Forecasting Models Project

04/11/2022

A L'ATTENTION DE MONSIEUR LAIB Naamane

AIT SI HAMMOU Mourad & EL BOUHI Ali & OUMLALA Omar
3ème année d'école d'ingénieur mathématiques appliquées, en spécialité Data Science. | Année scolaire 2022/2023.

Sommaire :

1 Introduction générale

Un processus stochastique stationnaire est un processus dont la distribution de probabilité varie plus ou moins constamment sur une certaine période de temps. En d'autres termes, une série de nombres peut sembler (et être) chaotique mais prendre des valeurs dans une fourchette limitée. Grâce à ces informations, des modèles peuvent être créés pour tenter de prédire la variable. Les rendements quotidiens d'un actif financier sont un exemple de processus stochastiques stationnaires.

L'une des méthodes de prévision de séries temporelles les plus répandues est la méthode ARIMA. ARIMA signifie : AutoRegressive Integrated Moving Average. Il s'agit d'un modèle qui prédit les valeurs futures d'une série temporelle sur certains aspects de la structure statistique de la série observée.

Le modèle ARIMA est une généralisation, pour les séries non-stationnaires, du modèle ARMA qui est lui-même la composition des modèles AR (auto-régressif) et MA (Moyennes Glissante ou *Moving Average*) où le I de ARIMA signifie *« integrated »* et indique qu'il faut différencier la série originale afin d'éliminer un caractère non-stationnaire éventuel.

Dans le cadre de ce projet nous allons donc effectuer des analyses sur un jeu de données afin de déterminer le modèle ainsi que les paramètres les plus adaptés. Le jeu de données que nous allons utiliser concerne le GDP (Gross domestic product) autrement dit le produit intérieur brut des états unis de 1947 à 2022. Lien du jeu de données

2 Caractéristique de la série

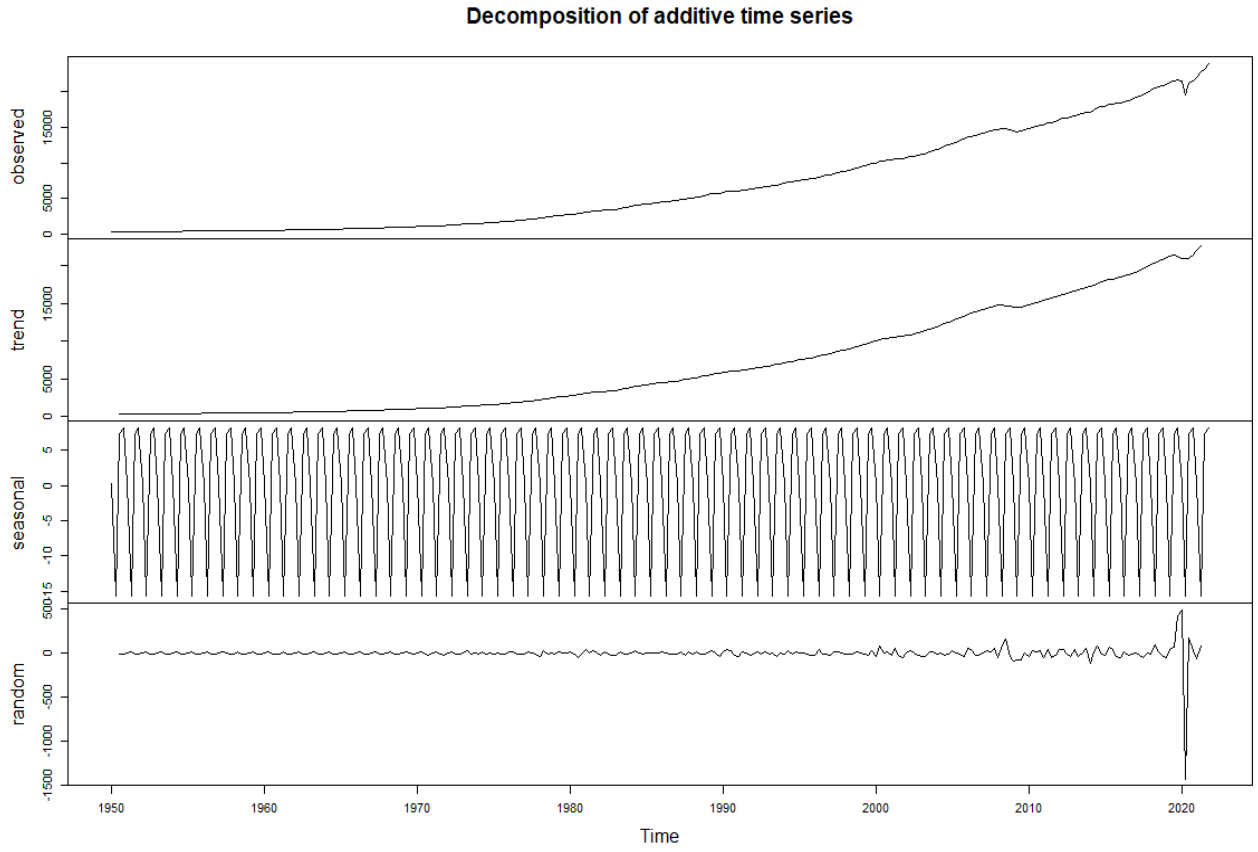
2.1 Description et stationnarité

Comme nous l'avons mentionné ci-dessus, notre jeu de donnée représente le PIB (Produit intérieur brut) de manière trimestrielle des années 1947 à 2022.

```
df <- read.csv('GDP_USA.csv',sep=',')

gdp <- ts(df$GDP,start=c(1947,1,1),frequency = 4)
gdp_ts <- window(gdp,start=c(1950,1),end=c(2021,4))
gdp_ts_test <- window(gdp,start=c(2022,1))
```

Afin d'exploiter ce jeu de données, nous l'avons d'abord converti en série temporelle avant d'en extraire la tendance et la saisonnalité, puis nous avons obtenus la décomposition suivante :

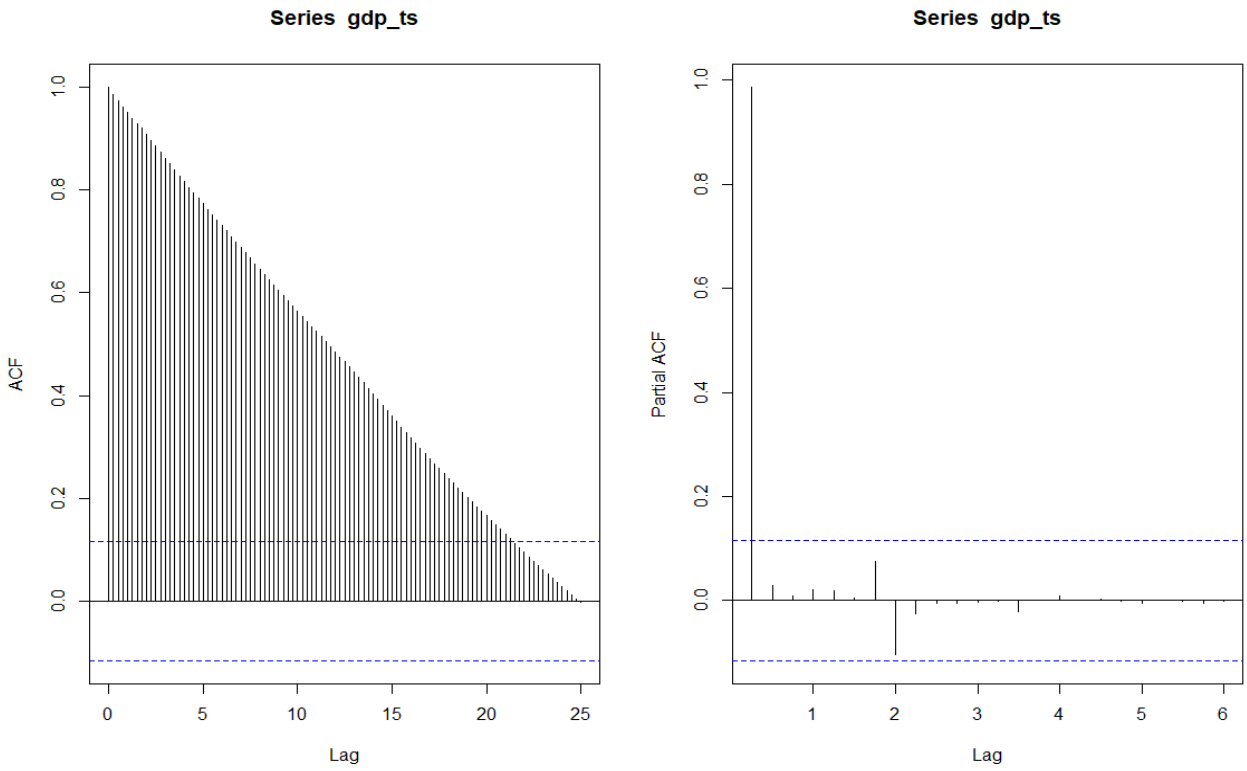
```
gdp_ts_component <- decompose(gdp_ts)
plot(gdp_ts_component)
```



Comme le montre la figure ci-dessus, les deux variables suivent approximativement une forme exponentielle avec une tendance générale à la hausse. En d'autres termes, la séquence originale du PIB n'est pas stationnaire.

Cependant, afin de confirmer cette conclusion, nous avons tracé la fonction d'autocorrélation (ACF) et la fonction d'autocorrélation partielle (PACF) en fonction des décalages, connus sous le nom de corrélogramme des deux séries.

```
par(mfrow=c(1,2))
acf_gdp<- acf(gdp_ts,lag.max=100)
pacf_gdp <- pacf(gdp_ts)
```




```
Box.test(gdp_ts, lag=20, type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  gdp_ts
## X-squared = 4665.5, df = 20, p-value < 2.2e-16
```

La figure montre clairement que la caractéristique la plus frappante de ce corrélogramme est que les coefficients d'autocorrélation aux différents retards sont élevés, jusqu'à un retard de 20 pour le PIB. De plus, le test de Ljung-Box au 20e retard a une valeur de probabilité de 0,000 sous (H0). Le Test de Ljung-Box est un test statistique qui teste l'auto-corrélation d'ordre supérieur à 1

- L'hypothèse nulle (H0) stipule qu'il n'y a pas auto-corrélation des erreurs d'ordre 1 à 20.
- L'hypothèse de recherche (H1) stipule qu'il y a auto-corrélation des erreurs d'ordre 1 à 20.

Nous pouvons donc valider l'hypothèse (H1), par suite nous pouvons conclure que le PIB est non stationnaire, nous devons donc transformer la série pour qu'elle soit stationnaire.

2.2 Transformation de la série

Pour éliminer la non-stationnarité, nous avons pris le logarithme naturel du PIB obtenant ainsi une nouvelle variable nommée LGDP. Variable à laquelle nous avons appliqué le test de **Dickey-Fuller augmenté** (ADF) qui a une valeur de probabilité de 0.99 sous (H0).

```
lgdp=log(gdp_ts)
adf.test(lgdp)
```

```
## Warning in adf.test(lgdp): p-value greater than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  lgdp
## Dickey-Fuller = 0.34807, Lag order = 6, p-value = 0.99
## alternative hypothesis: stationary
```

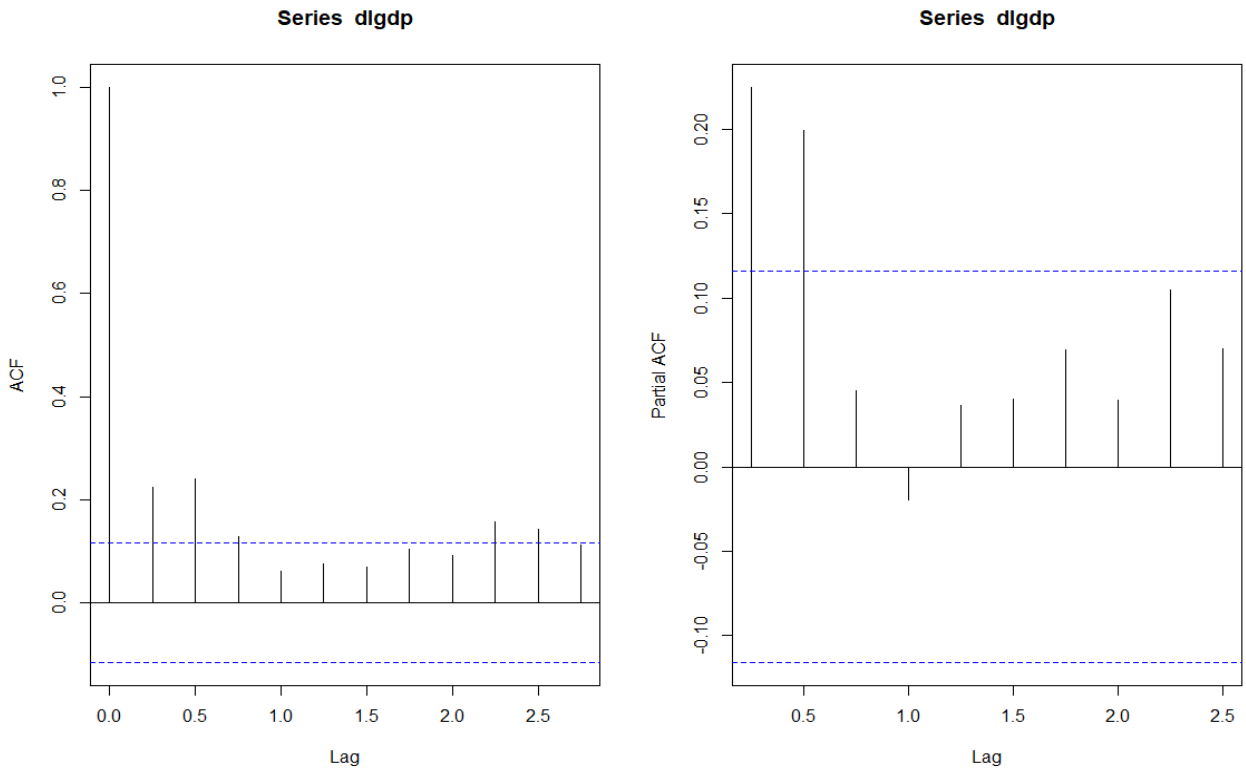
Le test augmenté de Dickey-Fuller est un test statistique qui vise à savoir si une série temporelle est stationnaire c'est-à-dire si ses propriétés statistiques (espérance, variance, auto-corrélation) varient ou pas dans le temps.

- L'hypothèse nulle (H0) stipule que les données ne sont pas stationnaires
- L'hypothèse nulle (H0) stipule que les données sont stationnaires

Nous pouvons donc valider l'hypothèse (H1), par suite nous pouvons affirmer que le LGDP était toujours **non stationnaires**.

Par conséquent, la différence de premier ordre a été prise, pour la série produisant une nouvelle variable DLGDP. Cette variable c'est également avérée **non stationnaires** comme le confirme le résultat de la corrélation **ACF** et **PACF**.

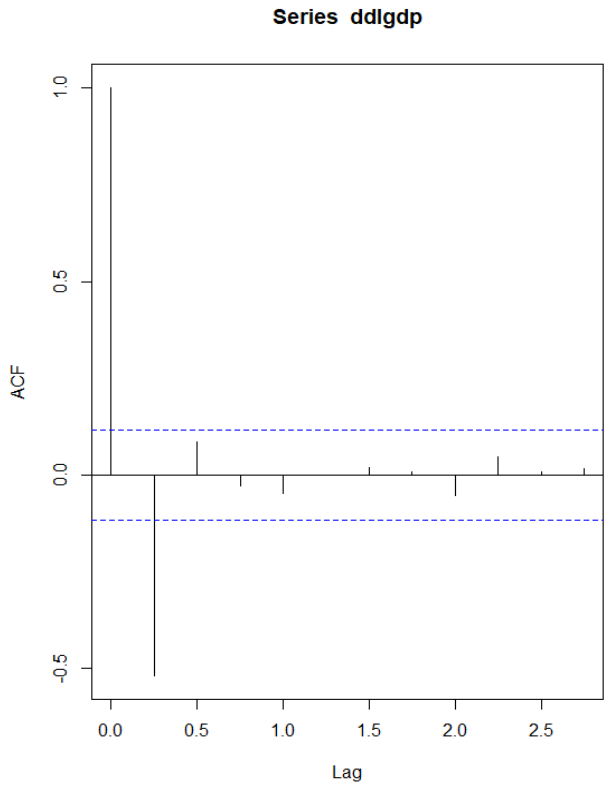
```
dlgdp <- diff(lgdp)
par(mfrow=c(1,2))
acf(dlgdp,lag.max=11)
pacf(dlgdp,lag.max=10)
```

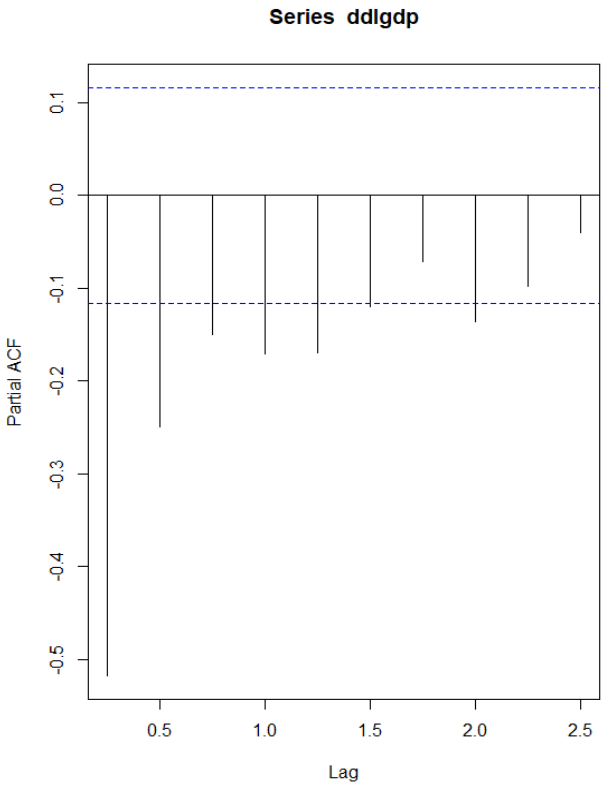


En conséquence, la différence de second ordre a été utilisée afin de produire une nouvelle variable DDLGDP.

```
ddlgdp <- diff(dlgdp)
adf.test(ddlgdp)

## Warning in adf.test(ddlgdp): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  ddlgdp
## Dickey-Fuller = -9.7442, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary

par(mfrow=c(1,2))
acf(ddlgdp,lag.max=11)
pacf(ddlgdp,lag.max=10)
```

Cette fois-ci, la variable s'est avérée **stationnaires**, comme le confirment les résultats du test ADF.





3 Modèle

3.1 Identification

Identifier le modèle revient à déterminer les valeurs des paramètres p et q dans le modèle ARIMA (**p,d,q**). Selon le corrélogramme de la série stationnaire (**DDLGDP**) :

- **p** étant le décalage auquel le **PACF** se coupe.
- **q** étant le décalage auquel le **ACF** se coupe

Etant donné que le modèle est stationnaire aux secondes différences, la paramètre **d** prend comme valeur 2 (d = 2).

En se référant à l'**ACF** précèdent, nous pouvons conclure que le coefficient d'autocorrélation pour la série **DDLGDP** est significativement non nul lorsque l'ordre de décalage est (0,1), ce qui représente les valeurs que **q** peut prendre.

Quant au coefficient d'autocorrélation partielle pour la série **DDLGDP**, il est significativement non nul lorsque l'ordre de décalage est (1,2,3,4,5,6,8), donc p=1,2 et q=1,2,3,4,5,6,8 et ces valeurs peuvent être prises en considération.

3.2 AIC minimum

Afin de prédire le PIB, cette combinaison peut être faite automatiquement, et une combinaison optimale peut être choisie sur la base de l'AIC minimum.

Le critère d'information d'Akaike, ou AIC est une mesure de la qualité d'un modèle statistique.Lorsque l'on estime un modèle statistique, il est possible d'augmenter la vraisemblance du modèle en ajoutant un paramètre. Le critère d'information d'Akaike, permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie. On choisit alors le modèle avec le critère d'information d'Akaike le plus faible.

```
model1 <- arima(lgdp,order=c(0,2,3))
model2 <- arima(lgdp,order=c(1,2,8))
model3 <- arima(lgdp,order = c(3,2,3))
```

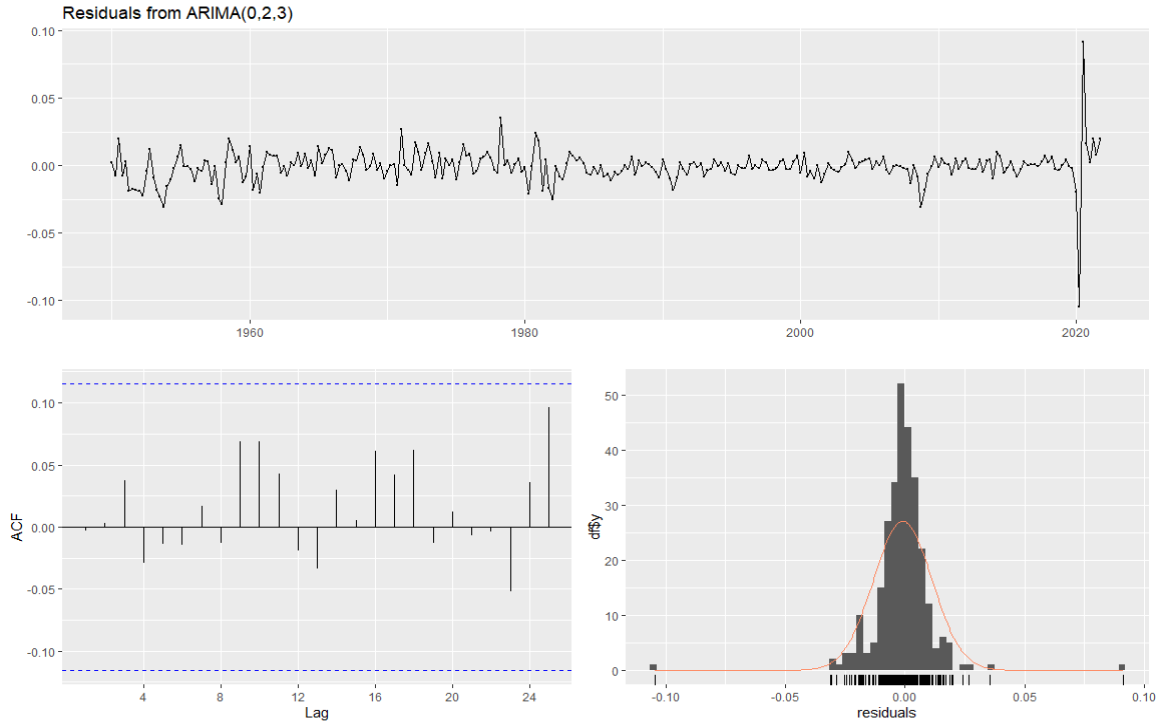
Modèle	Ordre (c)	AIC
1	0,2,3	-1693.83
2	1,2,8	-1683.72
3	3,2,3	-1688.22

Les résultats indiquent que la fonction ARIMA (0, 2, 3) est la mieux adaptée. De plus la fonction auto.arima de R confirme les paramètres présélectionnés.

```
ar_model <- auto.arima(lgdp,stepwise = FALSE,approximation = FALSE)
ar_model
```

```
## Series: lgdp
## ARIMA(0,2,3)
##
## Coefficients:
##           ma1     ma2      ma3
##       -0.8534  0.0648  -0.1487
## s.e.   0.0585  0.0823   0.0611
##
## sigma^2 = 0.0001533:  log likelihood = 850.91
## AIC=-1693.83   AICc=-1693.68   BIC=-1679.2
```

```
residual_plot<- checkresiduals(model1)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,2,3)
## Q* = 0.93381, df = 5, p-value = 0.9677
##
## Model df: 3.   Total lags used: 8

residual_plot

##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,2,3)
## Q* = 0.93381, df = 5, p-value = 0.9677
```

Le graphique ACF des résidus du modèle ARIMA (0,2,3) montre que toutes les autocorrélations se situent dans les limites du seuil, ce qui indique que les résidus se comportent comme un bruit blanc. Un test de Ljung-Box renvoie une grande probabilité sous (H0), ce qui suggère également que les résidus sont des bruits blancs.

3.3 Prédictions

L'un des modèles entrainés et choisis qui sont résumés dans la figure ci-dessus est ARIMA (0,2,3) et peut être appliqué à DDLGDP afin de prévoir les valeurs de LGDP. Il a été estimé avec les résultats suivants :

$$\boldsymbol{DDLGDP = -0.8534\epsilon_{t-1} + 0.0648\epsilon_{t-2} - 0.1487\epsilon_{t-3}}$$

Nous allons comparer les valeurs prédites pour l'ensemble de test qui est se compose de 2022Q1 et 2022Q2. Ci-dessous, nous pouvons voir nos points de prévision et notre intervalle de confiance :

- **Lo80** nous retourne la borne inférieure avec une précision de **80%**
- **Hi80** retourne la borne supérieure avec une précision de **95%**
- **Lo95** retourne la borne inférieure avec une précision de **95%**
- **Hi95** retourne la borne supérieure avec une précision de **95%**

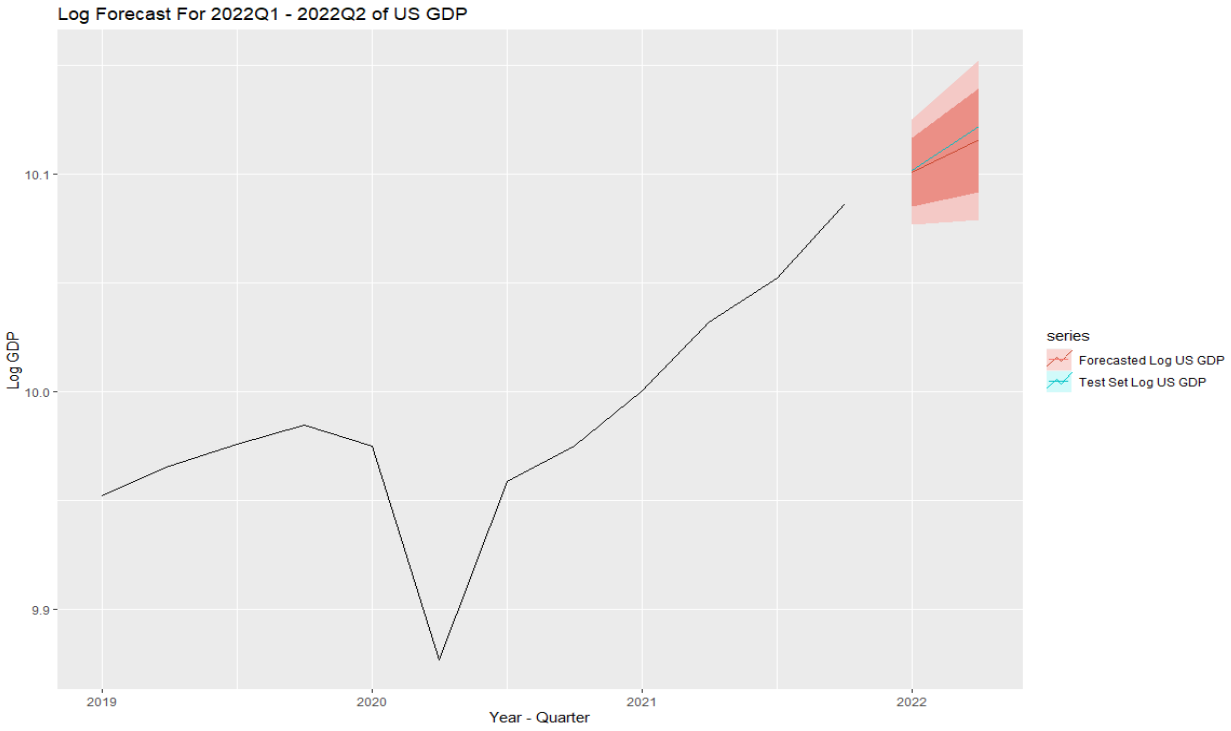
	Point Forecast	Lo80	Hi80	Lo95	Hi95
2022 Q1	10.10066	10.08489	10.11644	10.07654	10.12478
2022 Q2	10.11549	10.09150	10.13949	10.07879	10.15219

```
l_forecasted = forecast(model1,h=2)
acc_l_forecasted <- accuracy(object = l_forecasted,x = log(gdp_ts_test))
acc_l_forecasted

##                        ME       RMSE          MAE          MPE       MAPE
## Training set -0.0009317779 0.01227294 0.007225039 -0.01366347 0.09448783
## Test set      0.0037862685 0.00462470 0.003786269  0.03741770 0.03741770
```

Nous avons constaté que notre erreur est faible lorsqu'elle est comparée aux valeurs prévues pour l'ensemble de test. La métrique MAPE est couramment utilisée car elle est facile à interpréter et à expliquer. Notre valeur MAPE de 0,037% signifie que la différence moyenne entre la valeur prévue et la valeur réelle est de 0,037%. Ce qui est très faible, nous pouvons affirmer que notre modèle est très précis concernant le LGDP.

Pour une meilleure visibilité nous allons projeter notre série temporelle à partir de 2019.

```r
sub_ts <- window(lgdp,start=c(2019))
autoplot(sub_ts,main = 'Log Forecast For 2022Q1 - 2022Q2 of US GDP',xlab = 'Yea
r - Quarter',ylab = 'Log GDP')+
  autolayer(l_forecasted, "Forecasted Log US GDP")+
  autolayer(log(gdp_ts_test),series = 'Test Set Log US GDP')
```



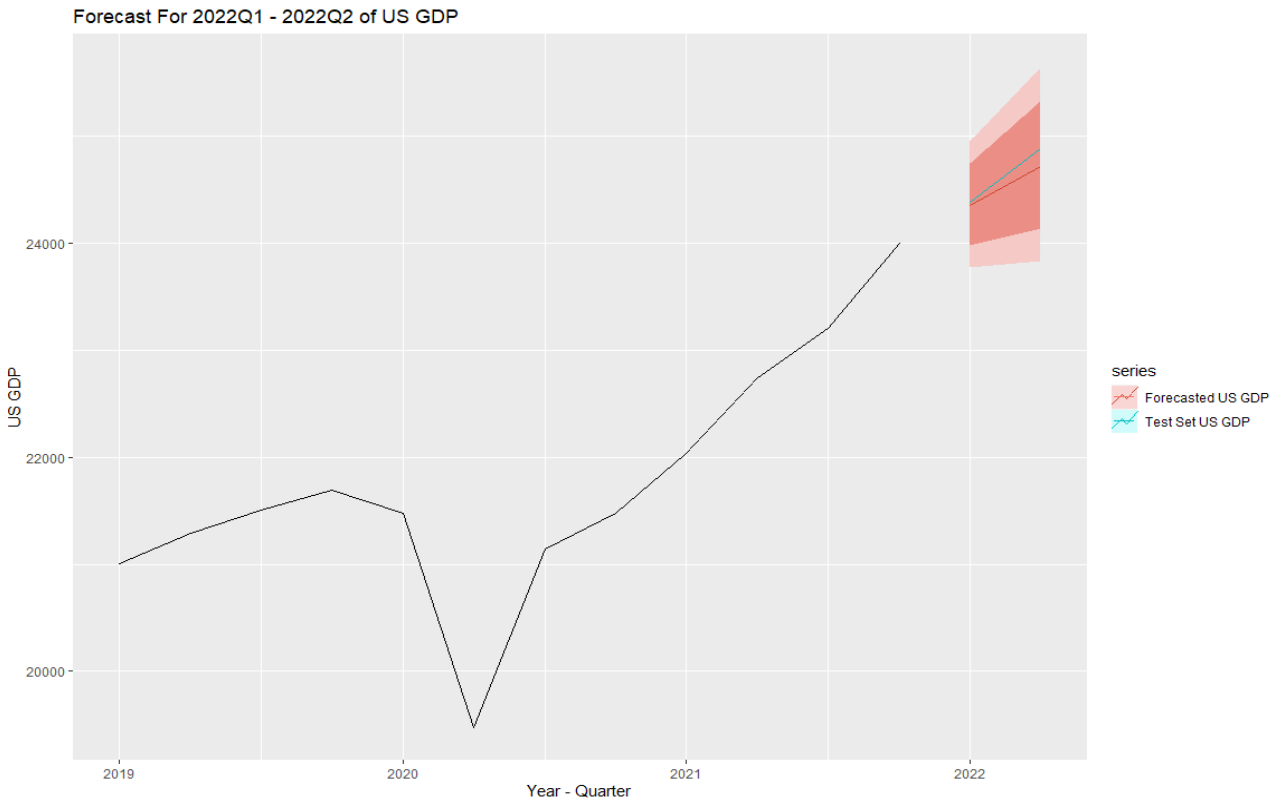
Passons maintenant à la prédiction du PIB en appliquant l'exponentielle à notre série.

	Point Forecast	Lo80	Hi80	Lo95	Hi95
2022 Q1	24359.18	23978.00	24746.41	23778.64	24953.89
2022 Q2	24723.10	24136.91	25323.53	23832.24	25647.26

```
forecasted=l_forecasted
forecasted$mean=exp(l_forecasted$mean)
forecasted$lower=exp(l_forecasted$lower)
forecasted$upper=exp(l_forecasted$upper)
```

```
acc_forecasted <- accuracy(object = forecasted,x = gdp_ts_test)
acc_forecasted

##                          ME         RMSE          MAE          MPE       MAPE
## Training set -0.0009317779   0.01227294  0.007225039 -0.01366347 0.09448783
## Test set     93.6673810402 114.64706304 93.667381040  0.37755970 0.37755970
```

Nous avons constaté que la valeur MAPE est passée à 0,378%, ce qui signifie que la différence moyenne entre la valeur prévue et la valeur réelle est de 0,378%. Cette valeur reste très faible, nous pouvons donc affirmer que notre modèle est très précis concernant le PIB.



4 Conclusion

La méthodologie de la technique Box - Jenkins (ARIMA), qui est une méthode de prévision des séries temporelles relativement avancée, est appliquée dans ce projet dans le but de prévoir le PIB en Amérique pour les deux prochains trimestres (2022Q1, 2022Q2). Après avoir testé la stationnarité des données en utilisant le test ADF et PP, les séries étaient stationnaires à la deuxième différence après avoir calculé le logarithme des données.

A partir du corrélogramme de l'ACF, nous avons déterminé le nombre approprié de termes autorégressifs (p). A partir du corrélogramme du PACF, nous avons déterminé le nombre approprié de termes de moyenne mobile (q). Sur la base de l'AIC minimum, nous avons choisi le modèle ARIMA optimal qui est ARIMA (0,2,3) pour prévoir le PIB pour les trimestres 2022Q1, 2022Q2 comme 24359.18, 24723.10 respectivement.