Utilizing Auto-Regressive Integrated Moving Average to Predict Newly Coronavirus Cases in Libya

Mansour Alssager * **Zulaiha Ali Othman**

[1]*Faculty of Information Technology, Sebha University, Sebha, Libya*

Abstract

Currently, Coronavirus is a major worldwide threat. It has affected millions of people around the world, resulting in hundreds of thousands of deaths. It is indeed important to forecast the number of new cases in aims to assist in disease prevention and healthcare service readiness. Many researchers used different mathematical and machine learning methods to forecast the pandemic's future trend. This research proposes an autoregressive integrated moving average model to forecast the estimated daily new cases in Libya over the next three months. The total number of confirmed cases is pre-processed and used to forecast the virus's spread. The cumulative number of confirmed cases is pre-processed and used to forecast the virus's spread. Based on the result obtained from the experiment, the number of cases expected to rise in the near future, reaching up to 1250 new cases every day. This research would help the government and medical staff members to plan for the upcoming conditions, as a result, increase the readiness of healthcare systems.

Keywords: COVID19, Coronavirus, Pandemic, ARIMA model, Epidemic, forecast, Libya

استخدام نموذج الانحدار الذاتي والمتوسط المتحرك للتنبؤ بحالات الإصابة بفيروس كورونا الجديد في ليبيا

منصور الصغير[1]* زوليخا علي عثمان[2]*

[1] كلية تقنية المعلومات، جامعة سبها، سبها، ليبيا

كلية التكنولوجيا وتقنية المعلومات، الجامعة الوطنية الماليزية، سلانجور، ماليزيا

الملخص

في وقتنا الحالي، يعد وباء كورونا تهديدًا عالميًا كبيرًا. فلقد أثر على حياة الملايين من الناس حول العالم، وأدى إلى وفاة مئات الآلاف. بناء على ذلك من المهم التنبؤ بعدد الحالات الجديدة بهدف المساعدة في الوقاية من المرض وكذلك لمساعدة الرعاية الصحية في الاستعداد المبكر لأي طارئ. استخدم العديد من الباحثين طرقًا مختلفة لتعلم الآلة للتنبؤ بالاتجاه المستقبلي للوباء. في هذا البحث تم اقتراح استخدام نموذج الانحدار الذاتي والمتوسط المتحرك للتنبؤ بالحالات الجديدة اليومية في ليبيا على مدى الأشهر الثلاثة المقبلة. حيث تمت معالجة العدد الإجمالي للحالات المؤكدة مسبقًا واستخدامها للتنبؤ بانتشار الفيروس. حيث تم معالجة العدد التراكمي للحالات المؤكدة واستخدامها للتنبؤ بمدي انتشار الفيروس. وبناءً على النتيجة التي تم الحصول

* man.essgaer@sebhau.edu.ly

عليها من التجربة، من المتوقع أن يرتفع عدد الحالات في المستقبل القريب ليصل إلى 1250 حالة جديدة كل يوم. سيساعد هذا البحث أعضاء الطاقم الطبي والحكومي على التخطيط للظروف القادمة، مما يزيد من جاهزية نظام الرعاية الصحية في البلاد.

Introduction

Coronavirus disease (COVID-19) is a new and serious disease that emerged in China in December 2019. It has rapidly spread across the world, with more than 150 million people infected and 3 million deaths as of May 2021. This disease has a high case fatality rate, especially among the elderly population, with underlying co-morbidities ranging from 2% to 18% [1]. Up to date, the most infected country are the USA, Brazil, India, France, and Russia. In the African continent, Libya ranked the seventh based on the infected people rate, which reaches up to (156,849), and the death toll of (2618) [2]. Currently, Libya is in a serious stage of epidemic spread, because of years of violence leaving its healthcare system highly vulnerable. With limited equipment for testing, there is very little protective gear and there is a severe shortage of medical workers.

Although the Libyan governments have attempted to impose different precautionary measures, such as raising public awareness about the virus's general behavior, and preventive mechanisms such as sanitization of public governmental facilities, quarantine of suspected and infected cases, closure of schools, and mosques. Because most of the people negligent preventive measures such as social distancing and mass gathering, the transmission has remained rapid.

In that regard, it is critical to building models that are both computationally competent and practical, so that they can assist in terms of outbreak management and resource planning. Modeling the disease to determine its evolution and inflection point, as well as providing a future forecast of a possible number of daily cases. In return, that could assist the medical system to get prepare for the new patients. The statistical prediction models can be used to forecast and monitor the global disease challenge. Time series models such as Auto-Regressive Integrated Moving Average (ARIMA) and sensual ARIMA (SARIMA) have been widely used for perfecting infectious disease trends [3]. Such as kidney disease [4], cardiovascular disease [5], malaria incidence [6], Dengue fever [7], and Tuberculosis [8], and have ascertained the effectiveness of the model.

ARIMA models are appropriate for fitting a variety of trajectories and for investigating the short-term consequences of disease trends [9, 10]. ARIMA models are preferred over SARIMA in the current situation because the data points span less than two years, negating the effect of seasonality. Several studies using ARIMA models were proposed to predict the outbreak.

ARIMA models are appropriate to fit several trajectories and suitable for investigations into short-term effects of disease trends In the current situation, ARIMA models are preferred over SARIMA since the data points span less than two years, negating the effect of any seasonality. Several studies based on ARIMA models proposed to forecast the outbreak in the USA, Italy, Brazil, Saudi Arabia, and South Korea [11]. However, to date, there are no studies to forecast the COVID-19 outbreak in Libya.

Therefore, this study proposed an ARIMA model for predicting Covid-19 cases in Libya over the period after April 4, 2021. The model developed based on using daily-confirmed cases to predict the number the cases for the next three months.

Matrial and Methods

Data Descriptions

The data collected from the official website of the Johns Hopkins University, from 24 March 2020 to 1 April 2021: *github.com/CSSEGISandData/COVID-19*. Python language was used to perform the statistical data analysis on the confirmed COVID-19 cases. The Figure 1 shows the total number of cumulative confirmed cases in (black) and daily confirmed cases (in blue) of COVID-19 cases in Libya.

Data forecasting method

Time series models are intended to predict future values by analyses of past and present [12]. It processes the historical data and tries to extract and describe some incidents and forecast future values. In recent years, many researchers have tried to predict the COVID-2019 trend and final size using different approaches [13-15]. The ARIMA model is one of them [1, 11, 12]. It is considered one of the more used prediction models for time series forecasting.

Auto-Regressive Integrated Moving Average model

ARIMA is one of the popular and applicable time series models in terms of taking into account changing trends, periodic changes, and random disturbances in the time series [12, 16]. Moreover, It applicable for all types of data, which include non-stationary time series, even if there is no systematic change in mean (no trend), no systematic change invariance, and periodic variations have been removed [17]. Practically, most time series are non-stationary in practice, several studies recommend eliminating all non-stationary sources of variance in time series results. Applying regular differencing and log transformation to the original time series is a common technique used in the literature (x_t). If differencing a series d times makes it into a stationary and then series said to follows an autoregressive integrated moving average process, denoted by ARIMA (p, d, q) and can be written as:

$$\emptyset(\beta)\nabla^d x_t = \theta(\beta)w_t \qquad (1)$$

where:

The autoregressive operator can be expressed as $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$

Moving average operator $(\theta(B))$: $\theta(B) = 1 + \theta_1 B + \varphi_2 B + \cdots + \varphi_q B$

Differencing operator $\nabla^d = (1 - \beta)^d$, it is the expression of dth consecutive differencing to make series stationery.

w_t is a Gaussian white noise series with mean zero and variance (σ2w).

Testing for stationary: Before analysing, time series must become station-based on mean and variance constant over time [17]. In this study, three method for analysis of data stationary are used:

1. Correlogram test: the correlation coefficient between two values in a time series, known as the Autocorrelation Function (ACF). That is, the degree of similarity between a certain sequence of time series and a lagged version of itself over successive time intervals. If the sample of ACF declines slowly in non-seasonal and seasonal lags [18], differences for non-stationary series are required.

2. The augmented Dickey–Fuller test: used in a time series sample to test the null hypothesis that a unit root is present. On other hand, is different depending on which version of the test is used, but is usually stationarity or trend-stationarity [19].

3. Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test: which used to test of the null hypothesis that an observable series is trend-stationary (stationary around a deterministic trend) [20].

Building ARIMA model : Several steps should taken to create an ARIMA model for a certain time series, namely: Model identifying, parameter estimates, diagnostic checking for the identified model, Application of the model (forecasting) [19]:

1. Check the stationarity of the observation data. If the sequence is not stationary, perform a difference or logarithmic transformation until it becomes a stationary time series;

2. Calculate the ACF and PACF of the stationary sequence, and use ARIMA model to identify preliminary values of the autoregressive order, p, the order of differencing, d, and the moving average order, q.

3. Perform model tests, including the significance test of the model and the significance test of the parameters.

4. To predict the epidemic situation in the 3 months.

Results and discussions

Descriptive analysis results

Figure 1 shows the daily-confirmed cases (in blue), which found to show a high degree of variability (noise), and skewedness that was determined by estimating the variance and test of normality, the main reason for the noise is due to the zero reported cases on Friday by the Libyan authorities. To remove the noise component, the data averaged on a moving window using a 5-Days Moving Average (5DMA) smoother (in red). The 5DMA smoother found to have a lower variance, while still maintaining the general observed trend of the original time series as shown in figure. Hence, the 5DMA smoother chosen subsequently as the dependent variable for the forecast model.

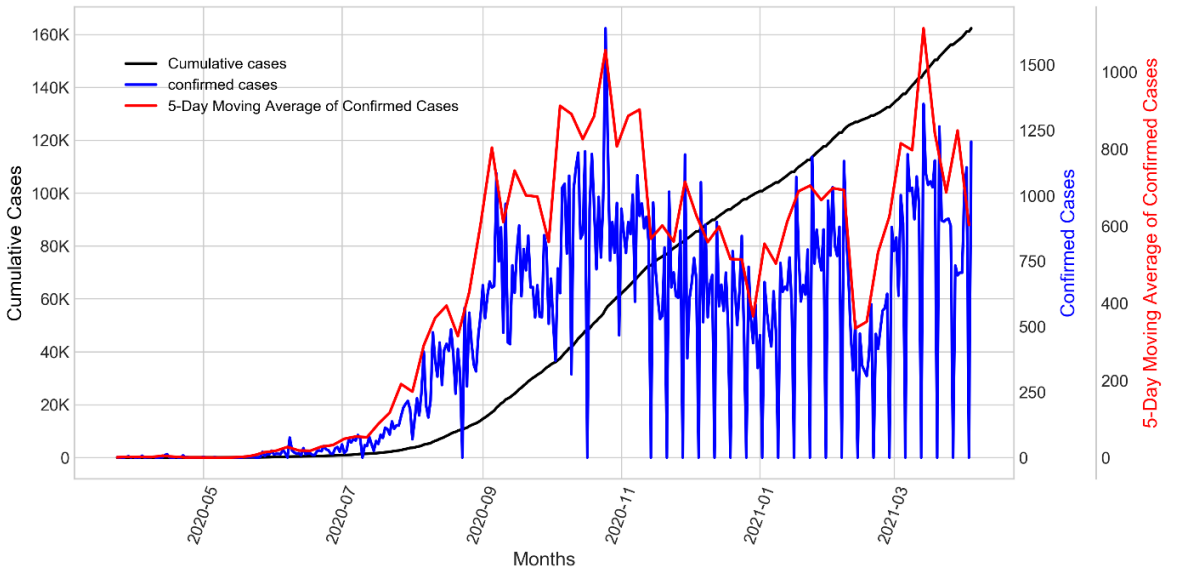

Figure -1 The cumulative (in black), daily confirmed (in blue), 5DMA (in red) cases in Libya as a function of days from March 24, 2020 to April 1, 2021.

Figure 2, shows the 5DMA (in blue) and the monthly moving average (in black) of the confirmed COVID-19 cases. Based on the 5DMA series Libya reported its first confirmed case on 24 March 2020. The government of Libya has immediately taken different measures to prevent and control the pandemic, but the case has been rapidly increasing and widely distributing. Starting from July 2020, the country confirmed cases gradually increase in rapid base, we can clearly see that it took almost 4 months after the first confirmed case to exceed 100 confirmed cases, until reach to the first wave peak with 1639 daily reported cases on 25 October 2020, with about fifty six thousand cumulative cases (56013) reported. The second wave peak occurred in January 2021 with (1148) confirmed new cases, followed by the third wave peak, which take place on March 2021 with (1350) confirmed new cases. However, between the second and the third peak, there was a fluctuation in the number of reported cases ranging from 300 to 900. It worth to mentioned that before the first peak, the governments closes all the schools, mosques, and decrees the governmental staff member up to 30%. Unexpectedly, the number of confirmed cases too much enlarged as the figure shows.

Moreover, the monthly moving average curve is divided into three category as shows in the figure 2, the red colour zone indicate new cases increase, green colour zone indicate stable new cases curve, and yellow zone indicate new cases decline. The period between Aprils to July 2019 is having the lowest confirmed new cases, the curve is slightly increase in the mentioned period, which indicate the success of governmental applied policy in preventing the spread of the disease. Despite the fact that on July to October is the summer season in Libya, however, it is the most period that the monthly-confirmed cases curve is in a sudden leap. That may be because of school holidays that leads to more traffic between cities. Whereas, between Octobers to December the winter season in Libya, the monthly confirmed case on decline as the yellow area shows. Moreover, there is another green zone from December 2019 to February 2021, whereby the curve is almost steady. In addition, there is another period where the new cases curve gradually increase which from February to April 2021.

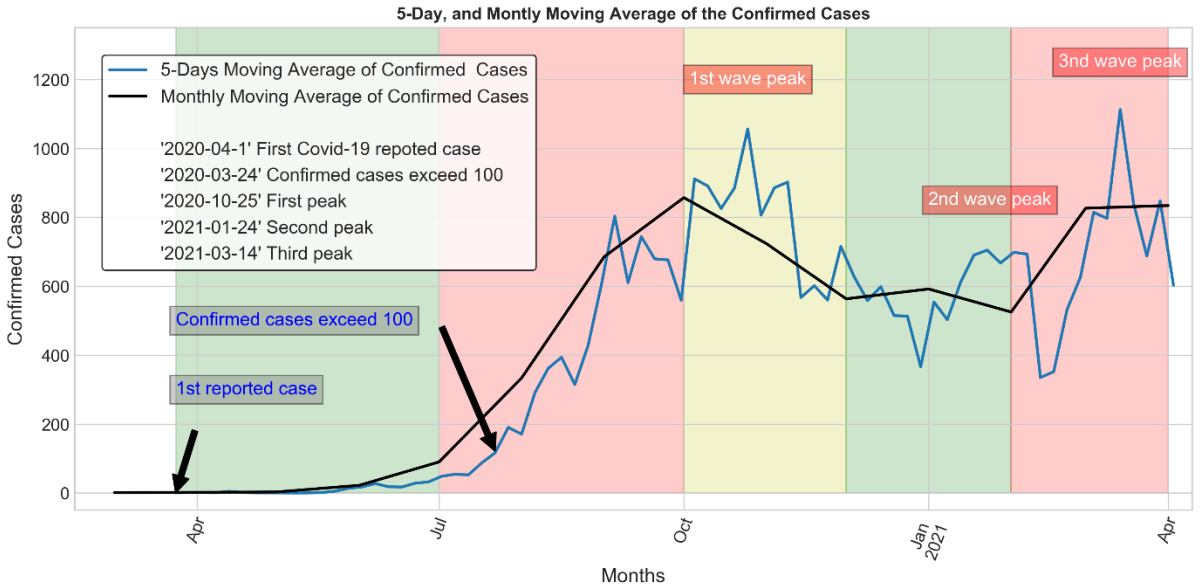

Figure -2 Distribution of 5DMA, and monthly moving average (from 2020-03-24 to 2021-03-27).

Result from ARIMA modelling forecasting on COVID-19 confirmed cases

Stationary test result:

The first step in any time series analysis like ARIMA modelling is to determine whether the time series is stationary or not [11]. For this purpose, three method used in this study ACF, ADF, and KPSS. Based on the Figure 1, which shows that there have been an obvious systematic upward change over the mean for the 5DMA series.

According to [18], if sample autocorrelation functions plot decay slowly in non-seasonal lags and took long lag to fall inside 95% significant line, it is an indication for non-seasonally non-stationary of the series. As it observed from Figure 3, sample autocorrelation function is also supporting evidence for the presence of regular linear trend behaviour for the considered series, so as it shows decaying slowly in non-seasonal lags to fall down inside 95% significant line.

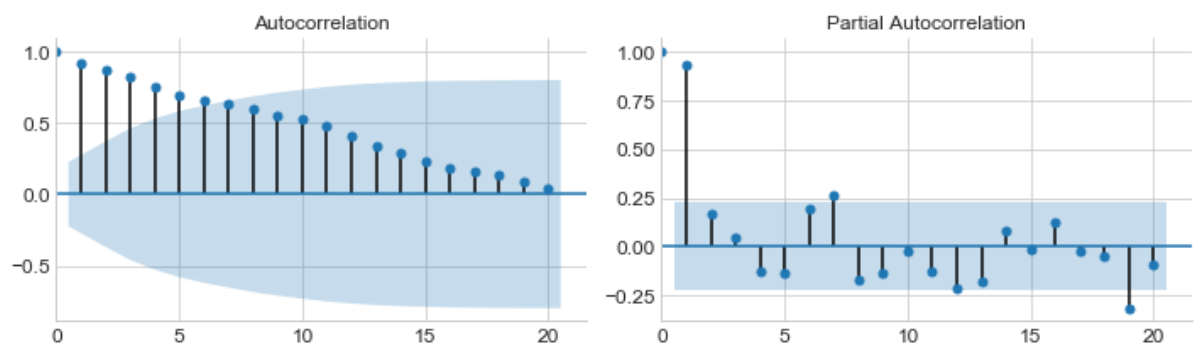

Figure- 3 Auto Correlation Function for 5DMA series for 20 lags.

To stabilize the variance, we used 5DMA of the confirmed cases per day. For investigating the stationarity of time series, we take the support of the KPSS and ADF test, and results shown in Table (1). Based on a 5% significance level, the KPSS test that is

testing the time series for stationary around a deterministic trend reject the hypothesis of stationarity of time series without making any difference. While the ADF, which is an augmented model, do not reject null hypothesis. After the KPSS agree on the stationarity of time series, a candidate ARIMA models recommended predicting the future COVID-19 cases in Libya.

Table-1 Summary of ADF and KPSS test

Series Level	ADF test statistic (p-value)	KPSS test statistic (p-value)	Decision: ADF Null at unit root (non-stationary)	Decision: KPSS Null at unit root (non-stationary)
5DMA	-1.38 (0.58)	0.53 (0.03)	Don't Reject null hypothesis	Reject null hypothesis

Key: 5DMA = 5-days Moving Average

Building ARIMA Model

The complete time series data divided into two split: a training set between 24 March 2020 to 1 December 2020 for fitting the model, and a validation set from 2 December 2020 to 1 April 2021 to validate and test the model. The validation phase performed to decide which model is sufficiently appropriate for future data. A model has been found suitable for the data observed when there are small and random differencing between the observed and the forecasted series (i.e. the residual cases) [13-15].

According to [17, 18], All the models that fulfilled and passed all residual tests (normality, independence, and homogeneity) and the parameters significantly differ from zero must be included and selected as candidate model for prediction. Then, all candidate models are compared their AIC and BIC, and the one which has least will be selected as the best model for prediction model. Table 2 shows the AIC values for different p, d, and q parameters. We have tested different values of p, d, and q parameters ranging from zero to four.

Accordingly, among the candidate models, ARIMA (2, 1, 2) chosen as the best models for predicting COVID19 cases of Libya, since it's the least AIC value. This was because it fulfils all diagnostic tests (independence, normality and homogeneity) and has least AIC and BIC (See Table 3),

Table 2- Comparison of tested ARIMA models.

Candidate model	Selection Criterion (AIC)	Best Model
ARIMA (0, 1, 0)	790.39	ARIMA(2, 1, 2)
ARIMA (0, 1, 1)	788.02	
ARIMA (0, 1, 2)	789.98	
ARIMA (0, 1, 3)	790.17	
ARIMA (1, 1, 0)	788.89	

ARIMA (1, 1, 1)	790.04
ARIMA (1, 1, 2)	791.86
ARIMA (1, 1, 3)	792.47
ARIMA (2, 1, 0)	789.17
ARIMA (2, 1, 1)	791.34
ARIMA (2, 1, 2)	764.96
ARIMA (2, 1, 3)	789.39

Best model: Candidate model with least AICc value

The best model estimated in Table 3 and eventually employed for forecasting the daily spread series of COVID19. The forecast is available in Table 4.

Table 3- Parameters of the Best ARIMA models

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	-0.4354	0.089	-4.886	0.000	-0.610	-0.261
ar.L2	-0.9638	0.065	-14.793	0.000	-1.091	-0.836
ma.L1	0.2850	0.134	2.130	0.033	0.023	0.547
ma.L2	0.8606	0.122	7.082	0.000	0.622	1.099
ar.S.L12	-0.2616	0.446	-0.586	0.558	-1.136	0.613
ar.S.L24	-0.5445	0.234	-2.329	0.020	-1.003	-0.086
ma.S.L12	-0.4740	0.662	-0.716	0.474	-1.772	0.824
sigma2	1.136e+04	3264.255	3.482	0.000	4966.744	1.78e+04

Table 4- Prediction of total confirmed cases of COVID-19 for the next three months according to ARIMA models with 95% confidence interval

Date	**Forecast Lower limit**	**Upper limit**	**Mean value**
2021-03-24	712.0	1131.0	922.0
2021-03-29	655.0	1203.0	929.0
2021-04-03	731.0	1376.0	1054.0
2021-04-08	631.0	1394.0	1013.0
2021-04-13	575.0	1433.0	1004.0
2021-04-18	575.0	1496.0	1035.0
2021-04-23	583.0	1576.0	1080.0
2021-04-28	560.0	1635.0	1097.0

2021-05-03	728.0	1862.0	1295.0
2021-05-08	584.0	1768.0	1176.0
2021-05-13	444.0	1695.0	1069.0
2021-05-18	436.0	1746.0	1091.0
2021-05-23	409.0	1794.0	1102.0

The forecasted COVID-19 confirmed new cases presented in Table 4 with 95% confidence interval (CI) plotted in Figure 4. According to the forecast, the number of confirmed COVID-19 cases expected to increase considerably in the coming three months. Based on the red line which is 1000 confirmed cases limit, the forecast cases is almost exceed the 1000 cases for all the three months period. The increase is highly suspected by the researchers to be associated with the attitude of the people.

Some Libyans believe COVID-19 not real, and therefore do not adhere to follow the rules and regulations put in place by health experts. People also engage in social gatherings and communicate with still infected people. negligence on the side of the few people who did not follow the suggested 14 days of isolation following their return from abroad may also might be consider as another reasons. Another factor contributing to the rise in the number of the confirmed cases is the belief that a lack of confidence in government institutions confirmed the COVID-19 findings.

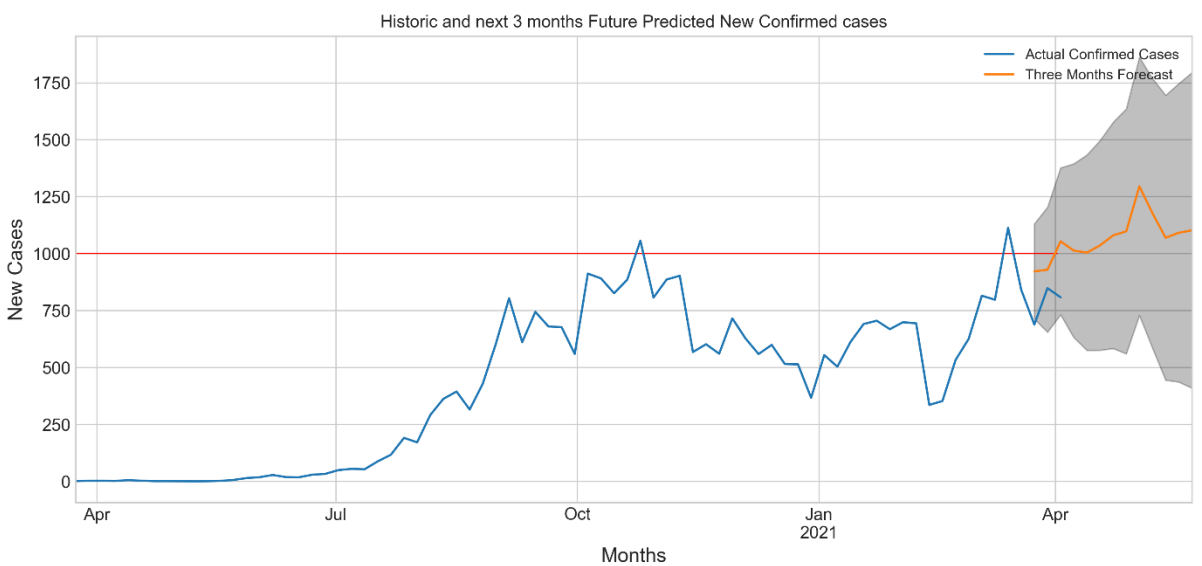

Figure -4 Three-month future confirmed cases predictions

Conclusion:

In order to support the prevention of the disease, and aid in the healthcare service planning and preparation, we have conducted this study to examine the best model for the prediction of confirmed COVID-19 infection cases, and to use that model for forecasting future confirmed cases in Libya. Based on the model forecast, with all other things being

equal, the confirmed cases expected to reach to thousand conformed cases in the coming three months. However, Libya can still control the situation if the prevention measures such as quarantine and city sanitization strictly followed. This study has several implications to practice especially the government. Specifically, government can use the result of this study to design measures that will curtail the uprising trend of COVID-19. The upward trend of confirmed COVID-19 cases can prevented when the people take appropriate preventive measures. The prediction models come up in this study will help the government and medical workforce to prepared for the upcoming situations and have more readiness in healthcare systems. The Libyan Healthcare system known to be weak and a laxity in taking stringent measures to curtail the disease will spell result into greater loss of lives to the virus.

References

[1] G. Onder, G. Rezza, and S. Brusaferro, "Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy," *Jama,* vol. 323, pp. 1775-1776, 2020.

[2] F. B. Hamzah, C. Lau, H. Nazri, D. V. Ligot, G. Lee, C. L. Tan*, et al.*, "CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction," *Bull World Health Organ,* vol. 1, 2020.

[3] R. Allard, "Use of time-series analysis in infectious disease surveillance," *Bulletin of the World Health Organization,* vol. 76, p. 327, 1998.

[4] J. Coresh, H. J. Heerspink, Y. Sang, K. Matsushita, J. Arnlov, B. C. Astor*, et al.*, "Change in albuminuria and subsequent risk of end-stage kidney disease: an individual participant-level consortium meta-analysis of observational studies," *The lancet Diabetes & endocrinology,* vol. 7, pp. 115-127, 2019.

[5] C. Anderson, K. Teo, P. Gao, H. Arima, A. Dans, T. Unger*, et al.*, "Renin-angiotensin system blockade and cognitive function in patients at high risk of cardiovascular disease: analysis of data from the ONTARGET and TRANSCEND studies," *The Lancet Neurology,* vol. 10, pp. 43-53, 2011.

[6] S.-s. Zhou, F. Huang, and Y.-z. Shen, "Application of ARIMA model on prediction of malaria incidence," *J Pathogen Biol,* vol. 2, pp. 284-6, 2007.

[7] S. Polwiang, "The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017)," *BMC infectious diseases,* vol. 20, pp. 1-10, 2020.

[8] Q. Liu, Z. Li, Y. Ji, L. Martinez, U. H. Zia, A. Javaid*, et al.*, "Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses," *Infection and drug resistance,* vol. 12, p. 2311, 2019.

[9] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renewable and Sustainable Energy Reviews,* vol. 74, pp. 902-924, 2017.

[10] W.-Y. Chang, "A literature review of wind forecasting methods," *Journal of Power and Energy Engineering,* vol. 2, p. 161, 2014.

[11] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, and H. Perez-Meana, "Forecasting of COVID19 per regions using ARIMA models and polynomial functions," *Applied Soft Computing,* vol. 96, p. 106610, 2020.

[12] R. Takele, "Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries," *Infectious Disease Modelling,* vol. 5, pp. 598-607, 2020.

[13] M. Batista, "Estimation of the final size of the COVID-19 epidemic," *MedRxiv,* 2020.

[14] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in China, Italy and France," *Chaos, Solitons & Fractals,* vol. 134, p. 109761, 2020.

[15] R. Gupta and S. K. Pal, "Trend Analysis and Forecasting of COVID-19 outbreak in India," *MedRxiv,* 2020.

[16] S. T. Nassir, A. B. Khamees, and W. T. Mousa, "Estimation the Missing Data of Meteorological Variables In Different Iraqi Cities By using ARIMA Model," *Iraqi Journal of Science,* vol. 59, pp. 792-801, 04/29 2018.

[17] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer, *Time series analysis and its applications* vol. 3: Springer, 2000.

[18] G. C. Reinsel, *Elements of multivariate time series analysis*: Springer Science & Business Media, 2003.

[19] G. Ding, X. Li, Y. Shen, and J. Fan, "Brief Analysis of the ARIMA model on the COVID-19 in Italy," *medRxiv,* 2020.

[20] A. Jadevicius and S. Huston, "ARIMA modelling of Lithuanian house price index," *International Journal of Housing Markets and Analysis,* 2015.