Projet

Justine LOUARN - Lucie Raimbault - Marion Moussay

4/29/2022

Contents

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 des données

```
data <- read.csv("hungary_chickenpox.csv")
data$Date <- dmy(data$Date)
mean(colMeans(data[-1]))
```

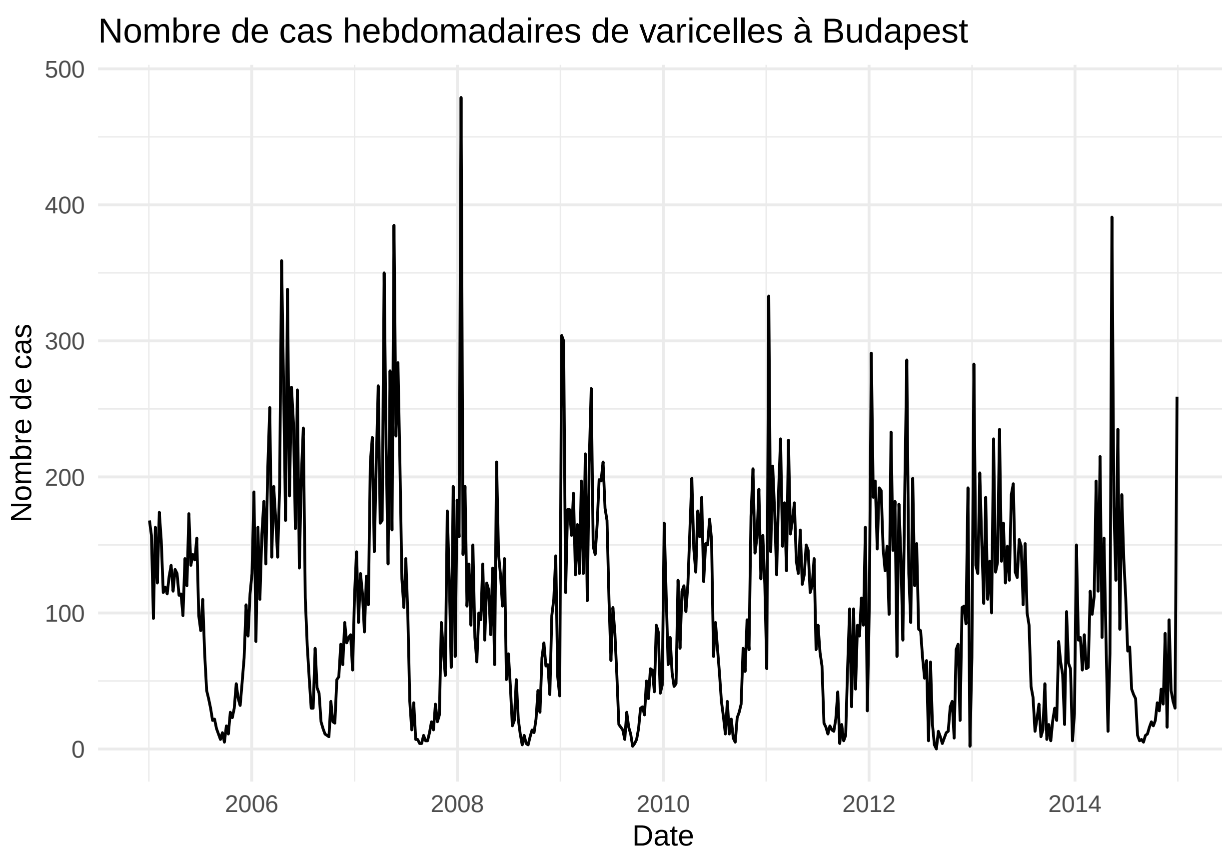
```
## [1] 38.84282
```

```
colMeans(data[-1])
```

```
##  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
```

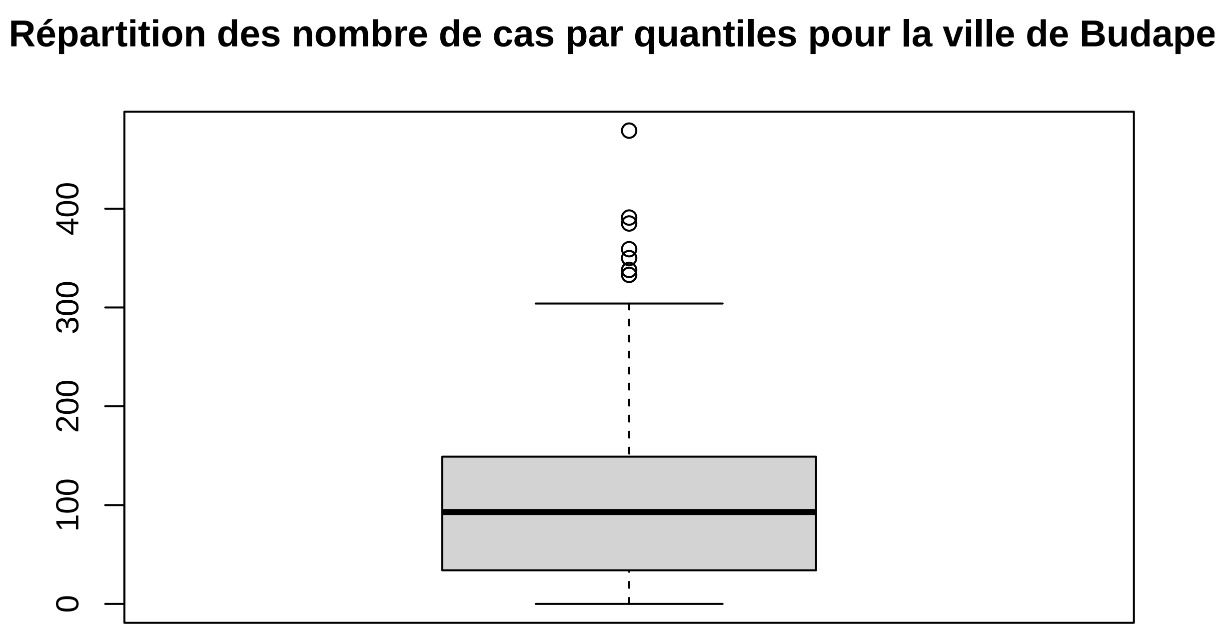
La moyenne générale des nombres cas de varicelles est de 38.84 de toutes les villes du jeu de données. La moyenne de Budapest est trois fois supérieure à la moyenne générale, ce qui semble tout à fait logique puisque c'est juste une question de proportion d'habitants par ville.

```
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
```



De cette première représentation, nous remarquons directement une forte saisonnalité d'un an, soit 52 semaines dans notre 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.

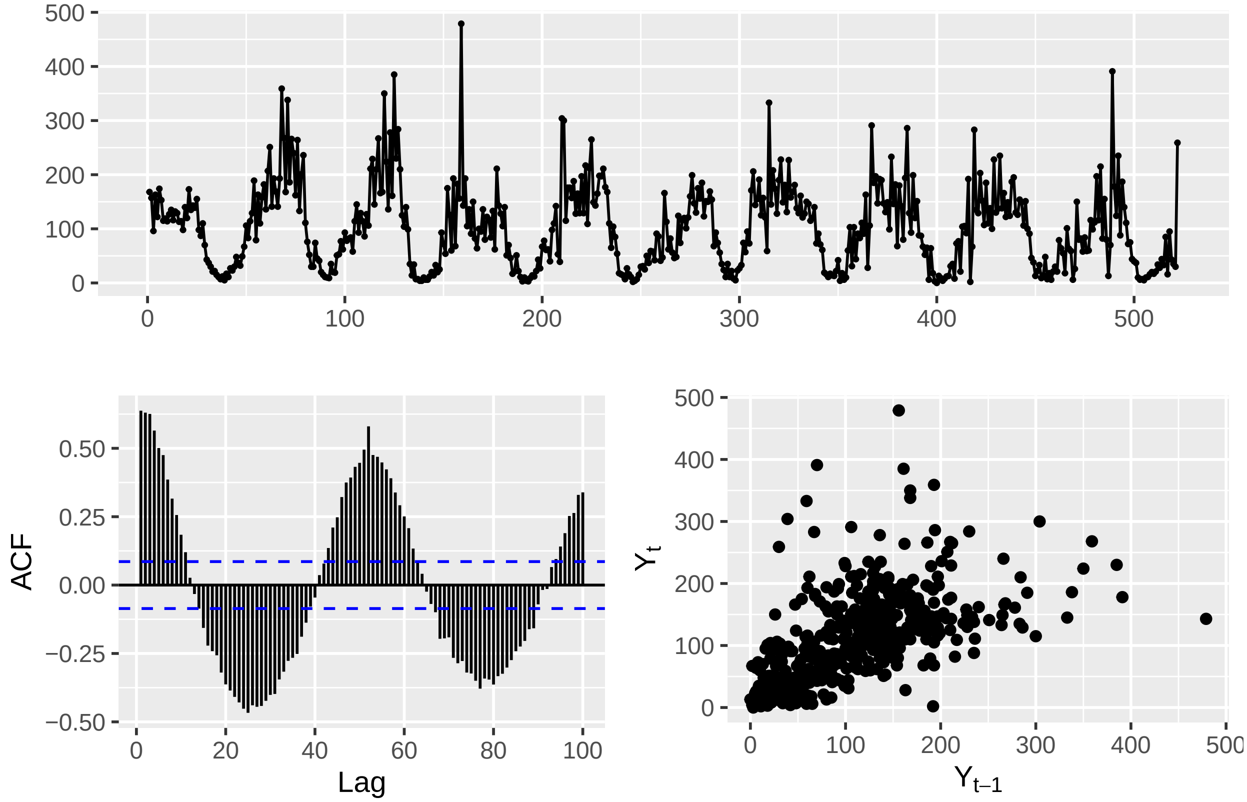
```
boxplot(budapest$nb, main="Répartition des nombre de cas par quantiles pour la ville de Budapest")
```



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.

Première approche

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

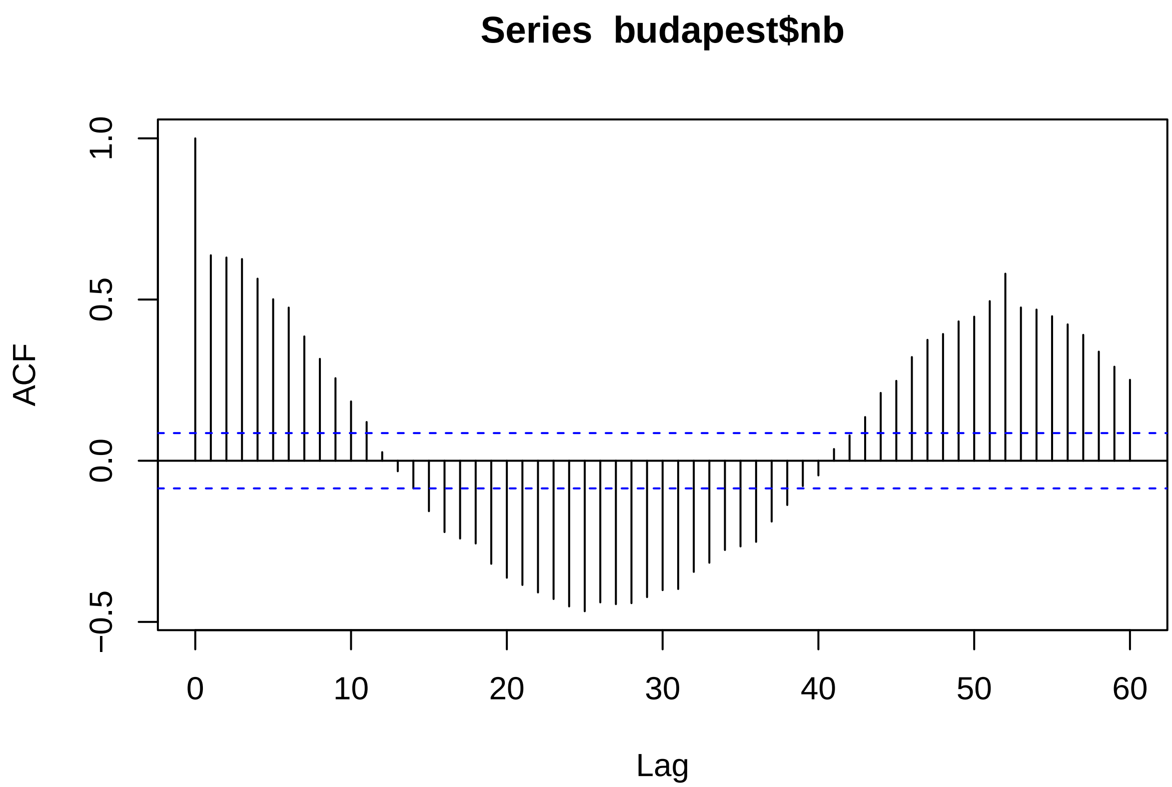


Pour mieux déterminer la saisonnalité nous allons observer la fonction d'autocorrélation. Elle semble périodique, ce qui indique une périodicité dans la série temporelle. La ligne pointillée bleue indique le niveau en-dessous duquel la corrélation n'est plus statistiquement significative.

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".

Observons l'auto-corrélation de plus près :

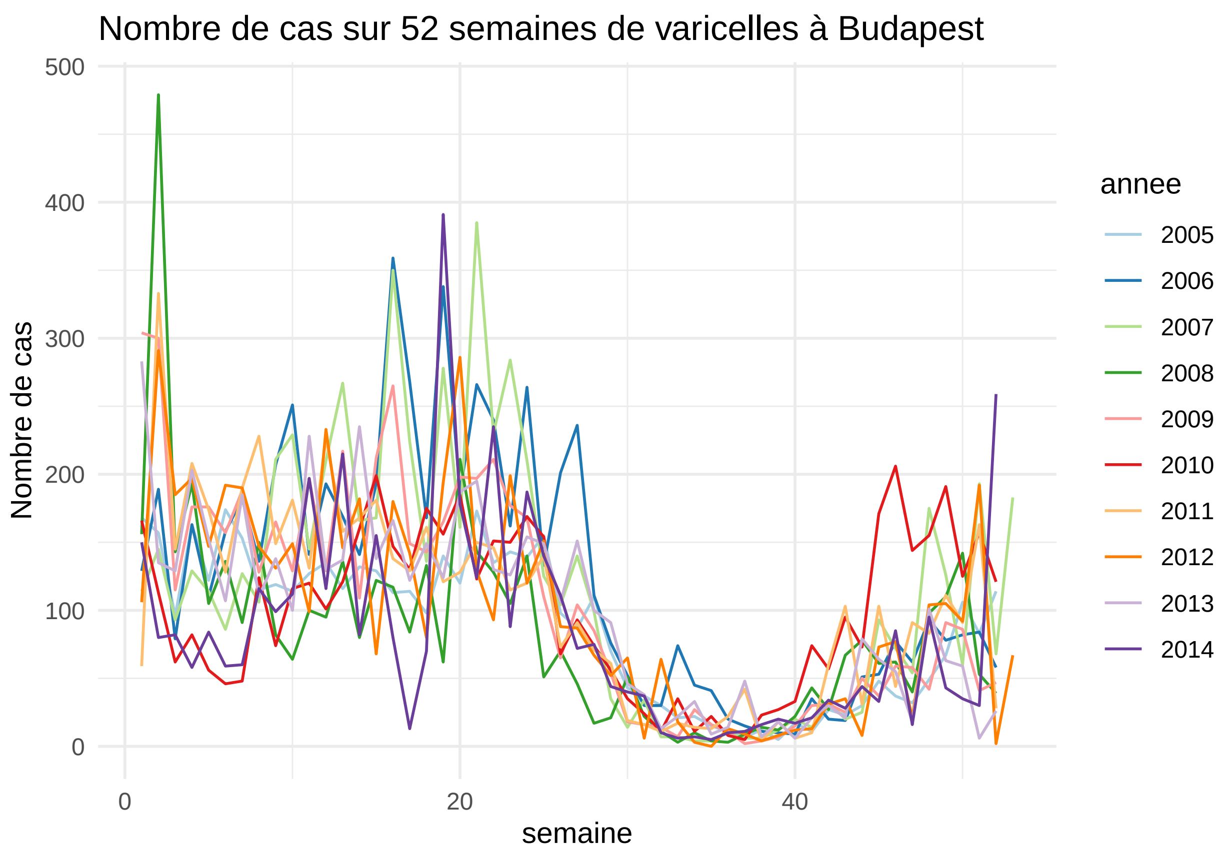
```
acf(budapest$nb, lag.max = 60)
```



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.

On regarde l'évolution pour chaque année.

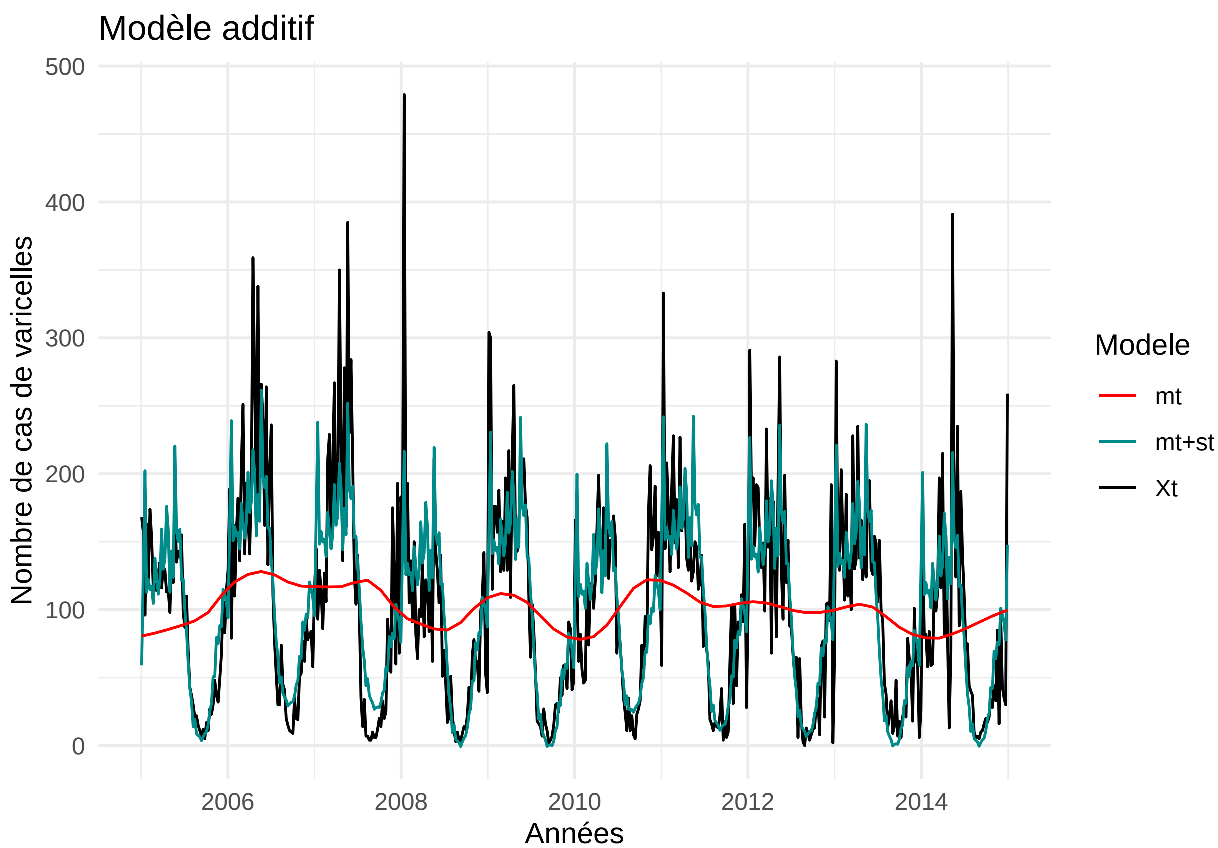
```
budapest$annee<-factor(year(budapest$date))
budapest$semaine<-week(budapest$date)


ggplot(budapest,aes(x = semaine, y = nb,group=annee,colour=annee)) + geom_line()+scale_color_brewer(pal
```



Avec cette représentation on observe bien la saisonnalité, le même schéma se reproduit tous les ans. On remarque qu'il y a plus de cas de varicelle au début de l'année (en hiver et au printemps). Durant l'été, il y a peu de cas de varicelles, puis à la fin de l'année le nombre de cas de varicelle augmente. Cela semble logique, d'après santé publique france, la varicelle est une maladie saisonnière, on observe chaque année une hausse des cas au printemps. Il s'agit d'une augmentation attendue de la maladie.

Décomposition de la série

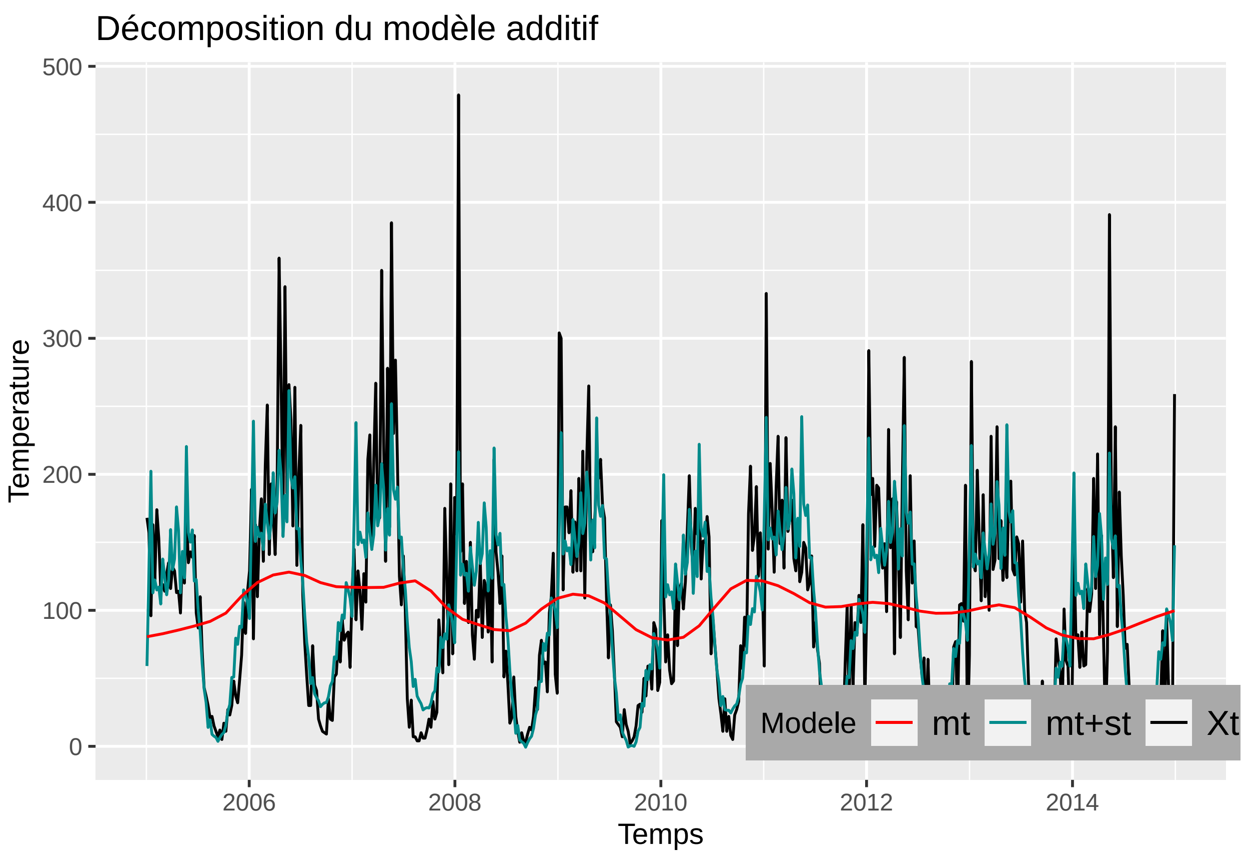
```
temp.ts <- ts(budapest$nb, start=c(2005,1,3), frequency=52)
mod_stl_add <- stl(temp.ts, s.window = "periodic")

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", "cyan4", "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()
```



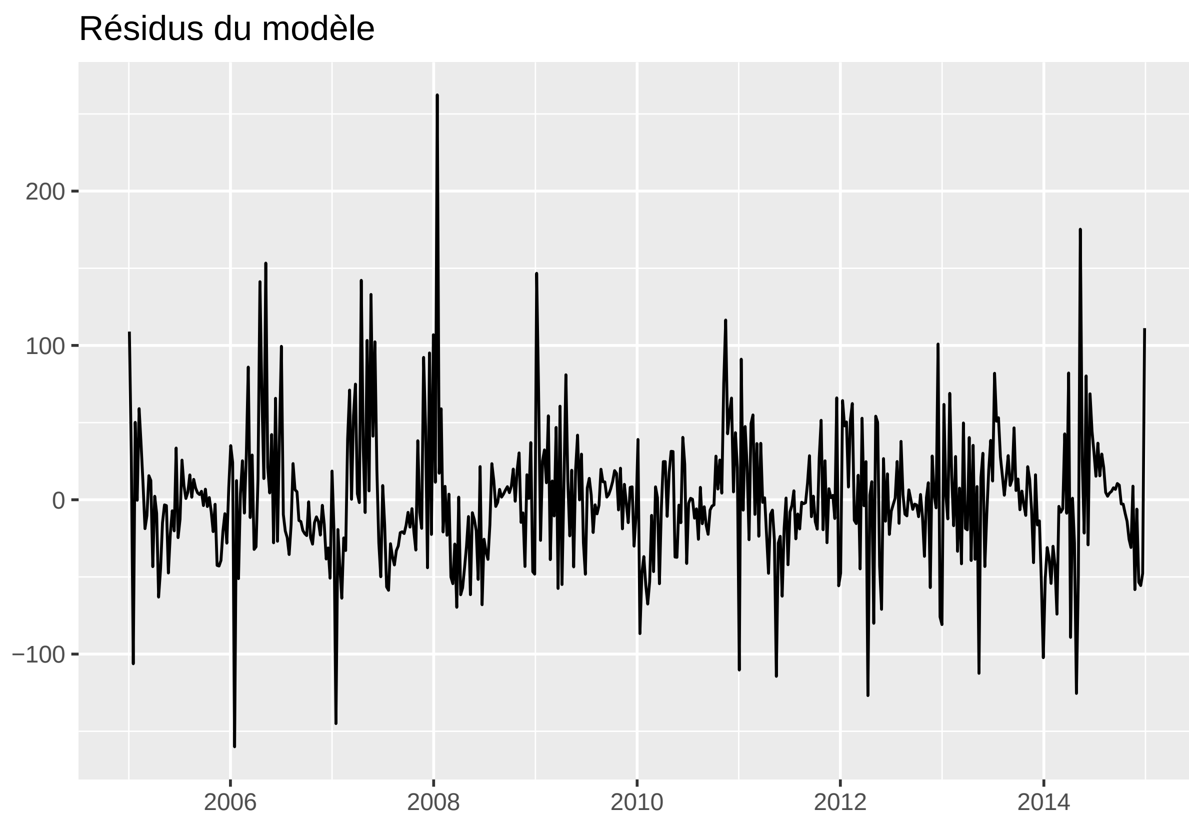
Elimination de la saisonnalité

Puisqu'on a conclut qu'il n'y avait pas de tendance apparente, on utilise une différence de 1 uniquement dans un premier temps puis on inclut le lag en plus dans un second temps. On utilise la fonction *stl()* mais on aurait pu utiliser la méthode des différences ou la fonction *décompose* qui utilise le principe des moyennes mobiles. Ces deux méthodes sont illustrées dans l'annexe (i) et (ii).

```
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", "cyan4", "black")) +
  theme(legend.position = c(0.8, 0.08), legend.direction = "horizontal") +
  labs(colour = "Modele") + ggtitle("Décomposition du modèle additif") +
  xlab("Temps") + ylab("Temperature") +
  theme(
    legend.background = element_rect(fill = "darkgray"),
    legend.text = element_text(size = 13) #+ theme_economist()
  )
```

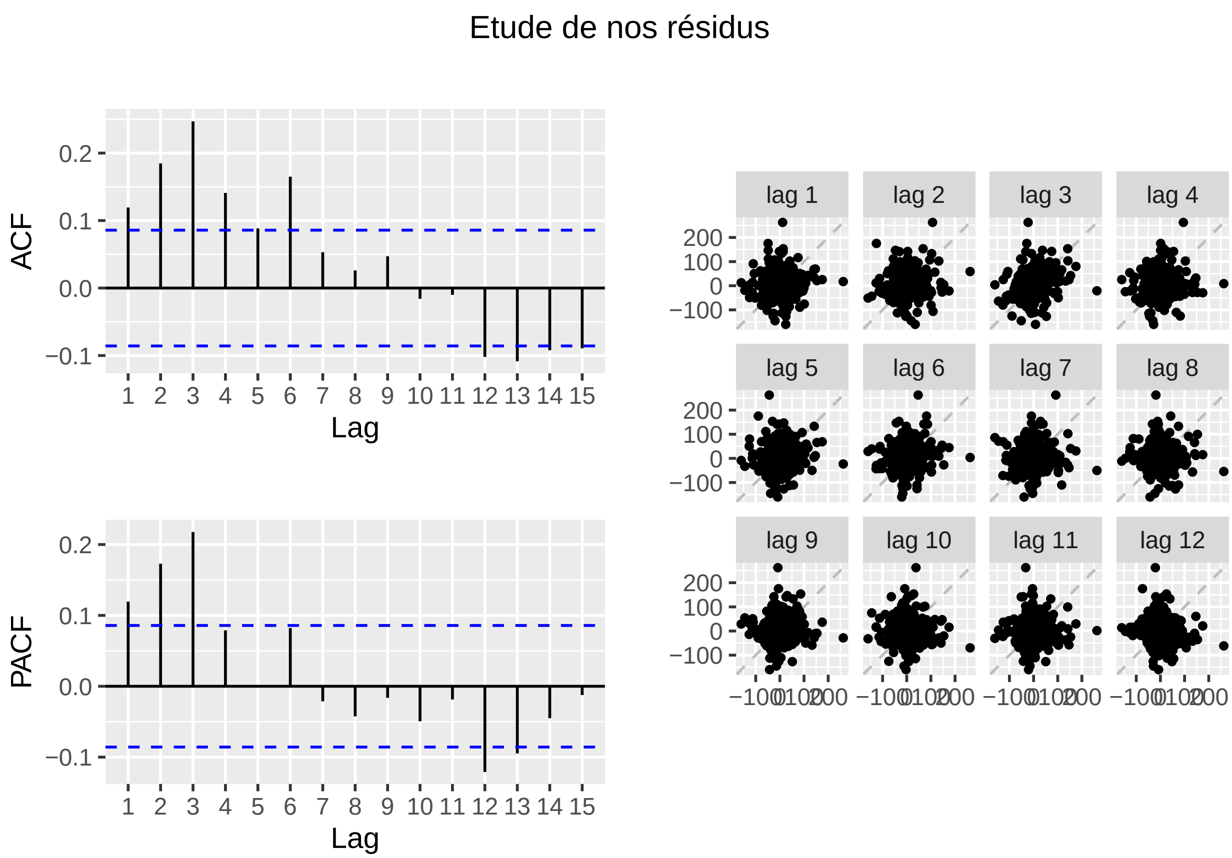



```
donnee %>% ggplot() + aes(x=date, remainder) + geom_line() +  xlab("") + ylab("") + ggtitle("Résidus du
```



Les résidus sont bien centrés en 0.

```
p1 <- gglagplot(donnee$remainder, do.lines = FALSE, set.lags = 1:12, colour = FALSE)
p2 <- ggAcf(donnee$remainder, lag.max = 15) + ggtitle(" ")
p3 <- ggPacf(donnee$remainder, lag.max = 15) +  ggtitle(" ")

grid.arrange(top = "Etude de nos résidus", p2,p3, p1, layout_matrix = rbind(c(1,3),c(2,3)))
```



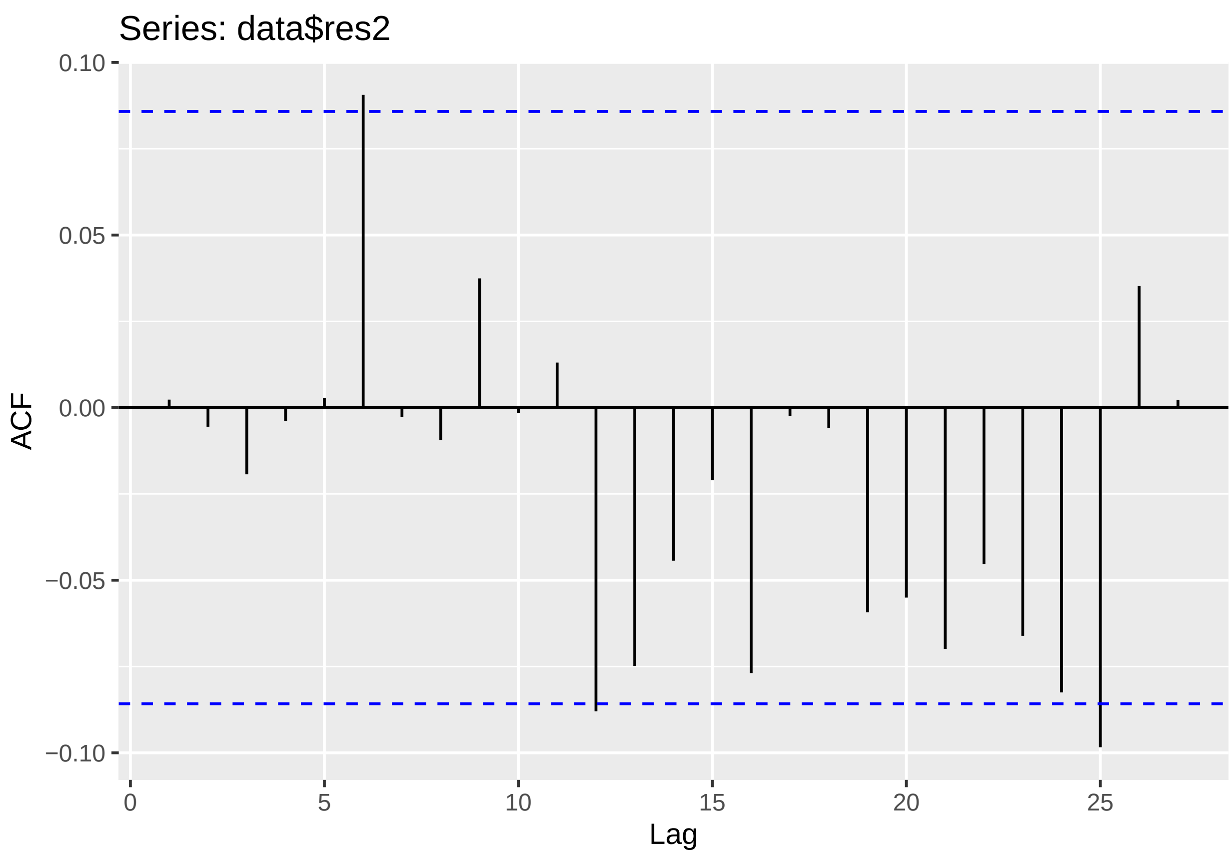
Sur la PACF, on observe une auto-corrélation nulle au lag 4. On choisirerait un processus auto-régressif de paramètre p=4 : *AR(4)*. De plus, la fonction *auto.arima()* nous permet de conforter notre choix. Elle est disponible en annexe iii.

Modélisation par un processus auto-régressif

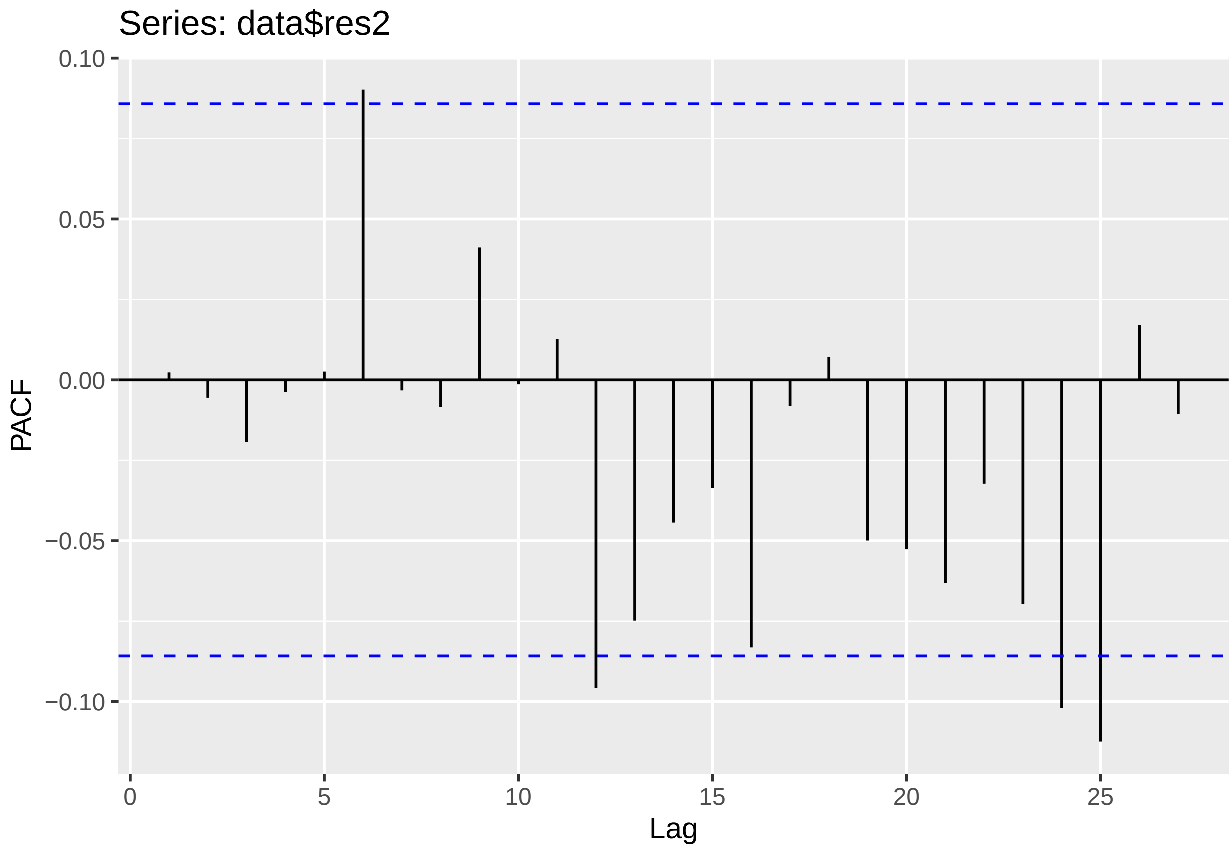
```
AR <- Arima(donnee$remainder, c(4,0,0))
data <- donnee %>% mutate(res2 = AR$residuals)
```

ACF et PACF

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



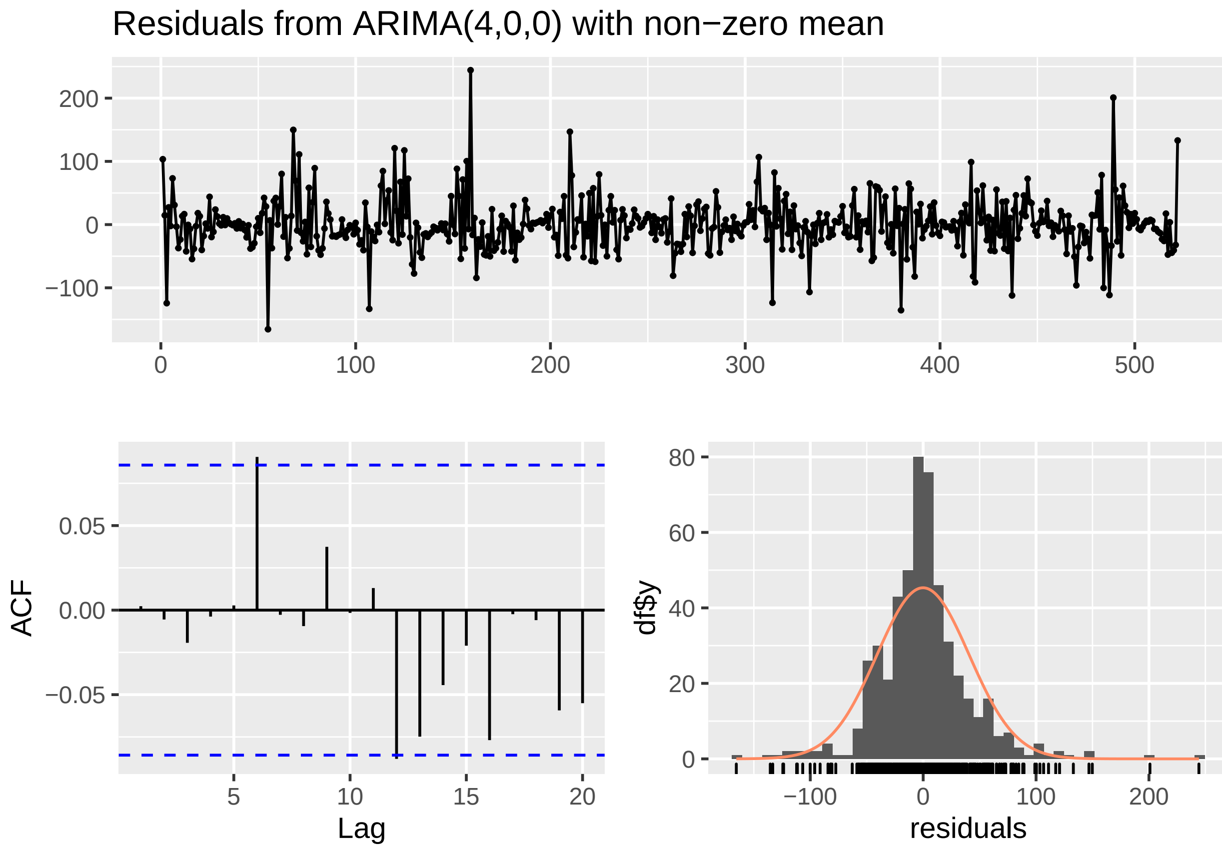

```
ggPacf(data$res2)
```



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

Analyse des résidus

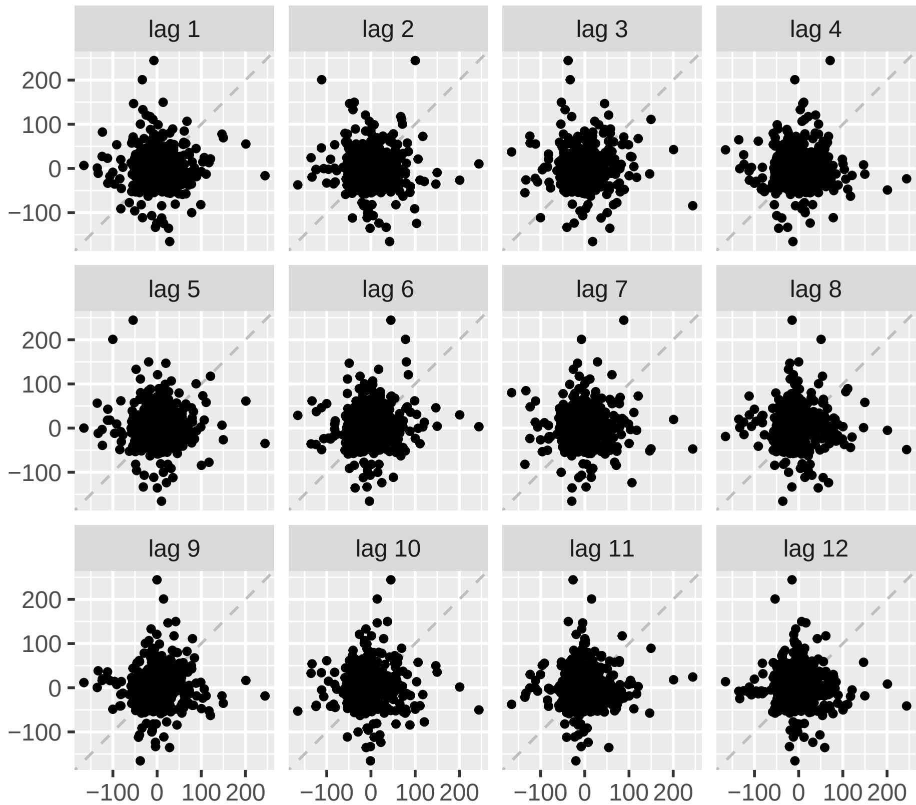
```
checkresiduals(AR, lag.max=20)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(4,0,0) with non-zero mean
## Q* = 5.3789, df = 5, p-value = 0.3714
##
## Model df: 5.   Total lags used: 10
```

On ne perçoit pas de struture particulière sur l'ACF, de plus la densité des résidus semblent être gaussiennes.

```
gglagplot(AR$residuals, do.lines = FALSE, set.lags = 1:12, colour = FALSE)
```



Les résidus ne semblent pas corrélés, les nuages de points aux différents lags sont plutôt arrondis.

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

```
##
##  Box-Pierce test
##
## data:  data$res2
## X-squared = 20.127, df = 18, p-value = 0.3257
```

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

```
##
##  Box-Ljung test
##
## data:  data$res2
## X-squared = 20.708, df = 18, p-value = 0.2944
```

Les 2 test de Box-Pierce et de Box-Ljung, renvoie une p-value > 5 %, alors on ne rejette pas l'hypothèse HO selon laquelle les auto-corrélations sont nulles. Les résidus sont décorréllés.

L'estimateur de la moyenne nous indique bien que les résidus sont centrés. De plus, l'estimateur de la variance nous montre une valeur de sigma étant de 1675.

```
mean(AR$residuals)
```

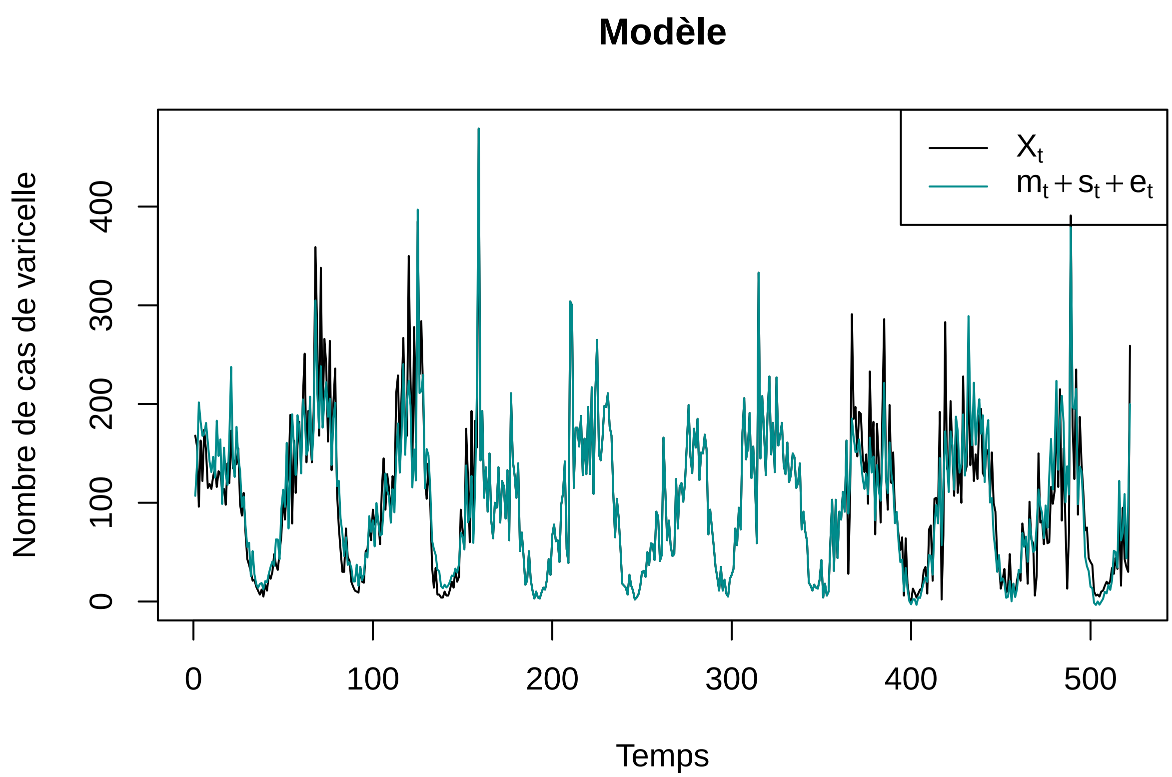
```
## [1] -0.09981575
```

```
var(AR$residuals)
```

```
## [1] 1675.004
```

Modèle de régression avec la décomposition de Fourrier

```
t <- 1:nrow(budapest)
st <- t%o%c(rep(1:182,2))*pi/182
st[,1:182] <- cos(st[,1:182])
st[,183:364] <- sin(st[,183:364])
names(st) <- c(paste("cos",1:182),paste("sin",1:182))
df <- data.frame(st, t, x=budapest$nb)

mod <- lm(data = df, x~. )

plot(budapest$nb,xlab='Temps',ylab="Nombre de cas de varicelle",main='Modèle', type = "l")
points(mod$fitted.values, type='l',col='cyan4')
legend('topright',c(expression(X[t]),expression(m[t]+s[t]+e[t])),col=c(1,"cyan4"),lty=1)
```



On voit alors que le modèle que nous avons construit est en totale adéquation avec les données que nous avons.

Conclusion

Nous avons un modèle additif et un modèle AR(4) sur nos résidus.

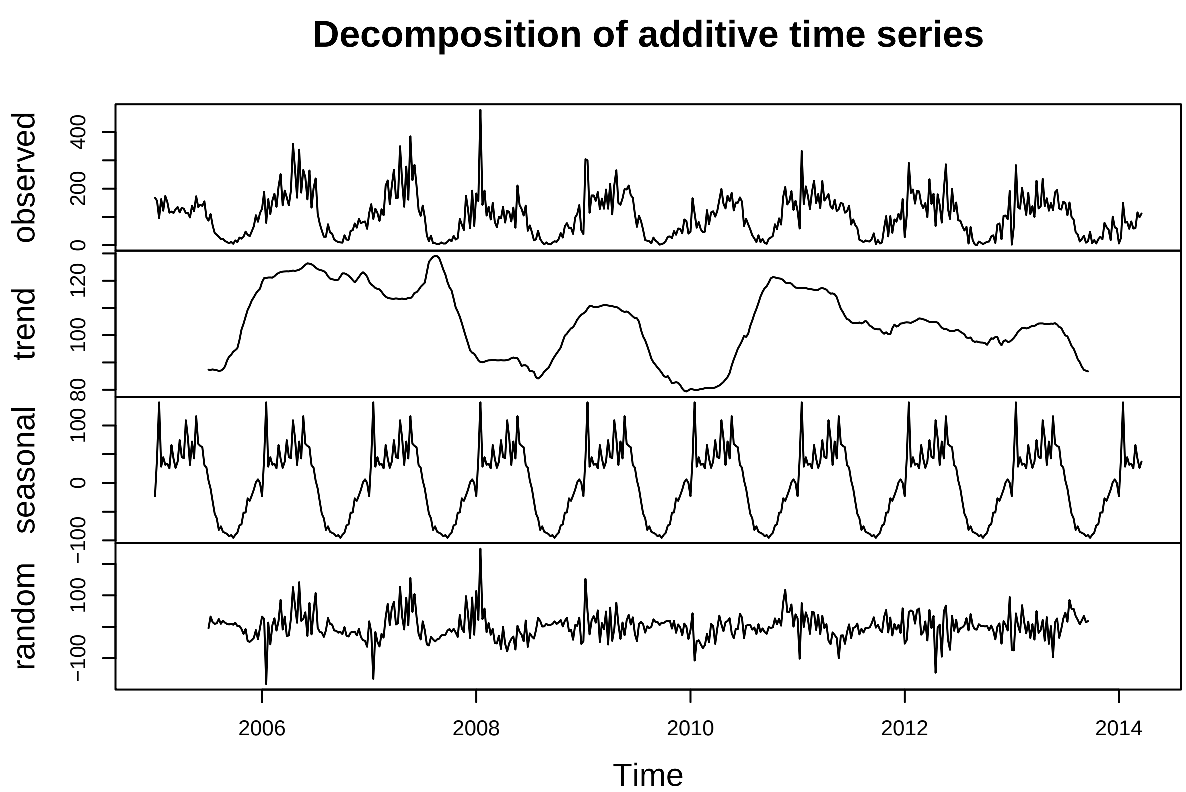
$$X_t = \hat{m}_t + \hat{s}_t + n_t + BB(\sigma^2 = 1675)$$
$$n_t = \hat{a}_1 n_{t-1} + \hat{a}_2 n_{t-2} + \hat{a}_3 n_{t-3} + \hat{a}_4 n_{t-4}$$

un AR(4).

Annexe

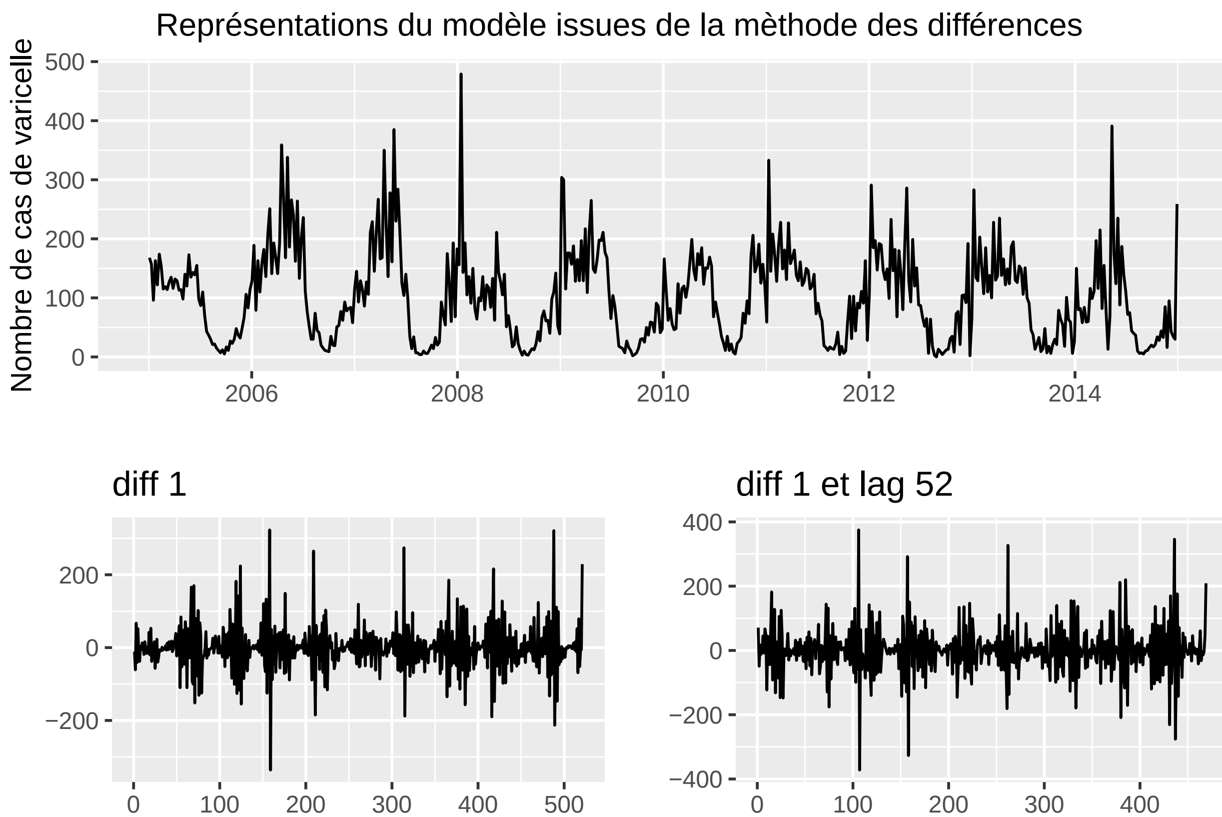
i. Fonction *decompose* : moyennes mobiles

```
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)
```



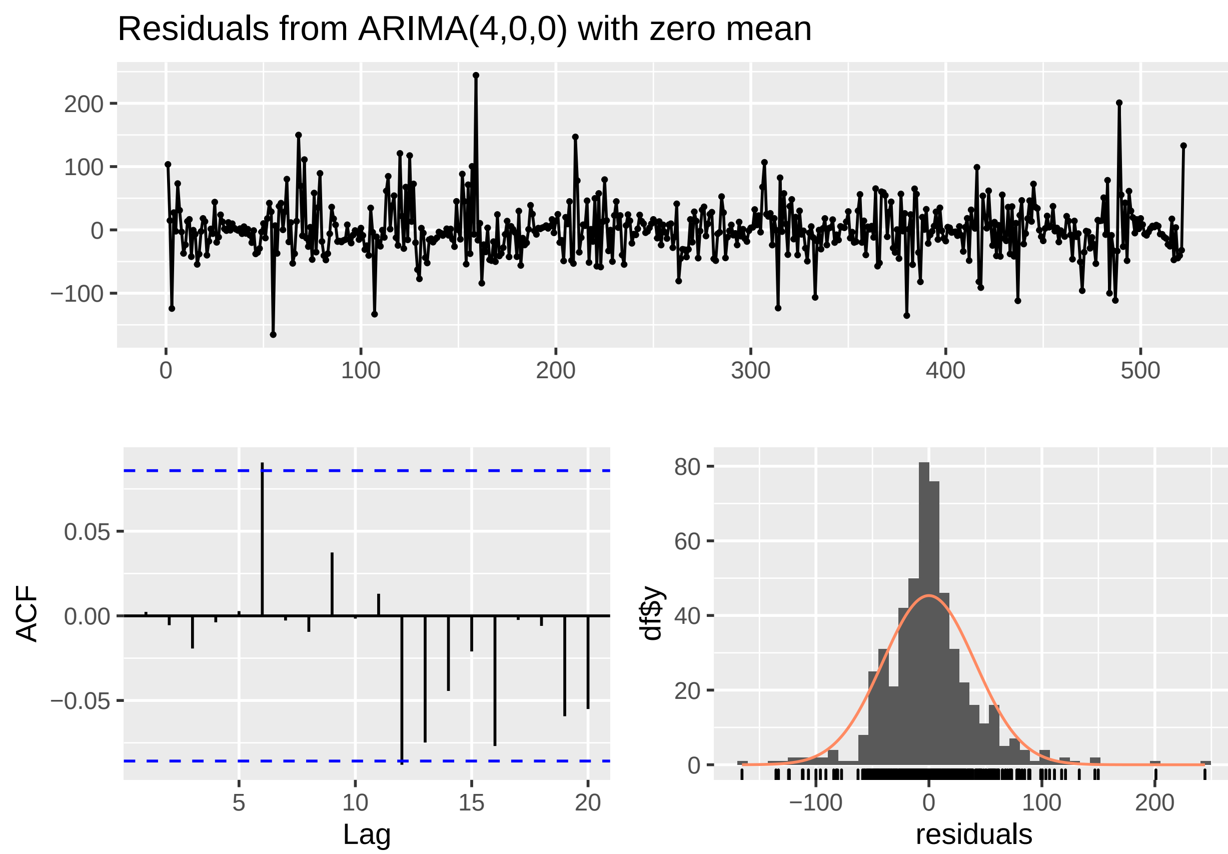
ii. Méthode des différences

```
T=52

xt_diff = diff(budapest$nb, differences = 1)
xt_diff_lag = diff(diff(budapest$nb, differences = 1), lag = T)

p1 <- donnee %>% ggplot() + aes(x=date, y=nb) + geom_line() +  xlab("") + ylab("Nombre de cas de varicel
p2 <- ggplot() + aes(x=1:521, y=xt_diff) + geom_line() + xlab("") + ylab(" ") + ggtitle("diff 1")
p3 <- ggplot() + aes(x= 1:469, y=xt_diff_lag) + geom_line() +  xlab("") + ylab(" ") + ggtitle("diff 1 e

grid.arrange( top = "Représentations du modèle issues de la mèthode des différences", p1, p2, p3, layou
```



iii. auto.arima()

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