)
Processing Time Series
transformation (logarithm)
First difference for a time series
Estimate And Selection Model
MAS Criterion
BIC Criterion
Residual Test
Model ACF
l-jung Box Test
Forecasting
Set Future Forecast Period
Simulate Forecast Data
Plot The Forecast Time Series

Fig. 9. Flowchart (the steps for forecasting)

## VII. Recommendations

A comparative study between Box - Jenkins models and artificial neural network models applying the same error criteria can be taken place. It can also be used to determine the most accurate method for predicting the numbers of people with COVID-19 disease. It is also vital to ensure forced homestay for all individuals by the Iraqi military and police forces and raise people's perception of the COVID -19 infection control protocol.

## References

[1] Y.R. Guo, Q.D. Cao, Z.S. Hong, Y. Y. Tan, S. D. Chen, H.J. Jin, K.S. Tan, D.Y. Wang, “The Origin, Transmission and Clinical Therapies on Coronavirus Disease 2019 (COVID-19) Outbreak–an Update on The Status”, *Military Medical Research*, Vol.7, no.1, pp.1-10, December 2020.

[2] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, R. Di Napoli, “Features, Evaluation and Treatment Coronavirus (COVID-19)”, Stat Pearls, 8 March 2020.

[3] H.A. Rothan, S.N. Byrareddy, “The Epidemiology and Pathogenesis of Coronavirus Disease (COVID-19) Outbreak”, Journal of Autoimmunity, Vol. 109, p.102433, 26 Feb 2020.

[4] J.S.Armstrong , “ Introduction” in Principles of Forecasting: A Handbook for Researchers and Practitioners. Springer Science & Business Media, Ch.1, p [6] S.Sharmin, I.Rayhan , “Modelling of Infectious Diseases for Providing a Signal of Epidemics: A Measles Case Study in Bangladesh”and nutrition, Vol 29, no. 6, pp.567-573.29 December 2011.

[5] U. Helfenstein, “Box& Jenkins Modelling of Some Viral Infectious Diseases”. Statistics in Medicine”, Vol.5, no.1, pp.37-47. January 1986.

[6] E.A.Frah, A.A.Alkhalifa , “ Tuberculosis Cases in Sudan; Forecasting Incidents 2014-2023 Using Box & Jenkins ARIMA Model”. American Journal of Mathematics and Statistics.Vol.6, no. 3, pp.108-114, 2016.

[7] P.Newbold , C.W.Granger , “Experience with Forecasting Univariate Time Series and the Combination of Forecasts” , Journal of the Royal Statistical Society , Vol 137 , no.2 ,pp.131-146 , March1974 .

[8] D.J.Bartholomew , “Time Series Analysis Forecasting and Control ” , Journal of the Operational Research Society , Vol .22 , no.2 , pp.199-201, Jun1971 .

[9] A.V.Metcalfe, P.S.Cowpertwait , “ Forecasting Strategies” in Introductory time series with R. Springer-Verlag New York,Ch. 3 , pp. 45-66 , 2009. p.1-12 ,2001.

[10 J.Makhoul , "Linear prediction: A tutorial review." Proceedings of the IEEE, Vol .63no.4, pp 561-580, 1975.

[11] M.T. Hagan, S.M. Behr, “The Time Series Approach to Short Term Load Forecasting”, IEEE Transactions on Power Systems. Vol 2, no.3, pp.785-791, August 1987.

[12] W.S. Hopwood, J.C. McKeown, P. Newbold, “Time Series Forecasting Models Involving Power Transformations”, Journal of Forecasting,Vol.3, no.1, pp.57-61,January1984 .

[13] G. Box . “Box and Jenkins: Time Series Analysis, Forecasting and Control”. In A Very British Affair, pp. 161-215, 2013.

[14] U. Helfenstein , “Box-Jenkins Modelling in Medical Research”, Statistical Methods in Medical Research, Vol.5 , no .1, pp.3-22 , March 1996 .

[15] D.Pena , Tiao GC, Tsay RS. “Univariate Time Series: Autocorrelation, linear prediction, Spectrum and state space Model” in A Course in Time Series Analysis, John Wiley & Sons, Ch. 2, pp. 25-53, 25 January 2011.

[16] S.Makridakis , M. Hibon , ARMA Models And The Box–Jenkins Methodology , Journal of Forecasting , Vol.16 , no. 3 ,pp.147-163, May 1997.

[17] R. Rauf Ibrahim, H. Oluwakemi Oladipo, "Forecasting the spread of COVID-19 in Nigeria using Box-Jenkins Modeling Procedure ", medRxiv, 2020.

[18] Q.Yang, J. Wang, Hongli Ma, Xihao Wang, "Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy", Journal of Infection and Public Health, 2020.

[19] Harun Yonar, A. Yonar, M. Agah Tekindal, Melike Tekindal, "Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods", EJMO, 2020;4(2):160–165.

[20] A. S. Hamood, S. B Sadkhan, “Keywords Sensitivity Recognition of Military Applications in Secure CRNs Environments”, 2017 Second Al-Sadiq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), pp: 96- 101

[21] S. B Sadkhan, S. J Mohammed, M. M Shubbar, “Fast ICA and JADE Algorithms for DS-CDMA”, 2017 Second Al-Sadiq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), pp: 325- 329.

[22] A. Alkhayyat, S. B Sadkhan, Q. H Abbasi, “Multiple Traffics Support in Wireless Body Area Network over Cognitive Cooperative Communication”, 2019 2nd International Conference on Electrical, Communication, Computer, Power and Control Engineering (ICECCPCE), pp: 199- 203.

[23] S. B Sadkhan, S. T. Hasson, M. T Gaata, “Image Quality Assessment by Combining Fuzzy Similarity Measures using Neural Network”, 2012 International Symposium on Photonics and Optoelectronics (SOPO 2012), China.