Projet

Justine LOUARN - Lucie Raimbault - Marion Moussay

4/29/2022

Contents

Introduction **1**

Importation en package 1

Importation des données 2

Première approche 4

Décomposition de la série 8

Elimination de la saisonnalité **9**

Methode avec opérateur différence 9

Fonction *decompose* : moyennes mobiles 13

DETECTION MA(q), AR(p) ? **16**

INTERPRETATION 17

STATISTIQUES DESCRIPTIVES 18

ACF ET PACF 18

Test de BOX PIERCE 22

Introduction

La varicelle est une maladie infantile extrêmenment contagieuse, elle est responsable d'une éruption de boutons. Elle guérit en une dizaine de jours. Dans cette étude nous allons nous intéresser au nombre de cas hebdomadaires de varicelle en Hongrie de 2005 à 2015. Nous avons choisis de nous interesser seulement à la ville de Budapest, capitale et plus grande ville de Hongrie (1 752 286 habitants).

Dans ce projet nous avons pour objectif de déployer les outils vus en cours pour essayer d'ajuster un modèle aux données.

Dans un premier temps nous allons analyser notre série brutes puis nous allons déceller ou non une saisonnalité et une tendance, puis nous étudierons les résidus afin de créer un modèle qui s'ajuste à nos données.

Importation en package

```
library(zoo)
library(xts)
library(ggplot2)
library(dplyr)
library(lubridate)
library(forecast)
library(gridExtra)
```

Importation des données

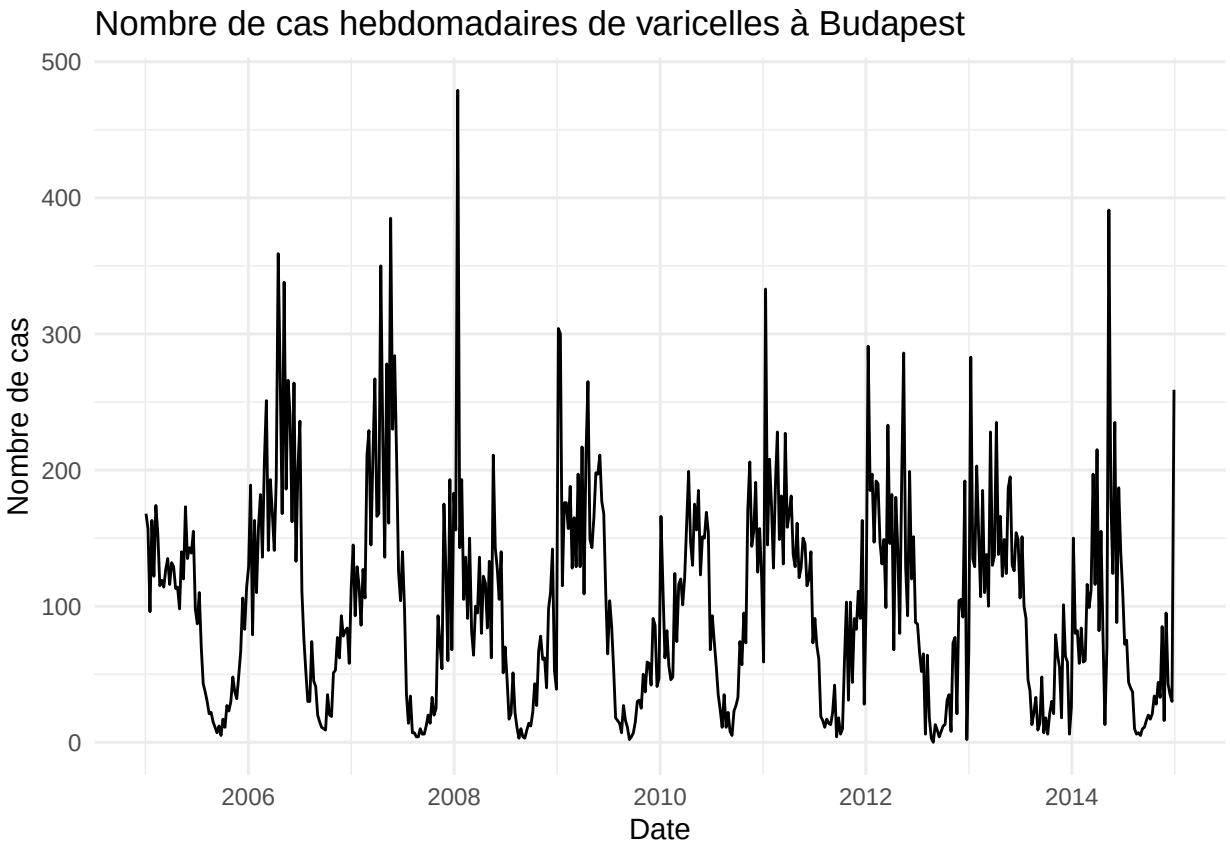
```
data <- read.csv("hungary_chickenpox.csv")
data$Date <- dmy(data$Date)
mean(colMeans(data[-1])) # Moyenne générale de cas de varicelle de 38.84282
```

```
## [1] 38.84282
```

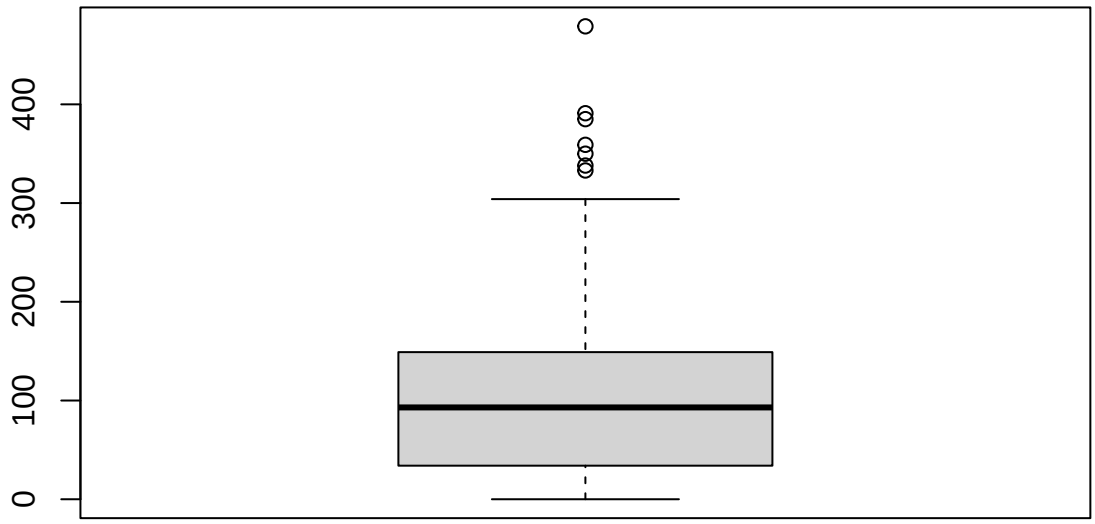
```
colMeans(data[-1]) # Budapest bien supérieure à la moyenne, top 1 (logique car capitale)
```

```
##  BUDAPEST   BARANYA      BACS     BEKES    BORSOD  CSONGRAD     FEJER      GYOR
## 101.24521  34.20498  37.16667  28.91188  57.08238  31.48851  33.27203  41.43678
##     HAJDU     HEVES      JASZ   KOMAROM    NOGRAD      PEST    SOMOGY  SZABOLCS
##  47.09770  29.69157  40.86973  25.64368  21.85057  86.10153  27.60920  29.85441
##     TOLNA       VAS  VESZPREM      ZALA
##  20.35249  22.46743  40.63602  19.87356
```

```
budapest <- data[,1:2]
colnames(budapest) <- c("date", "nb")
budapest %>% ggplot() + aes(x=date, y=nb) + geom_line() + ggtitle("Nombre de cas hebdomadaires de varic
```



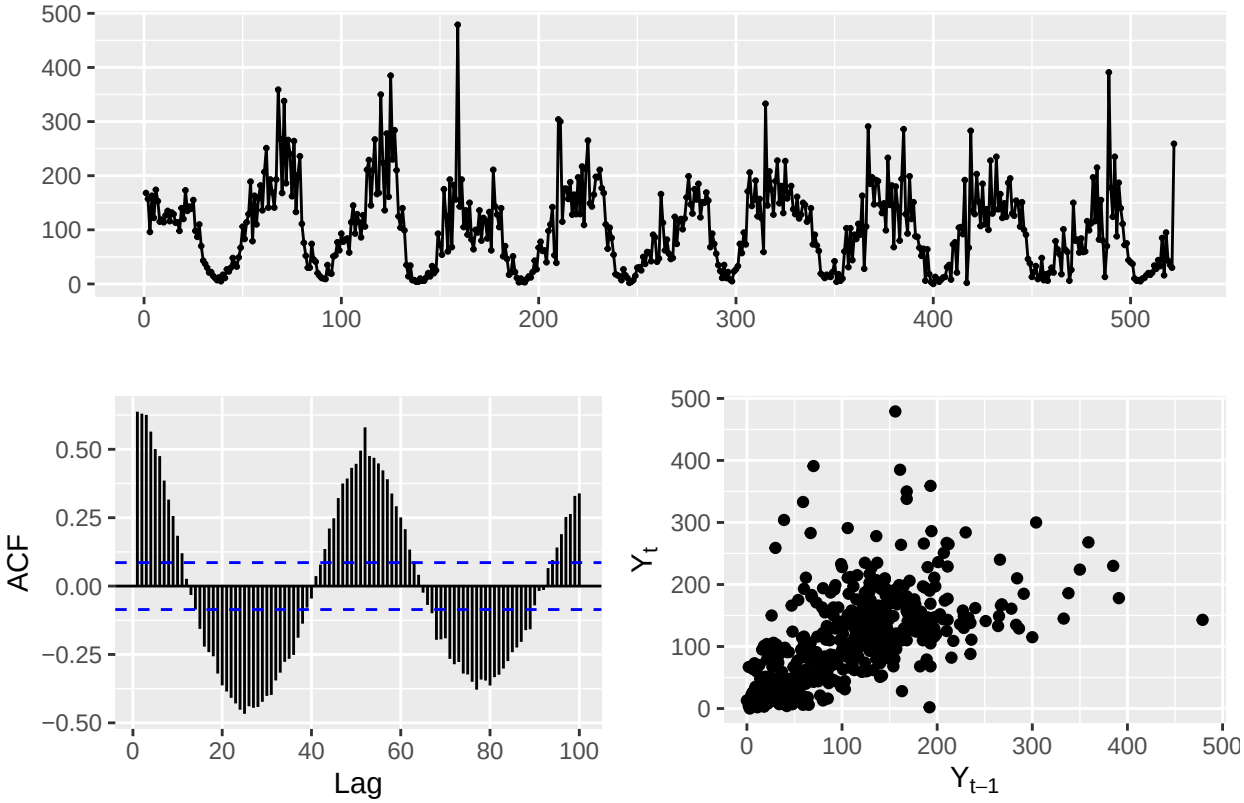

```
boxplot(budapest$nb)
```



Sur le boxplot des données hebdomadaires de varicelles à Budapest, on observe qu'il y a des données hautes, après vérification on ne les considère pas comme abbérantes. Les données étant très propre, nous n'avons pas eu de modifications à faire sur le jeu de données. De cette première représentation, nous remarquons directement une forte saisonnalité d'un an, soit 52 semaines dans notres cas. Nous n'observons pas vraiment de tendance ou alors une légère décroissance mais cela reste difficile à dire avec ce graphique. De plus, on imagine un modèle additif puisque l'on voit une amplitude plutot constante.

Première approche

```
budapest %>% select(nb) %>%
  ggtsdisplay(
    plot.type = "scatter",
    lag.max=100
    )
```

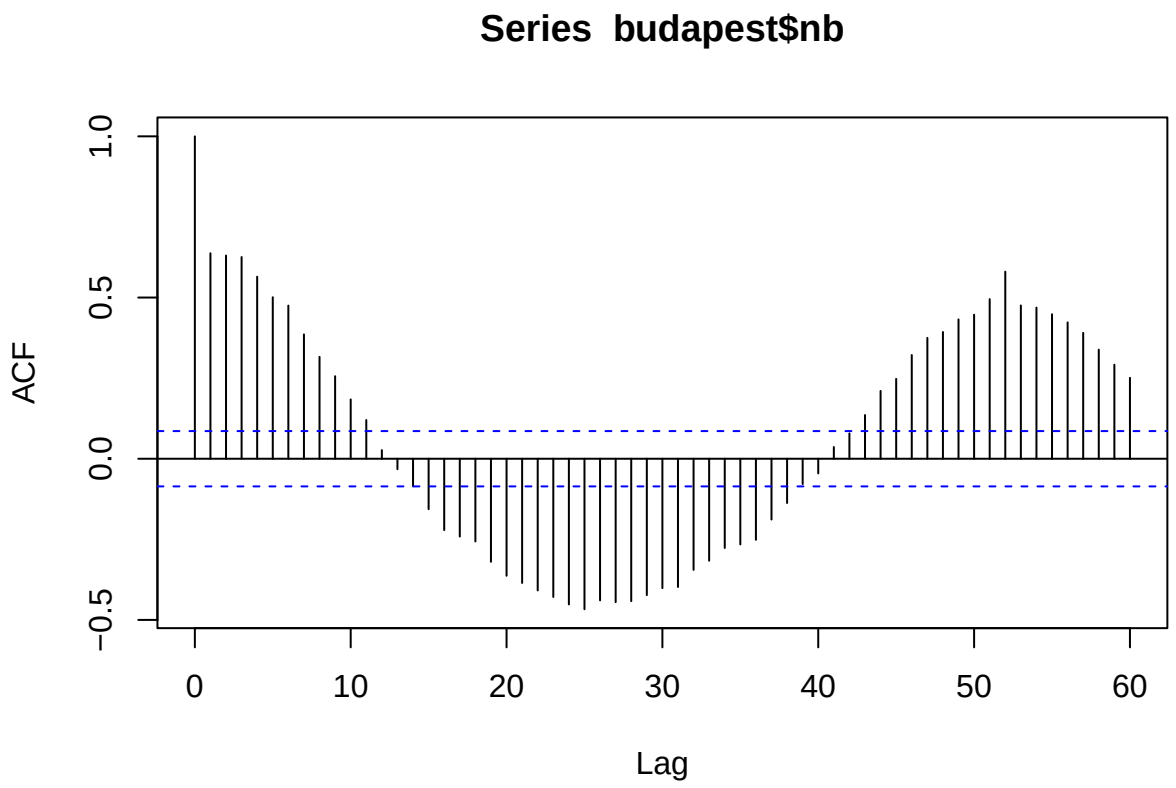


Pour mieux déterminer la saisonnalité nous allons observer la fonction d'autocorrélation. Elle semble périodique, ce qui indique une périodicité dans la série temporelle. La ligne pointillée bleue indique le niveau en-dessous duquel la corrélation n'est plus statistiquement significative.

Le nuage de point permet de visualiser l'auto-corrélation d'ordre 1, soit le quotient des covariances empiriques par la variance empirique. Plus le nuage de points est arrondis plus l'auto-corrélation est proche de 1. Ici on ne distingue rien de "remarquable".

Focus sur l'auto-corrélation :

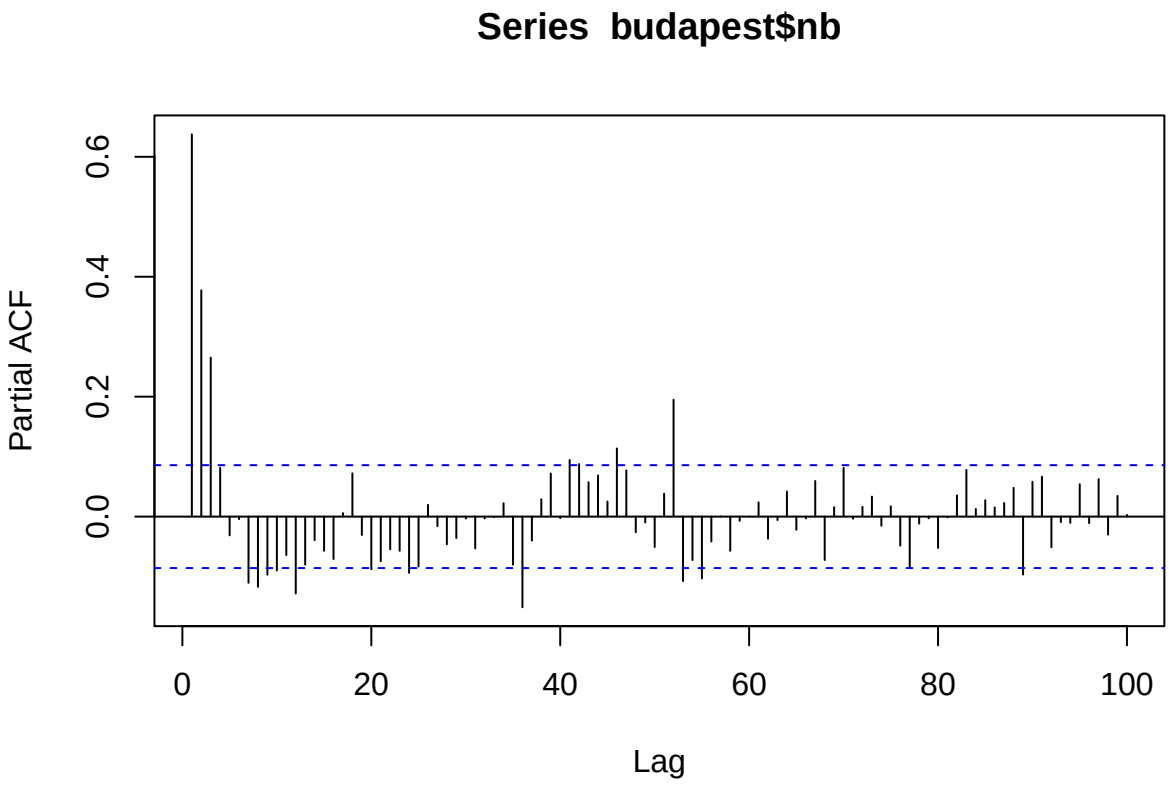
```
acf(budapest$nb, lag.max = 60)
```



On voit ici que la périodicité est de 52 semaines. En effet chaque donnée est espacée de 7 jours. **T= 52** semaines donc environ 12 mois, soit 1 an.

Le corrélogramme indique donc des fortes autocorrélations qui se répètent de manière périodique, vérifions qu'il ne s'agit pas d'un effet résiduel avec la fonction *pacf* qui mesure l'autocorrélation partielle. Elle permet de mesurer l'autocorrélation d'un signal pour un décalage k "indépendamment" des autocorrélations pour les décalages inférieurs.

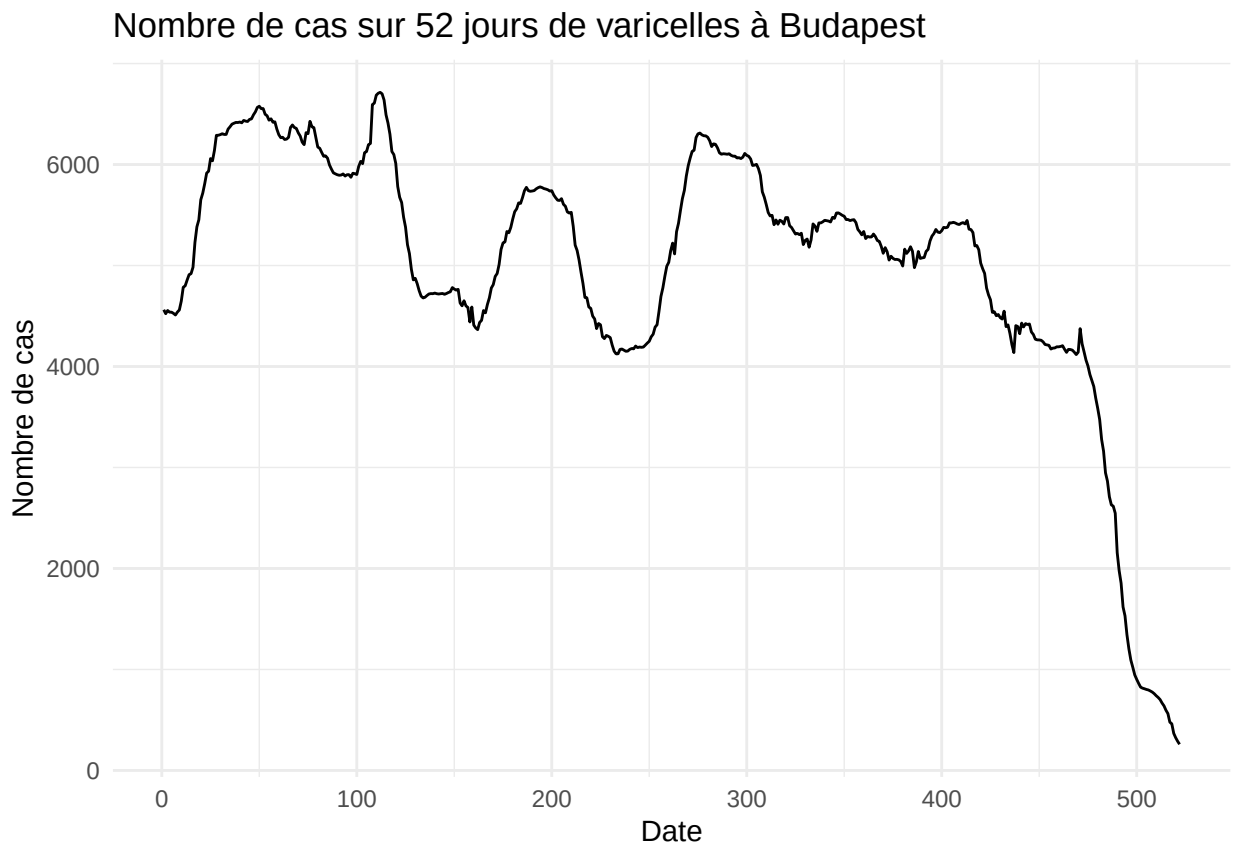
```
pacf(budapest$nb, lag.max = 100)
```



On voit bien que l'auto-corrélation atteint les mêmes valeurs, il n'y a donc pas d'effet résiduel.

On regarde l'évolution tout les 52 jours sur l'année :

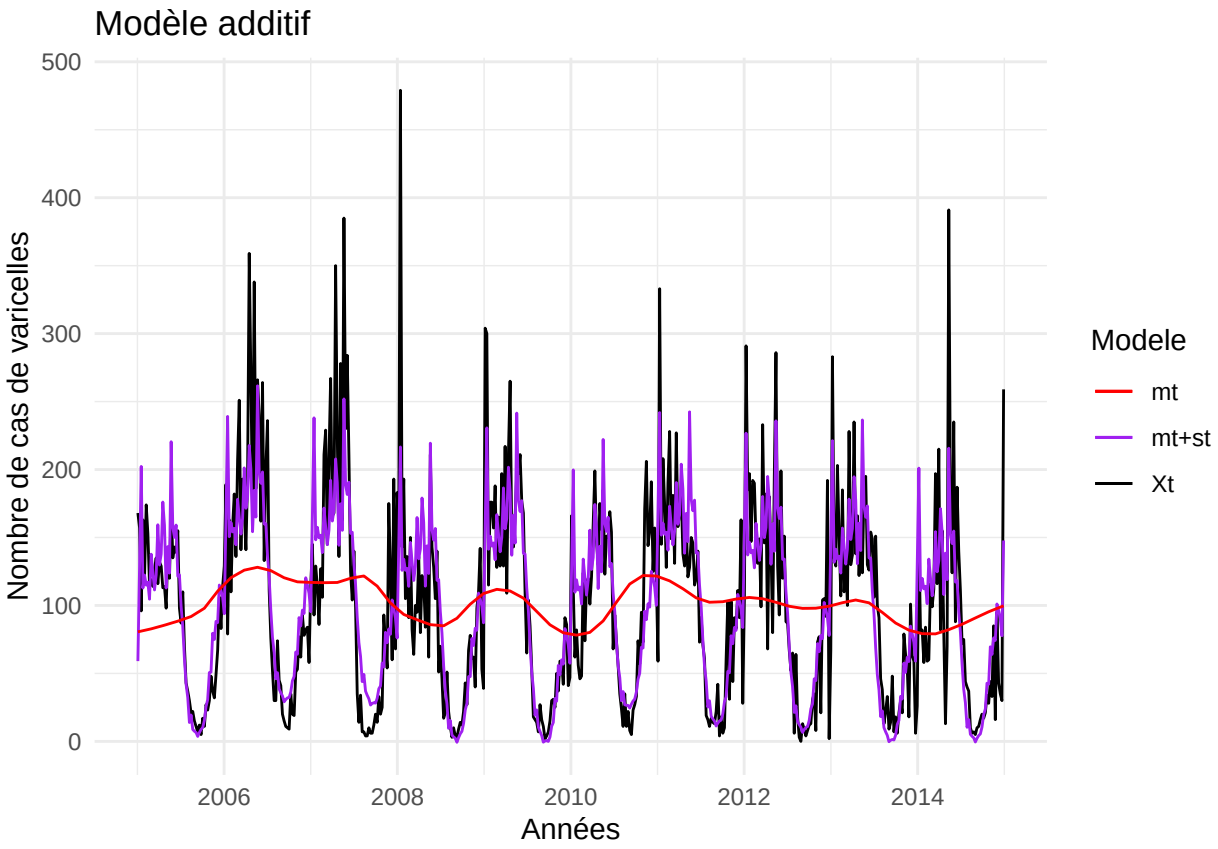
```
v=c()
for (i in 1:522){
  cp<-window(budapest$nb,start=c(i,1),end=c(i,52))
  v=c(v,sum(cp))
}
data.frame(x=1:522, y=v) %>% ggplot() + aes(x=x, y=y) + geom_line() + ggtitle("Nombre de cas sur 52 jou
```



On ne décèle pas de tendance évidente si ce n'est qu'une décroissance vers la fin de la série.

Décomposition de la série

```
temp.ts <- ts(budapest$nb, start=c(2005,1,3), frequency=52)
mod_stl_add <- stl(temp.ts, s.window = "periodic")

budapest_decomp <- cbind(budapest,as.data.frame(mod_stl_add$time.series))

budapest_decomp %>% ggplot() +
  geom_line(aes(x = date, y=nb, color="Xt")) +
  geom_line(aes(x=date, y=trend+seasonal, color="mt+st")) + geom_line(aes(x=date, y=trend, color="mt")) +
  scale_color_manual(values = c("red", "purple", "black")) +
  theme(legend.position = c(0.8, 0.08), legend.direction = "horizontal") +
  labs(colour = "Modele") + ggtitle("Modèle additif") +
  xlab("Années") + ylab("Nombre de cas de varicelles") + theme_minimal()
```



Elimination de la saisonnalité

Methode avec opérateur différence

```
T=52

x1= diff(diff(budapest$nb, lag = T), differences = 1)
x2=diff(diff(budapest$nb, lag = T), differences = 2)
x3=diff(diff(budapest$nb, lag = T), differences = 3)

print(mean(x1))
```

```
## [1] 0.5799574
```

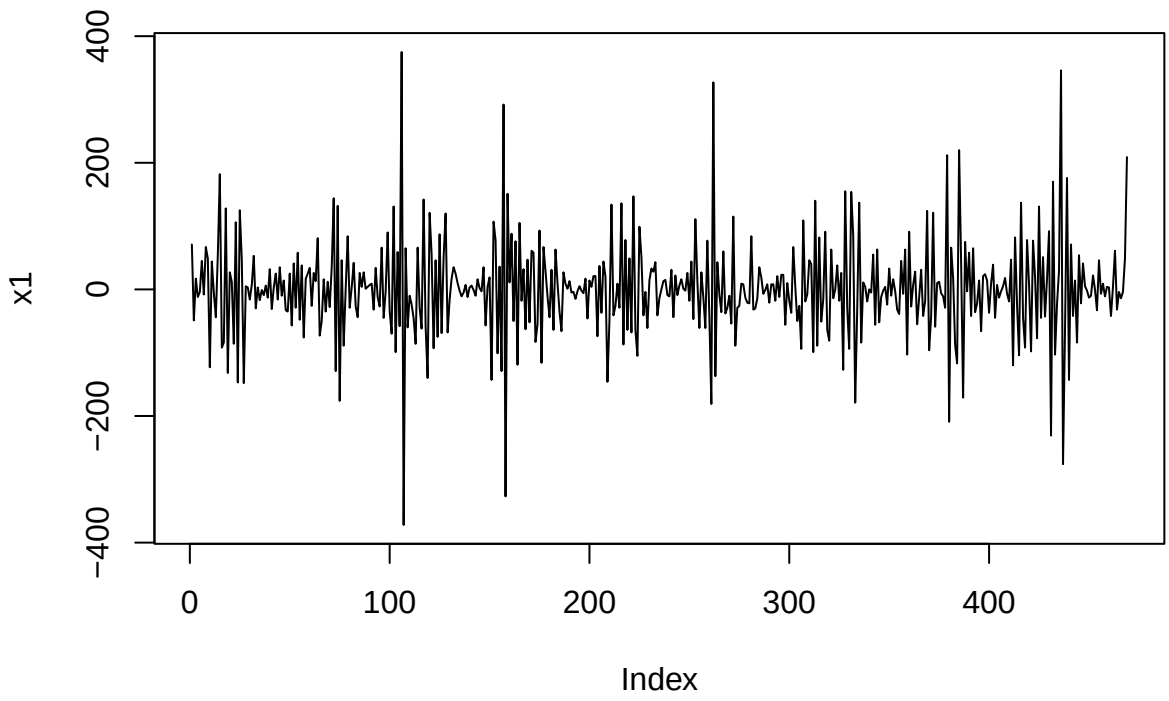
```
print(mean(x2))
```

```
## [1] 0.2948718
```

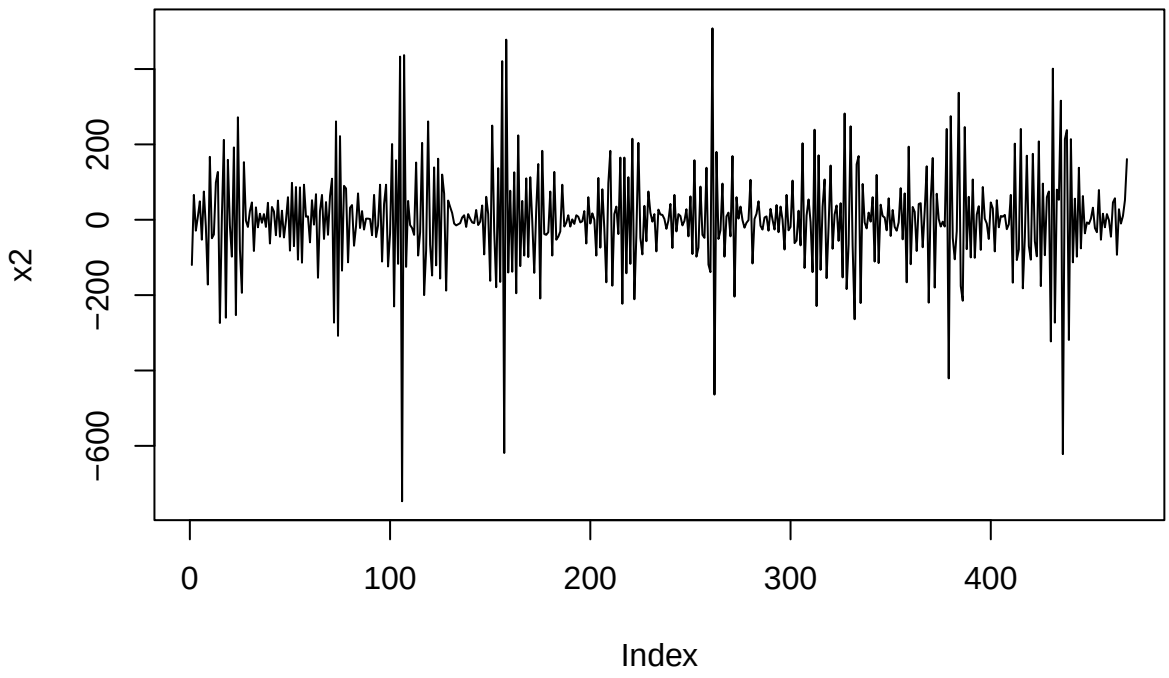
```
print(mean(x3))
```

```
## [1] 0.6017131
```

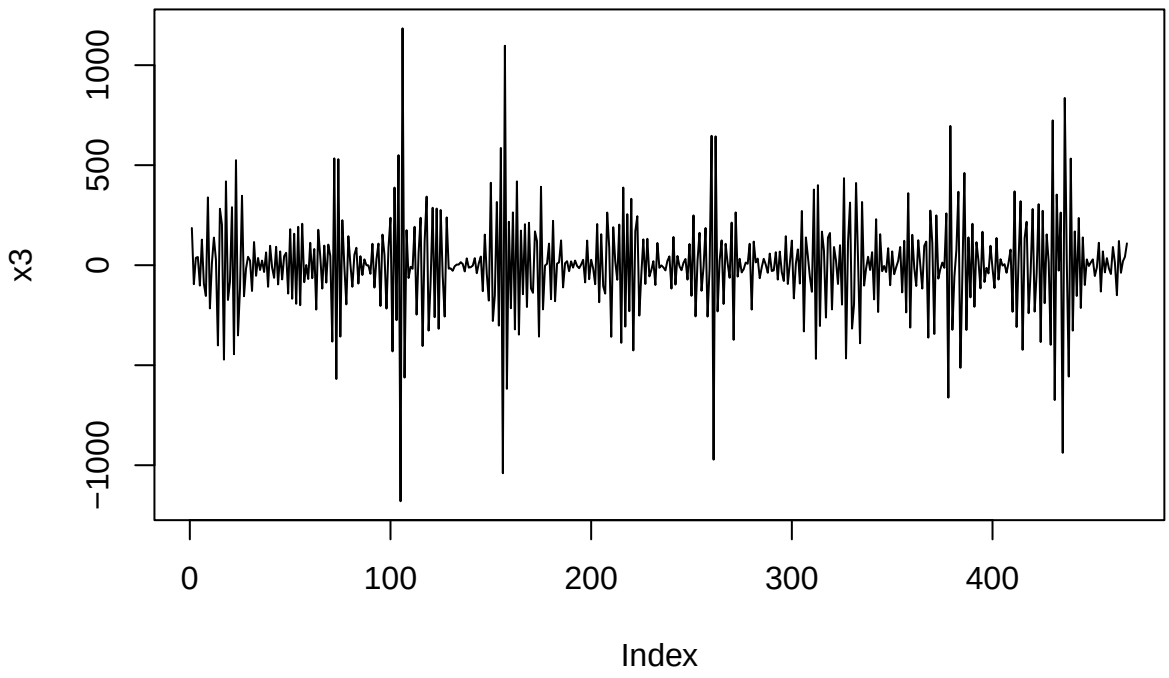
```r
# mean(x2) et mean(x3) sont du même ordre donc on en déduit que x2 est de moyenne nulle, donc la tendan
plot(x1, type='l')
```




```r
plot(x2, type='l')
```



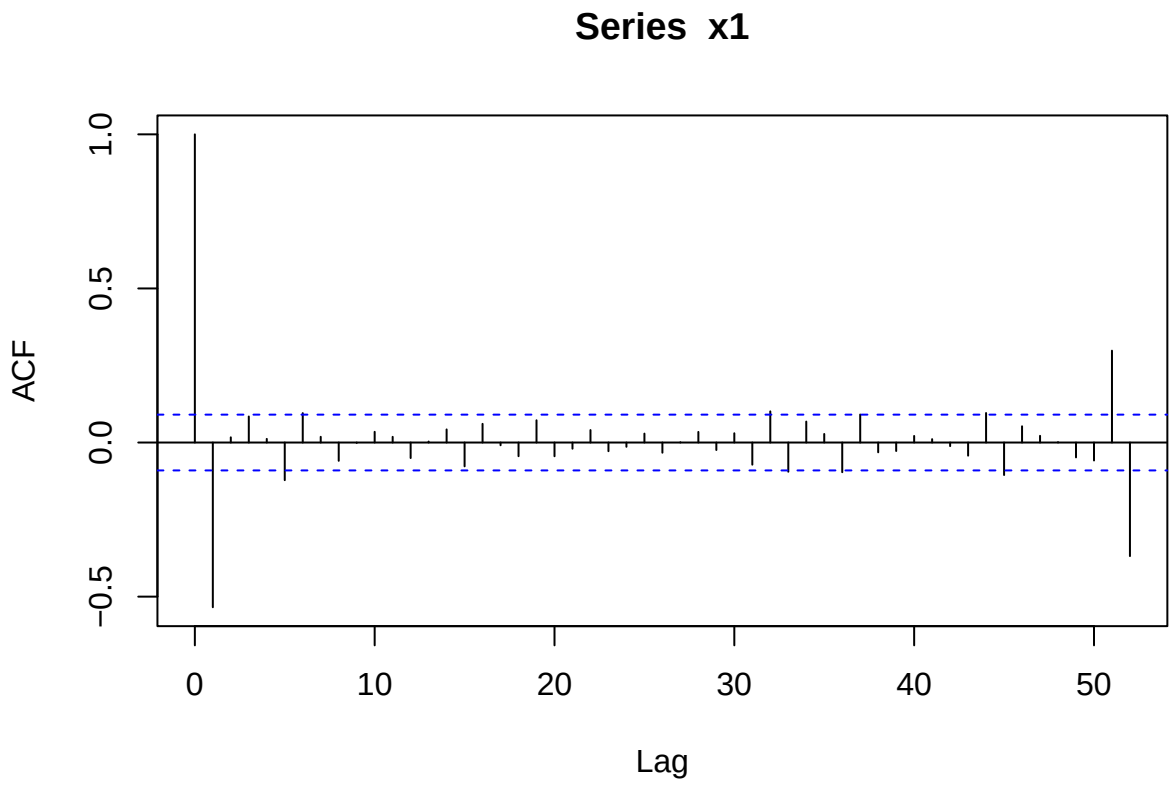

```
plot(x3, type='l')
```




```
Box.test(x1,lag=52)
```

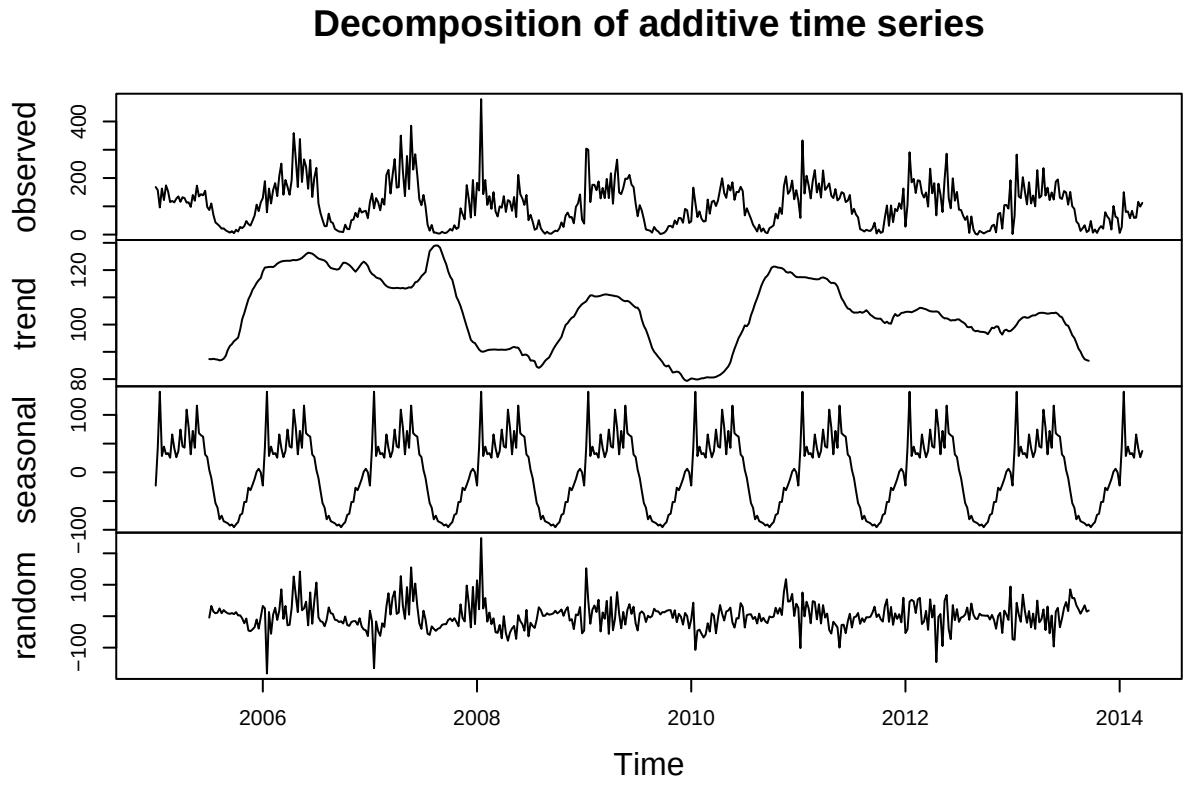
```
##
##  Box-Pierce test
##
## data:  x1
## X-squared = 309.13, df = 52, p-value < 2.2e-16
```

```
#On voit que la valeur de la statistique est en dessous du seuil (0,05) donc on rejette l'hypothèse "br
acf(x1,lag.max=52,type=c("correlation"))
```

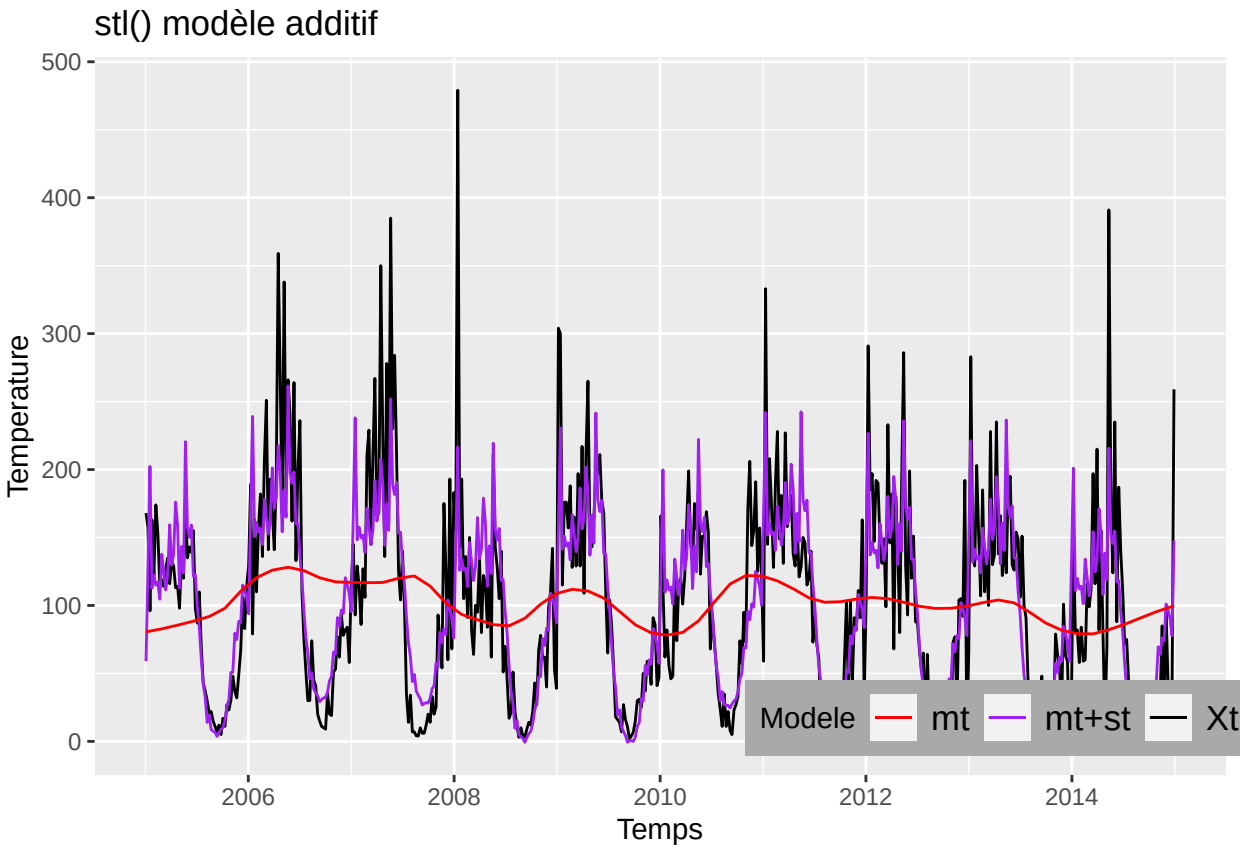


Fonction *decompose* : moyennes mobiles

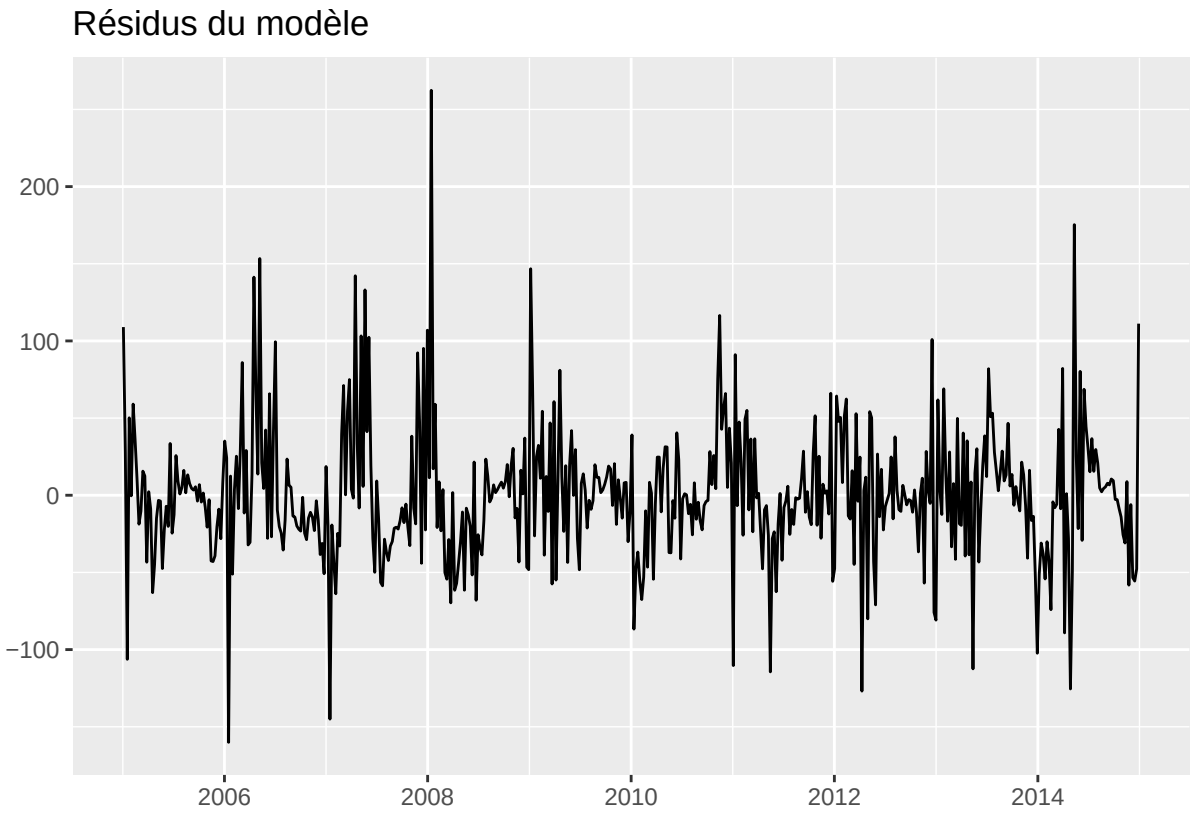
```
budapest.ts <- ts(budapest$nb, start=c(2005,1,3),end =c(2014,12,29),  frequency=52)
budapest_dec <- decompose(budapest.ts, type = "additive")

plot(budapest_dec)
```




```
temp.ts <- ts(budapest$nb, start=c(2005,03,01), frequency=52)
mod_stl_add <- stl(temp.ts, s.window = "periodic")

donnee <- cbind(budapest,as.data.frame(mod_stl_add$time.series))

donnee %>% ggplot() +
  geom_line(aes(x = date, y=nb, color="Xt")) +
  geom_line(aes(x=date, y=trend+seasonal, color="mt+st")) + geom_line(aes(x=date, y=trend, color="mt")) +
  scale_color_manual(values = c("red", "purple", "black")) +
  theme(legend.position = c(0.8, 0.08), legend.direction = "horizontal") +
  labs(colour = "Modele") + ggtitle("stl() modèle additif") +
  xlab("Temps") + ylab("Temperature") +
  theme(
    legend.background = element_rect(fill = "darkgray"),
    legend.text = element_text(size = 13) #+ theme_economist()
  )
```



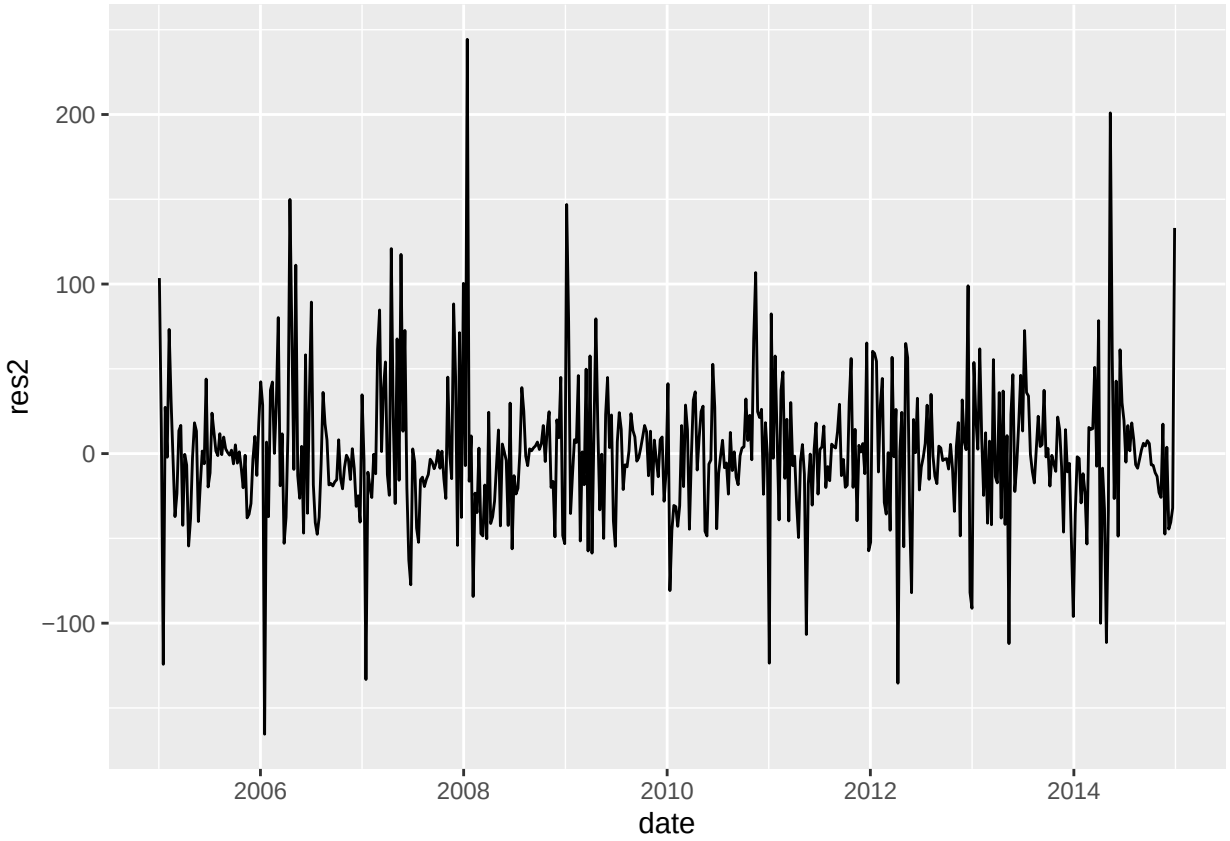

```
donnee %>% ggplot() + aes(x=date, remainder) + geom_line() +  xlab("") + ylab("") + ggtitle("Résidus du
```



DETECTION MA(q), AR(p) ?

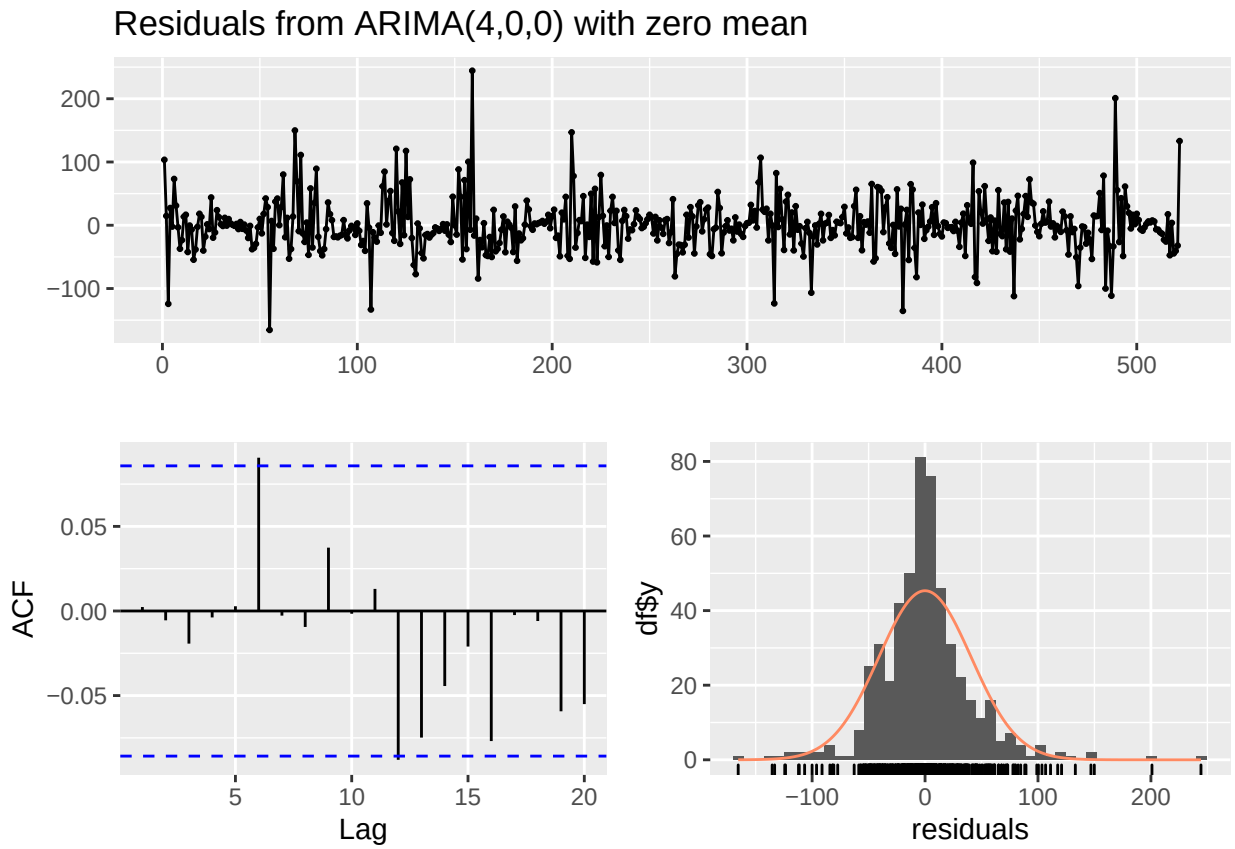
```
arima <- auto.arima(donnee$remainder)

data <- donnee %>% mutate(res2 = arima$residuals)

data %>% ggplot() + aes(x=date,y=res2) + geom_line()
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



INTERPRETATION

```
checkresiduals(arima, lag.max=20)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(4,0,0) with zero mean
## Q* = 5.3789, df = 6, p-value = 0.4962
##
## Model df: 4.   Total lags used: 10
```

STATISTIQUES DESCRIPTIVES

```
mean(data$res2)
```

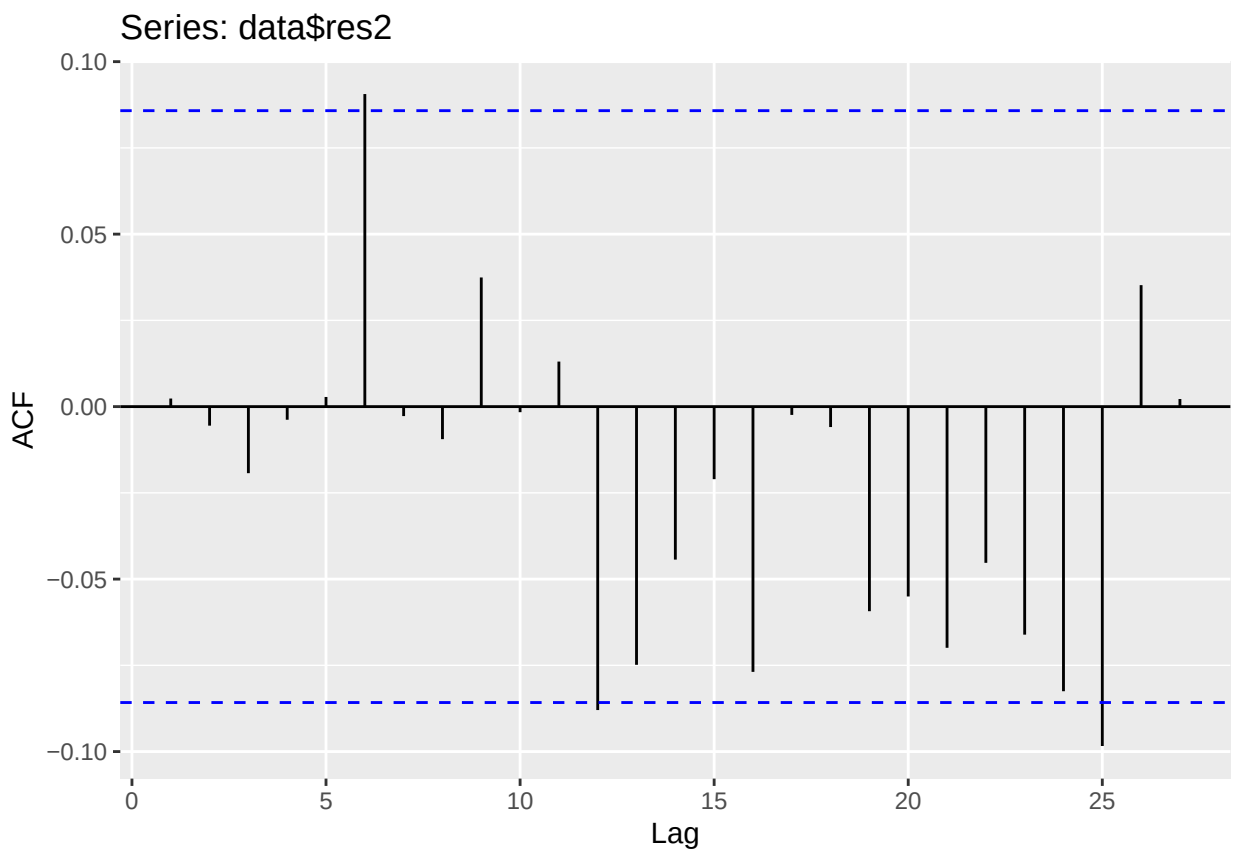
```
## [1] -0.03906448
```

```
var(data$res2)
```

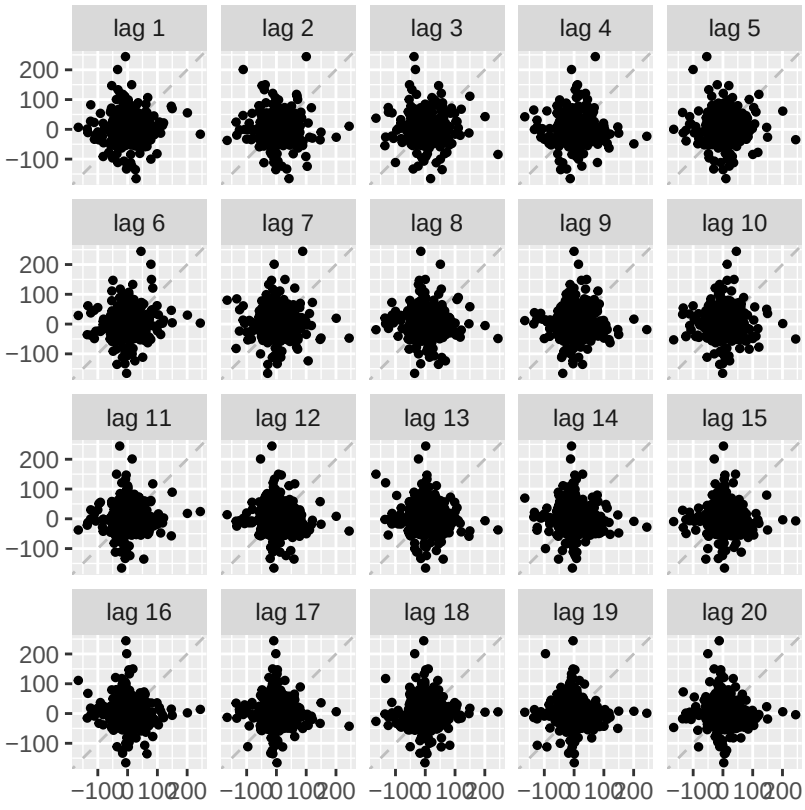
```
## [1] 1675.017
```

ACF ET PACF

```
ggAcf(data$res2)
```

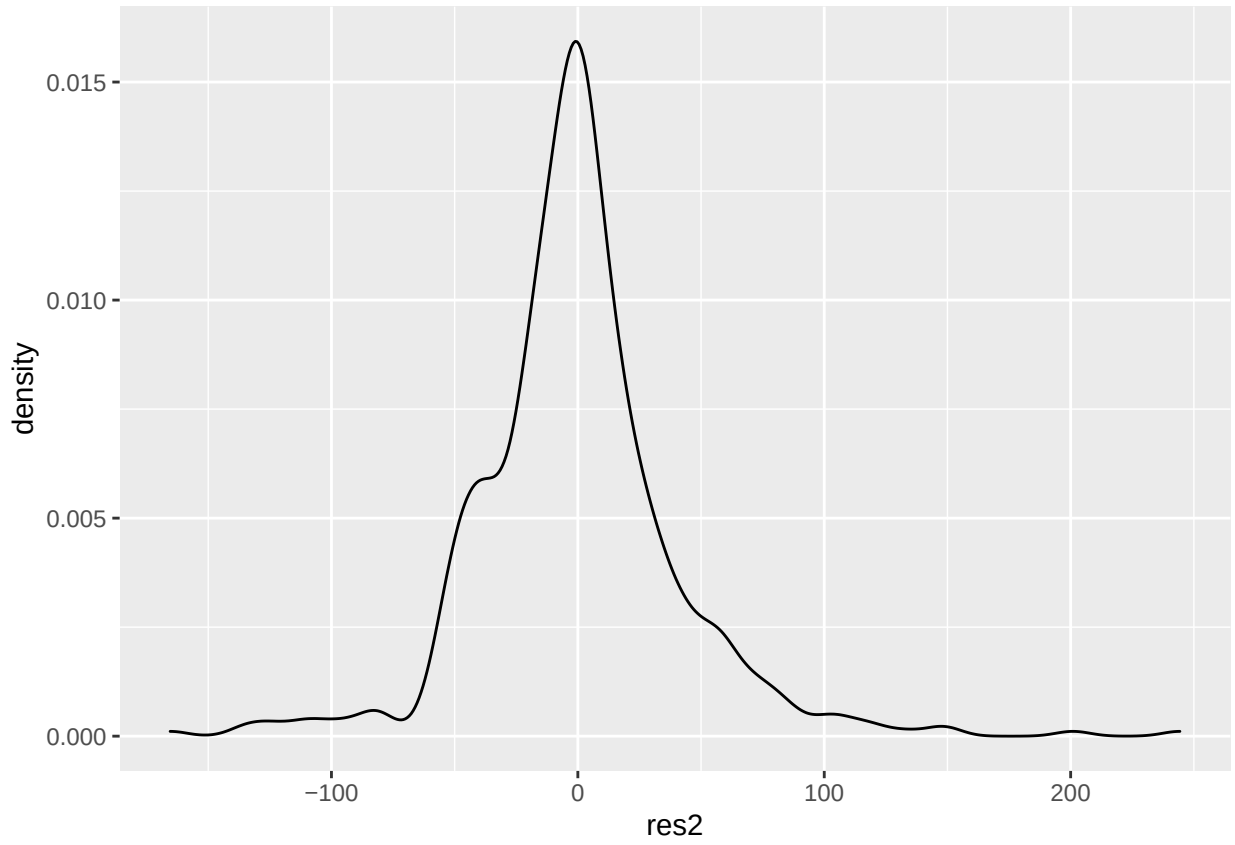



```
gglagplot(data$res2, do.lines = FALSE, set.lags = 1:20, colour = FALSE)
```

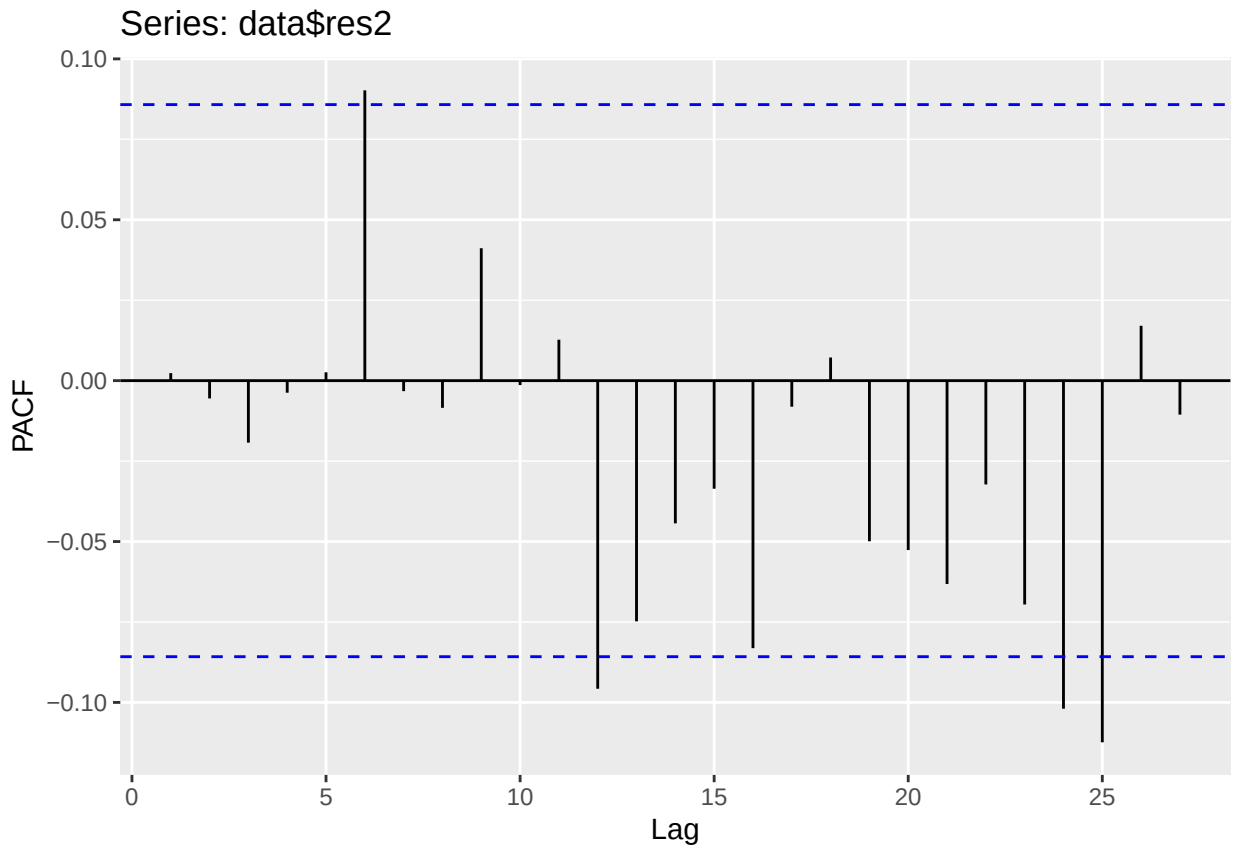



```
data %>% ggplot() + aes(x=res2) + geom_density()
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```




```
ggPacf(data$res2)
```



On remarque que l'ACF et la PACF sont similaires, ce qui reflète une absence d'effet résiduel.

Test de BOX PIERCE

```
Box.test(data$res2, lag = 20, type = "Box-Pierce", fitdf = 2)
```

```
##
##  Box-Pierce test
##
## data:  data$res2
## X-squared = 20.13, df = 18, p-value = 0.3256
```

```
Box.test(data$res2, lag = 20, type = "Ljung-Box", fitdf = 2)
```

```
##
##  Box-Ljung test
##
## data:  data$res2
## X-squared = 20.711, df = 18, p-value = 0.2942
```

Les 2 test de Box-Pierce et de Box-Ljung, renvoie une p-value > 5 %, alors on ne rejette pas l'hypothèse HO. Les résidus sont décoréllés.