Modeling and forecasting the number of coronavirus infections in Togo: an ARIMA model approach with R software

Kadanga, Mayo Takémsi Norris and Togbenu, Fo-Kossi Edem

24 September 2021

Modélisation et prévision du nombre d'infections au coronavirus au Togo: une approche par un modèle ARIMA avec le logiciel R

Edem Togbenu et Mayo Norris Kadanga

Cabinet de recherche et d'études Statistique, Economique et de Gestion (Ca-SEG)

Septembre, 2021

Abstract:

In this paper, we attempt to propose a short-term prediction model of the number of new cases of coronavirus infections in Togo using the R software. From the original daily data, a new weekly database containing 80 observations was constructed. After splitting this new database into training and test samples in order to select the appropriate model, the database was then used to build our forecasting model, the ARIMA(2,1,2) model. This model was used to make forecasts for the next four weeks. The findings show that Togo can expect approximately 1200 infections in average every week if suitable measures are not adopted in order to stop the rapid spread of the virus in the country.

Keywords: Coronavirus, COVID-19, Forecast, ARIMA

Classification JEL : C53

Introduction

La maladie de COVID-19, détectée pour la première fois en Chine en décembre 2019 dans la ville de Wuhan a été causée par le virus SRAS-CoV-2 et s'est rapidement propagée dans le monde entier. Selon Kibala (2020), le taux de propagation de cette pandémie dans le monde et dans le temps a un rythme d'une fonction exponentielle et a attiré l'attention des autorités dans le monde sur le temps imparti pour l'éradiquer et éviter le pire.

A l'heure actuelle, aucun continent n'est épargné. Le virus est maintenant présent dans plus de 190 pays sur cinq continents. En date du 11 septembre 2021, 224 910 918 cas de COVID-19 ont été confirmés et plus de 4 635 260 personnes en sont décédées.

Le continent africain qui au départ semblait être à l'abri de cette pandémie ne l'est plus. A l'heure actuelle, l'Afrique enrégistre près de 8 011 248 cas d'infections au coronavirus avec plus de 202 613 décès.

Le 6 mars 2020, le Togo annonce son premier cas de Covid-19, et depuis cette date le nombre d'infetions ne cessent d'augmenter. Quatre mois après, c'est-à-dire en Juillet 2020, le Togo enrégistrait au total 720 cas et un an plus tard (à la fin Juillet 2021), ce chiffre est passé à plus de 15 000 personnes infectés. Il a juste fallu un mois pour que le pays ne faranchisse le cap de 21 000 infections (soit plus de mille infections toutes les semaines) avec un total de 197 décès liés à cette pandémie de COVID-19 et ceci malgré toutes les mesures de préventions initiées par le gouvernement (mesures barrières, campagne de vaccination…).

Ces chiffres inquiètent le gouvernement qui ne cesse de multiplier les actions pour la prévention.

Il serait donc judicieux en se basant sur les chiffres actuelles, d'effectuer des prévisions futures afin d'aider à la prise des décisions par les autorités pour venir à bout de cette "rapide" propagation du virus à laquelle nous assistons.

L'objectif poursuivit est donc de proposer un modèle de prévision du nombre de nouveaux cas d'infections à court terme au Togo.

Pour cela, nous disposons des données quotidiennes du nombre nouveaux cas au Togo recueillies sur la plate forme " Our World in Data" ***(Disponible ici)*** que nous avons transformées en données hebdomadaires. Ces données couvrent la période de mars 2020 à septembre 2021. Notre modèle est donc construit à partir de ces données hebdomadaires et une prévision a été effectuée pour les quatre périodes suivantes.

Présentation du modèle ARIMA

Les deux méthodes les plus utilisées pour la prévision à court terme des séries chronologiques sont le lissage exponentiel et les modèles ARIMA, qui offrent des approches complémentaires du problème. Si le lissage exponentiel est basé sur une description de la tendance et de la saisonnalité des données, le modèle ARIMA permet de décrire les autocorrélations entre les données.

La méthode ARIMA est une méthode d'analyse en série temporelle proposée par Box-Jenkins (1970) et a ouvert une nouvelle page dans les méthodes de prévision. La série temporelle peut s'expliquer par les comportements présents, passés, les retards et les éléments contingents (Nguyen, 2002). L'utilisation de la méthode proposée par Box-Jenkins nécessite une étape primordiale: *l'étude de la stationnarité* de la série temporelle. Dans le cas de non stationnarité de la série, celle-ci devra être differenciée afin d'obtenir la stationnarité. Plusieurs tests tel que celui de stationnarité KPSS développé par Kwiatkowski-Phillips-Schmidt-Shin, permettent d'étudier la stationnarité d'une chronique . Dans le cadre de ce article, nous avons utilisé ce dernier.

Une chronique y_t admet une représentation *ARIMA(p,d, q)* si elle vérifie l'équation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \ (1)$$

Le modèle ARIMA peut être subdivisé en trois parties :

- La partie autoregressive (AR) qui constitue un modèle permettant de prévoir la variable d'intérêt en utilisant une combinaison linéaire des valeurs passées de la variable elle-même. Le terme autorégression indique qu'il s'agit d'une régression de la variable sur elle-même. Le modèle se présente sous la forme :

$$AR(p): y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \ (2)$$

.

- La partie intégrée (I) qui représente la nombre de différenciation nécéssaire des observations brutes pour rendre la série chronologique stationnaire, c'est-à-dire que les valeurs des données sont remplacées par la différence entre les valeurs des données et les valeurs précédentes.

- La partie moyenne mobile (MA) prend en compte la dépendance entre une observation et une erreur résiduelle à partir d'un modèle à moyenne mobile appliqué à des données décalées.

$$MA(q): y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \ (3)$$

La spécification du modèle consiste à :

- L'identification du modèle (identifier les p, d et q) :
    - ordre de la partie autorégressive **(p)**

- nombre de différenciations **(d)**
- ordre de la partie moyenne mobile **(q)**;

- L'estimation du modèle ;
- La réduction du modèle ;
- La validation du modèle, ce qui consiste à sélectionner le modèle le plus approprié (le test le plus simple est le test de la stationnarité des résidus. Si les résidus sont stationnaires, le modèle est jugé est acceptable (Nguyen, 2002 : 127)) ;
- La prévision des valeurs futures.

Préparations des données

Présentation des données brutes

- **Importation des données**

```
TOGO_Article<-read.csv("Base_Togo_covid.csv", header = T, sep = ";", dec =
".")
TOGO_Article$date<-dmy(TOGO_Article$date)
# Les 15 premières lignes et les 6 prmières colonnes
Base<-as.data.frame(head(TOGO_Article[,1:6],15))
 Base %>% flextable()
```

iso_code	continent	location	date	total_cases	new_cases
TGO	Africa	Togo	2020-03-04	0	0
TGO	Africa	Togo	2020-03-05	0	0
TGO	Africa	Togo	2020-03-06	1	1
TGO	Africa	Togo	2020-03-07	1	0
TGO	Africa	Togo	2020-03-08	1	0
TGO	Africa	Togo	2020-03-09	1	0
TGO	Africa	Togo	2020-03-10	1	0
TGO	Africa	Togo	2020-03-11	1	0
TGO	Africa	Togo	2020-03-12	1	0
TGO	Africa	Togo	2020-03-13	1	0
TGO	Africa	Togo	2020-03-14	1	0

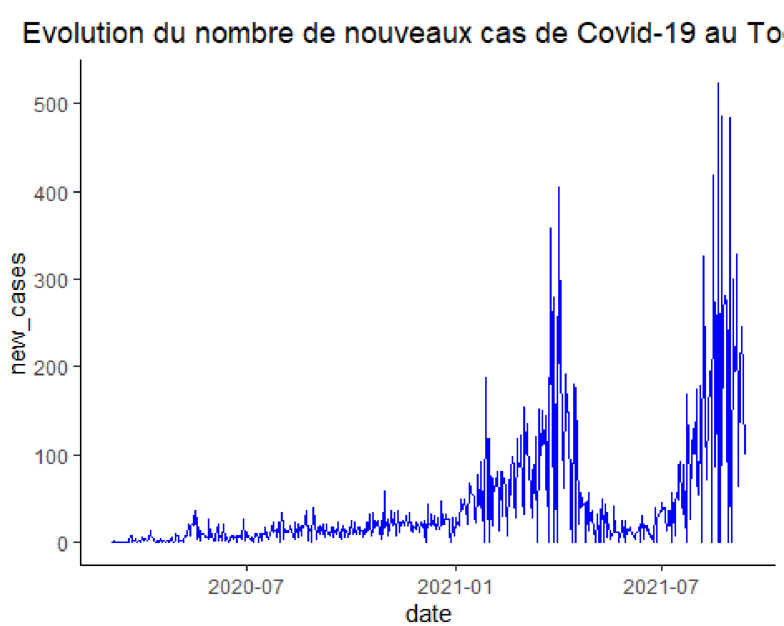
iso_code	continent	location	date	total_cases	new_cases
TGO	Africa	Togo	2020-03-15	1	0
TGO	Africa	Togo	2020-03-16	1	0
TGO	Africa	Togo	2020-03-17	1	0
TGO	Africa	Togo	2020-03-18	1	0

Source: Our World in Data

- **Evolution du nombre de nouveaux cas par jour**

Le graphique ci-dessous décrit l'évolution journalière du nombre de nouvelles infections au coronavirus.

```
#  Graphique des données journalières avec la fonction ggplot() du package
ggplot2
p<-ggplot(TOGO_Article,aes(date,new_cases)) + geom_line(colour ="blue") +
 ggtitle("Evolution du nombre de nouveaux cas de Covid-19 au Togo") +
 theme(panel.grid = element_blank(),
    panel.background = element_rect(fill="white"), axis.line.y =
element_line(colour= "black",

size=0.2),
    axis.line.x = element_line(colour= "black", size=0.2)) +
theme_classic() + theme(plot.title = element_text(hjust=0.6))
p
```



Transformations en données hebdomadaires

Puisque notre modélisation concerne le nombre de nouveaux cas toutes les semaines, nous avons uniquement considéré la variable **new_cases**.

```
# Programme permettant  les données journalières en données hebdomadaires
TOGO_Week<-data.frame(Date=TOGO_Article$date
,New_cases=TOGO_Article$new_cases)
TOGO_Week<-TOGO_Week%>%tq_transmute(select     = New_cases,
                                    mutate_fun = apply.weekly,
                                    FUN        = sum)

data.frame(head(TOGO_Week,15))%>% flextable()
```

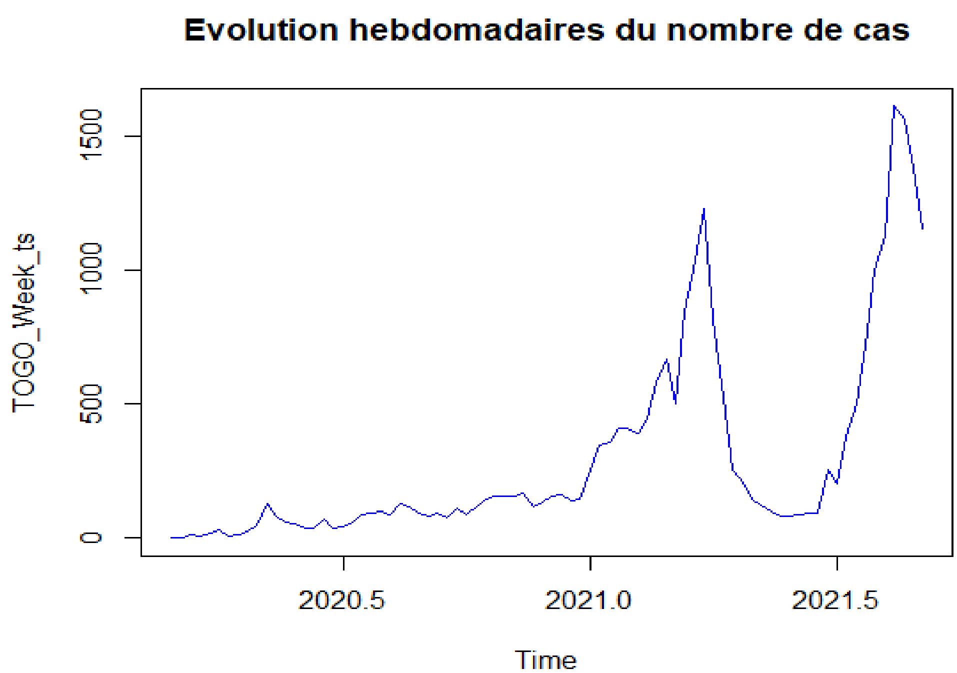
Date	New_cases
2020-03-08	1
2020-03-15	0
2020-03-22	15
2020-03-29	9
2020-04-05	19
2020-04-12	32
2020-04-19	8
2020-04-26	14
2020-05-03	26
2020-05-10	50
2020-05-17	127
2020-05-24	80
2020-05-31	61
2020-06-07	53
2020-06-14	35

```
summary(TOGO_Week)

##       Date              New_cases
##  Min.   :2020-03-08   Min.   :   0.0
##  1st Qu.:2020-07-24   1st Qu.:  81.5
##  Median :2020-12-09   Median : 128.5
##  Mean   :2020-12-09   Mean   : 296.5
##  3rd Qu.:2021-04-26   3rd Qu.: 382.0
##  Max.   :2021-09-12   Max.   :1614.0
```

```r
# Déclaration sous forme de série temporelles

TOGO_Week_ts<-ts(TOGO_Week$New_cases, start =c(2020,9), end = c(2021,36),
frequency = 52)
data.frame(tail(TOGO_Week_ts))%>% flextable()
```

tail.TOGO_Week_ts.
1004
1126
1614
1567
1381
1156

```r
# Grapique des données hebdomadaires

ts.plot(TOGO_Week_ts, col="blue", type = "l",main = "Evolution hebdomadaires
du nombre de cas")
```



Choix de l’échantillon d’apprantissage

Nous disposons de 80 observations (80 semaines). L’échantillon d’apprantissage considéré prend en compte 79 observations et couvre la période du 08 mars 2020 au 05 septembre 2021, la dernière consitue notre échantillon test.

```
TOGO_Week_ts_Appr<-TOGO_Week_ts[1:79]
```

Construction du modèle

Etude de la stationnarité de la série

Test de stationnarité KPSS

```
(Test_resul_Appr<-kpss.test(TOGO_Week_ts_Appr))

##
##  KPSS Test for Level Stationarity
##
## data:  TOGO_Week_ts_Appr
## KPSS Level = 0.97302, Truncation lag parameter = 3, p-value = 0.01
```

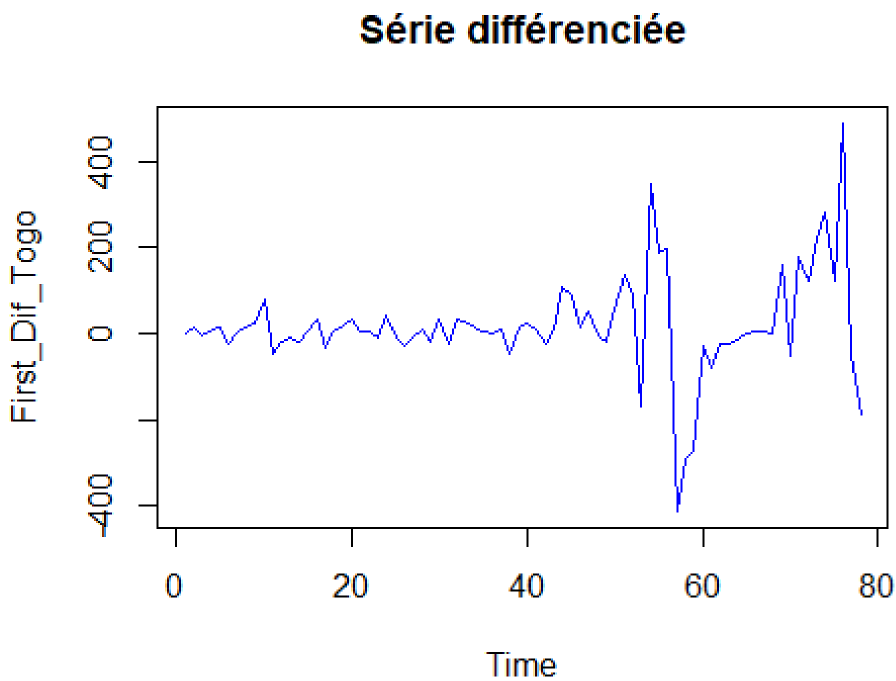
La série est non stationnaire.

Nous allons différencier maintenant une fois la série et étudiée la stationnarité de celle-ci. Cela nous permet de déterminer l’ordre d’intégration d.

```
First_Dif_Togo<-diff(TOGO_Week_ts_Appr,differences = 1)
kpss.test(First_Dif_Togo)

##
##  KPSS Test for Level Stationarity
```

```
##
## data:  First_Dif_Togo
## KPSS Level = 0.15546, Truncation lag parameter = 3, p-value = 0.1
```

```
ts.plot(First_Dif_Togo, main="Série différenciée", col="blue")
```

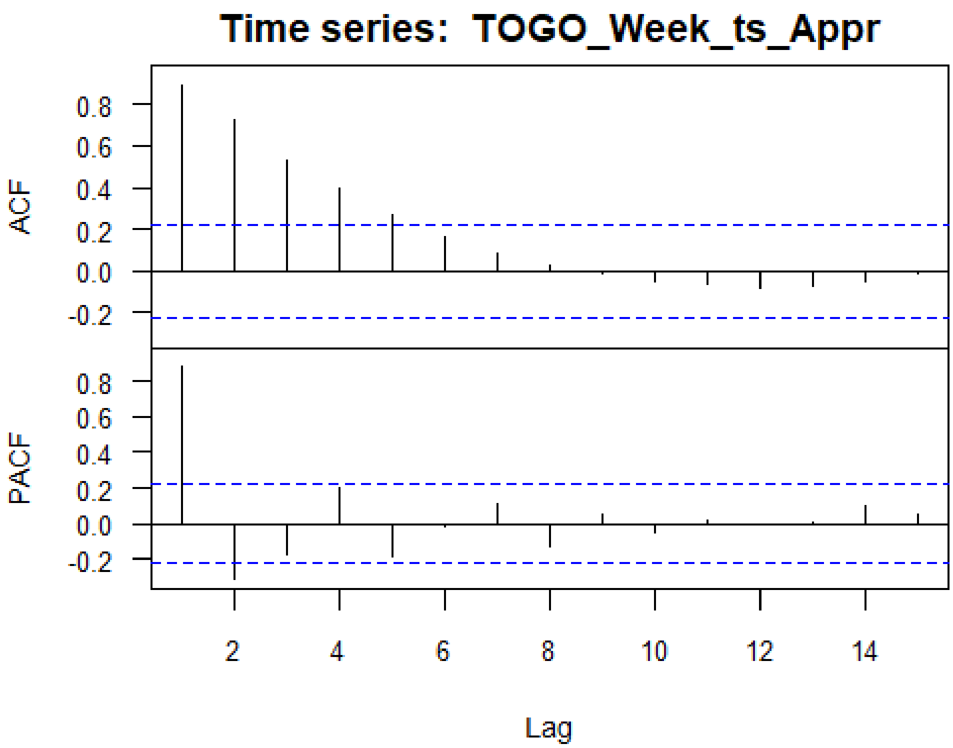


Première difference stationnaire. La série est donc I(1).

Identification des dégrés p et q et estimations des paramètres

- **Fonctions d'autocorrélations et d'autocorrélations partielles.**

```
acf2y(TOGO_Week_ts_Appr, lag.max=15)
```




```
##       LAG        ACF1         PACF
##  [1,]   1  0.88763757  0.887637572
##  [2,]   2  0.72249883 -0.308353462
##  [3,]   3  0.53233862 -0.172064080
##  [4,]   4  0.39527202  0.195595989
##  [5,]   5  0.26463378 -0.183961980
##  [6,]   6  0.15977331 -0.019782719
##  [7,]   7  0.08360945  0.106740952
##  [8,]   8  0.02375275 -0.129082927
##  [9,]   9 -0.01246699  0.053330741
## [10,]  10 -0.04899226 -0.050956425
## [11,]  11 -0.06359257  0.017035830
## [12,]  12 -0.07656233 -0.005746697
## [13,]  13 -0.07336659  0.002125450
## [14,]  14 -0.05462813  0.093946859
## [15,]  15 -0.01471743  0.047833654
```

On choisit p=2 et q=2

- **Estimation**

Afin de pouvoir estimer le modèle, nous avons utilser le fonction *auto.arima() du package forecast* qui permet d'effectuer une modélisation automatique. En précisant les arguments **trace=T** et **ic=aic**, nous avons donner la main au logiciel R de sélectionner le meilleur modèle sur la base du critère *AIC*.

```
Model_arima<-auto.arima(TOGO_Week_ts_Appr,trace=T, ic="aic")
```

```
##
##  ARIMA(2,1,2) with drift         : 959.7146
##  ARIMA(0,1,0) with drift         : 971.1702
##  ARIMA(1,1,0) with drift         : 965.4981
##  ARIMA(0,1,1) with drift         : 967.4571
##  ARIMA(0,1,0)                    : 970.8697
##  ARIMA(1,1,2) with drift         : 961.0042
##  ARIMA(2,1,1) with drift         : 967.2355
##  ARIMA(3,1,2) with drift         : 961.7016
##  ARIMA(2,1,3) with drift         : 961.7054
##  ARIMA(1,1,1) with drift         : 967.1774
##  ARIMA(1,1,3) with drift         : 960.9303
##  ARIMA(3,1,1) with drift         : 966.9768
##  ARIMA(3,1,3) with drift         : Inf
##  ARIMA(2,1,2)                    : 958.3791
##  ARIMA(1,1,2)                    : 959.3705
##  ARIMA(2,1,1)                    : 965.7668
##  ARIMA(3,1,2)                    : 960.3787
##  ARIMA(2,1,3)                    : 960.3788
##  ARIMA(1,1,1)                    : 965.8339
##  ARIMA(1,1,3)                    : 959.6311
##  ARIMA(3,1,1)                    : 965.7078
##  ARIMA(3,1,3)                    : 962.3621
##
##  Best model: ARIMA(2,1,2)
```

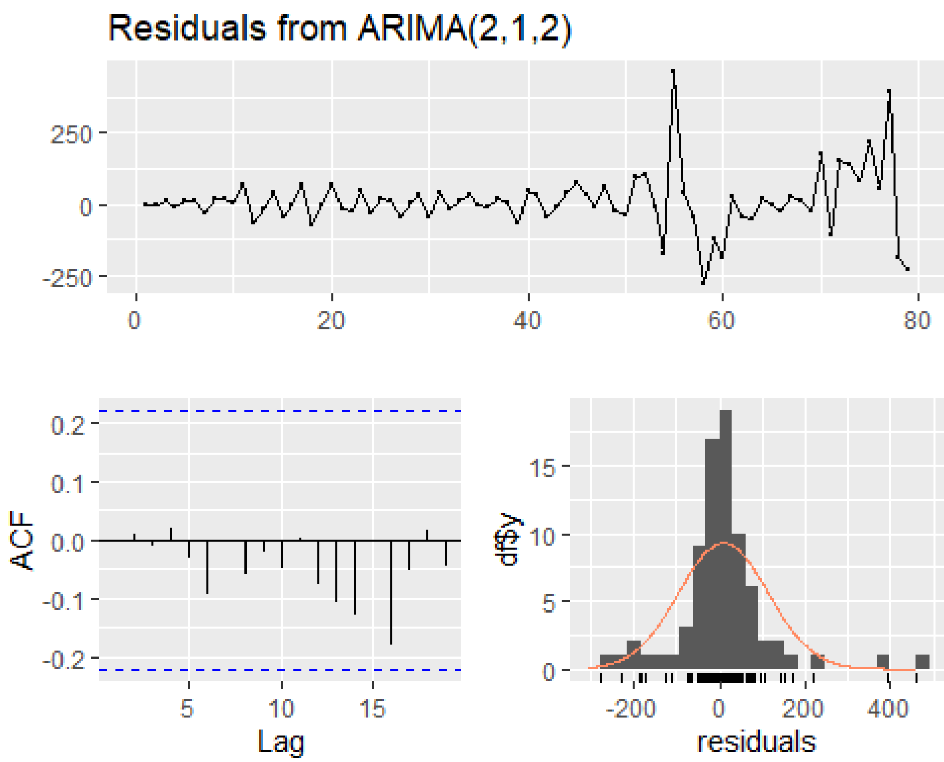
Modèle identifier : ARIMA(2,1,2)

```
summary(Model_arima)

## Series: TOGO_Week_ts_Appr
## ARIMA(2,1,2)
##
## Coefficients:
##           ar1      ar2     ma1     ma2
##       -0.3774  -0.3278  0.8049  0.8509
## s.e.   0.1644   0.1588  0.1044  0.0916
##
## sigma^2 estimated as 11538:  log likelihood=-474.19
## AIC=958.38   AICc=959.21   BIC=970.16
##
## Training set error measures:
##                    ME    RMSE      MAE  MPE MAPE      MASE         ACF1
## Training set 9.345671 103.961 62.48917 -Inf  Inf 0.9203466 -0.005461558

t_stat(Model_arima)

##              ar1       ar2      ma1      ma2
## t.stat -2.295206 -2.063842 7.711824 9.286293
## p.val   0.021721  0.039033 0.000000 0.000000
```

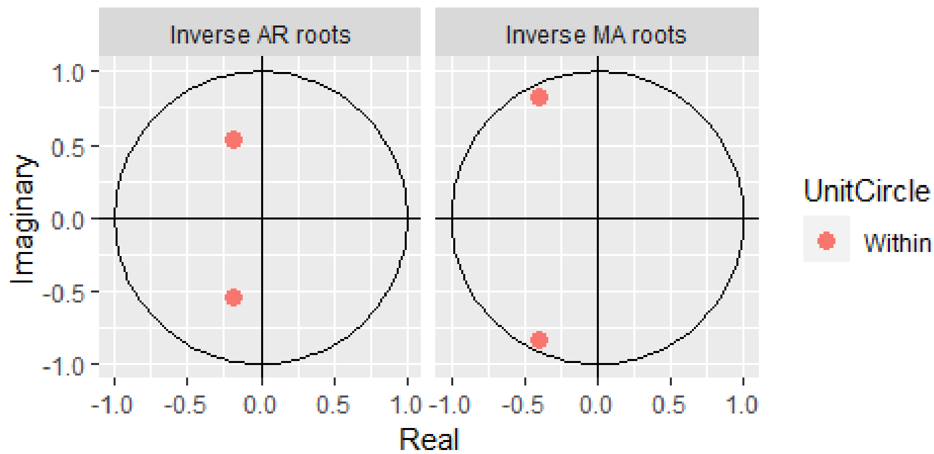
Le modèle n'est pas simplifiable.

Validation du modèle (validation interne)

```
Box.test(Model_arima$residuals,lag = 10, type = "Ljung-Box")

##
##  Box-Ljung test
##
## data:  Model_arima$residuals
## X-squared = 1.4785, df = 10, p-value = 0.999

checkresiduals(Model_arima)
```

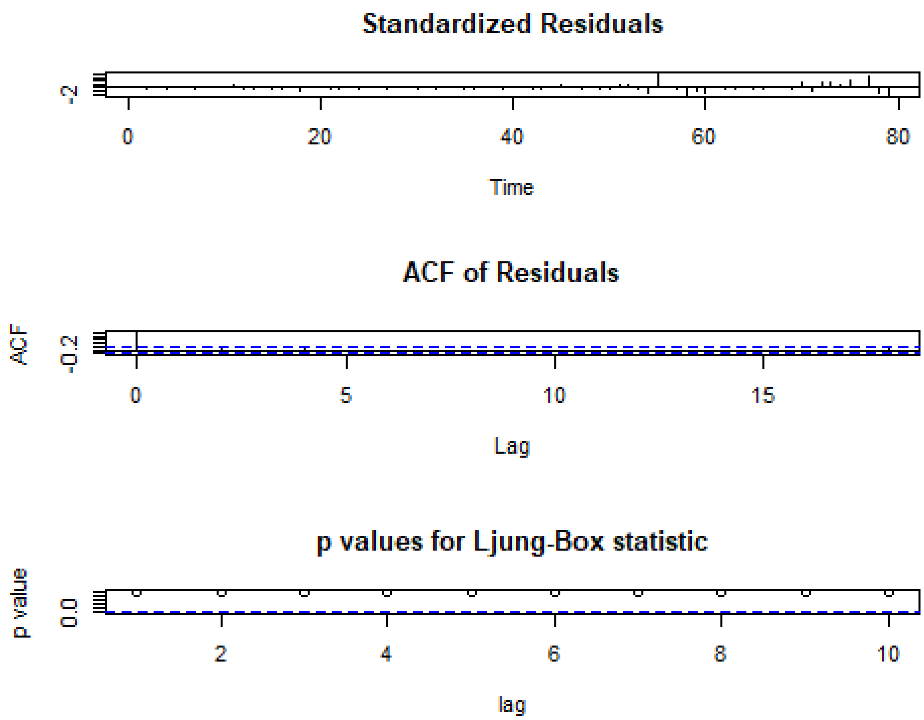



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,2)
## Q* = 1.4785, df = 6, p-value = 0.9609
##
## Model df: 4.   Total lags used: 10

autoplot(Model_arima)
```

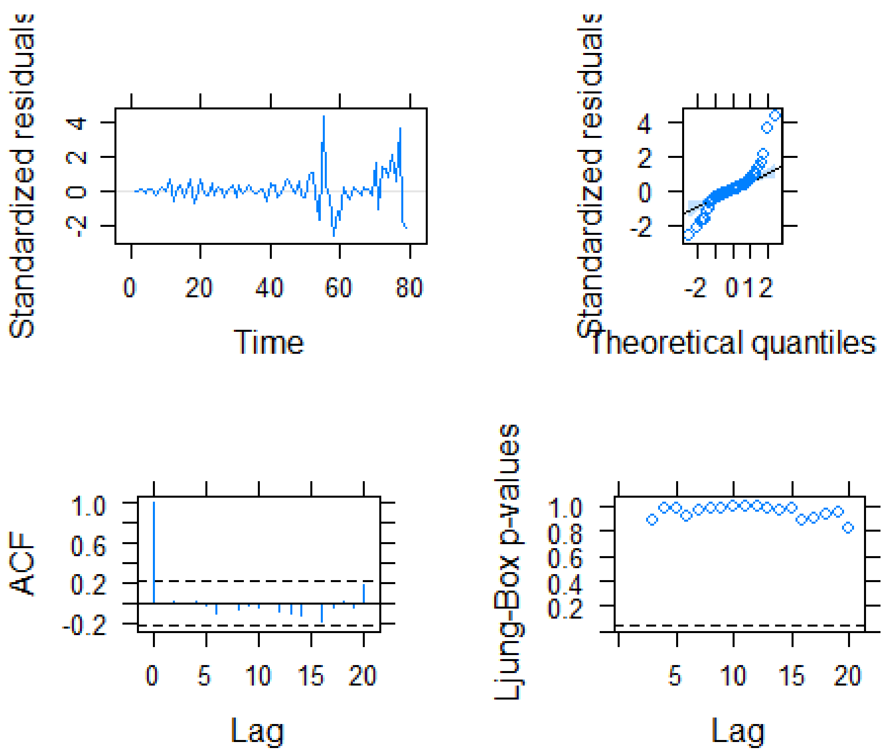


La p-valeur du test de Ljung-Box indique le modèle est acceptable.

```
# Diagnostics poussés avec le package "tactile"
library(tactile)
tsdiag(Model_arima)
```



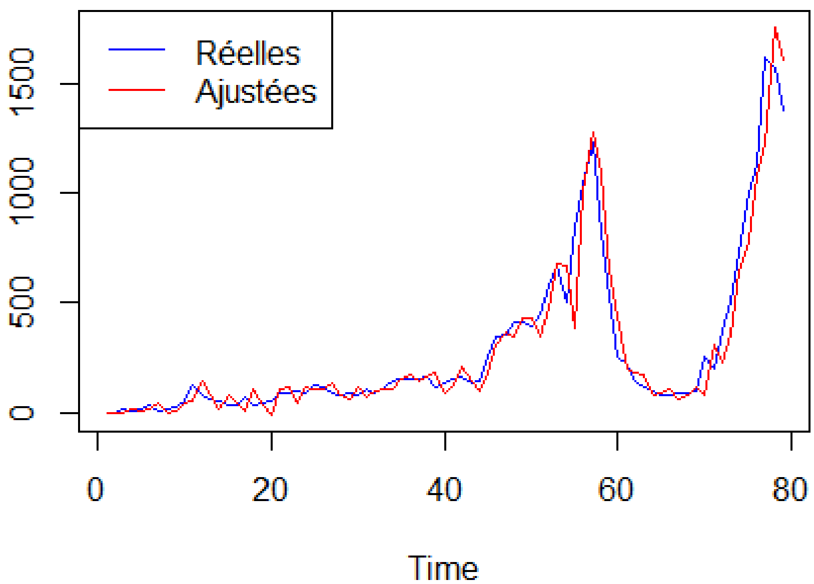

```
xyplot(Model_arima)
```



Le modèle est jugé acceptable après le diagnostic des résidus. Les résidus sont gaussiens, on peut donc inclure les intervalles de prévisions.

Modèle ajusté

```
Mod_estime<-fitted(Model_arima)
# Graphique données réelles et ajustées
ts.plot(TOGO_Week_ts_Appr,Mod_estime,gpars = list(col=c("blue","red")))
legend("topleft",legend = c("Réelles", "Ajustées"),col = c("blue","red"), lty
= 1)
```

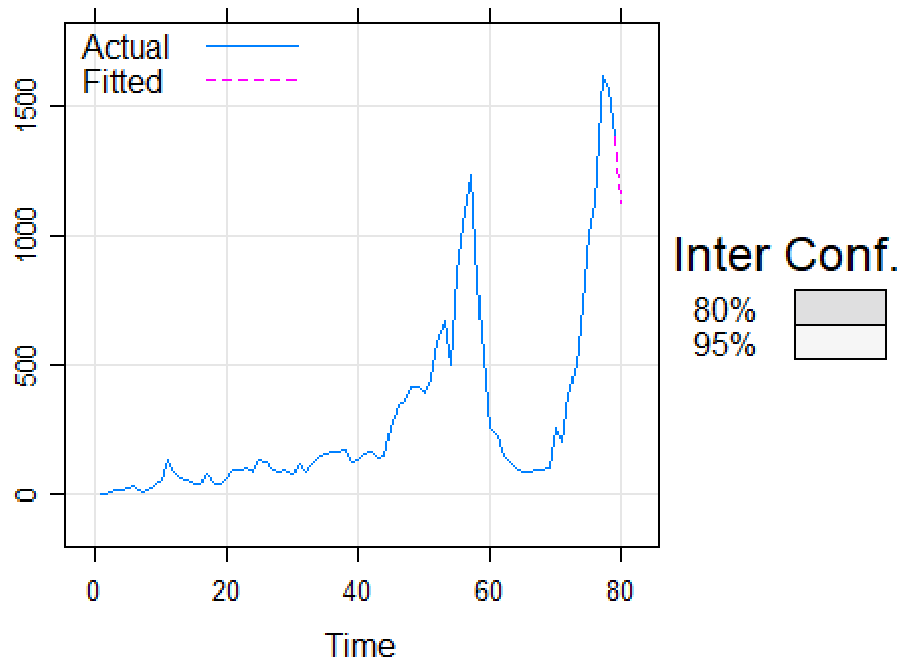


Validation externe

Le modèle étant jugé acceptable, nous avons effectué une prévision pour l'observation restante, cette valeur prédite sera comparée à la donnée réelle afin d'apprécier le pouvoir prédictif de notre modèle.

```
(Prev1<-forecast(Model_arima,1))
```

```
##     Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
## 80       1122.487 984.8278 1260.146 911.9556 1333.018

xyplot(Prev1,grid = T,auto.key=list(corner=c(0,0.99)),ci_key=
list(title="Inter Conf."))
```



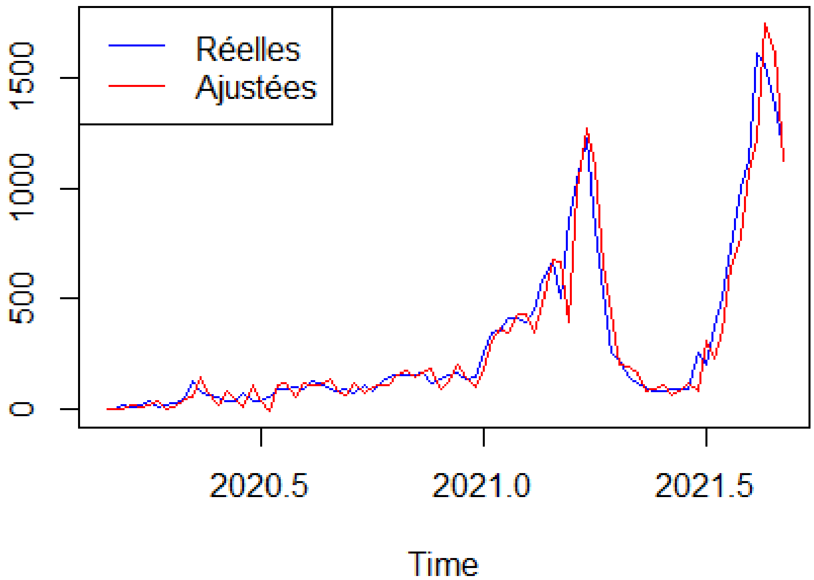

```
tail(TOGO_Week_ts,4)

## Time Series:
## Start = c(2021, 33)
## End = c(2021, 36)
## Frequency = 52
## [1] 1614 1567 1381 1156

BB<-data.frame(Reels=tail(TOGO_Week_ts,1),Prev1)
BB%>% flextable()
```

Reels	Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95
1156	1122.487	984.8278	1260.146	911.9556	1333.018

Le modèle a réuissi à prévoir l'observation de l'échantillon test. En effet, la valeur réelle observée (1156) se trouve dans l'intervalle de confiance à 95% prédit par le modèle.

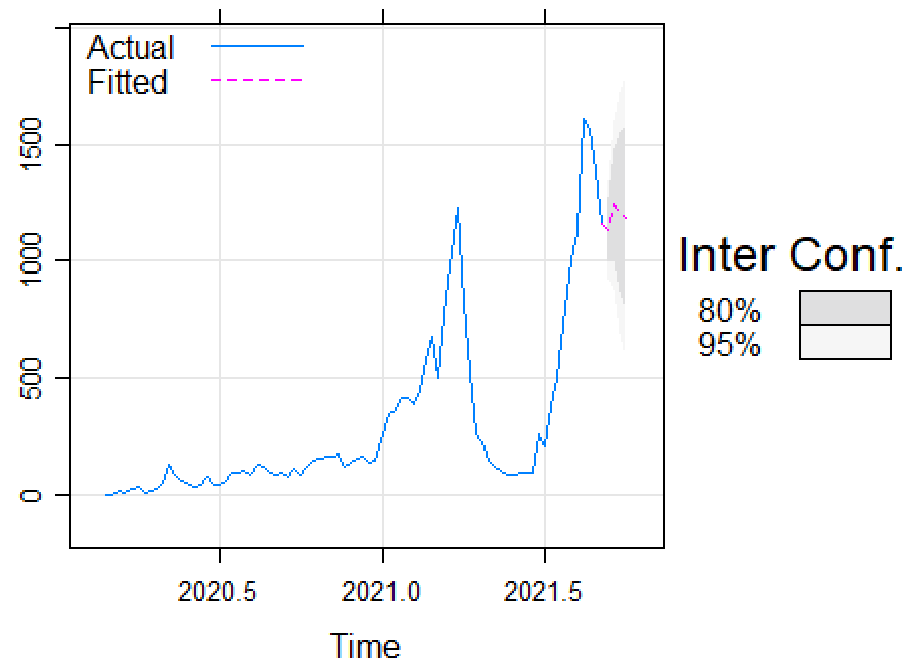
Nous allons maintenant procéder à la modélisation en considérant toutes les observations et ensuite faire des prévisions des 4 prochaines semaines.

```
Model_arima_covid_Togo<-auto.arima(TOGO_Week_ts,trace=T, ic="aic")

##
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,1,2)            with drift         : 961.9885
##  ARIMA(0,1,0)            with drift         : 976.9887
##  ARIMA(1,1,0)            with drift         : 970.3986
##  ARIMA(0,1,1)            with drift         : 972.1908
##  ARIMA(0,1,0)                               : 976.124
##  ARIMA(1,1,2)            with drift         : 962.9574
##  ARIMA(2,1,1)            with drift         : 971.9013
##  ARIMA(3,1,2)            with drift         : 964.9855
##  ARIMA(2,1,3)            with drift         : Inf
##  ARIMA(1,1,1)            with drift         : 972.1295
##  ARIMA(1,1,3)            with drift         : 962.6276
##  ARIMA(3,1,1)            with drift         : 971.3111
##  ARIMA(3,1,3)            with drift         : Inf
##  ARIMA(2,1,2)                               : 960.9197
##  ARIMA(1,1,2)                               : 961.418
##  ARIMA(2,1,1)                               : 970.2073
##  ARIMA(3,1,2)                               : 963.9496
##  ARIMA(2,1,3)                               : 962.9182
##  ARIMA(1,1,1)                               : 970.4993
##  ARIMA(1,1,3)                               : 961.3738
##  ARIMA(3,1,1)                               : 970.0266
##  ARIMA(3,1,3)                               : 965.939
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(2,1,2)                               : 969.5938
##
##  Best model: ARIMA(2,1,2)

Mod_estime_covid_Togo<-fitted(Model_arima_covid_Togo)
# Graphique données réelles et ajustées
ts.plot(TOGO_Week_ts,Mod_estime_covid_Togo,gpars = list(col=c("blue","red")))
legend("topleft",legend = c("Réelles", "Ajustées"),col = c("blue","red"), lty
= 1)
```



Prévisions pour les quatres prochaines semaines

```
Prev_fut<-forecast(Model_arima_covid_Togo,4)
xyplot(Prev_fut,grid = T,auto.key=list(corner=c(0,0.99)),ci_key=
list(title="Inter Conf."))
```


Actual
Fitted
1500
1000
500
0
2020.5
2021.0
2021.5
Time
Inter Conf.
80%
95%

Ci-dessous se trouve le tableau qui résume les résultats des prévisions.

Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95
1134.751	997.8940	1271.609	925.4460	1344.057
1241.854	1004.5557	1479.152	878.9376	1604.770
1207.291	868.7497	1545.832	689.5368	1725.045
1184.452	789.4225	1579.481	580.3066	1788.597

Conclusion

L'objectif visé dans cette étude était de proposer un modèle de prévision à court terme du nombre de nouveaux cas d'infection au coronavirus en utilisant le logiciel R. Pour cela, nous avons subdivisé en deux sous échantillons (étchantillon d'apprantissage et échantillon test) l'échatillon initial. Nous avons tout d'abord construit le modèle à partir de l'échantilon d'apprantissage et grâce à ce modèle, nous avons réussi à prévoir l'intervalle dans lequel ce situe l'obervation test. Ce resultat satisfaisant nous a ensuite motivié à proposé un modèle de prévision à court terme. Des prévisions pour les quatres prochaines semaines ont été effectués.

Le modèle prévoit approximativement un total de 4769 nouvelles inffections au coronavirus au Togo pour les quatre prochaines semaines soit en moyenne 1193 nouveaux cas toutes les semaines. Les résultats montrent que le Togo doit adopter des mesures adéquates afin de stoper cette rapide propagation du virus dans le pays.

RÉFÉRENCES BIBLIOGRAPHIQUES

E. E. Holmes, M. D. Scheuerell, and E. J. Ward. [2021] *Applied Time Series Analysis for Fisheries and Environmental Sciences*

R. J. Hyndman and G. Athanasopoulos [2018] *Forecasting: Principles and Practice (2nd ed) Monash University, Australia*

SETScholars: Coding and Math resources *Modeling and forecasting population in Bangladesh using ARIMA modelling approach in R*