International Journal of Statistics and Probability; Vol. 9, No. 6; November 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Statistical Modeling and Forecast of the Corona-Virus Disease (Covid-19) in Burkina Faso

VICTORIEN F. KONANE[1] & Ali TRAORE[2]

[1] Département de Matématique, Université Joseph KI-ZERBO, 03 BP 7021 Ouagadougou 03, BFA, Burkina Faso[1]

[2] Laboratoire de Mathématiques et d'Informatique, 03 BP 7021 Ouagadougou 03, BFA/Zone, Burkina Faso

Correspondence: VICTORIEN F. KONANE, Dpartement de Matmatique, Universit Joseph KI-ZERBO, 03 BP 7021 Ouagadougou 03, BFA, Burkina Faso. E-mail: kfourtoua@gmail.com

Received: September 8, 2020 Accepted: Ocotber 9, 2020 Online Published: October 19, 2020

doi:10.5539/ijsp.v9n6p76 URL: https://doi.org/10.5539/ijsp.v9n6p76

Abstract

In this paper, we present and discuss a statistical modeling framework for the coronavirus COVID-19 epidemic in Burkina Faso. We give a detailed analysis of well-known models, the ARIMA and the Exponential Smoothing model.

The main purpose is to provide a prediction of the cumulative number of confirmed cases to help authorities to take better decision.

We made prediction of the cumulative number of cases from 4^{th} may to 2^{nd} June.

Keywords: COVID-19, ARIMA models, exponential smoothing models, forecasting

1. Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by a new virus that has never been identified in humans before. This virus causes respiratory illness with symptoms like cough, fever and, in the most severe cases, pneumonia. The new COVID-19 is mainly spread through contact with an infected person, when they cough or sneeze, or through droplets of saliva or nasal secretions. The virus appeared for the first time on December 2019 in Wuhan, China. In less than four months, it has spread to more than 210 countries around the world. Africa got its first case of COVID-19 the 14th of February in Egypt and the first confirmed case in sub-Saharan Africa was in Nigeria.

In Burkina Faso, the first cases appear the 9^{th} of March. Up to the date of 9 April, Burkina Faso was one of the West African countries most affected by the pandemic with 443 cases including 146 cured and 19 deaths.

After the declaration of the first cases of COVID-19 in Burkina Faso, the leaders and the people of the country were troubled, schools were closed up one week later. Foreign radio and TV channels, predicted millions of confirmed cases in Africa. Face with all these statistics, we become concerned about the case of Burkina Faso.

Since the start of the pandemic, scientists all around the world have carried out several studies in various fields in several countries. See for instance ([Ivorra and Ramos, 2020], [Chen et al., 2020], [Jia et al., 2020],[Tang et al., 2020], [Liu et al., 2020], [Chen et al., 2020], [Maleki et al., 2020], [Khan and Gupta, 2020]). ([Ivorra and Ramos, 2020]) studied the validation of the forecasts by using a Be-CoDis mathematical model. ([Liu et al., 2020]) tried to understand the dynamic of the COVID-19 through the understanding of the unreported cases. They have developed a compartmental model to predict the behavior of the disease. ([Chen et al., 2020]) developed a Bats-Hosts-Reservoir-People transmission network model for simulating the potential transmission from the infection source to the human infection. As a result, they computed the reproduction number R_0. ([Khan and Gupta, 2020]) used times series to forecast the confirmed and recovered cases of COVID-19. More precisely, they used the family of Autoregressive time series models based on two-piece scale mixture normal distributions, called TPCSMNCAR models to analyze the real data of con?rmed and recovered COVID-19 cases. ([Maleki et al., 2020]) have adopted uni-variate time series models to predict the number of COVID-19 infected cases that can be expected in upcoming days in India. The ARIMA and the Nonlinear AutoRegressive Neural Network (NAR) models were used in their work.

In the present paper we review several approaches to mathematical modeling of the COVID-19 disease and develop these ideas further with an emphasis on the analysis of the dynamics of the cumulative number of confirmed cases and estimation of the parameters of the models. We focus on models which use fewer parameters, rather than a detailed description of the disease.

[1]Unité de Formation en Sciences Exactes et Appliquées, Département de Mathématiques, LAboratoire de Mathématiques et Informatique

We use these models to predict the cumulative number of confirmed cases of COVID-19. More precisely, we use ARIMA models to fit the available data and then predict the cumulative number of confirmed cases.

Data used in this work are cumulative number of confirmed cases of COVID-19, recorded from March 09 to May 03, 2020.

Dates	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Total Cases	2	2	2	2	3	7	15	20	24	27	33	40	64	75	99	114

25	26	27	28	29	30	31	1	2	3	4	5	6	7	8	9
146	152	180	207	222	246	261	282	288	302	318	345	364	384	414	443

10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
448	484	497	515	528	542	546	557	565	576	581	600	609	616	629	629

26	27	28	29	30	1	2	3
632	635	638	641	645	649	652	662

The reminder of this paper is organized as follows. In section 2, we introduce the ARIMA and ExponenTial Smoothing models (ETS). We start by defining the information criterion on which we base our models choice. Then, we get our first model of prediction base on the auto.arima package of R. In subsection 2.3, we look at closely the ETS model and compare it to the ARIMA model got previously. In subsection 2.4, we build an ARIMA model by using the Box-Jenkins method. In section 3, we make prevision using the chosen model. We end our works with a conclusion.

2. ARIMA Models, Exponential Smoothing Model

In this section, we use the AutoRegressive Integrated Moving Average model got through auto.arima of the package forecast, Exponential smoothing method to predict the cumulative number of cases of COVID-19. Next, we construct our own ARIMA model and again make prevision. ARIMA models and Exponential smoothing models are the most widely used approaches to time series forecasting, and provide complementary approaches to the problem. The motivation to use these approaches is that the infection chain of the COVID-19 is autocorelated and has a certain trend. Exponential smoothing models focused on description of the trend and seasonality in the data, while ARIMA models focused on describing the autocorrelations in the data.

2.1 Information Criterion (IC)

Modeling growth often involves comparing several models of different equations on the same data set. This comparison allows the choice of the model that best fits the data ("goodness of fit"). To compare these models, we will look at information criterion like Akaike Information Criterion (AIC) (cf.[Burnham et al.,]), Bayesian Information Criterion (BIC), corrected AIC,... These criteria measure the quality of a statistical proposed model. When it is estimated that a statistical model, it is possible to increase the likelihood of the model in one or more parameters. The AIC, BIC and AICc make it possible to penalize the models as a function of the number of parameters in order to satisfy the criterion of parsimony. We then choose the model with the weakest information criterion, and thus keeping only the parameters of main interest. The formula of each one of the criteria is written as follows:

$$\begin{aligned} AIC &= -2log(L) + 2(p+q+k+1) \\ BIC &= AIC + [\log(T) - 2](p+q+k+1) \\ AICc &= AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}, \end{aligned}$$

with $k = 1$ if there is a drift and $k = 0$ otherwise. T is the total number of observations, p is the order of the autoregressive part, q the order of the moving average part and L the maximum value of the likelihood function of the model (see [Akaike, 1974]). Of two the models, the better is the one with the lesser information criterion. It is also possible to compare the residuals of the different models and choose the one for which the values were in the residual matrix are the least variable.

2.2 Automatic ARIMA Modelling

Automatic ARIMA modelling consists of the use of Hyndman-Khandakar algorithm. For more details about the algorithm, see ([Akaike, 1974]). The function auto.arima of the package forecast of the software R combines unit root tests, minimization of the AICc and Maximum Likelihood Estimation to obtain an ARIMA model that fit the data available.

For the choice of the best ARIMA model that fit the data very well, we explore several models and choose based on the

Table 1. Models with information criterion

Models	Information Criterion
ARIMA(2,2,2)	371.8959
ARIMA(0,2,0)	409.7988
ARIMA(1,2,0)	375.628
ARIMA(0,2,1)	380.0589
ARIMA(1,2,2)	371.9165
ARIMA(2,2,1)	374.1523
ARIMA(3,2,2)	369.7954
ARIMA(4,2,2)	372.439
ARIMA(3,2,3)	372.4368
ARIMA(4,2,1)	375.5399

Table 2. Estimation of the parameters of ARMA(3,2,2) model

ar1	ar2	ar3	ma1	ma2
-0.6801	-0.7908	-0.4462	-0.2889	0.8821
0.1506	0.1547	0.1570	0.0906	0.1237

minimal Information Criteria (i.e min($AIC, BIC, AICc$)). So Table 1 gives the different model with their IC: The best model that fits the data is the ARIMA (3,2,2). For this model, the equation of the model is given by

$$Y_t = (1 - L)^3 X_t,$$

where Y_t describes an ARMA(2.2) process.

To estimate the parameters of the model, we use the maximum likelihood estimation (MLE) method. The aim of using this method is to find the values of parameters that maximize the probability of obtaining the available data. Table 2 provides the estimates of the parameters.

Figure 1 gives the forecast of the cumulative number of the confirmed cases.

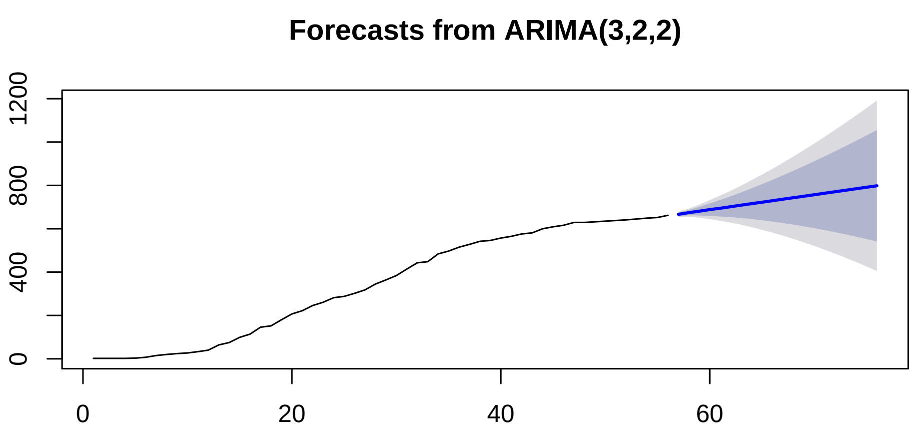

Figure 1. Forecasts from ARIMA model fitted to the available covid data

Remark 1. *The ARIMA(3,2,2) model goodly captures all the dynamics in the data as the residuals seem to be white noise (see Figure 2). The test of Ljung-Box applied to the residuals from ARIMA(3,2,2) gives a* $p - value = 0.9657$*, which*

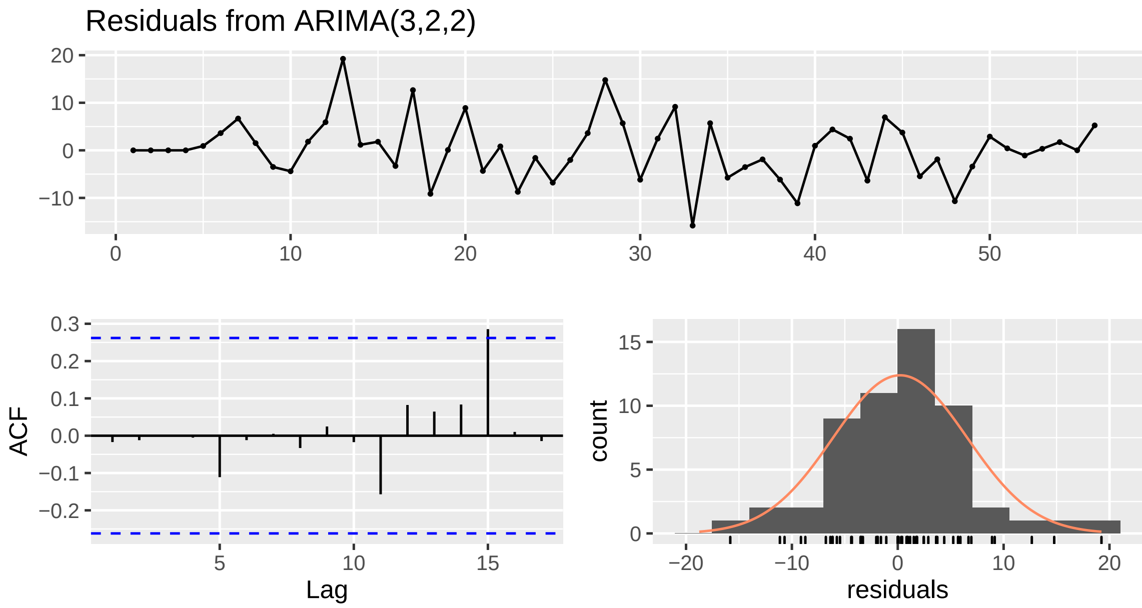

Figure 2. Residual diagnostic plots for the ARIMA model fitted to the cumulative number of confirmed cases of COVID-19

confirms that the residuals are white noise.

2.3 Exponential Smoothing Model

Exponential smoothing appeared around 1959 (cf.[Brown, 1959]) and has motivated some successful forecasting methods.

Here, based on the information criterion, we model the cumulative number of confirmed cases by the Holt's linear method with additive errors. For this model the equation of the model is given by

$$\begin{cases} X_t &= L_{t-1} + B_{t-1} + \varepsilon_t \quad \text{Forecast equation} \\ L_t &= L_{t-1} + B_{t-1} + \alpha\varepsilon_t \quad \text{Level equation} \\ B_t &= B_{t-1} + \beta\varepsilon_t \quad \text{Trend equation,} \end{cases}$$

where L_t is the level (or the smoothed value) of the series at time t, B_t is the trend component, α, β are smoothing coefficients of the model having the following constraints $0 < \alpha < 2$ and $0 < \beta < 4 - 2\alpha$ (see [Akaike, 1974],chapter 10).

To estimate the smoothing parameter, we use the MLE method and obtain the following system.

$$\begin{cases} X_t &= L_{t-1} + B_{t-1} + \varepsilon_t \\ L_t &= L_{t-1} + B_{t-1} + 0.569\varepsilon_t \\ B_t &= B_{t-1} + 0.569\varepsilon_t, \end{cases} \tag{1}$$

with initial states $L_0 = 3.2247$, $B_0 = 0.056$.

Figure 3 shows the forecasts of cumulative number of confirmed cases of COVID-19 from the ETS model.

Remark 2. *This model capture very well the dynamics in the data, since the residuals appear on Figure 4 to be white noise. That is confirmed by the test of Ljung-Box on the residuals of the ETS(A, A, N), where the* $p - value = 0.2823$.

Table3 give the prevision in term of the confidence interval of both models the ARIMA(3,2,2) and the ETS(A, A, N) models over thirty days. To our knowledge, all the predictions given meet the real data.

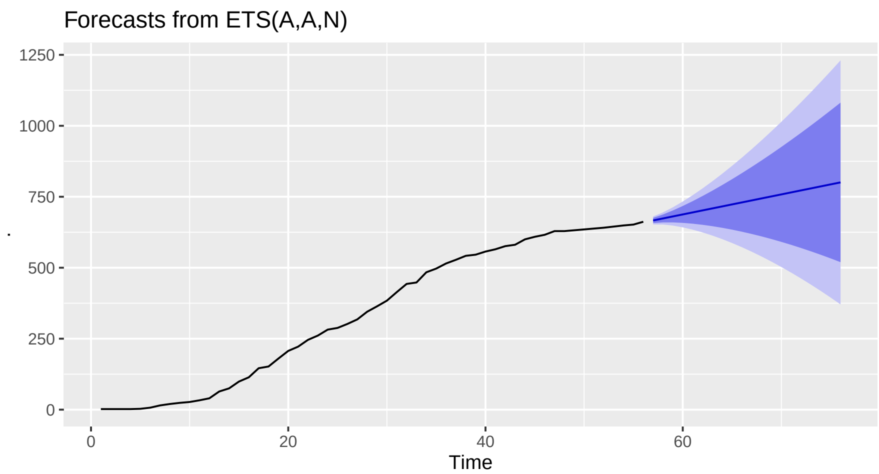

Figure 3. Forecasts from Exponential smoothing models fitted to the available covid data

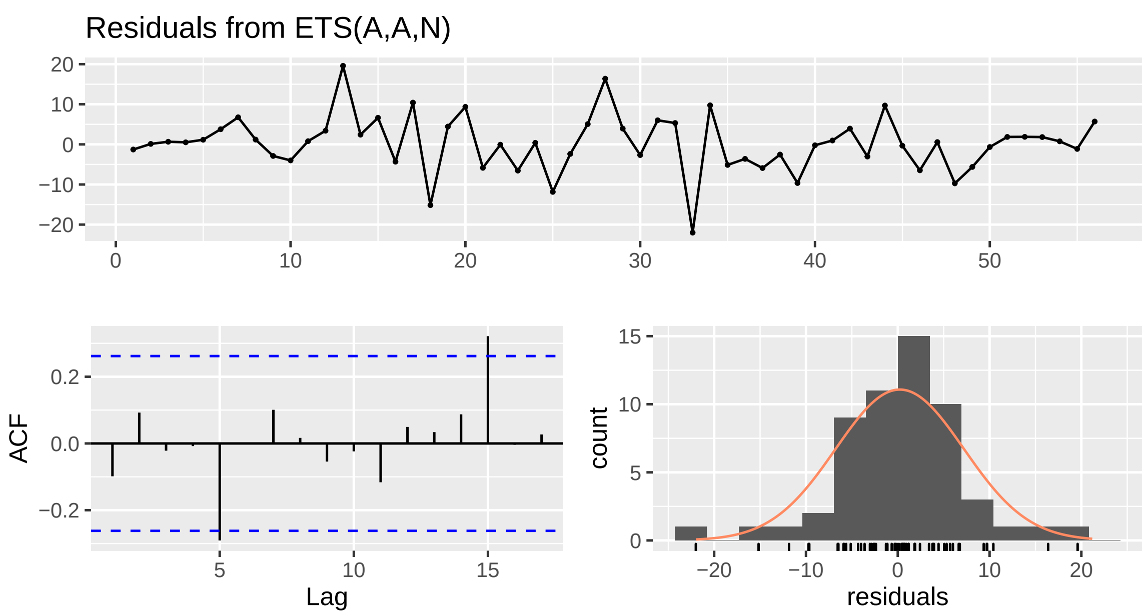

Figure 4. Residual diagnostic plots for the ETS model fitted to the cumulative number of confirmed cases of COVID-19

Table 3. Prediction of the cumulative numnber of confirmed cases of COVID-19 in Burkina Faso till the 2^{nd} of Jun

	ARIMA(3,2,2)		ETS(A,A,N)	
Dates	Forecast	Lo 95 Hi 95	Forecast	Lo 80 Hi 80
04/05/2020	667	654 679	667	657 676
05/05/2020	674	656 692	674	660 688
06/05/2020	681	652 710	681	660 702
07/05/2020	688	647 730	688	658 717
08/05/2020	695	639 751	695	655 734
09/05/2020	702	632 772	702	651 752
10/05/2020	709	623 795	709	646 771
11/05/2020	716	612 819	716	641 791
12/05/2020	722	601 844	723	634 812
13/05/2020	729	589 869	730	627 833
14/05/2020	736	576 896	737	619 855
15/05/2020	743	562 924	744	610 878
16/05/2020	750	547 953	751	601 902
17/05/2020	757	532 982	758	591 926
18/05/2020	764	516 1012	765	581 950
19/05/2020	771	499 1043	772	569 975
20/05/2020	778	481 1074	780	558 1001
21/05/2020	784	463 1106	787	546 1028
22/05/2020	791	444 1139	794	533 1054
23/05/2020	798	424 1173	801	520 1082
24/05/2020	805	403 1207	808	506 1110
25/05/2020	812	382 1242	815	492 1138
26/05/2020	819	360 1277	822	477 1167
27/05/2020	826	338 1314	829	462 1196
28/05/2020	833	315 1350	836	446 1226
29/05/2020	840	292 1387	843	430 1256
30/05/2020	847	268 1425	850	413 1287
31/05/2020	853	243 1464	857	396 1318
01/06/2020	860	218 1503	864	379 1349
02/06/2020	867	192 1542	871	361 1381

2.4 Choosing Our Own Model

According to ([Hyndman and Athanasopoulos, 2018]) the automatic arima modeling technique uses a variation of Hyndman-Khandakar algorithm, which combines unit root tests, minimization of the AICc and MLE to obtain an ARIMA model. Our purpose in this subsection is to use a general procedure for forecasting using an ARIMA model. The modeling procedure used in the following is based on the one in ([Hyndman and Athanasopoulos, 2018] p. 321), which can be summarized by Figure 5.

Plot of the data The curve in Figure 6 shows the evolution of the cumulative number of COVID-19 cases in Burkina Faso. We notice on this graph that the number of cases is increasing regularly. Figure 7 shows the scatterplots of the COVID4. We can notice the randomness of the data, but no clear seasonality. Figure 8 shows the autocorrelation function of the cumulative number of confirmed cases of COVID-19. The autocorrelations for small lags tend to be large and slowly decrease as the lags increase. Therefore the time series has a trend. Moreover, The data are strongly autocorrelated positive.

Box-Cox transformation In this paragraph, we proceed to the transformation of the data using the Box-Cox transformation. Indeed, the Standard Normal Homogeneity Test (SNHT), the test of Buishand and the test of von Neumann confirmed that the data are not homogeneous; the variance is not constant over time.

The analysis of the Box-Cox transformation reveals that the serial is not from a normal distribution. Moreover, the three tests, KPSS test, Phillips-Perron test and ADF test of Dickey-Fuller show that the process is non-stationary. Table 4 gives the results of tests of stationarity.

Differentiation We look at the differentiation of the Box-Cox process in this paragraph. The first differentiation of the process is non-stationary. But the differentiation of order 2 is stationary. Figure 10 shows the two-times difference process. We can observe the stationarity of the process. Figure 10 shows the differentiation of order two of the Box-Cox

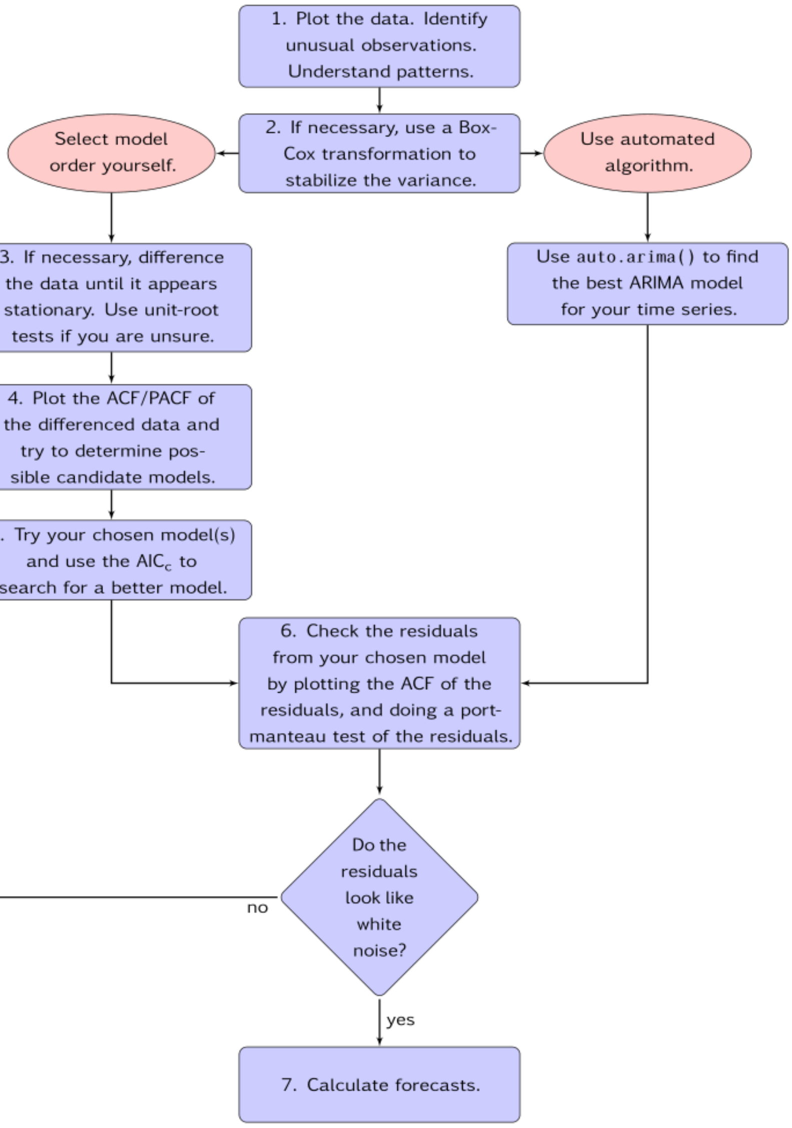

Figure 5. General process for forecasting using an ARIMA model

Table 4. result of the test of stationarity

Tests	p-values
ADF	< 0.0001
Phillips-Perron	0.960
KPSS	< 0.0001

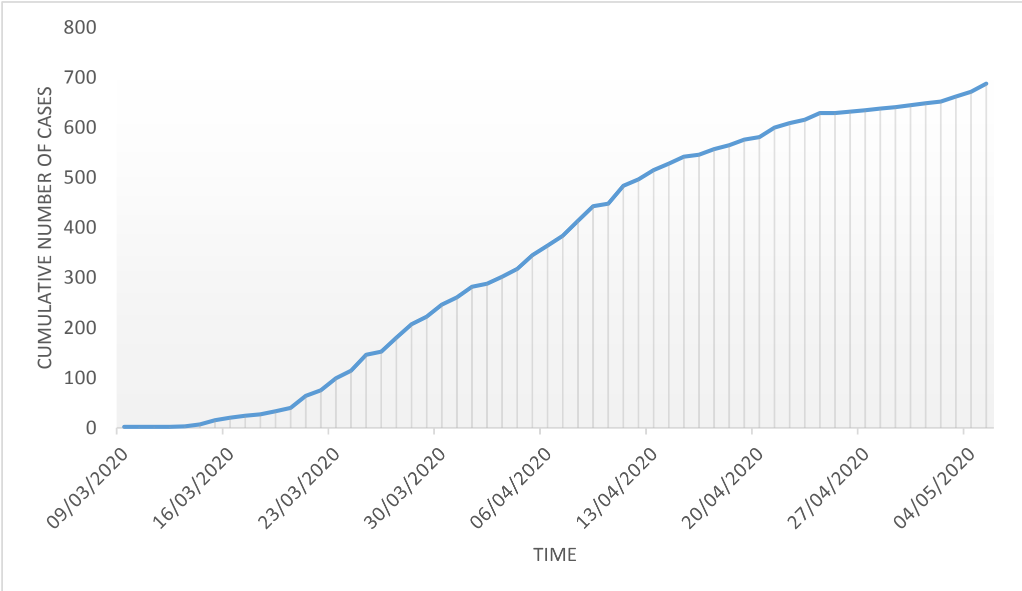

Figure 6. Cumulative number of confirmed cases

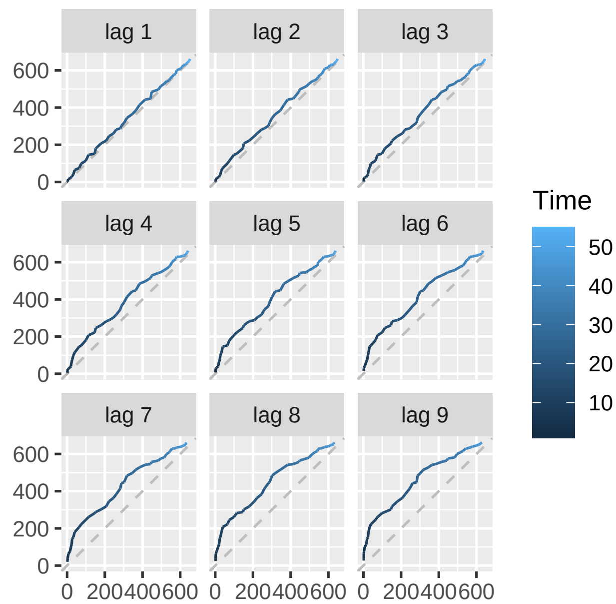

Figure 7. Lagged scatterplots of the cumulative number of cases of COVID-19

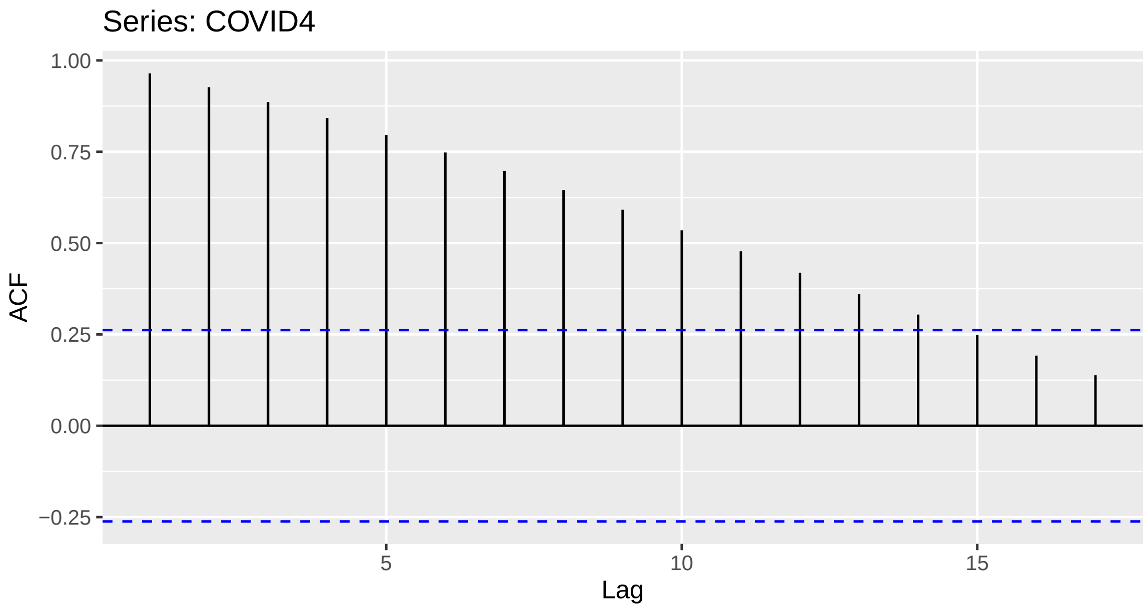

Figure 8. Autocorrelation function of the cumulative number of confirmed cases

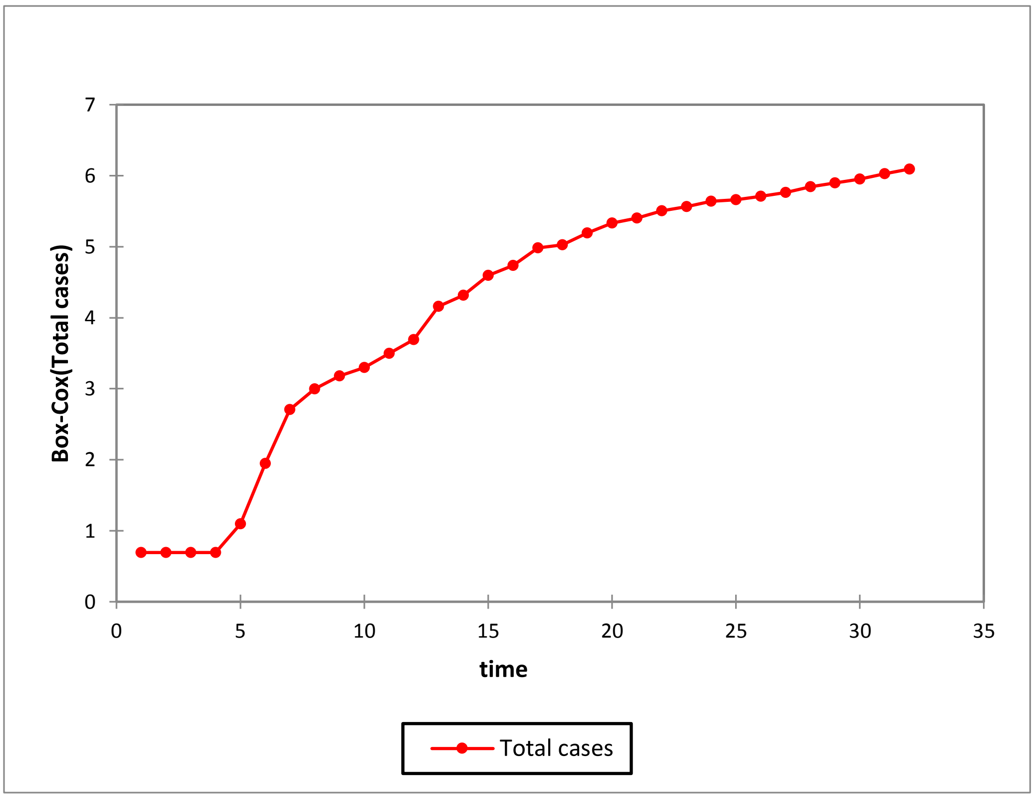

Figure 9. Box-Cox of the cumulative number of confirmed cases

Table 5. Result of the normality and white noise tests

Statistique	Valeur	p-value
Box-Pierce	0,011	0,915
Ljung-Box	0,013	0,911
McLeod-Li	1,154	0,283
Box-Pierce	0,120	0,942
Ljung-Box	0,137	0,934
McLeod-Li	2,572	0,276

process.

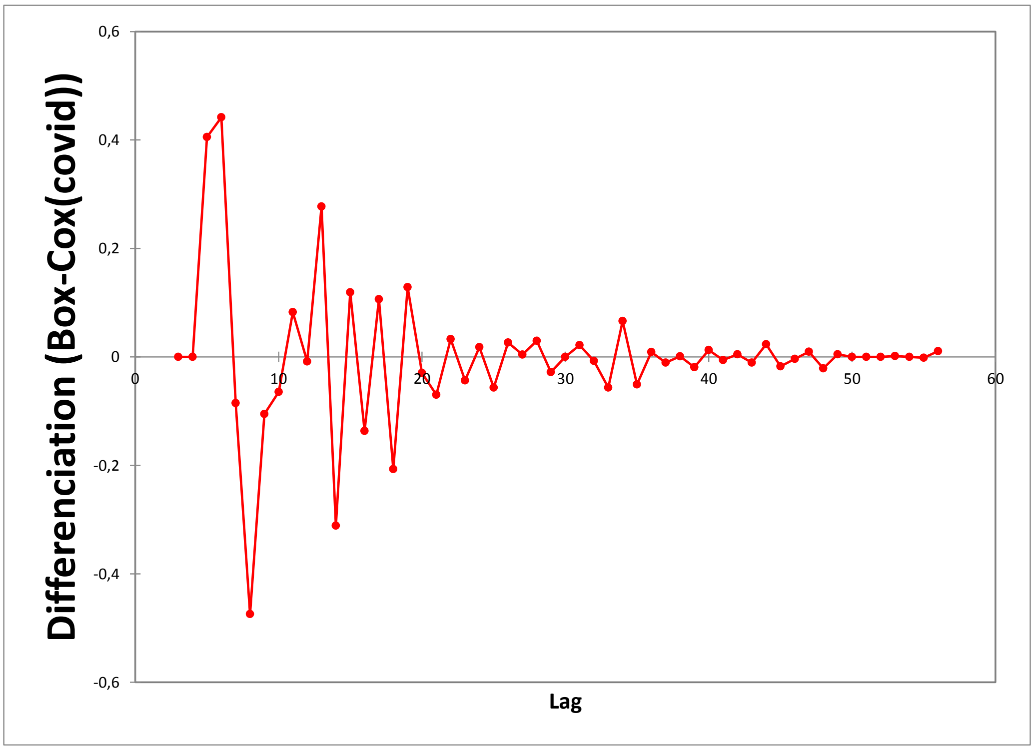

Figure 10. Differentiation of the Box-Box transformation process

In Figure 11, The ACF of the residuals from the ARMA(1,2,1) model indicates that the residuals are white noise.

Moreover, the analysis of the white noisiness of the residuals shows that the process has a normal distribution and is stationary (cf. Table 5).

We can therefore says that the differentiation Box-Cox process is a Gaussian white noise.

Analysis of the ACF and PACF Here now, we analyze the Auto Correlation Function (ACF) and the Partial Auto Correlation Function (PACF) of the 2-order Differentiation Box-Cox process. Figure 12 gives both functions ACF and PACF. The PACF and ACF functions in Figure 12 are suggestive of an ARMA(1,1) model.

Use of the AICc for searching the the better ARIMA model We fit the ARIMA(1,2,1) model along with variation including ARIMA(3,2,1), ARIMA(3,2,2), ARIMA(2,2,0), ARIMA(1,2,0). Among them, the ARIMA(1,2,1) has a slightly smaller "BIC".

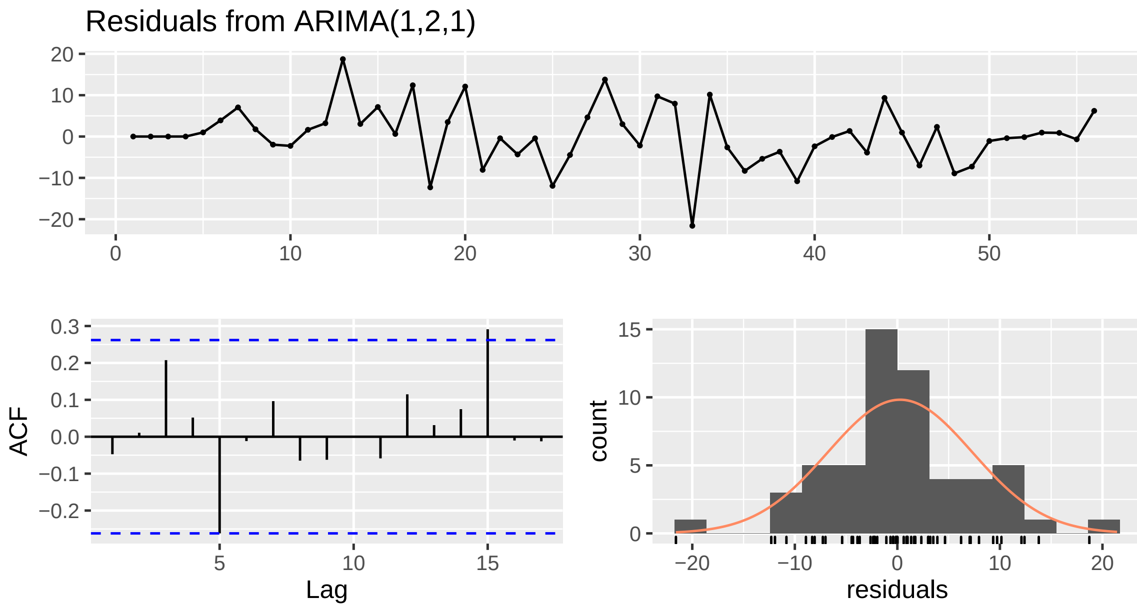

Figure 11. Residual plots for the ARIMA(1,2,1) model

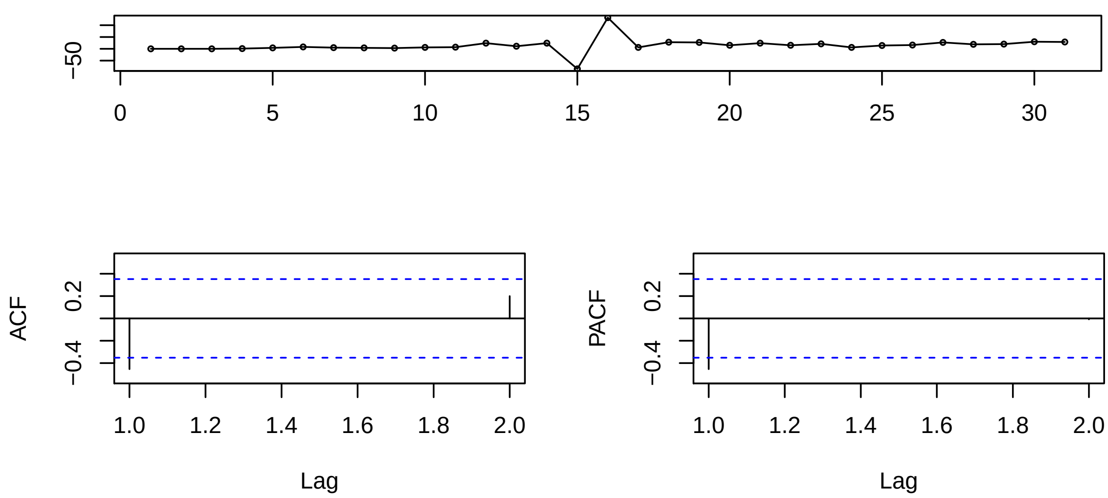

Figure 12. AutoCorrelation and Partial AutoCorrelation Function

3. Forecast From the Chosen Model

Now that we have our model, In this section we predict the cumulative number of confirmed cases. Figure 13 gives an overview of the prediction of twenty days from the 3rd May.

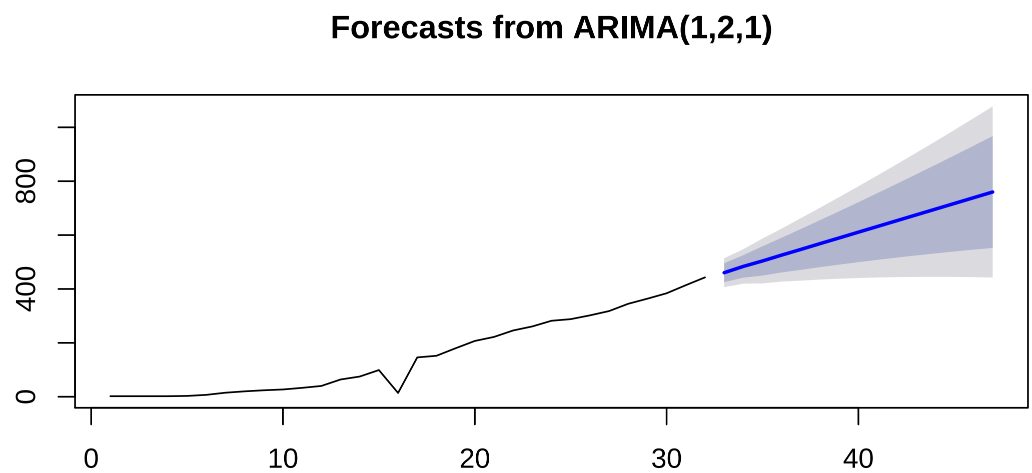

Figure 13. Forecast for 15 days from ARIMA(1,2,1)

Moreover, Table 6 gives the prevision within thirty days from the 3th of May in term of confidence interval.

Table 6. Forecast of the cumulative number of confirmed cases of COVID-19 from ARIMA(1,2,1) model

Date	Forecast	Lower 95	Higher 95
04/05/2020	666	652	680
05/05/2020	673	652	694
06/05/2020	679	647	710
07/05/2020	685	642	728
08/05/2020	691	636	746
09/05/2020	697	629	765
10/05/2020	703	620	786
11/05/2020	709	611	807
12/05/2020	715	602	829
13/05/2020	721	591	852
14/05/2020	727	580	875
15/05/2020	733	567	899
16/05/2020	739	555	924
17/05/2020	745	541	950
18/05/2020	752	527	976
19/05/2020	758	512	1003
20/05/2020	764	497	1030
21/05/2020	770	481	1059
22/05/2020	776	464	1087
23/05/2020	782	447	1116
24/05/2020	788	430	1146
25/05/2020	794	412	1176
26/05/2020	800	393	1207
27/05/2020	806	374	1238
28/05/2020	812	354	1270
29/05/2020	818	334	1303
30/05/2020	824	313	1335
31/05/2020	830	292	1369
01/06/2020	836	270	1402
02/06/2020	842	248	1437

Table 7. Accuracy evaluation of the ARIMA(3,2,2) model, ARIMA(1,2,1) model and the ETS model

		RMSE	MAE	MAPE	MASE
ARIMA(3,2,2)	Training set	6.271917	4.622425	5.020457	0.3852021
	Test set	531.393498	514.728935	3463.153828	42.8940779
ETS	Training set	6.93007	4.969203	7.502859	0.4141003
	Test set	534.47391	518.548964	3463.052249	43.2124137
ARIMA(1,2,1)	Training set	6.995422	4.999534	5.061874	0.4166278
	Test set	512.160714	491.242912	3443.821265	40.9369094

Remark 3. *The value of the order of difference d has an effect on the prediction intervals ł the higher the value of d, the more rapidly the prediction intervals increase in size. So one should take that into account for the model to choose for the prediction.*

Remark 4. *We notice that, in one hand, the automatic arima model fits the training data slightly better than the ETS model. However, ETS model out performs the ARIMA(1,2,1) model. On the other hand, the ARIMA(1,2,1) model provides more accurate forecasts on the test set than the automatic arima model ARIMA(3,2,2), which in turn outperforms the ETS model. Table 7 below gives an insight of what we say.*

Likewise, when we use time series cross-validation to compare the three models, based on the Mean Squared Error, the ARIMA(1,2,1) model has a lower tsCV statistic, then come the automatic arima model and finally the ETS model.

4. Conclusion

The main contribution of this paper is the daily prediction of the cumulative number of confirmed cases using a number of times series models. The ARIMA(1,2,1) gives good predictions than the others.

It is important to point out that we haven't developed new statistical methods, but used existing simple ones to show their usefulness and practicability.

Acknowledgments

The authors would like to thank International Science Program for their partial support, the editor for accepting that we submit our paper in this journal and the referees for their helpful comments.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw/Hill.

Burnham, K. et al. dr anderson. (2002). model selection and multimodel inference: a practical information-theoretic approach. *Ecological Modelling. Springer Science & Business Media, New York, New York, USA*.

Chen, T.-M., Rui, J., Wang, Q.-P., Zhao, Z.-Y., Cui, J.-A., & Yin, L. (2020). A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious Diseases of Poverty, 9*(1), 1-8.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Ivorra, B., & Ramos, A. M. (2020). Application of the be-codis mathematical model to forecast the international spread of the 2019–20 wuhan coronavirus outbreak. *ResearchGate http://dx. doi. org/10.13140/RG. 2.2*, 31460.

Jia, L., Li, K., Jiang, Y., & Guo, X., et al. (2020). Prediction and analysis of coronavirus disease 2019. *arXiv preprint arXiv:2003.05447*.

Khan, F. M., & Gupta, R. (2020). Arima and nar based prediction model for time series analysis of covid-19 cases in india. *Journal of Safety Science and Resilience, 1*(1), 12-18. https://doi.org/10.1016/j.jnlssr.2020.06.007

Liu, Z., Magal, P., Seydi, O., & Webb, G. (2020). Understanding unreported cases in the covid-19 epidemic outbreak in wuhan, china, and the importance of major public health interventions. *Biology, 9*(3), 50. https://doi.org/10.3390/biology9030050

Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K.-H. (2020). Time series modelling to forecast the confirmed and recovered cases of covid-19. *Travel Medicine and Infectious Disease*, page 101742.

Tang, B., Bragazzi, N. L., Li, Q., Tang, S., Xiao, Y., & Wu, J. (2020). An updated estimation of the risk of transmission of the novel coronavirus (2019-ncov). *Infectious disease modelling, 5*, 248-255.