Projet Série chronologie

Group_D

2023-01-03

Introduction, présentation et pré-traitement des données:

Le présent jeu des données renseigne sur la consommation d'électricité en Allemagne du 1 janvier 2016 au 31 décembre 2017, qui peut être téléchargé sur la site kaggle. Il contient un tableau contenant la date, la consommation d'électricité, la production par l'énergie éolienne, la production par l'énergie solaire, la somme de l'énergie éolienne et l'énergie solaire, toutes sont présentées par GWh. Dans un premier temps nous allons chargé les données puis effectuer des pré-traitements des données. les données sont disponible sur le lien dessus

https://www.kaggle.com/datasets/mvianna10/germany-electricity-power-for-20062017

Vérifion qu'on a bien chargé les données

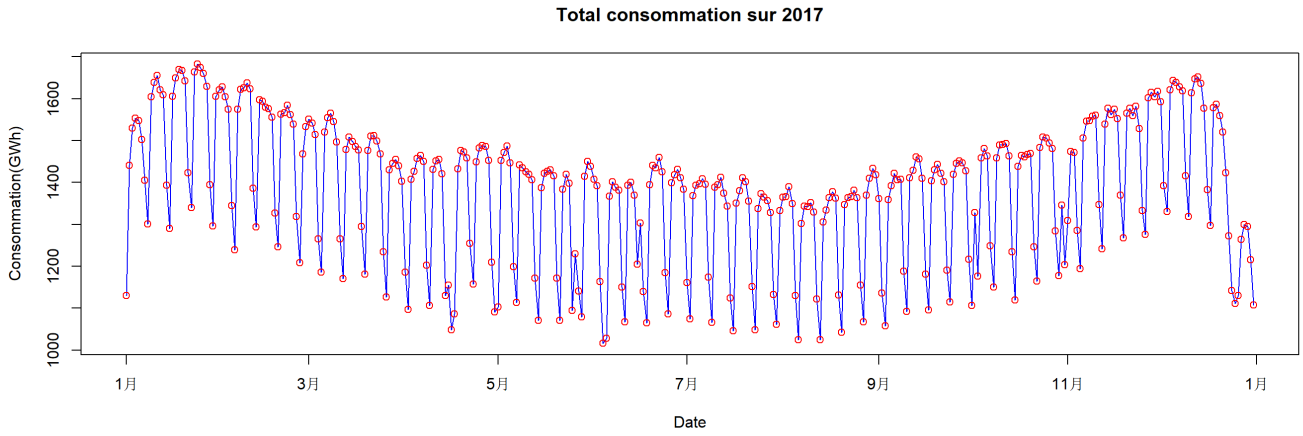
```
##         Date Consumption Wind Solar Wind.Solar
## 1 2006-01-01    1069.184   NA    NA         NA
## 2 2006-01-02    1380.521   NA    NA         NA
## 3 2006-01-03    1442.533   NA    NA         NA
## 4 2006-01-04    1457.217   NA    NA         NA
## 5 2006-01-05    1477.131   NA    NA         NA
## 6 2006-01-06    1403.427   NA    NA         NA
```

Comme pour l'année du 2006 au 2012, la production de l'énergie éolienne et solaire sont des valeurs **na**, nous allons étuidier que les données de l'année 2017 pour diminuer le temps d'exécution et nous allons focaliser dans l'attribut consommation. Voici dessous nous donne des impressions pour cette des données.

```
##            Date Consumption    Wind  Solar Wind.Solar
## 4019 2017-01-01    1130.413 307.125 35.291    342.416
## 4020 2017-01-02    1441.052 295.099 12.479    307.578
## 4021 2017-01-03    1529.990 666.173  9.351    675.524
## 4022 2017-01-04    1553.083 686.578 12.814    699.392
## 4023 2017-01-05    1547.238 261.758 20.797    282.555
## 4024 2017-01-06    1501.795 115.723 33.341    149.064
```

```
##            Date Consumption    Wind  Solar Wind.Solar
## 4378 2017-12-26    1130.117 717.453 30.923    748.376
## 4379 2017-12-27    1263.941 394.507 16.530    411.037
## 4380 2017-12-28    1299.864 506.424 14.162    520.586
## 4381 2017-12-29    1295.088 584.277 29.854    614.131
## 4382 2017-12-30    1215.449 721.247  7.467    728.714
## 4383 2017-12-31    1107.115 721.176 19.980    741.156
```

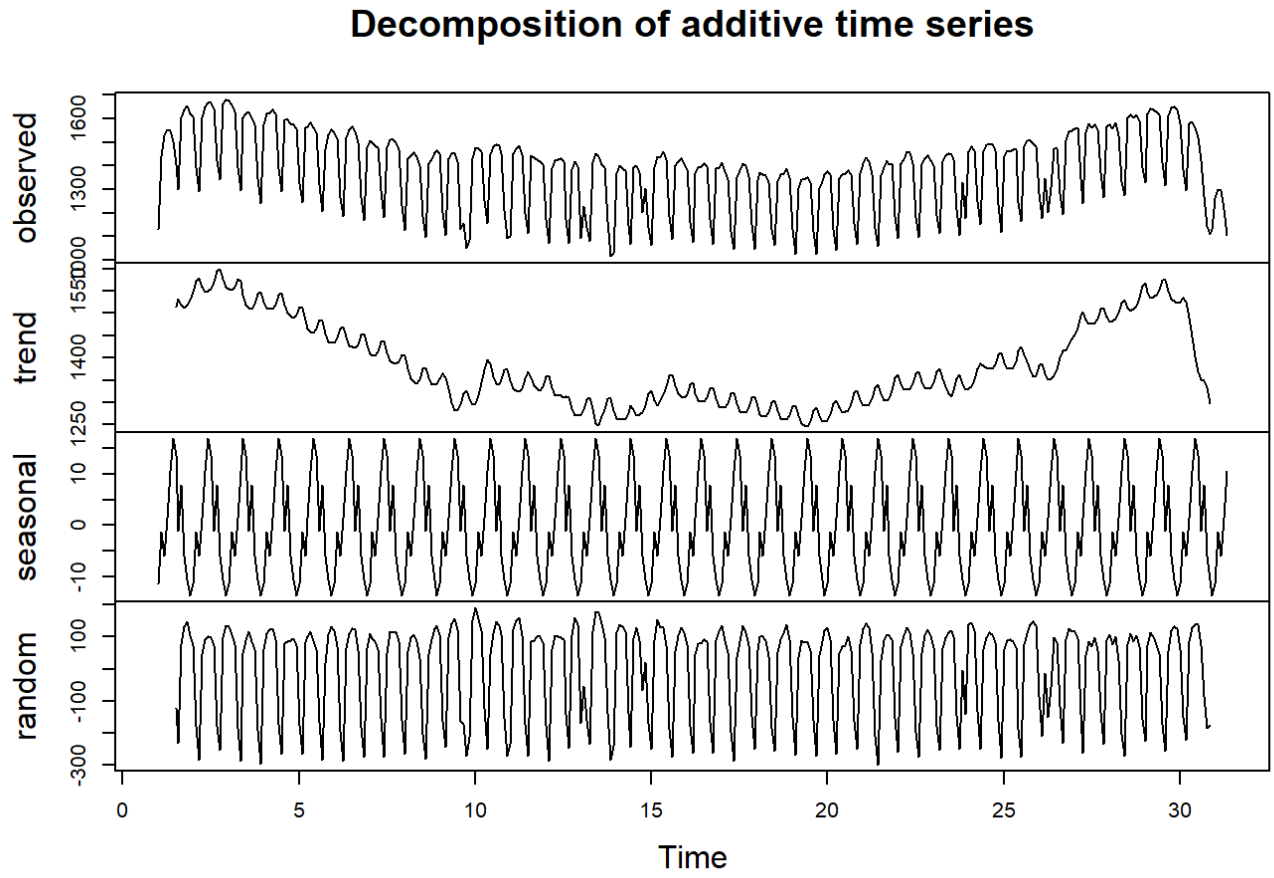
Courbes: La consommation



De 1er janvier 2017 à 31 décembre 2017, la consommation d'électricité est élevé du janvier au mars et du octobre au décembre, et est moins élevé pour la reste. Ceci est raisonnable parce que pendant hiver, les gens utilisent plus d'électricité pour chauffage.

Traitement de la série chronologie:

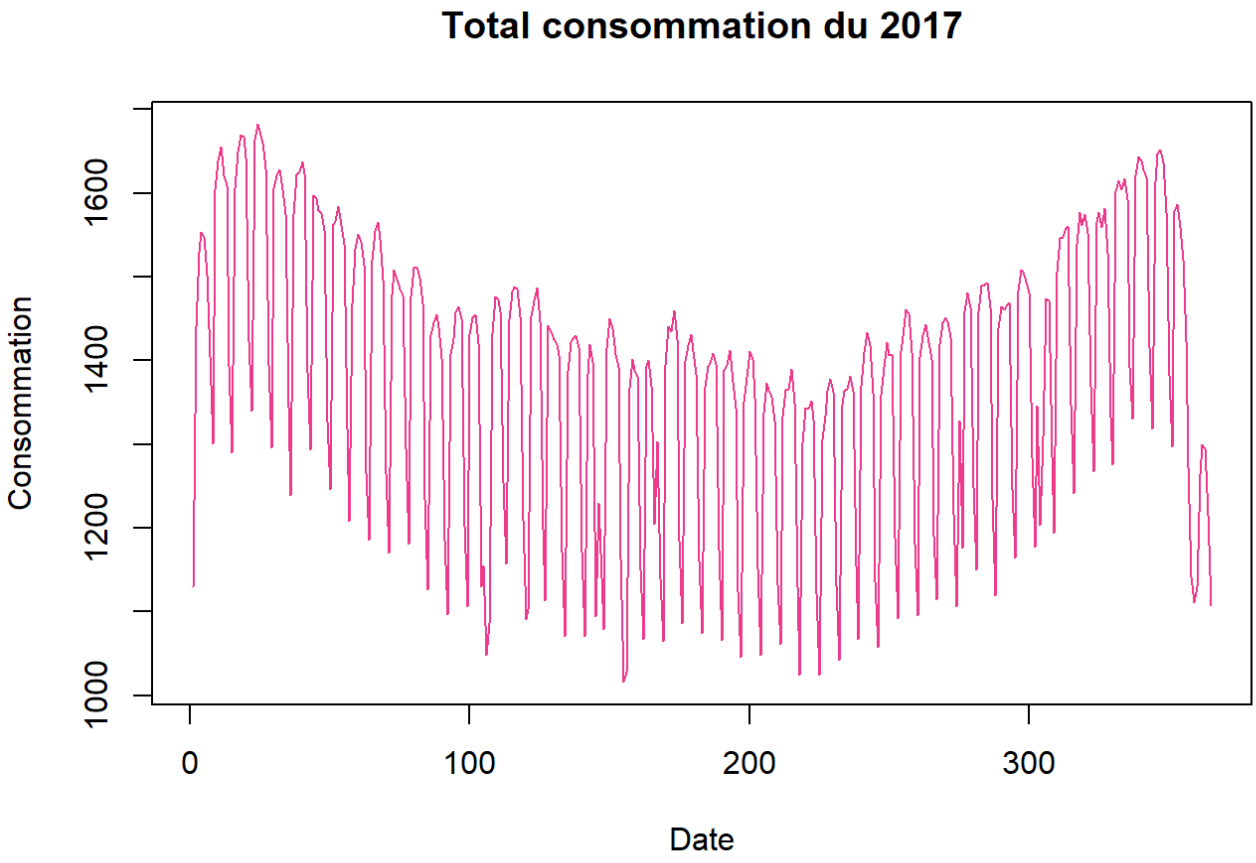
On va traiter les données comme une année entière, donc on va mettre le **frequency** = 12 ce qui présente 12 mois, on dans les figures ci-dessous on peut visualiser une tendance et une périocité. Et que le bruit ne suit pas la loi normale.

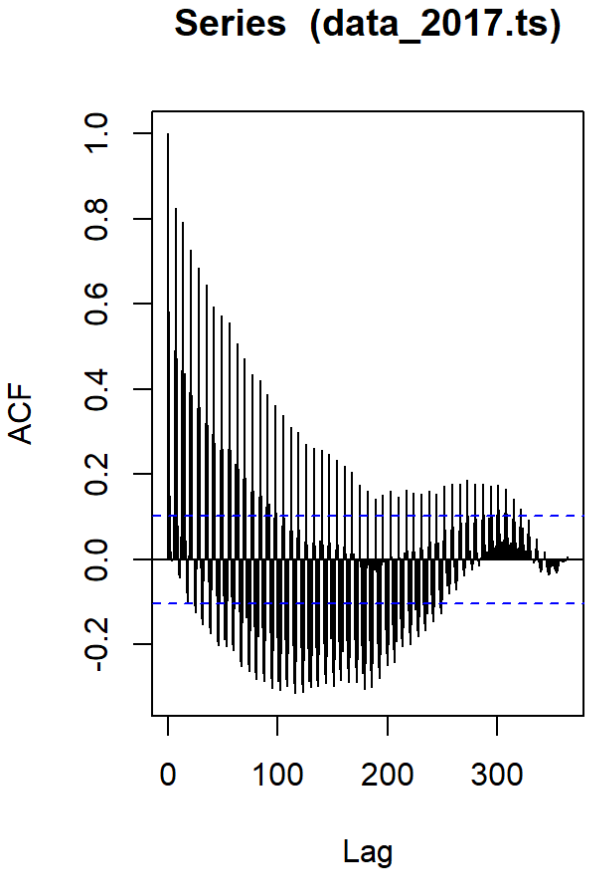


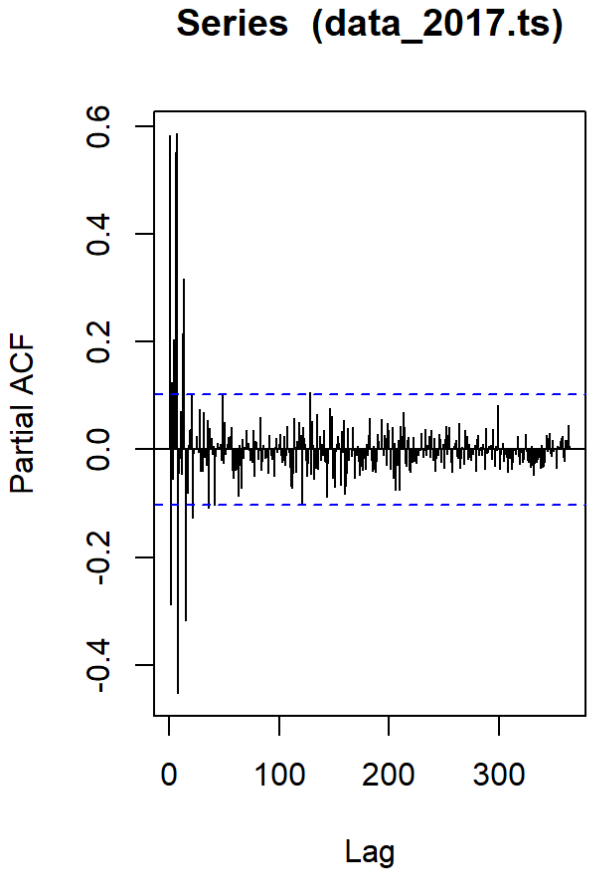
La tendance

Regardons d'abord pour les graphes ci-dessous, on remarqu'il présente une tendance. Mais on ne sait pas trop quelle type de la tendance c'est. Peut-être c'est le type **cos** ou **polynomial** , mais on ne peut pas être

sûre. Donc ce que l'on peut faire c'est tout d'abord faire la différentiation de l'ordre 1 pui on peut encore faire une différentiation de l'ordre 7 pour essayer d'annuler l'effet de la saison. Et on va refaire les même figures et faire du test **Augmented Dickey-Fuller Test** pour tester si la série temporelle après le traitement soit une série temporelle stationnaire ou pas.







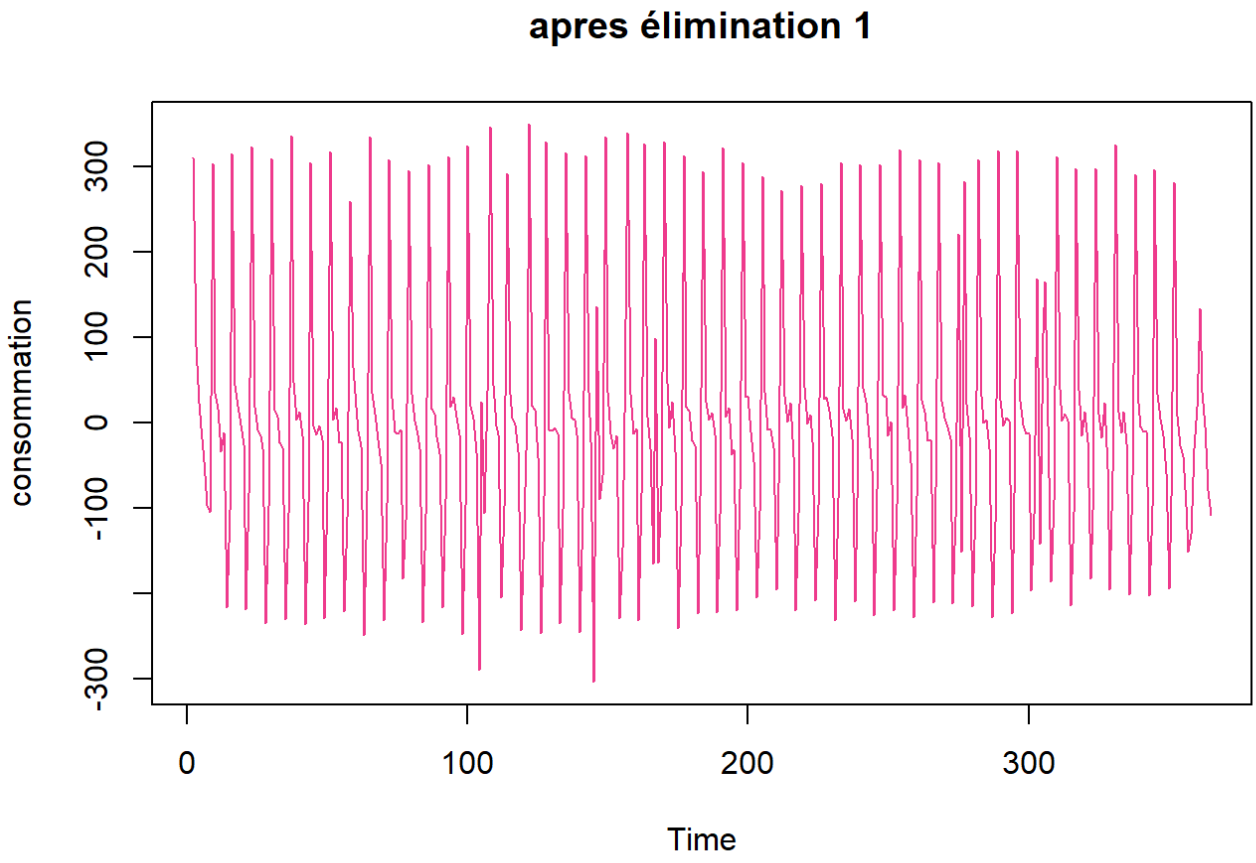

```
##
##  Augmented Dickey-Fuller Test
```

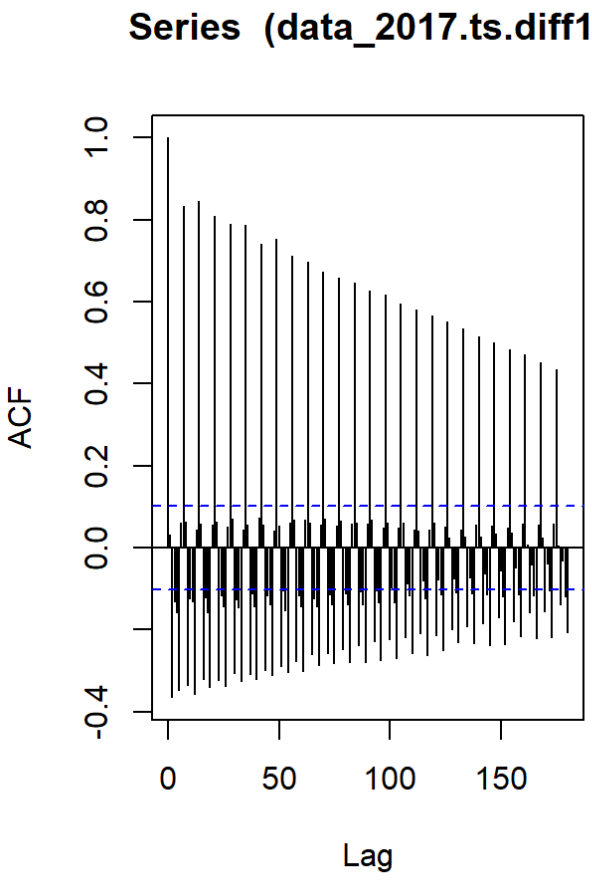
```
##
## data:  data_2017.ts
## Dickey-Fuller = -2.2434, Lag order = 7, p-value = 0.4742
## alternative hypothesis: stationary
```

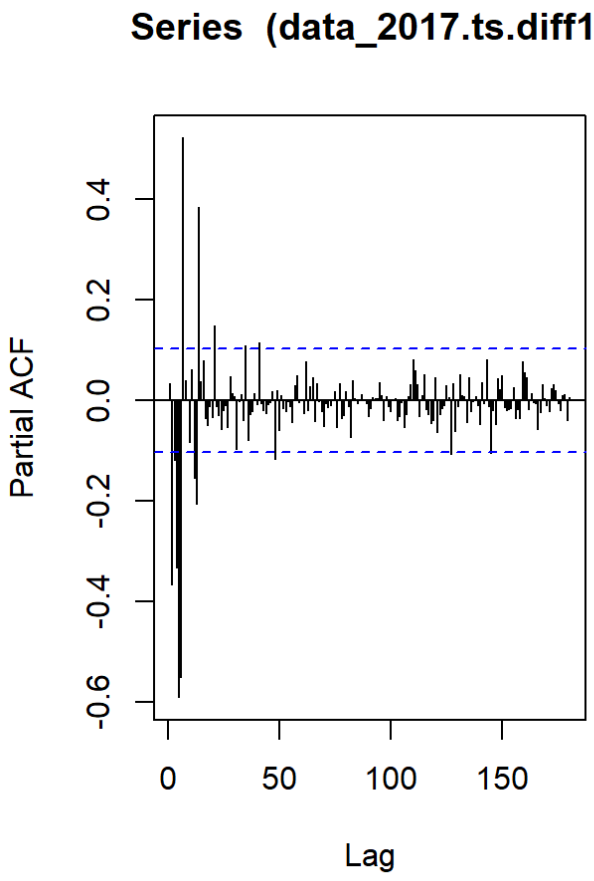
Et le **p-value** ci-dessus nous montre que ce n'est pas une série temporelle stationnaire.

Elimination de la tendance:

on va donc utiliser l'opérateur de différentiation à l'ordre 1 pour essayer d'éliminer la tendance.

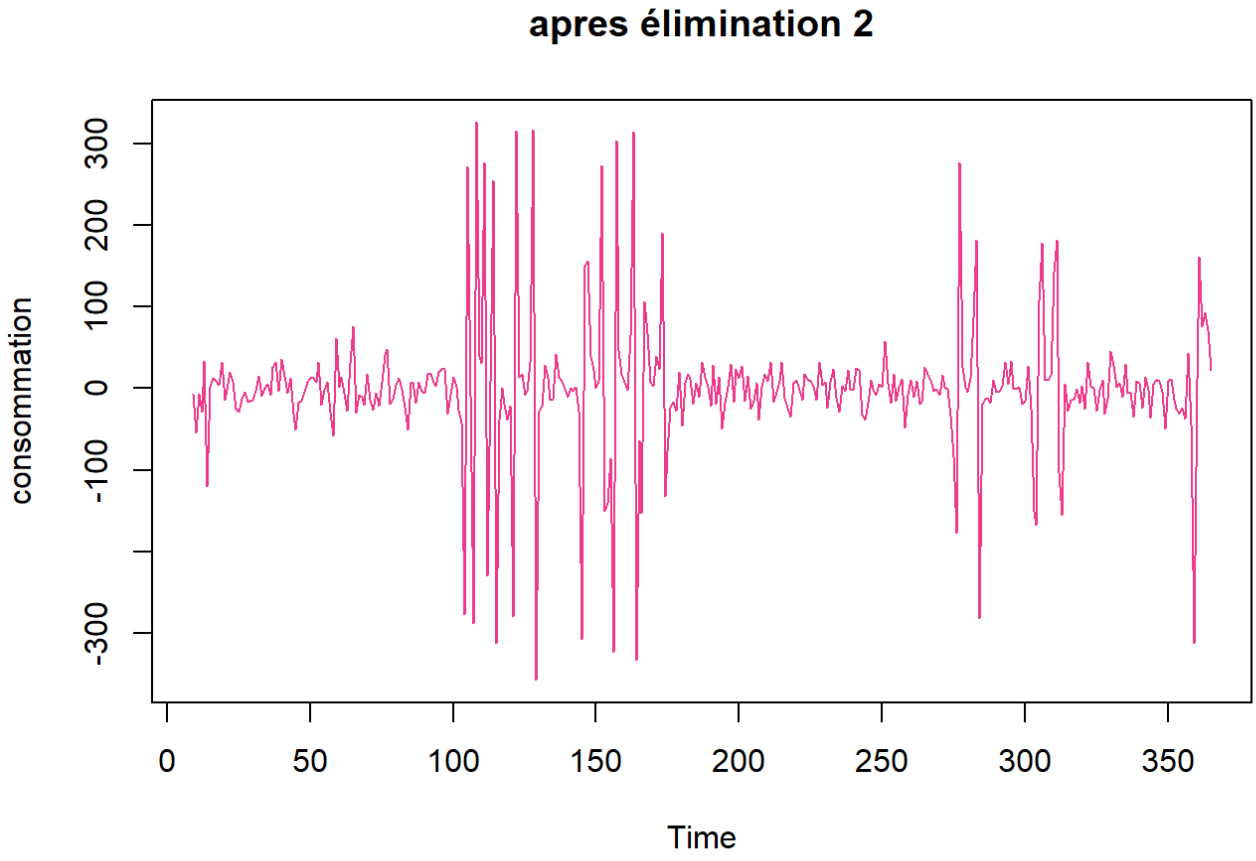


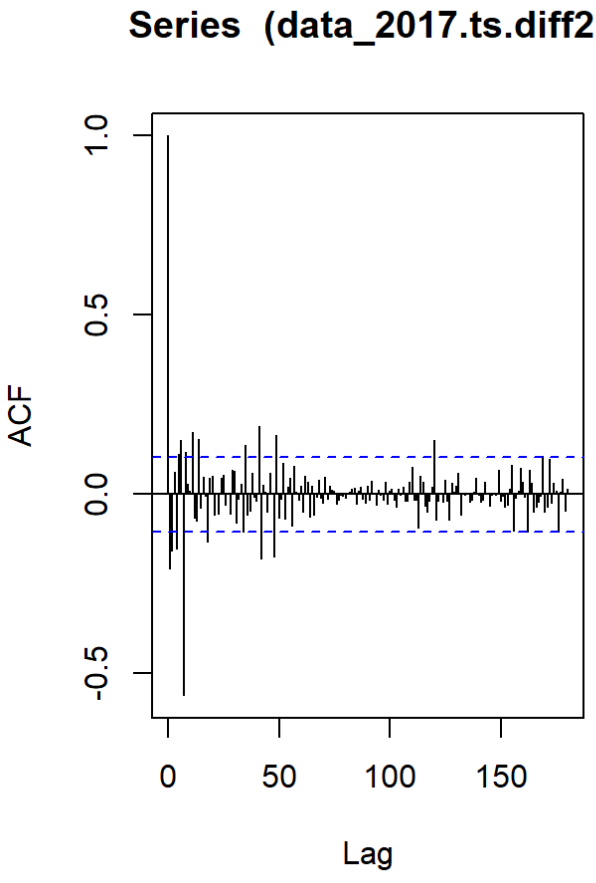


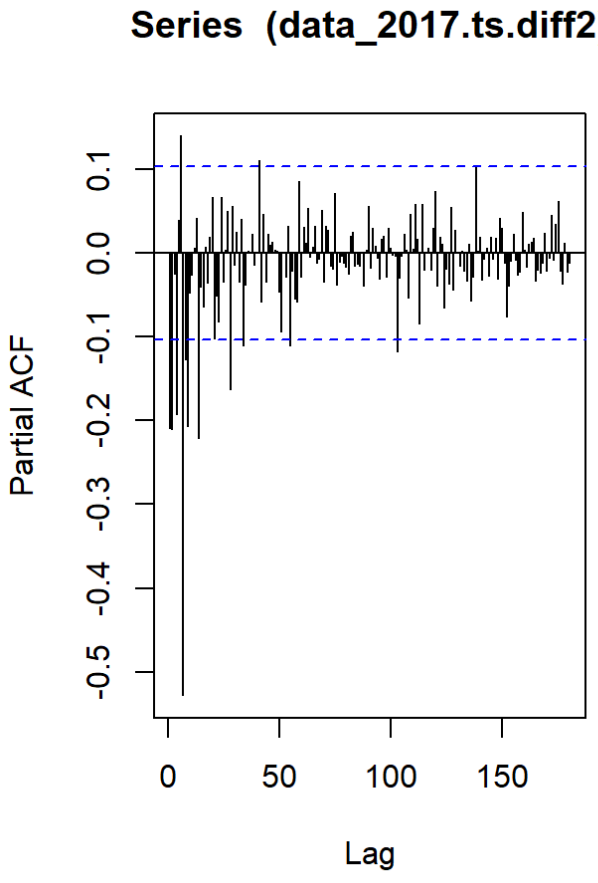


Ce que l'on peut constater que la tendance a été bien éliminée, et on peut constater à partir des graphes **ACF** et **PACF** qu'il y a une période de 7, donc on va essayer de l'éliminer encore par la méthode de la différenciation.

Elimination de la périodicité







On note alors la périocité est disparu, L'**ACF** ressemble bien à celui d'un modèle **ARMA**. En revanche en regardant la série et les pics qui dépasse de L'**ACF**, on pourrait tenter de modéliser notre série par **ARMA**

Dans la suite on essaie de trouver un processus **ARIMA** qui modélise bien nos données. on va utiliser la fonction **auto.ARIMA** pour modéliser le modèle qui nous trouve une meilleure modèle.

Et on peut confirmer par le test de DICKEY-FULLER augmenté comme ci-dessous. Le **p-value** est 0.01, donc notre série temporelle est stationnaire.

TEST de DICKEY-FULLER augmenté

```
## Warning in adf.test(data_2017.ts.diff2): p-value smaller than printed p-value
```

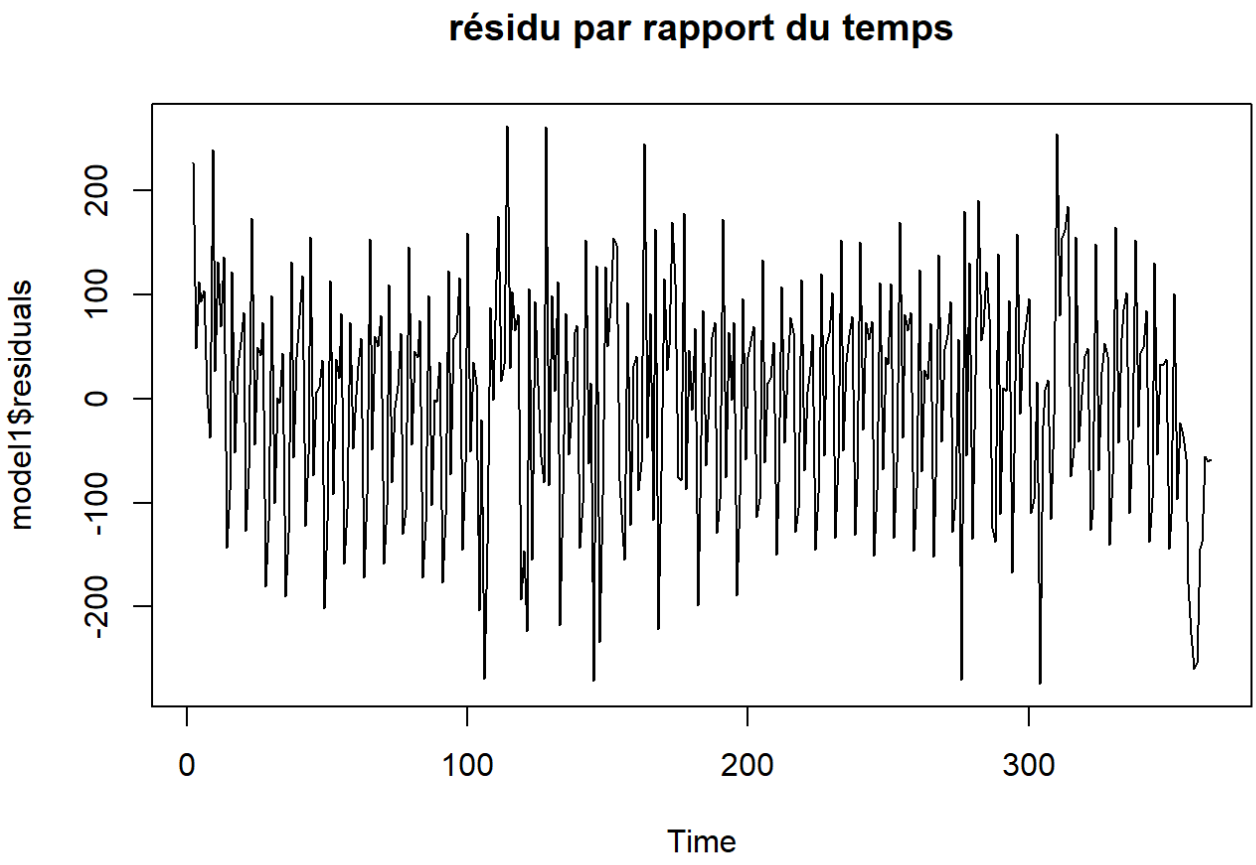
```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_2017.ts.diff2
## Dickey-Fuller = -12.342, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

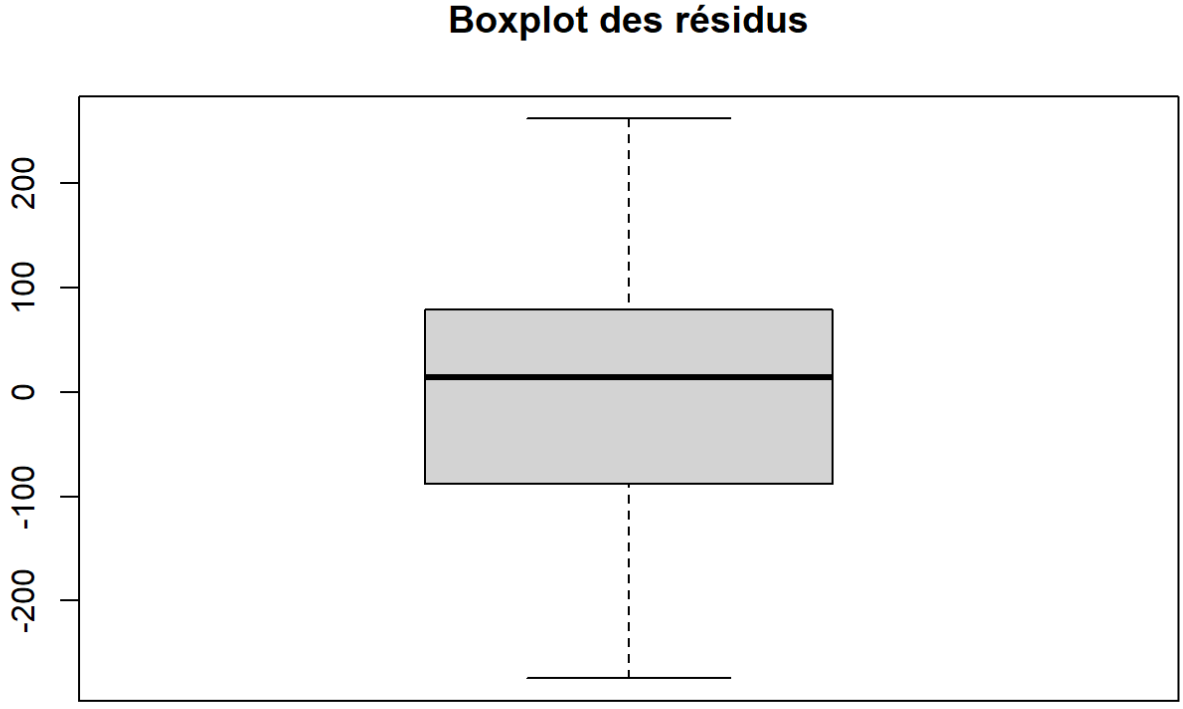
L'**ACF** et le **PACF** des résidus ressemblent à un bruit blanc.

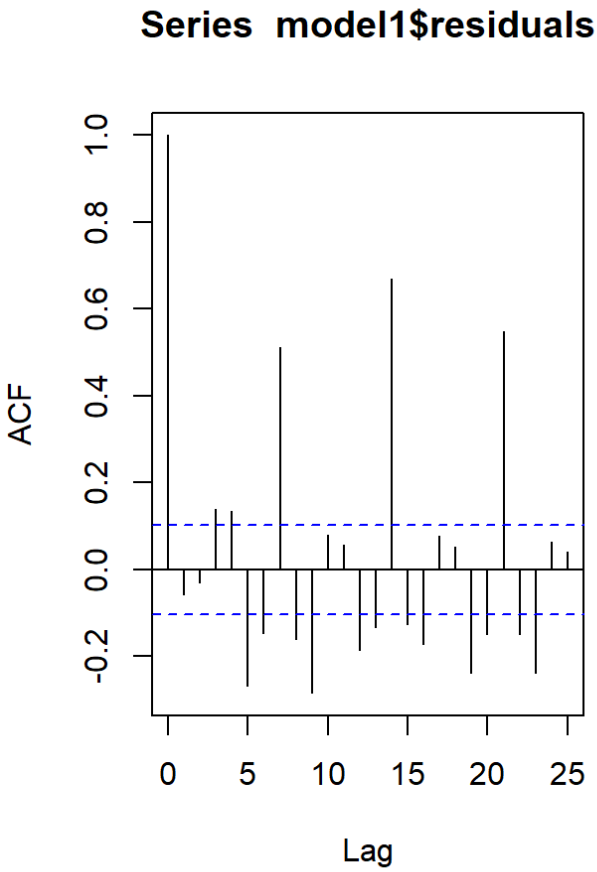
Choix du modèle

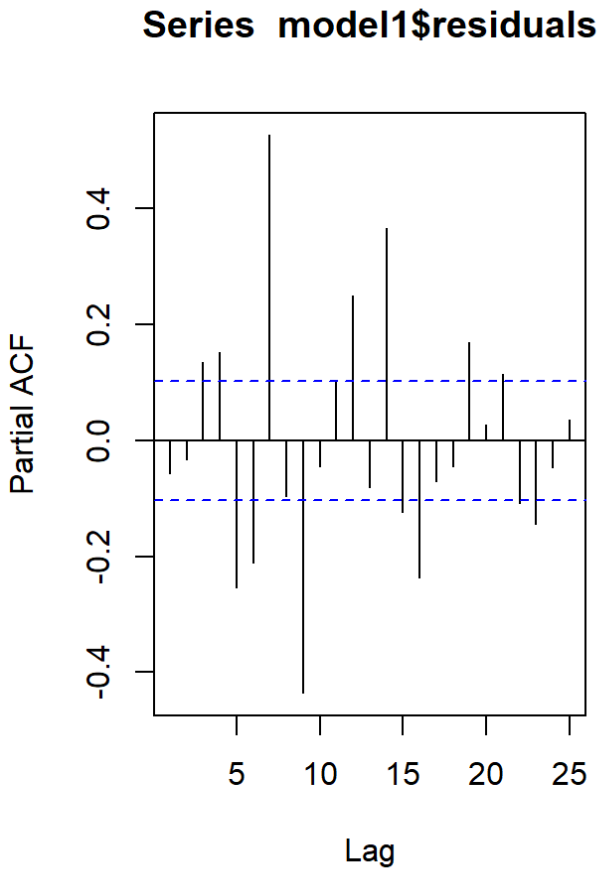
Comme notre série est stationnaire, on peut essayer d'appliquer le modèle ARIMA puis SARIMA parce que c'est une série saisonnière. On va tout d'abord utiliser la fonction auto.arima pour déterminer les paramètres du modèle pour avoir notre premier modèle puis on va ajouter les paramètres saisonnière pour comparer les 2 modèle et voir lequel est le meilleur.

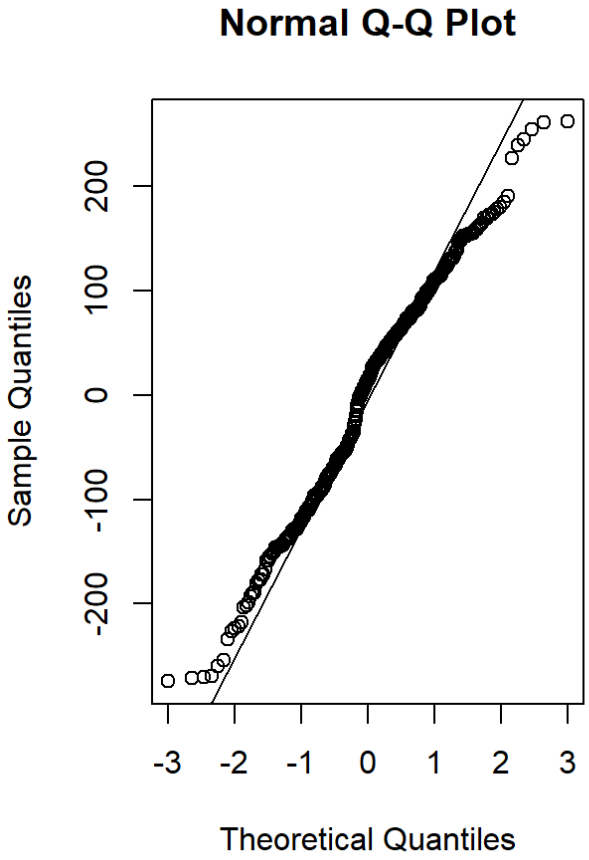
```
## Series: data_2017.ts.diff1
## ARIMA(3,0,3) with zero mean
##
## Coefficients:
##          ar1     ar2      ar3      ma1      ma2     ma3
##       0.3175  0.0990  -0.5025  -0.6418  -0.7125  0.6120
## s.e.  0.0783  0.0754   0.0694   0.0698   0.0623  0.0465
##
## sigma^2 = 12151:  log likelihood = -2226.51
## AIC=4467.02   AICc=4467.34   BIC=4494.3
##
## Training set error measures:
##                      ME     RMSE      MAE      MPE     MAPE      MASE
## Training set -1.767065 109.3206 92.31884 75.80565 437.7483 0.5900168
##                      ACF1
## Training set -0.05684642
```


résidu par rapport du temps
model1$residuals
200
100
0
-100
-200
0
100
200
300
Time


Boxplot des résidus
200
100
0
-100
-200



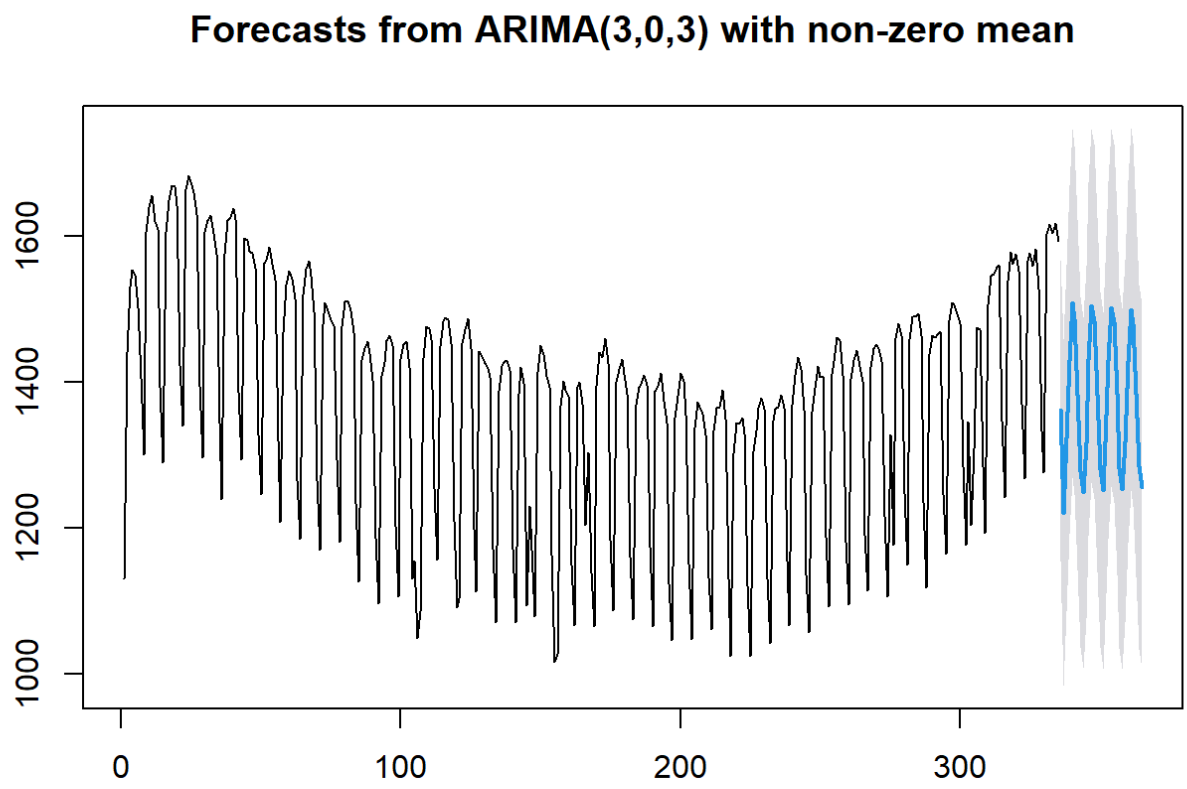




On obtient un modèle ARIMA de paramètre (3,0,3).

Nous pouvons constater que l'ACF et l'PACF nous ressemblent que ce n'est pas un bruit blanc. Le QQ-plot des résidues ne ressemble pas trop à la loi normale, on va tester quand même l'effet de la prédiction. Pour

la prédiction, on fixe les 30 derniers jours pour le tester, les autres données comme les données training.

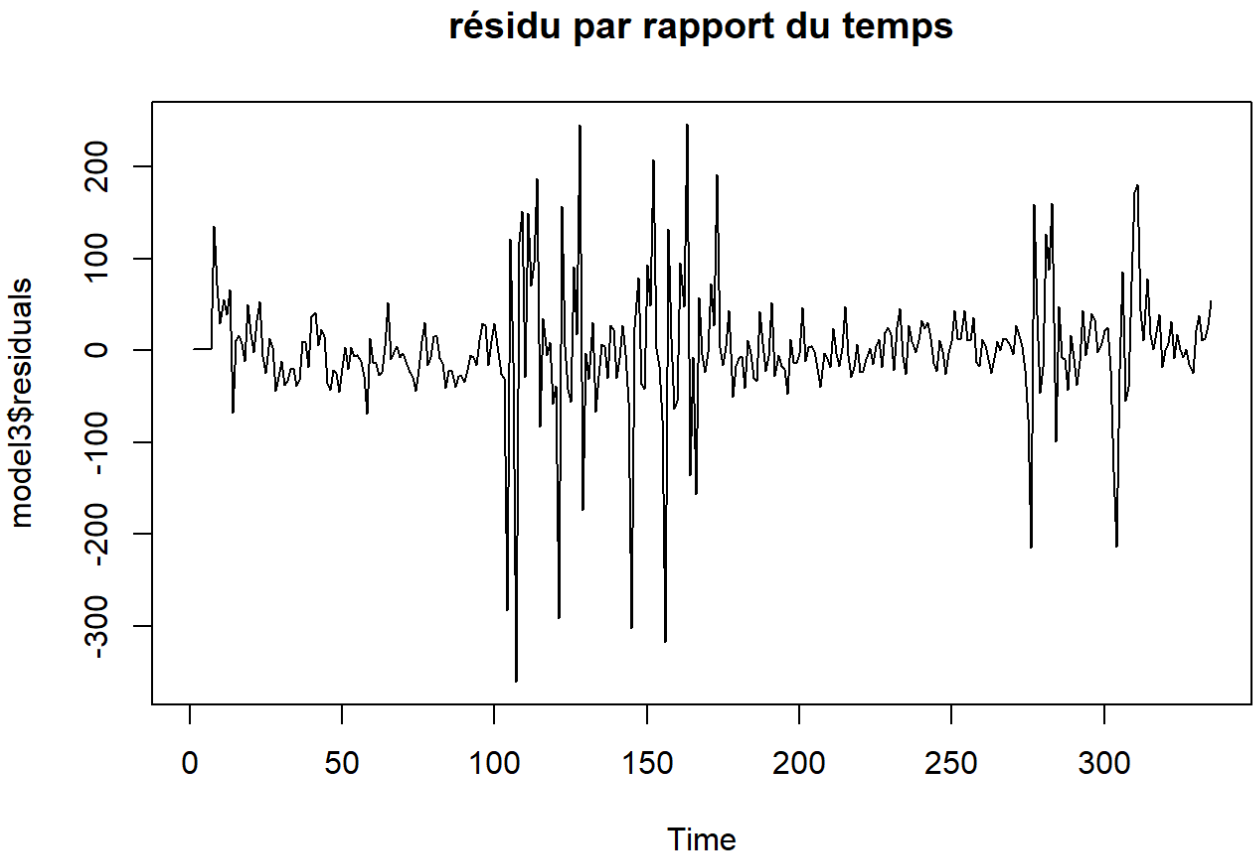


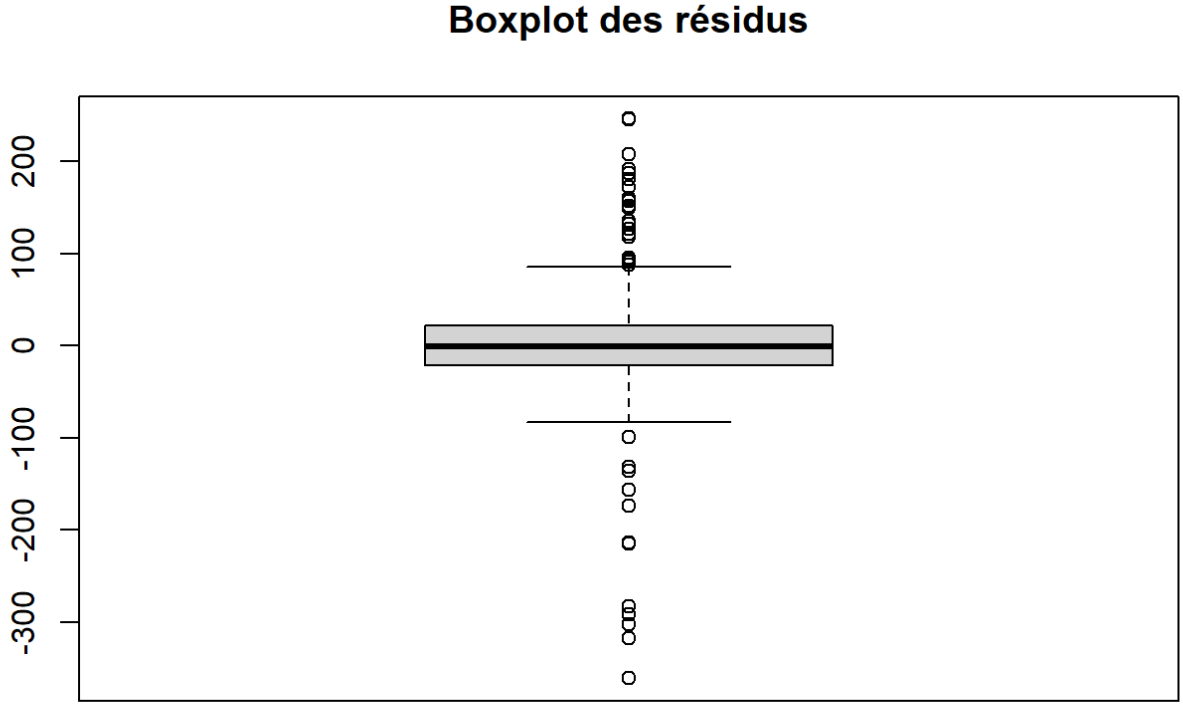
On peut remarquer qu'il reproduit les même pics et il est loin de figure originale, donc on essaie un autre modèle SARIMA.

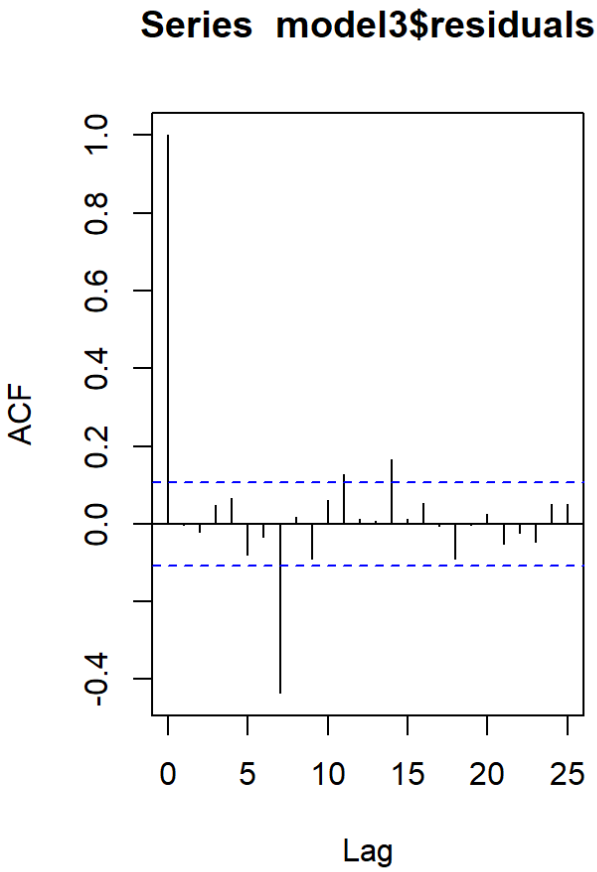
Modèle SARIMA

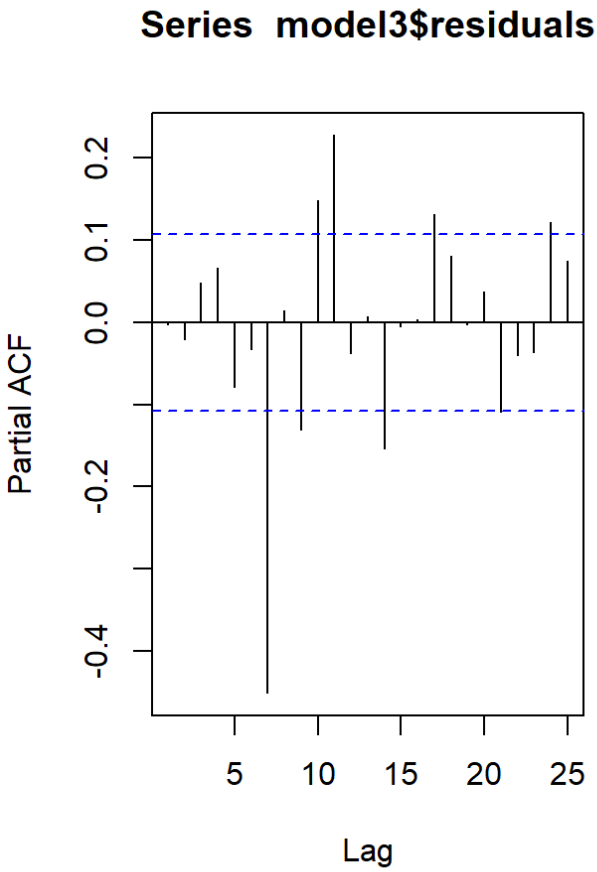
on peut utiliser la fonction **nsdiffs** et **ndiffs** pour déterminer les paramètre saisonière du modèle SARIMA, on obtient SARIMA(3,0,3)(0,1,0)7. Et on refait la même chose.

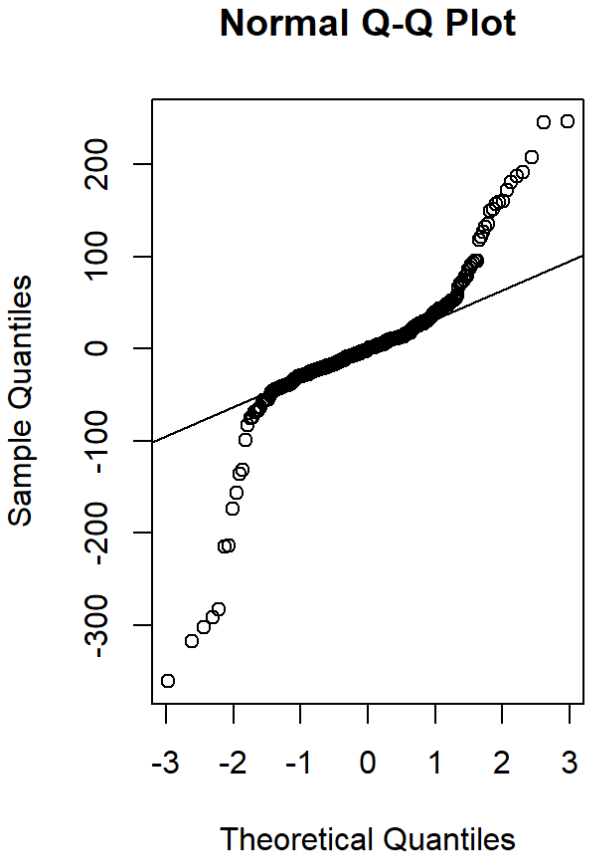
Prédiction


résidu par rapport du temps
model3$residuals
200
100
0
-100
-200
-300
0
50
100
150
200
250
300
Time


Boxplot des résidus
200
100
0
-100
-200
-300

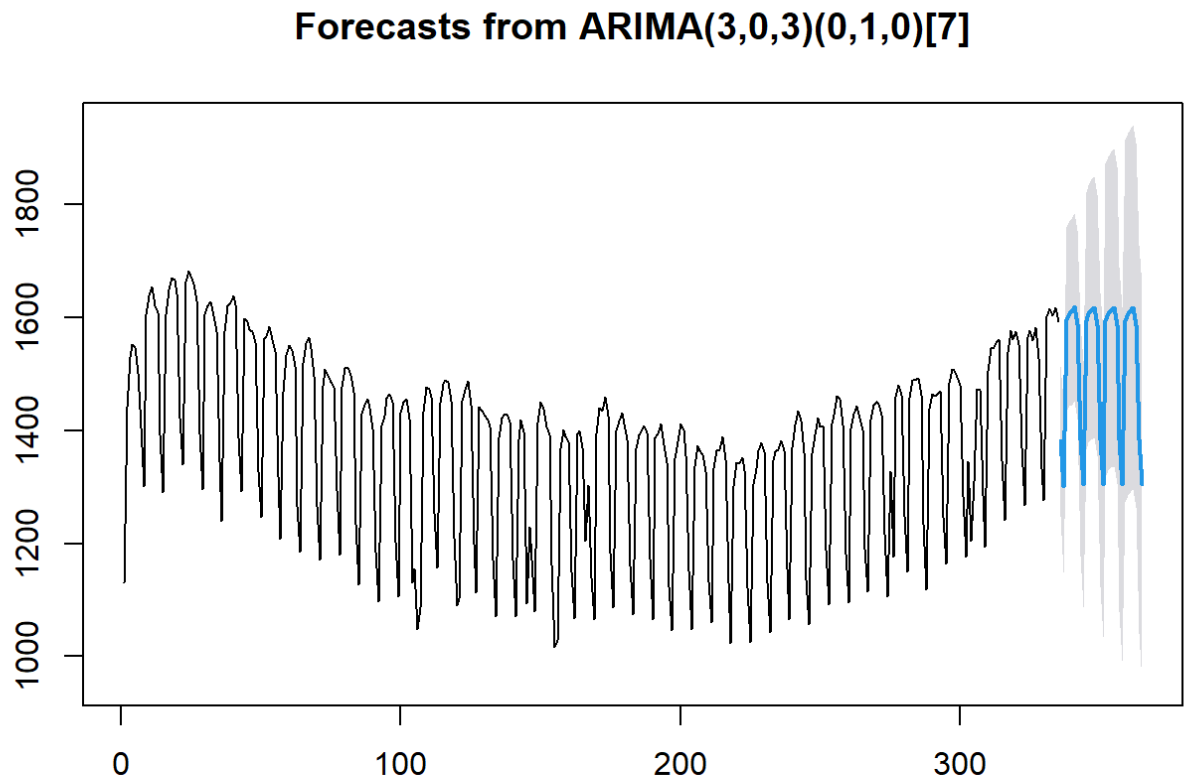






On peut constater depuis les graphes ACF et PACF que c'est mieux que le modèle précédent, et pour le qq-plot, on constate que la plupart du points résidus sont dans la droite. Mais c'est étonnant que dans le

queue et la tête des points sont éloignés. On va effectuer la prédiction pour ce modèle pour voir est-ce qu'il est mieux que le modèle précédent.



On remarque la prédiction est mieux que celui précédent.

Conclusion

Pour la conclusion, nous n'avons pas réussi à modéliser la consommation d'électricité parce que le modèle qu'on a choisit, les résidues ne suivent pas la loi normale, et nous avons essayé 2 modèles, ARIMA et SARIMA, le meilleur entre ces 2 est un peu loin de la réalité.

Peut-être c'est parce que dans cette base des données, le bruit n'est pas assez aléatoire. Ou peut-être on doit essayer d'autre modèle.

Bibliographie

1. https://www.themachinelearners.com/series-temporales-arima/
2. https://www.dataquest.io/blog/tutorial-time-series-analysis-with-pandas/