AJAS / SPAS

**African Journal of Applied Statistics**
Vol. 8 (1), 2021, pages 1507 - 1522.
DOI: http://dx.doi.org/10.16929/ajas/2021.1507.260

ISSN 2316-0861

# Short-term prediction model of daily COVID-19 reported positive cases in SENEGAL

**Aba Diop, Abdourahmane Ndao, Cheikh Tidiane Seck (*), Ibrahima Faye and Mouhamed Amine Niang**

Address of all authors: Equipe de Recherche en Statistique et Modèles Aléatoires (*ERESMA*) Université Alioune Diop, Bambey, SENEGAL

Received on April 13, 2021; Accepted on July 20, 2021


**Abstract.** In this work, we use an Auto-Regressive Integrated Moving Average (*ARIMA*) model to study the evolution of *COVID-19* disease in Senegal and then make short-term predictions about the number of people likely to be infected by *SARS-Cov-2* coronavirus. We are dealing with daily data provided by the Senegalese Ministry of Health during the period from march 02/2020 to march 02/2021. Our results show that the peak of the disease appears during the second wave and seems to be reached on February 12/2021. But they also reveal that the number of *COVID-19* infections will be around of 200 cases per day during the next 30 days, if the trend of the total number of performed tests is maintained.

**Key words:** auto-regressive integrated moving average; COVID-19; coronavirus; forecasting.
**AMS 2010 Mathematics Subject Classification Objects :** 37M10; 62P10; 62F03.

*Corresponding author: Cheikh Tidiane Seck (cheikhtidiane.seck@uadb.edu.sn)
Aba Diop: aba.diop@uadb.edu.sn
Abdourahmane Ndao: mame.ndaoensae@gmail.com
Ibrahima Faye: ibou.faye@uadb.edu.sn
Mouhamed Amine Niang: amine.niang@uadb.edu.sn

**Résumé.** Dans ce travail, nous analysons l'évolution de la pandémie du *COVID-19* au Sénégal en considérant la série des données sur le nombre de cas positifs journaliers. Plus précisément, nous utilisons un modèle *ARIMA* (Auto-Regressive Integrated Moving Average) pour établir des prévisions à court terme concernant le nombre de nouvelles contaminations au coronavirus. Nous travaillons avec les données journalières fournies par le Ministère de la Santé et de l'Action Sociale du Sénégal durant la période du 02 mars 2020 au 02 mars 2021. Nos résultats montrent que le pic de la maladie est apparu durant la deuxième vague et semble être atteint à la date du 12 février 2021. Ils révèlent également que le nombre de cas positifs au *COVID-19* tournerait en moyenne autour de 200 cas par jour durant les 30 prochains jours si la tendance actuelle du nombre total de tests effectués est maintenue.

**The authors**.
*Aba Diop*, Ph.D., is Assistant in Mathematics and Statistics at Alioune Diop University, SENEGAL.
**Abdourahmane Ndao**, M.Sc. degree in Statistics, is preparing a Ph.D thesis under the supervision of the third author at Alioune Diop University, SENEGAL,.
**Cheikh Tidiane Seck**, Ph.D., is Professor of Mathematics and Statistics at Alioune Diop University, SENEGAL.
**Ibrahima Faye**, Ph.D., Assistant in Mathematics and Statistics at Alioune Diop University, SENEGAL.
**Mouhamed Amine Niang**, Ph.D., Assistant in Mathematics and Statistics at Alioune Diop University, SENEGAL.

## 1. Introduction

*COVID-19* epidemic starts in Wuhan in December 2019, exactly in Hubei province in China Paules *et al.* (2020). It is caused by a new coronavirus (*SARS-Cov-2*) which is highly contagious. This explains its rapid spread from China to other countries throughout the world, with very large mortality and morbidity rates according to the World Health Organization (*WHO*). It was declared as pandemic on march 11 2020 by this organization, who set a number of recommendations in order to control the expansion of the disease.

Like many other countries, the Senegalese government has adopted these recommendations, namely the strict compliance with barrier measures. In addition, the Senegalese authorities have also closed schools and universities, banned gatherings in public places and decreed a night lock-down period throughout the whole country.

*COVID-19* remains a highly contagious disease with an average incubation period between 3 and 5 days. However, this period can last up to 14 days according to *WHO*. In this case, we may have the presence of lot of asymptomatic cases in the population ; and hence it might be difficult to control the disease, as it is accepted that many non-identified factors can also influence the infection rate such as :

promiscuity in households, behaviour in public places and lack of belief in the disease.

Research on *COVID-19* has gained a quick and considerable growth since the appearance of the virus in China. Many empirical studies have been proposed in order to analyze the evolution of the epidemic. Lui *et al.* (2020a) developped a differential equations model to predict the number of *COVID-19* cases from early reported case data in some regions throughout the world. Considering the Italian data, Bodini *et al.* (2020) proposed a stochastic *SIR* model, that should be adapted in fixed time-intervals in order to provide short-term predictions and updated assessment of the basic reproduction number. Fall *et al.* (2020) studied a compartmental model incorporating a time-dependent health care capacity, and analyzed the evolution of *COVID-19* in Senegal. Also, Dath *et al.* (2020) conducted an observational and peak survey study of the *COVID-19* epidemic in Senegal over a 100-day period. For more details in *COVID-19* modeling, we refer the interested reader to Lui *et al.* (2020b), Lui *et al.* (2020c), Wang (2020), Wang *et al.* (2020) and references therein.

The goal in this paper is to first analyze the evolution of the number of daily *COVID-19* positive cases during the period from march 02/2020 to march 02/2021, which enables us to detect the peak of the disease in terms of reported positive cases. Then, in a second step, we make a short-term forecast for the number of daily *COVID-19* positive cases over a 30-day period. Our approach makes use of time series tools and is essentially based on the methodology of Box and Jenkins (1970). More precisely, we deal with *ARIMA* (Auto-Regressive Integrated Moving Average) processes, which are used to describe, explain and predict observed phenomena over time (see Brockwell and Davis (2002), Gouriéroux and Monfort (1995) and Fuller (1976) for more details).

There are various applications dealing with Box and Jenkins methodology in epidemiology and public health. Promprou *et al.* (2006) have applied this method to predict dengue haemorrhagic fever cases in Southern of Thailand. In an epidemic situation, Earnest *et al.* (2005) have used *ARIMA* models to predict and monitor the number of occupied beds in a tertiary hospital in Singapore. For malaria modeling in Afghanistan, Anwar *et al.* (2016) have also dealt with *ARIMA* models to make prediction of the trend and incidence. Recently Fatih *et al.* (2020) applied *ARIMA* models to explain and predict the number of *COVID-19* cases in South Africa, Algeria and Egypt. For more details on *ARIMA* models, the interested reader is refereed to Brockwell and Davis (2002), Box and Jenkins (1970), Box and Pierce (1970) and references therein.

The rest of the paper is organized as follows : Section 2 describes the theoretical tools we employ to analyze the evolution of the daily *COVID-19* positive cases in Senegal. In Section 3, we present some graphical and numerical results and discuss them. Finally, Section 4 concludes the paper and points out some limits

and perspectives.

## 2. Materials and methods

### *2.1. Model presentation*

A stochastic process $(X_t)_{t\in\mathbb{Z}}$ has an *ARIMA* representation$(p,d,q)$ if it satisfies the following relationship:

$$\Phi(B)\Delta^d X_t = \Theta(B)\epsilon_t, \quad \forall t \in \mathbb{Z}, \tag{1}$$

where $(p,q) \in \mathbb{N}^2$ are respectively the orders of the autoregressive part and the moving average part; $d \in \mathbb{N}$ is the integration order (or differentiation order, that is the number of times it is necessary to differentiate the initial series to make it stationary); $(\epsilon_t)_{t\in\mathbb{Z}}$ is a centered white noise of variance $\sigma^2 > 0$ ; $B$ is the delay operator defined by $B^n X_t = X_{t-n}$ for all $n \geq 1$ and $\Delta$ is the difference operator defined by $\Delta X_t = X_t - X_{t-1}$ and $\Delta^d X_t = \underbrace{\Delta \dots \Delta}_{d \text{ times}} X_t$. Functions $\Phi$ et $\Theta$ are built from the autoregressive and moving average processes which are defined as follows:

- Auto-Regressive process of order $p$: *AR*$(p)$

$$\begin{aligned} \epsilon_t &= X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \dots - \phi_p X_{t-p} \\ &= X_t - \phi_1 B X_t - \phi_2 B^2 X_t - \dots - \phi_p B^p X_t \\ &= \underbrace{1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p}_{\Phi(B)} X_t \\ &= \Phi(B) X_t \qquad \phi_j \in \mathbb{R}, \quad j = 1, \dots, p \end{aligned} \tag{2}$$

- Moving Average process of order $q$: *MA*$(q)$

$$\begin{aligned} X_t &= \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \dots - \theta_q \epsilon_{t-q} \\ &= \epsilon_t - \theta_1 B \epsilon_t - \theta_2 B^2 \epsilon_t - \dots - \theta_q B^q \epsilon_t \\ &= \underbrace{1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q}_{\Theta(B)} \epsilon_t \\ &= \Theta(B) \epsilon_t \qquad \theta_j \in \mathbb{R}, \quad j = 1, \dots, q. \end{aligned} \tag{3}$$

***Remark 1****: Whenever $d = 0$, ARIMA$(p,0,q)$ processes are equivalent to ARIMA$(p,q)$ processes which are used to model stationary time series. In the presence of a non-linear or polynomial trend, an ARIMA$(p,d,q)$ process is more appropriate.*

### *2.2. Augmented Dickey-Fuller (ADF) test*

Dickey and Fuller (Dickey and Fuller (1979)) were the first to provide a set of formal statistical tools for detecting the presence of a unit root in a purely auto-regressive first-order process. This well-known test procedure is based on ordinary least squares estimation, under the alternative assumption of stationarity, of three first-order auto-regressive models with independent and identically distributed errors: the model without constant, the model with constant and the model with constant and trend. The simple Dickey and Fuller test assumes that the $\epsilon_t$ error term is not auto-correlated in the three previous models. But in practice, this assumption is not verified in most cases. Thus, for a choice of $p$ lags, corresponding to an auto-correlation of order $p+1$ of the innovations in an $AR(1)$ representation, the three models used to develop the *ADF* test are the following:

$$\Delta X_t = \phi X_{t-1} + \sum_{j=1}^{p} \phi_j \Delta X_{t-j} + \epsilon_t$$

$$\Delta X_t = \phi X_{t-1} + \sum_{j=1}^{p} \phi_j \Delta X_{t-j} + c + \epsilon_t$$

$$\Delta X_t = \phi X_{t-1} + \sum_{j=1}^{p} \phi_j \Delta X_{t-j} + c + \beta t + \epsilon_t$$

The testing hypotheses are defined as follows:

$$\begin{cases} H_0 : \ \phi = 1 & \text{(presence of unit root } \Leftrightarrow \text{ no stationarity)} \\ H_1 : \ |\phi| < 1 & \text{(no unit root } \Leftrightarrow \text{ stationarity)} \end{cases}$$

We reject the null hypothesis $H_0$ at level $\alpha \in ]0;1[$ if the p-value of the test is strictly less than $\alpha$.

***Remark 2**: The first parameter to choose is the order of differentiation $d$. Several methods are possible to detect non-stationarity (presence of unit root). The graphical representation of the data can show polynomial trends. We can also calculate the empirical auto-correlations functions and analyze the speed of decay towards zero of these functions. If this decay is slow (slower than an exponential decay) we can suspect a non-stationarity. Note that in practice the case $d > 2$ is rarely encountered. Indeed, one must be careful to not over-differentiate a process because it can lead to a non-invertibility of the associated ARMA processes.*

### *2.3. Model choice*

The model is chosen by finding the appropriate values of the orders $p$ and $q$ of the autoregressive and moving average processes, respectively. These orders

$p$ and $q$ of the *AR* and *MA* processes are obtained by looking at the graphs of partial auto-correlation and simple auto-correlation functions, respectively (Box and Jenkins (1970) and Fuller (1976)). Often, these orders are not obvious to find. But we can obtain upper bounds for $p$ and $q$ and then select a model by minimizing a penalized criterion of type *AIC* (Akaike (1974)) or *BIC* Schwarz (1978). These criteria are given by:

$$AIC = -2\log(L) + 2(p+q+1), \quad BIC = -2\log(L) + \log(n)(p+q+1)$$

where $L$ and $n$ represent respectively the log-likelihood of the model *ARIMA*$(p, d; q)$ and the sample size.

The following properties can be used to identify the orders $p$ and $q$ of the *AR* and *MA* processes:

1. For an *AR*$(p)$ process, the partial auto-correlation function is equal to zero for lags greater than $p$.
2. For a *MA*$(p)$ process, the auto-correlation function is equal to zero for lags greater than $q$.

After selecting the model, the estimation of the associated parameters is done by the maximum likelihood method. Then, the adequacy of the model with respect to the data is studied using the residuals of the model. If the model is adequate, the residuals should follow a Gaussian white noise.

### *2.4. Testing model adequacy*

After estimating the model, it is important to validate it by studying the residuals (voir Box and Pierce (1970)). For this purpose, different tests and diagnostics are used. We obtain an estimator of the error term $\epsilon_t$ of the *ARIMA* model by:

$$\hat{\epsilon}_t = [\hat{\Theta}(B)]^{-1}\hat{\Phi}(B)\Delta^d X_t, \quad \forall\, t \in \mathbb{Z} \tag{4}$$

We first start to estimate the empirical auto-correlation function of the residuals defined as follows:

$$\hat{\rho}(h) = \frac{\sum_{t=1}^{n-h} \hat{\epsilon}_t \hat{\epsilon}_{t+h}}{\sum_{t=1}^{n} \hat{\epsilon}_t^2} \quad \forall\, h \in \mathbb{N}^*, \quad h < n. \tag{5}$$

***Remark 3****: Ideally, we would like to obtain the auto-correlation function of a white noise (i.e. $\hat{\rho}(h) = 0, \forall h > 0$ ). If it is not the case, the choice of model orders or the correction of the trend must be reviewed (in this case, visualising the trajectory of errors can be instructive).*

**Box-Pierce test**: it is used to test the null hypothesis that the residuals of a series following an *ARMA* model are white noise, i.e.,

$$\begin{cases} H_0 : \ \rho(1) = \rho(2) = \dots \rho(h) = 0 & \text{(white noise)} \\ H_1 : \ \exists \ k \in \{1, 2, \dots, h\} \mid \rho(k) \neq 0 & \text{(not white noise)} \end{cases}$$

The Box-Pierce statistic is given by:

$$S_{BP} = n \sum_{j=1}^{h} [\hat{\rho}(j)]^2 \ \longrightarrow \ \chi^2_{h-m}, \ \text{ under } H_0, \text{ as } n \to +\infty,$$

where $m$ is the number of parameters in the model.

We accept $H_0$ at level $\alpha \in ]0; 1[$ if the observed value $S_{BP}^{\text{obs}}$ of $S_{BP}$ is greater than the quantile of order $1 - \alpha$ of $\chi^2_{h-m}$ (or if $\mathbb{P}(\chi^2_{h-m} > S_{BP}^{\text{obs}} | H_0) < \alpha$).

**Ljung-Box test**: it is a variant of the Box-Pierce test and is used to test the null hypothesis that the residuals of a series following an *ARMA* model are white noise, i.e.,

$$\begin{cases} H_0 : \ \rho(1) = \rho(2) = \dots \rho(h) = 0 & \text{(white noise)} \\ H_1 : \ \exists \ k \in \{1, 2, \dots, h\} \mid \rho(k) \neq 0 & \text{(not white noise)} \end{cases}$$

The Ljung-Box statistic is given by:

$$S_{LB} = n(n+2) \sum_{j=1}^{h} \frac{[\hat{\rho}(j)]^2}{n-j} \ \longrightarrow \ \chi^2_{h-m}, \ \text{ under } H_0, \text{ as } n \to +\infty$$

where $m$ is the number of parameters in the model.

We accept $H_0$ at level $\alpha \in ]0; 1[$ if the observed value $S_{LB}^{\text{obs}}$ of $S_{LB}$ is greater than the quantile of order $1 - \alpha$ of $\chi^2_{h-m}$ (or if $\mathbb{P}(\chi^2_{h-m} > S_{LB}^{\text{obs}} | H_0) < \alpha$).

### *2.5. Testing parameters significance (model comparison)*

The idea here is to compare nested *ARIMA*$(p, d, q)$ models to see whether it makes sense to reduce the orders $p$ and $q$. This results in an iterative algorithm that corrects or refines the previously determined orders (comparison of *ARIMA*$(p-1, d, q)$ or *ARIMA*$(p, d, q-1)$ with *ARIMA*$(p, d, q)$). Decreasing $p$ by one unit (respectively $q$) is equivalent to testing the significance of the parameter $\phi_p$ (respectively $\theta_q$). We can also use the *AIC* or *BIC* criteria to choose the best model.

## 3. Results

In this work, we deal with the data series of daily positive cases of *COVID-19* in Senegal during the period from march $02$, $2020$ to march $02$ $2021$ (data from the Ministry of Health and Social Action of Senegal).

### *3.1. Exploratory treatment of the series*

Figure 1 shows the evolution of the number of daily positive cases of *COVID-19* in Senegal during the period from march $02/2020$ to march $02/2021$. We observe an upward trend at the beginning of the infection which then starts to decline from the second half of august $2020$. This notable decrease would have led to a relaxation of the population. This situation is in favor of the virus which then spread rapidly among the population, causing an exponential growth in the number of community cases and a significant increase in the total number of positive cases. During this second wave the maximal number of *COVID-19* contaminations is observed on February $12$ $/2021$, which corresponds to the peak of the disease.

It is important to note that the number of daily positive *COVID-19* cases is strongly dependent on the total number of performed tests per day (see Figure 3). Indeed we observe a linear correlation coefficient equal to $0.855$ ($95\%$ confidence interval: $CI = ]0.825; 0.880[$. The p-value associated with this correlation test is less than $2.2 \times 10^{-16}$, which largely confirms the hypothesis of linear correlation of these two series).

We can also see that the data does not show a perfect seasonality but shows two observable cycles during the periods from march $02/2020$ to November $15/2020$ and from December $16/2020$ to march $02/2021$.

The series of daily *COVID-19* positive cases shows a variable trend which may increase or decrease over time. This suggest a non-stationarity of the data which should be confirmed by a Dickey-Fuller unit root test.

### *3.2. Testing stationarity*

We run the Augmented Dickey-Fuller test at level $5\%$ to study the stationarity of the number of daily positive cases of *COVID-19* in Senegal. The hypotheses of the test are defined as follows:

$$\begin{cases} H_0 : \ \phi = 1 & \text{(presence of unit root } \Leftrightarrow \text{ not stationarity)} \\ H_1 : \ |\phi| < 1 & \text{(absence of unit root } \Leftrightarrow \text{ stationarity)} \end{cases}$$

The p-value of the test is equal to $0.9308$ which is greater than $0.05$. We cannot reject the null hypothesis of non-stationarity of the series. It admits a unit root and a non-linear trend. The number of daily positive cases of *COVID-19* in Senegal cannot therefore be modelled by an $ARMA(p, q)$ model. We will instead use an

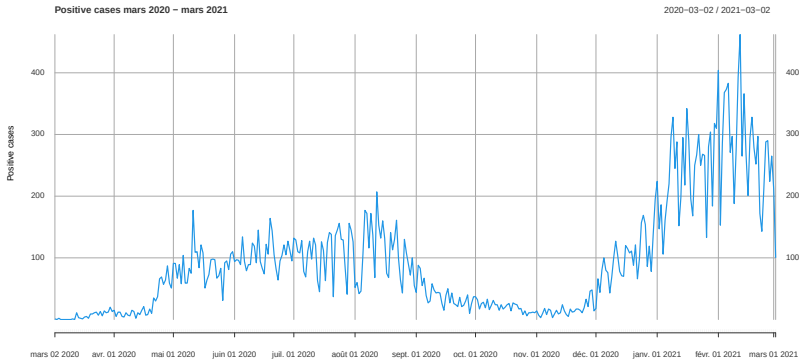


**Fig. 1.** Exploration of the number of daily positive cases of *COVID-19*.

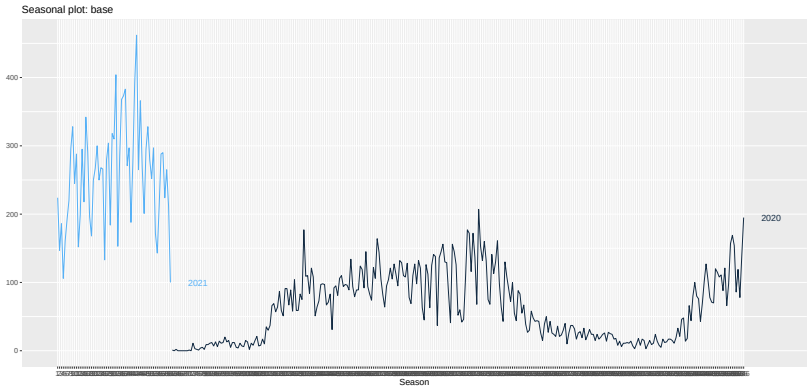


**Fig. 2.** Seasonality study

*ARIMA*$(p, d, q)$ model.

We run the *ADF* test again for the once differentiated serials of daily *COVID-19* positive cases and obtain a p-value of $0.01$ which is less than $0.05$. The once differentiated serial is thus stationary. Therefore, $d = 1$. In the next section we will choose the orders $p$ and $q$ of the *AR* and *MA* processes respectively in the model *ARIMA*$(p, 1, q)$.

### *3.3. Model choice*

Figure 4 shows the empirical auto-correlation function and the empirical partial auto-correlation function. The two properties defined in section 2.3 will guide us on the choices of $p$ and $q$.
Figure 4 shows, on the one hand, for the empirical partial auto-correlation function, significant peaks for $4$ lags (negative correlations) and $2$ lags (positive correlations), therefore a choice of $p = 4$ could suffice for the auto-regressive part.

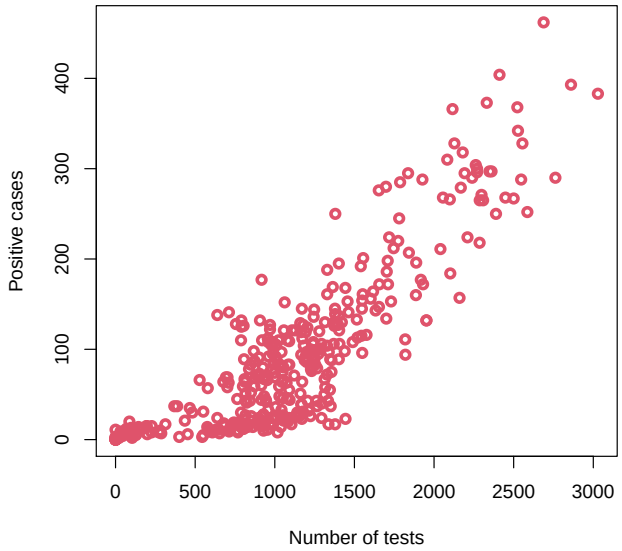


**Fig. 3.** Correlation between number of positive cases and number of daily tests.

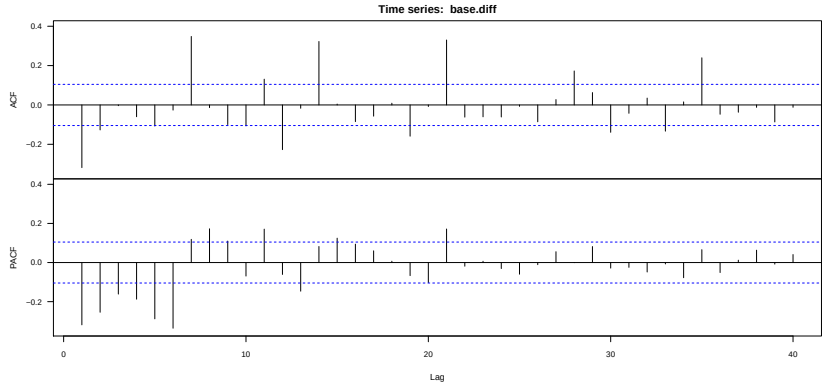


**Fig. 4.** Auto-correlation function

On the other hand, for the empirical auto-correlation function, it shows significant peaks for $3$ lags (positive correlations) and $2$ lags (negative correlations), therefore a choice of $q = 3$ could suffice for the moving average part. It should be noted that $95\%$ of the values of the simple and partial auto-correlations should be within the confidence interval. We thus obtain an *ARIMA*$(4, 1, 3)$ model.

***Remark 4****: Let us note that with the function* auto.*ARIMA*() *of the library* forecast *of the statistical software R, we can choose automatically the orders* $p$ *and* $q$ *which will be adequate for the data. Indeed, this function compares models evaluated on the data by choosing the values of* $p$ *and* $q$ *which minimise the AIC criterion with residuals which follow a white noise. We obtain the values of* $p = 4$ *and* $q = 3$*. Table 1 gives the results comparing the AIC values of the model ARIMA*$(4, 1, 3)$ *with other models for different choices of* $p$ *and* $q$*.*

**Table 1.** *ARIMA* models comparison

| Model | *ARIMA*(4,1,3) | *ARIMA*(4,1,2) | *ARIMA*(4,1,1) | *ARIMA*(4,1,0) | *ARIMA*(5,1,3) | *ARIMA*(6,1,3) |
|---|---|---|---|---|---|---|
| *AIC* | 3656.377 | 3629.992 | 3681.587 | 3720.307 | NA | NA |
| Model | *ARIMA*(3,1,3) | *ARIMA*(2,1,3) | *ARIMA*(0,1,3) | *ARIMA*(4,1,4) | *ARIMA*(4,1,5) | *ARIMA*(4,1,6) |
| *AIC* | 3671.307 | 3681.587 | 3696.155 | *NA* | 3626.172 | *NA* |

*NA: Note Available*

***Remark 5***

1. *The AR parts of the models ARIMA(5,1,3), ARIMA(6,1,3), ARIMA(4,1,4) and ARIMA(4,1,6) are not stationary. The algorithm do not converge.*
2. *The models ARIMA(4,1,2) and ARIMA(4,1,5) seem to be better than the model ARIMA(4,1,3) in terms of AIC. We will compare the behaviour of their residuals.*

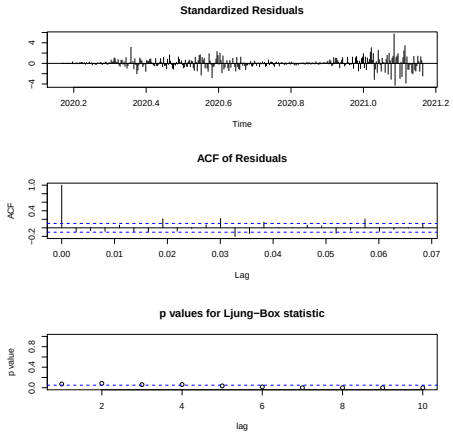


**Fig. 5.** *ARIMA*(4,1,2)

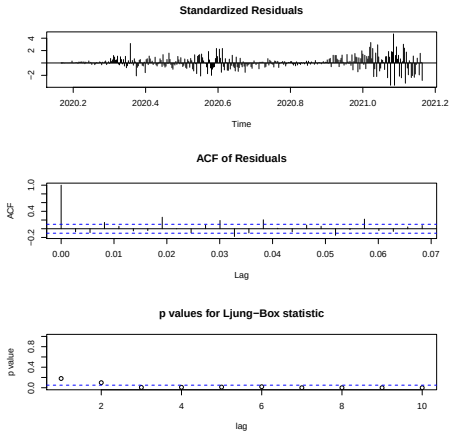


**Fig. 6.** *ARIMA*(4,1,5)

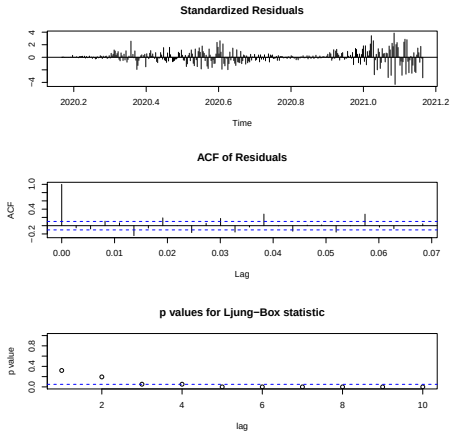


**Fig. 7.** *ARIMA*(4,1,3)

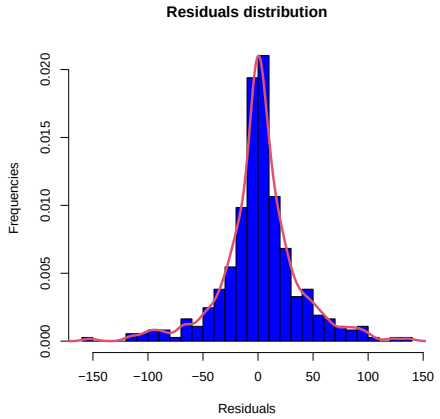


**Fig. 8.** Residuals distribution

By comparing the residuals of the three models presented in figures 5, 6 and 7, we can easily see that those of the model *ARIMA*$(4,1,3)$ satisfy the whiteness hypothesis (see p-values of Ljung-Box statistics) with some p-values exceeding the level $5\%$. This is not the case for the residuals of the models *ARIMA*$(4,1,2)$ and *ARIMA*$(4,1,5)$ for which the p-values of the Ljung-Box test are not significant at level $5\%$. In the sequel, we will then consider the model *ARIMA*$(4,1,3)$.

In Table 2 we present the estimation results of the model *ARIMA*$(4,1,3)$. We note that all the parameters are significant at level $5\%$ (all the observed values of the Student's statistics "t-stat" are greater in absolute values than $1.96$).

**Table 2.** Model *ARIMA*(4,1,3) estimation

| Parameter | $\hat{\phi}_1$ | $\hat{\phi}_2$ | $\hat{\phi}_3$ | $\hat{\phi}_4$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ |
|---|---|---|---|---|---|---|---|
| Estimate | -0.1656 | 0.3145 | -0.5389 | -0.3243 | -0.4619 | -0.6013 | 0.6619 |
| Std. error | 0.0740 | 0.0677 | 0.0681 | 0.0684 | 0.0634 | 0.0781 | 0.0566 |
| t-stat | -2.2378 | 4.6455 | -7.9134 | -4.7412 | -7.2855 | -7.6991 | 11.6943 |

### *3.4. Model adequacy*

In this section, we perform a $5\%$ level *ADF* test to evaluate the whiteness of the residuals of the model *ARIMA*$(4,1,3)$. The obtained p-value is equal to $0.3217$ which is greater than $5\%$. This indicates that the residuals of the model follow a white noise.

Figure 8 shows the distribution of the residuals of the model (which is similar to a centered Gaussian distribution). We can now turn to the model predictions.

### *3.5. Forecasts*

The previous results suggest us to apply *ARIMA*$(4,1,3)$ model to make a short-term forecast for a period of $30$ days, about the number of daily positive cases of *COVID-19* in Senegal. The results are given in Table 3:

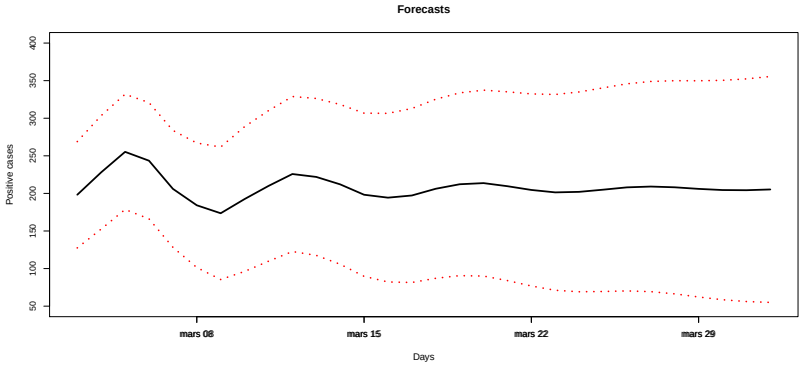


**Fig. 9.** Model forecasts for 30 days.

According to these results, we can see that the predicted values remain smaller than the peak observed on February $12/2021$ (which is $412$ positive *COVID-19* cases). This means that the trend of the number of daily *COVID-19* positive cases in Senegal is in a decreasing phase. Thus herd immunity might be reached if the

**Table 3.** Model forecasts for 30 days.

| Date | Forecasts | Lower | Upper |
|---|---|---|---|
| 03/03/2021 | 198.2185 | 128.93175 | 267.5053 |
| 04/03/2021 | 228.0651 | 154.12684 | 302.0034 |
| 05/03/2021 | 255.1046 | 180.00846 | 330.2007 |
| 06/03/2021 | 243.4073 | 167.63682 | 319.1779 |
| 07/03/2021 | 205.9130 | 129.76582 | 282.0602 |
| 08/03/2021 | 184.1921 | 103.07667 | 265.3076 |
| 09/03/2021 | 173.5312 | 87.20541 | 259.8570 |
| 10/03/2021 | 192.4632 | 98.12992 | 286.7965 |
| 11/03/2021 | 209.8398 | 111.73027 | 307.9493 |
| 12/03/2021 | 225.7056 | 124.70365 | 326.7076 |
| 13/03/2021 | 221.7989 | 119.55131 | 324.0466 |
| 14/03/2021 | 211.9323 | 107.83553 | 316.0290 |
| 15/03/2021 | 198.1524 | 91.84297 | 304.4619 |
| 16/03/2021 | 194.2909 | 84.43318 | 304.1487 |
| 17/03/2021 | 197.1803 | 83.81151 | 310.5492 |
| 18/03/2021 | 206.1128 | 89.43401 | 322.7916 |
| 19/03/2021 | 212.0921 | 93.01160 | 331.1727 |
| 20/03/2021 | 213.6066 | 92.48313 | 334.7301 |
| 21/03/2021 | 209.4859 | 86.49538 | 332.4764 |
| 22/03/2021 | 204.5256 | 79.37683 | 329.6744 |
| 23/03/2021 | 201.2957 | 73.73089 | 328.8606 |
| 24/03/2021 | 201.9999 | 71.81635 | 332.1835 |
| 25/03/2021 | 204.8768 | 72.20833 | 337.5453 |
| 26/03/2021 | 207.9711 | 73.04103 | 342.9011 |
| 27/03/2021 | 209.0315 | 72.08185 | 345.9811 |
| 28/03/2021 | 208.0504 | 69.15937 | 346.9414 |
| 29/03/2021 | 205.9460 | 65.08509 | 346.8068 |
| 30/03/2021 | 204.4110 | 61.47498 | 347.3470 |
| 31/03/2021 | 204.1881 | 59.12052 | 349.2556 |
| 01/04/2021 | 205.1944 | 58.01272 | 352.3761 |

population continue to strictly comply with the barriers.

## 4. Conclusion and perspectives

In this work, we modelled the evolution of daily positive cases data of *COVID-19* in Senegal using Auto Regressive Integrated Moving Average (*ARIMA*) processes. We were able to fit an $ARIMA(4,1,3)$ model to the data and make a prediction for a period of $30$ days. The predicted values fluctuate around an average trend of $200$ positive cases per day out of a total number of daily tests approximately equal to $2000$. It should be noted, however, that the data do not show any seasonality and the period during which we observed the lowest number of positive cases corresponds to the winter season. In order to limit damages caused by *COVID-19*, governmental authorities should commit to reinforcing the measures already taken and ensure their strict and rigorous application in order to avoid a new slackening of the population. This work may be extended in two directions:

- With more data (*e.g.* two to three years of data), we could set up a multivariate *ARIMA* model that would take into account the effect of seasonality and some explanatory factors such as: patients' age, medical history, gender, locality, etc. This would allow us to better explain and predict the evolution of the number of daily positive cases of *COVID-19* in Senegal.
- With the beginning of the vaccination campaign, we will also be able to consider an impact study of the anti-covid vaccine on the evolution of the number of daily positive cases of *COVID-19* in Senegal.

**Acknowledgment**. The authors thank a lot the reviewer for his detailed and constructive comments.

## References

Akaike H., 1974., A new look at the statistical model identification, *IEEE* Transactions on Automatic Control, 19 (6): 716–723, doi:10.1109/TAC.1974.1100705.

Anwar M.Y., Lewnard J.A., Parikh S., Pitzer V.E., 2016. Time series analysis of malaria in Afghanistan: using *ARIMA* models to predict future trends in incidence. Malaria journal, 15(1), 566.

Bodini A., Pasquali S., Pievatolo A. Rugerri R., 2020. A stochastic *SIR* model for the analysis of the *COVID-19* Italian. http://arxiv.org/abs/2102.07566v3 epidemic.

Brockwell P.J., Davis R.A., 2002, Introduction to Time Series and Forecasting, Springer.

Box G.E., Jenkins G., 1970. Time Series Analysis: Forecasting and Control. San-Francisco: Holden-Day.

Box G.E., Pierce D.A., 1970. Distribution of residual auto-correlations in auto-regressive-integrated moving average time series models. Journal of theAmerican statistical Association, 65(332), 1509-1526.

Dath C.A.B., Ndollane N., Ndao A., 2020. Observation de la *COVID-19* au Sénégal: Sondage de pic et de situation sans urgence à partir de cent jours d'urgence. Notes Africaines, Epidémiologie, p. 123-128.

Dickey D. A., Fuller W.A., 1979. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association. 74 (366): 427-431.

Earnest A., Chen M.I., Ng D., Sin L.Y., 2005. Using autoregressive in-tegrated moving average (*ARIMA*) models to predict and monitor the number ofbeds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMCHealth Services Research, 5(1), 36.

Fatih C., Hamimes A., Mishra P., 2020. A note on *Covid-19* Statistics, Strange trend and Forecasting of Total Cases in the most Infected African Countries: An *ARIMA* and Fuzzy Time Series Approaches. Afrika Statistika . Volume 7(2), 967-982. Doi: http://dx.doi.org/10.16929/ajas/2020.967.250.

Fall M., Ndiaye B.M., Seydi O., Seck D., 2020. Analysis of *COVID-19* evolution in Senegal : impact of health care capacity. arXiv:2011.06278v1.

Fuller W.A., 1976. Introduction to Statistical Time Series. New York: John Wileyand Sons. ISBN 0-471-28715-6.

Gourieroux C., Monfort A., 1995, Séries temporelles et modèles dynamiques, Economica.

Kwiatkowski D., Phillips P.C.B., Schmidt P., Shin Y., 1992. Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root, Journal of Econo-metrics, 54, 159-178.

Lui Z., Magal P., Seydi O., Webb G., 2020. A model to predict *COVID-19* epidemics with applications to South Korea, Italy, and Spain. medRxiv preprint doi: https://doi.org/10.1101/2020.04.07.20056945.

Liu Z., Magal P., Seydi O., Webb G., 2020. Understanding unreported cases in the 2019-nCov epidemic outbreak in Wuhan, China, and the importance of major public health interventions, *MPDI Biology,*, 9(3), 50.

Liu Z., Magal P., Seydi O., Webb G., 2020. Predicting the cumulative number of cases for the *COVID-19* epidemic in China from early data, *medRxiv,*.

Nesteruk I., 2020. Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China. medRxiv:2020.02.12.20021931. doi: 10.1101/2020.02.12.20021931.

Paules C.I., Marston H.D., Fauci A.S., 2020. Coronavirus infections-more than just the common cold. JAMA, 323 (8): 707-708.

Promprou S., Jaroensutasinee M., Jaroensutasinee K. (2006). Forecasting Dengue Haemorrhagic Fever Cases in Southern Thailand using *ARIMA* Models.

Schwarz G.E., 1978, Estimating the dimension of a model, Annals of Statistics, 6 (2): 461–464, doi:10.1214/aos/1176344136.

Wang N., Fu Y., Zhang H., Shi H., 2020. An evaluation of mathematical models for the outbreak of *COVID-19*. *Precision Clinical Medicine,* 3(2):85–93.

Wang Z., Zhang X., Teichert G., Carrasco-Teja M., Garikipati K., 2020. System inference for the spatio-temporal evolution of infectious diseases: Michigan in the time of *COVID-19*. *Computational Mechanics,* 66(5):1153–1176.

Wu J.T., Joseph T., Kathy L., Gabriel M. L., 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet. doi: https://doi.org/10.1016/S0140-6736(20)30260-9.