Série Chronologique - Projet

```
library(TSA)
```

```
##
## Attaching package: 'TSA'
```

```
## The following objects are masked from 'package:stats':
##
##     acf, arima
```

```
## The following object is masked from 'package:utils':
##
##     tar
```

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(forecast)
```

```
## Registered S3 methods overwritten by 'forecast':
##   method       from
##   fitted.Arima TSA
##   plot.Arima   TSA
```

```
library(ggplot2)
source(file = "functions.R")
```

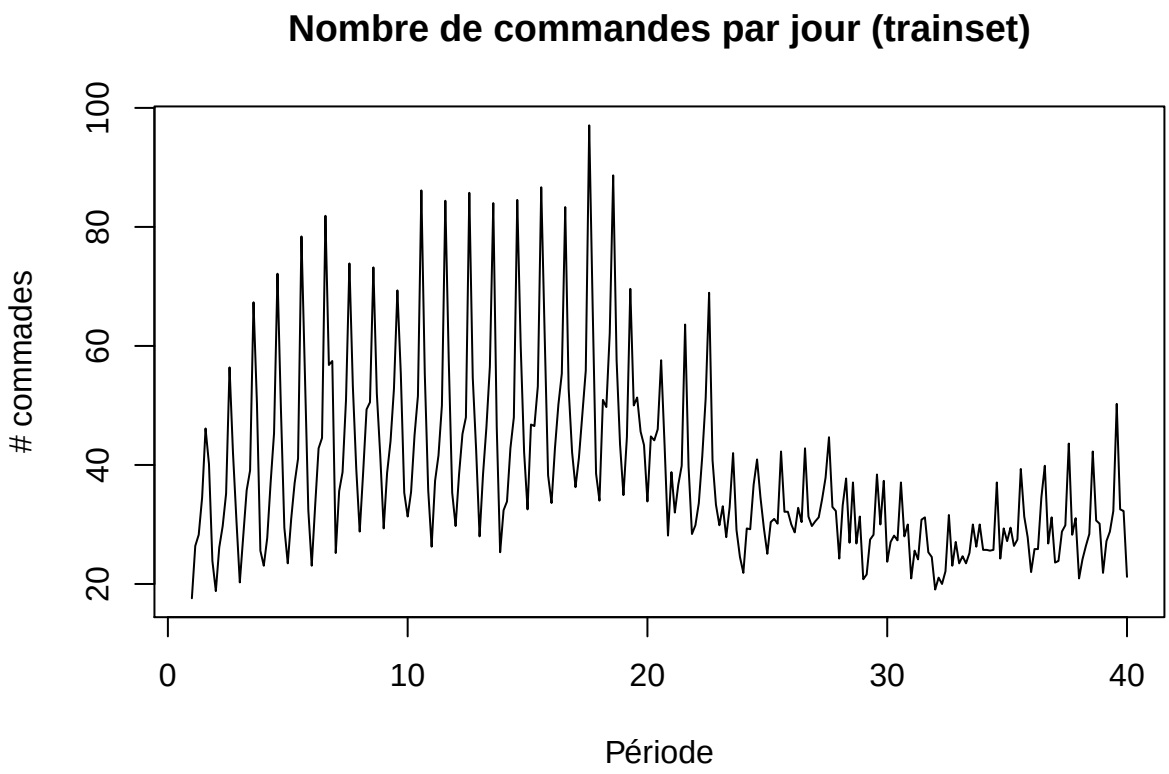
```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

Introduction

L'objet d'étude de ce projet est une série temporelle décrivant le nombre de commandes par jour de l'entreprise angevine de livraison de repas à domicile. Cette série est découpée en un jeu d'entrainement (train set) de 280 observations et un jeu de test (test set) de 21 observations. Notre objectif sera d'étudier cette série et de construire un modèle parmi les déclinaisons des modèles $ARMA$.

```
data = read.csv('orders.csv')
trainset = ts(data$n_orders, start=1, end=40, frequency=7)
testset = ts(data$n_orders, start=41, end=44, frequency=7)
plot(trainset, type='l', xlab="Période", ylab="# commades", main="Nombre de commandes par jour (trainse
```



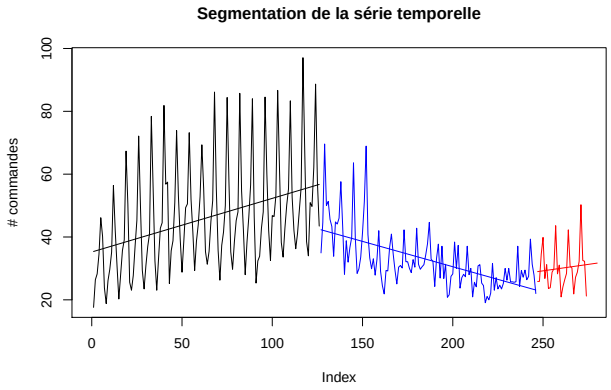
Cette série présente plusieurs caractéristiques remarquables.

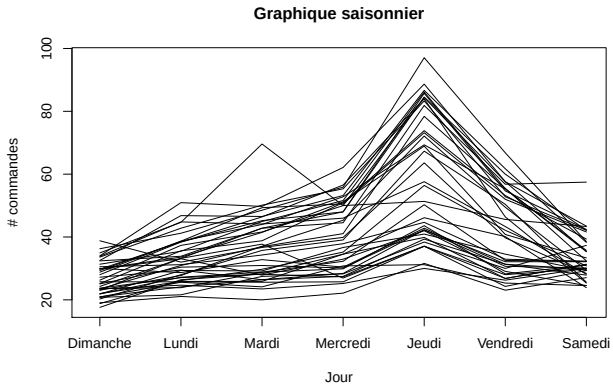
Il y a une forte saisonalité sur les jours de la semaine. En effet, le nombre de commandes croit du dimanche au vendredi puis décroît le samedi en moyenne.

La série semble composée de deux voir trois régime. Un premier lors des 18 premières semaines avant une tendance haussière, une saisonnalité très marquée et une variance relativement grande. Puis, vient 16 semaines de tendance baissière, avec une saisonnalité moins marquée et une plus faible variance. Enfin, les 5 dernières semaines semble reprendre quelque peu le schéma de premières semaine avec un tendance à la hausse, une saisonnalité marquée et une variance qui augmente.

```
layout(matrix(c(1, 2), nrow=1, ncol=2, byrow=TRUE))
# Segmentation
t1 = 1:126; X1 = trainset[t1]
t2 = 127:246; X2 = trainset[t2]
t3 = 247:280; X3 = trainset[t3]
lm1 = lm(X1~t1)
lm2 = lm(X2~t2)
lm3 = lm(X3~t3)
plot(t1, X1, type='l', xlab="Index", ylab="# commandes", main="Segmentation de la série temporelle", xli
lines(t2, X2, type="l", col="blue")
lines(t3, X3, type="l", col="red")
lines(t1, lm1$coefficients[1]+ t1 * lm1$coefficients[2])
lines(t2, lm2$coefficients[1]+ t2 * lm2$coefficients[2], col="blue")
lines(t3, lm3$coefficients[1]+ t3 * lm3$coefficients[2], col="red")

# Saisonalité
seasonplot(trainset, type="l", main="Graphique saisonnier", xlab="Jour", ylab="# commandes", season.labe
```





Les modèles $ARMA$ font l'hypothèse que la série temporelle (X_t) est stationnaire (du second ordre), c'est à dire que sa moyenne et sa covariance sont invariantes par translation dans le temps: - $\mathbb{E}[X_t] = m(t) = m$ pour tout $t \in \mathbb{Z}$. - $Cov(X_t, X_s) = \Gamma(t,s) = \Gamma(t+k, s+k)$ pour tout $(t,s) \in \mathbb{Z}^2$ et tout décalage temporelle $k \in \mathbb{Z}$.

Cette hypothèse est nécéssaire car ... .

En plus de l'aspect visuel, nous utiliserons les tests suivants pour étudier sur la stationarité : - Le test ADF avec $\mathcal{H}_0$: "la trajectoire est issue d'un processus non stationnaire" contre $\mathcal{H}_1 = \bar{\mathcal{H}_0}$. - Le test KPSS avec $\mathcal{H}_0$: "la trajectoire est issue d'un processus stationnaire" contre $\mathcal{H}_1 = \bar{\mathcal{H}_0}$.

```
adf.test(trainset)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  trainset
## Dickey-Fuller = -2.6144, Lag order = 6, p-value = 0.3172
## alternative hypothesis: stationary
```

```
kpss.test(trainset)
```

```
## Warning in kpss.test(trainset): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  trainset
## KPSS Level = 2.4713, Truncation lag parameter = 5, p-value = 0.01
```

Les tests s'accordent. En effet, le test ADF ne rejette pas l'hypothèse de stationarité (p-value»0.05) alors que le test KPSS rejette l'hypothèse de non-stationarité (p-value«0.05).

=> Remarque sur la stat visuellement, transformation (Cependant on voit bien que la variance ne reste pas constante, ce que les tests ADF et KPSS ne détectent pas.)

De manière général, lorsque l'on souhaite modéliser une série temporelle, on applique un modèle pour tenter d'obtenir des résidus centré et de même variance finie aussi appellé *bruit blanc*. Ainsi, on s'appuie en particulier sur le test l'auto-corrélation de Ljung-Box d'hypothèse nulle $\mathcal{H}_\prime$: "la série temporelle ne possède pas d'autocorrélation".

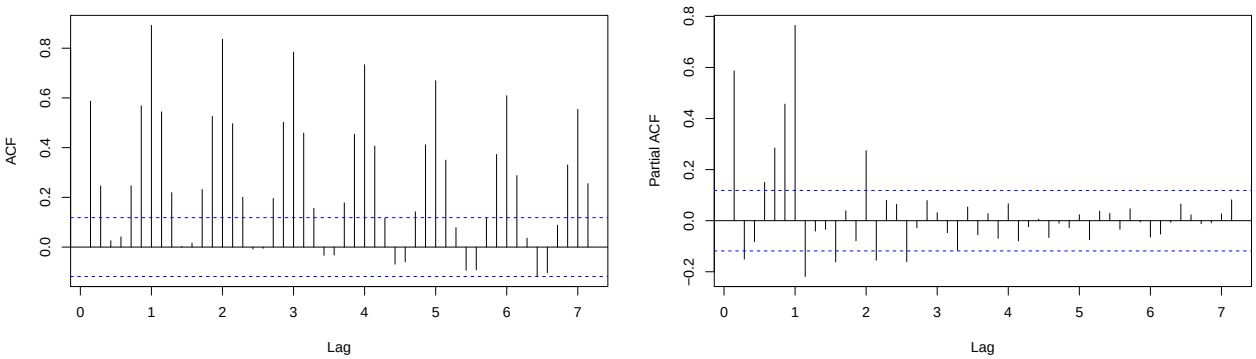
```
Box.test(trainset, type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  trainset
## X-squared = 95.418, df = 1, p-value < 2.2e-16
```

En ce qui concerne notre série temporelle, on rejette fortement cette hypothèse.

Nous présentons maintenant deux outils indispensable à l'étude des séries temporelles: l'ACF et la PAFC.

L'AFC représente les autocorrélation entre deux valeurs distantes de h dans le temps. Une autocorrélation nulle à partir d'un rang $q+1$ est caractéristique d'un processus $MA(q)$. L'autocorrélogramme PAFC représente quand à lui les corrélation "pur" entre deux valeurs distantes de h dans le temps, c'est à dire entre lesquelles on a supprimé l'influence linéaire des valeurs intermédiaires. Une PACF avec des pics dans le couloir de non-significativité à partir du rang $p+1$ est caractéristique d'un processus $AR(p)$.



L'ACF présente des pics périodiques significatifs majoritairement. On ne peut donc pas conclure que la série est stationnaire car elle possède une composante saisonnière.

La PACF possède des corrélations partielles significatives aux lags 1 et 2 et aux lags 5 à 8 ainsi qu'au lag 14. Ainsi, on en conclut que le nombre de livraisons est corrélé significativement au nombre de commandes lors de la semaine passée ainsi qu'au nombre de commande réalisée 14 jours auparavant.

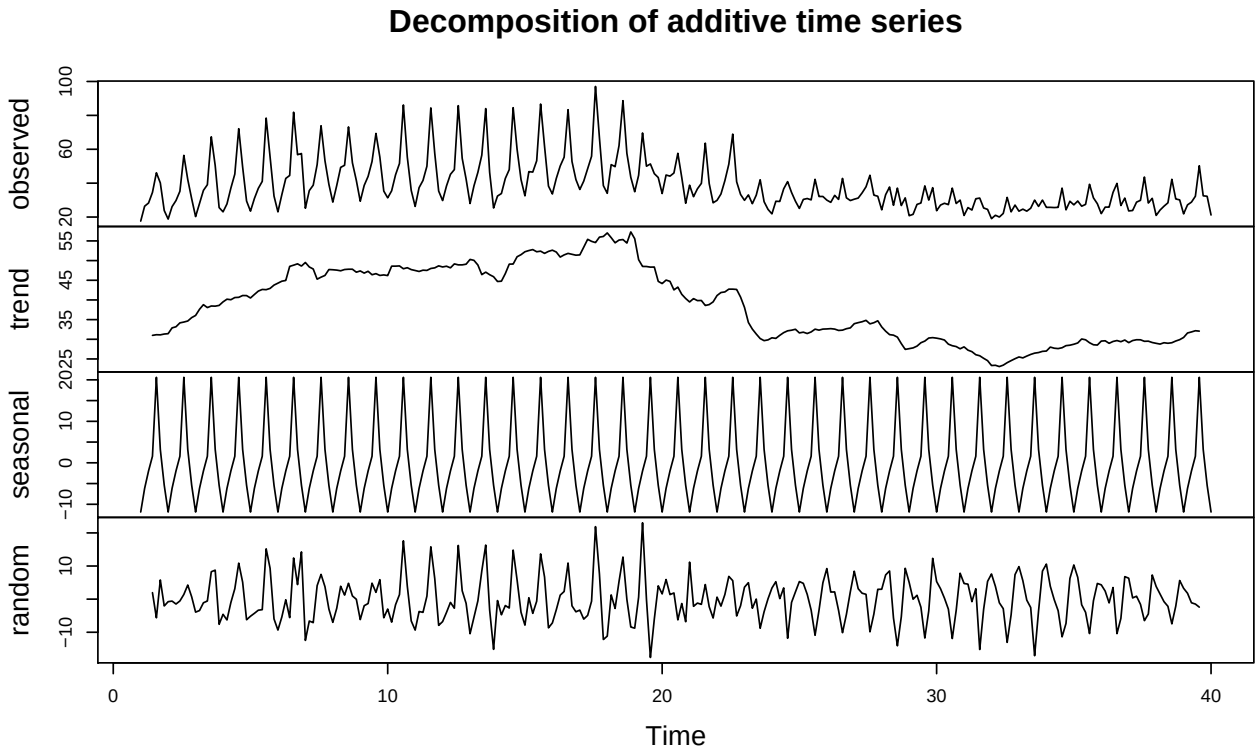
En conclusion, la série n'est pas stationnaire et présente une saisonalité hebdomadaire ainsi qu'une variance fluctuante au fil du temps.

Pour rendre cette série stationnaire, nous allons utiliser plusieurs approches tel que la décomposition (partie 2.1) et la différentiation (partie 2.2). Puis nous construirons et comparons nos modèle $ARIMA$ à l'aide des fonction `auto.arima` et `checkup_res` (partie 3) avant d'évaluer nos modèles sur le jeu de données test avec des critères prédictifs (partie 4).

Stationarité

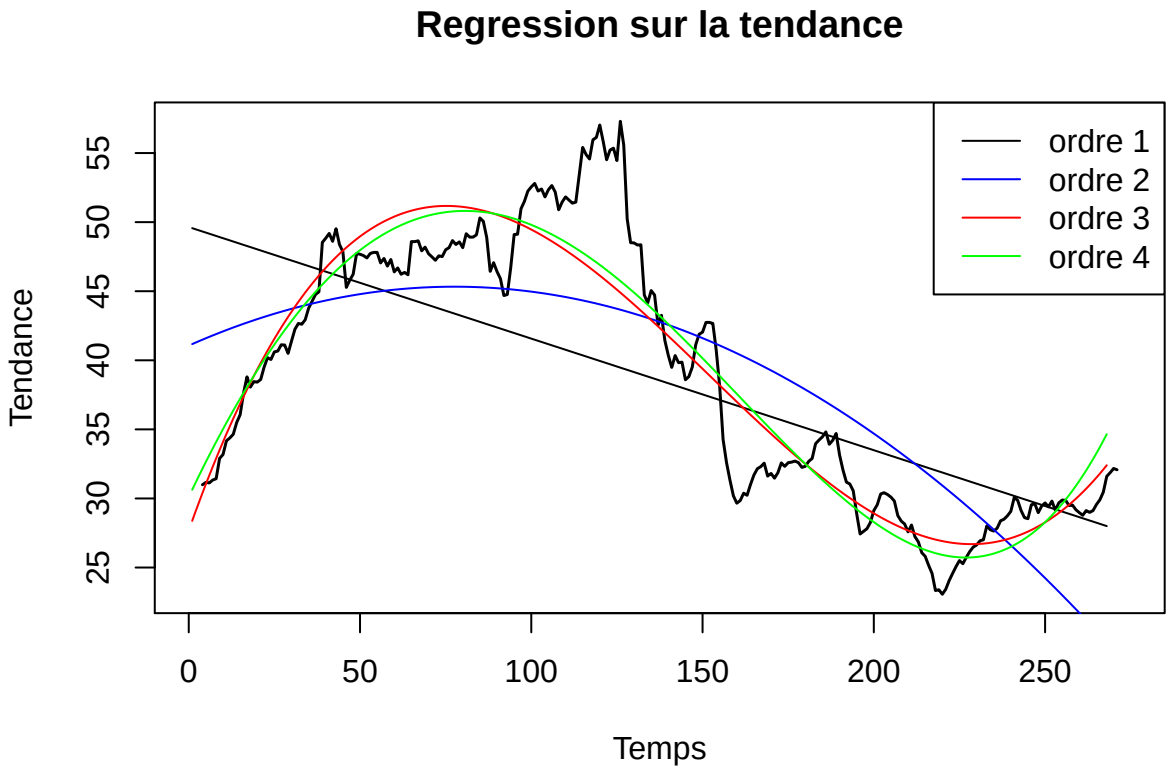
Decomposition

La fonction `decompose` renvoie une tendance, une composante saisonnière ainsi que des résidus.



La fonction `decompose` détermine d'abord la composante de tendance en utilisant une moyenne mobile, et la supprime de la série chronologique. Ici, on voit que l'on pourrait l'approximer par une tendance d'ordre 3. Ensuite, la figure saisonnière est calculée en faisant la moyenne, pour chaque unité de temps, de toutes les périodes. La figure saisonnière est ensuite centrée. Enfin, la composante d'erreur est déterminée en retirant la tendance et la figure saisonnière (dupliquée si nécessaire) de la série temporelle originale.

Pour pouvoir construire un modèle de prédiction sur la base de ces résultats, nous allons appliquer une regression linéaire sur la tendance puis y ajouter la composante saisonnière.



Visuellement, les regression d'ordre 3 et 4 donnent de très bon résultat et obtienent un R^2 ajusté de 0.88. Le principe de parcimonie nous pousse à garder la régression linéaire d'ordre 3.

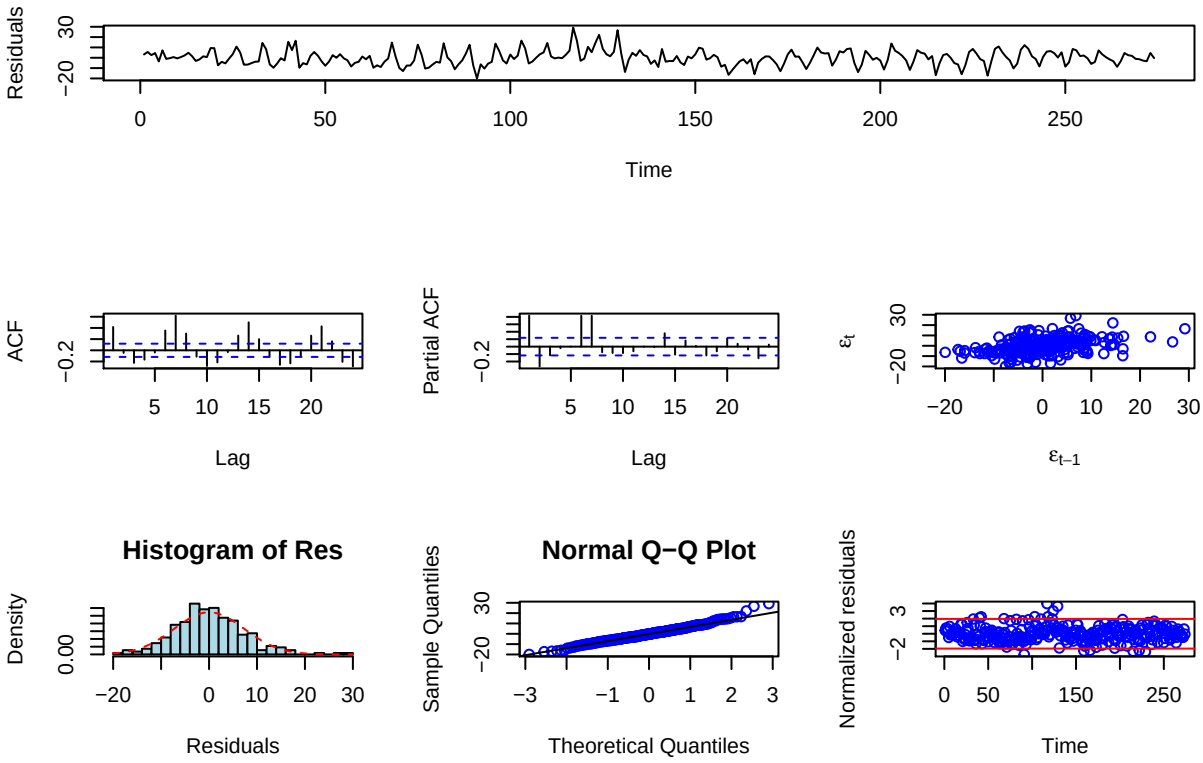
```
# retourne la tendance sommée avec la composante saisonnière pour le trainset et le testset.
decomposition = decomp(trainset, testset)
```

Analysons à présent les résidus de la décomposition.

```
decomp_residuals = trainset - decomposition$train
check_stat(decomp_residuals)
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 215.46, df = 7, p-value < 2.2e-16
```

```
## Warning in kpss.test(res): p-value greater than printed p-value
```




```
##
##  KPSS Test for Level Stationarity
##
## data:  res
## KPSS Level = 0.10925, Truncation lag parameter = 5, p-value = 0.1
##
##
##  Augmented Dickey-Fuller Test
##
## data:  res[which(!is.na(res))]
## Dickey-Fuller = -2.8368, Lag order = 6, p-value = 0.2235
## alternative hypothesis: stationary
##
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98438, p-value = 0.004363
```

Les résidus ne sont toujours pas stationnaires, ce dont on peut se rendre compte visuellement car on distingue une périodicité. L'ACF conforte cette hypothèse par la périodicité de ses pics significatifs. De plus, d'après l'histogramme, le QQ_plot et le test de Shapiro (p-value=0.05), les résidus ne semblent pas suivre une loi gaussienne.

Un modèle ARMA ne peut être appliqué car la série n'est pas stationnaire. La technique privilégier est la différenciation. Nous expliquons et appliquons cette technique dans la partie suivante.

Différenciation

Introduisons l'opérateur de retard B qui à une donnée chronologique donnée lui associe sa valeur précédante. On a alors $B^0 X_t = I X_t = X_t$, $B^k X_t = X_{t-k}$ et $B^{-k} X_t = X_{t+k}$.

Différencier une série temporelle à l'ordre d consiste à lui appliquer le filtre $(I - B^d)$. On regarde alors si $\Delta X_t = (I - B^d)X_t$ est stationnaire.

Différenciation sur les résidus de la décomposition

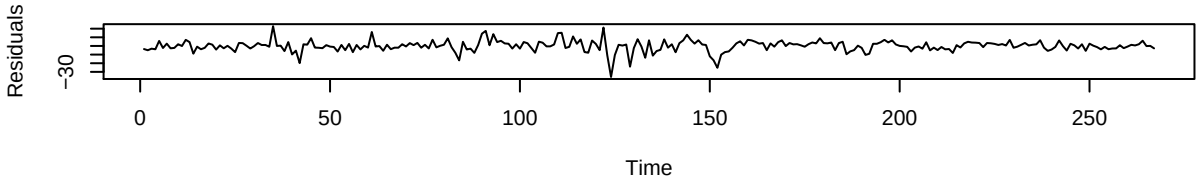
```
diff_decompose = diff(decomp_residuals, 7)
check_stat(diff_decompose)
```

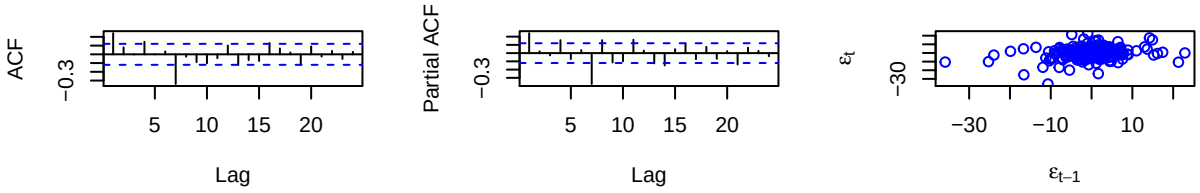
```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 55.138, df = 7, p-value = 1.4e-09
```

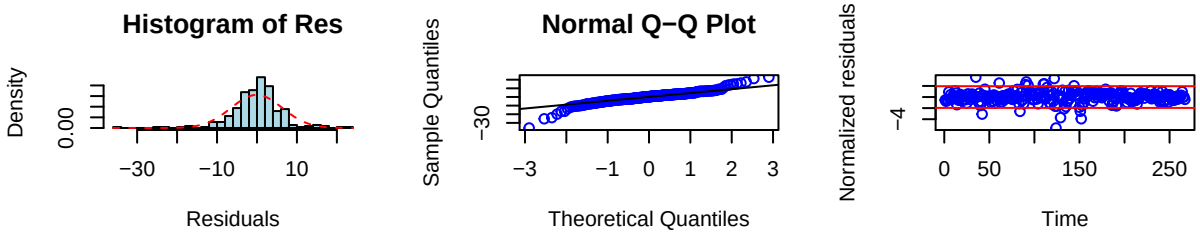
```
## Warning in kpss.test(res): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  res
## KPSS Level = 0.034531, Truncation lag parameter = 5, p-value = 0.1
```

```
## Warning in adf.test(res[which(!is.na(res))]): p-value smaller than printed p-
## value
```








```
##
##  Augmented Dickey-Fuller Test
##
## data:  res[which(!is.na(res))]
## Dickey-Fuller = -7.9117, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
##
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.91688, p-value = 4.971e-11
```

Différenciation sur la série original

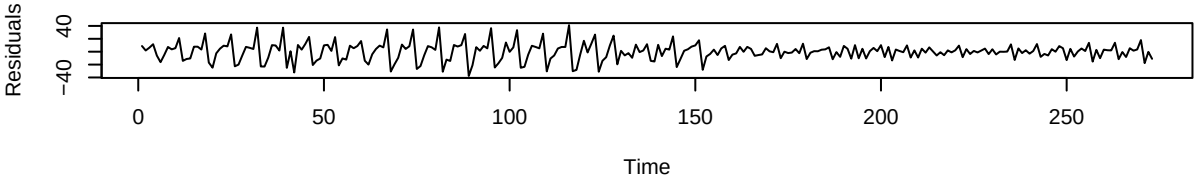
```
diff1 = diff(trainset)
check_stat((diff1))
```

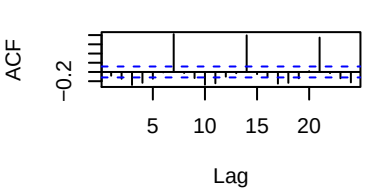
```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 241.24, df = 7, p-value < 2.2e-16
```

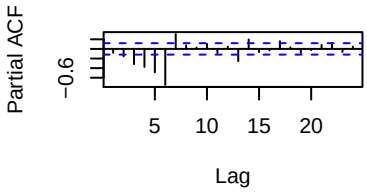
```
## Warning in kpss.test(res): p-value greater than printed p-value
```

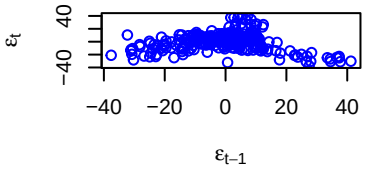
```
##
##  KPSS Test for Level Stationarity
##
## data:  res
## KPSS Level = 0.06673, Truncation lag parameter = 5, p-value = 0.1
```

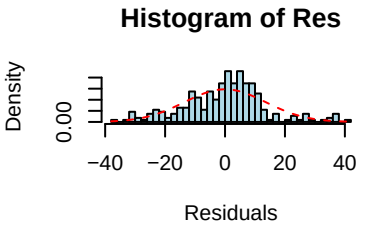
```
## Warning in adf.test(res[which(!is.na(res))]): p-value smaller than printed p-
## value
```

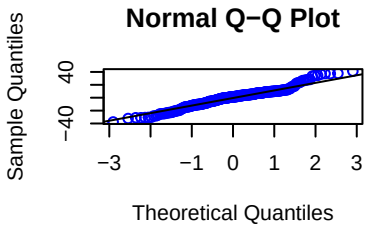


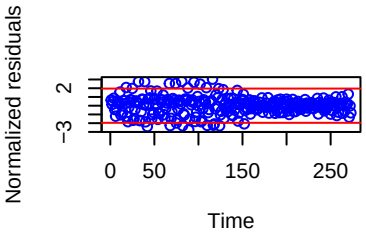













```
##
##  Augmented Dickey-Fuller Test
##
## data:  res[which(!is.na(res))]
```

```
## Dickey-Fuller = -11.875, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97117, p-value = 2.598e-05
```

Les tests KPSS et ADF obtiennent des p-value respective supérieur à 0.1 et inférieur à 0.01. On pourrait alors penser que cette série est stationnaire. Néanmoins cette différenciation n'est pas concluante car: - On observe une saisonalité de la série graphiquement - L'ACF montre aussi un saisonalité avec des pics très significatifs en 7, 14 et 21. - La PACF montre des pics de plus en plus significatifs jusqu'à 7. - Le graphique des obsrvations X_t en fonction de $X_t - 1$ montre visuellement une - Enfin, l'histogramme, le Q-Q plot, le graphique des residus normalisés ainsi que le test de shapiro donne la série non gaussienne.

Passons alors à une différentiation d'ordre 7.

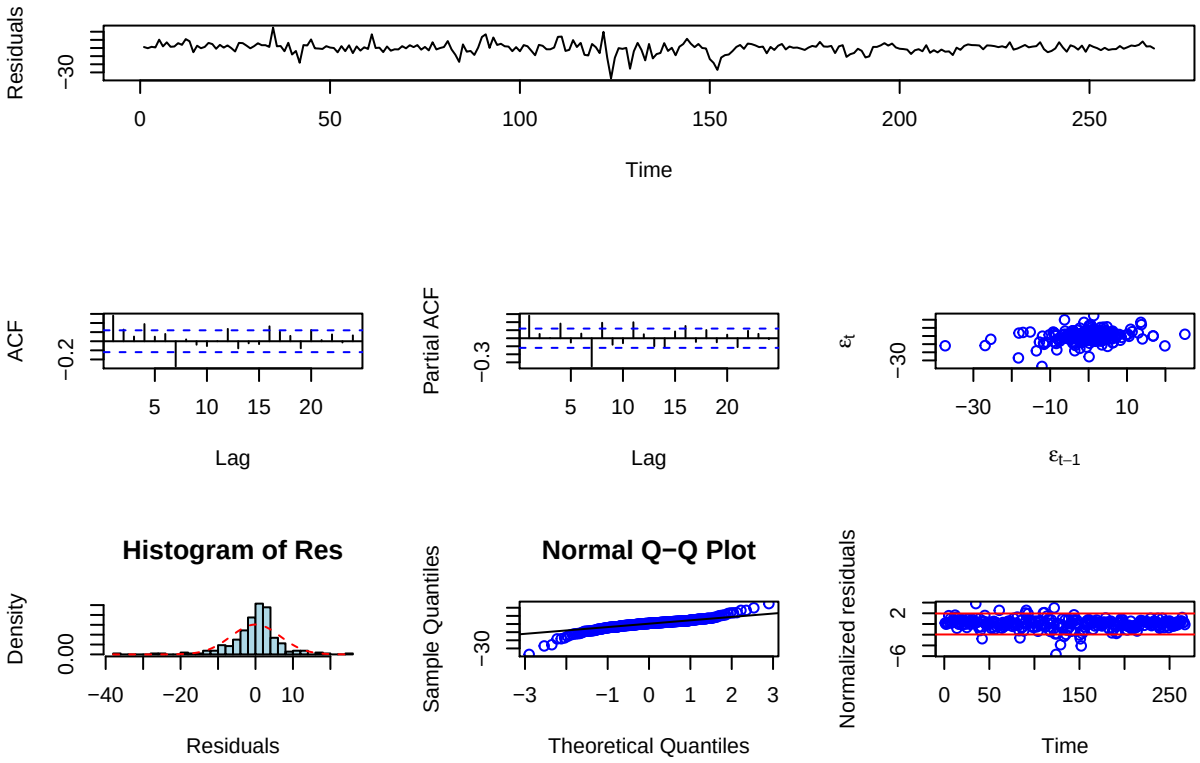
```
diff2 = diff(trainset, 7)
check_stat(diff2)
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 59.656, df = 7, p-value = 1.768e-10
```

```
## Warning in kpss.test(res): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  res
## KPSS Level = 0.33734, Truncation lag parameter = 5, p-value = 0.1
```

```
## Warning in adf.test(res[which(!is.na(res))]): p-value smaller than printed p-
## value
```




```
##
##  Augmented Dickey-Fuller Test
##
## data:  res[which(!is.na(res))]
## Dickey-Fuller = -7.2264, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.90327, p-value = 4.495e-12
```

Les tests ADF et KPPSS vont dans le sens de la stationnarité. L'ACF montre une décroissance rapide avec une légère périodicité et des pics significatifs en 4 et 7. La PACF possède des pics significatifs en 4, 7, 8 et 11 avec une légère périodicité. On retrouve une série dont l'allure est en accord avec la stationnarité. Néanmoins les graphiques et les tests montre que cette différanciation ne donne pas une série probablement gausienne.

Essayons de différencier à nouveau la série, c'est à dire d'appliquer le filtre $(I - B)(I - B^7)$ à X_t.

```
diff3 = diff(diff(trainset,7))
check_stat(diff3)
```

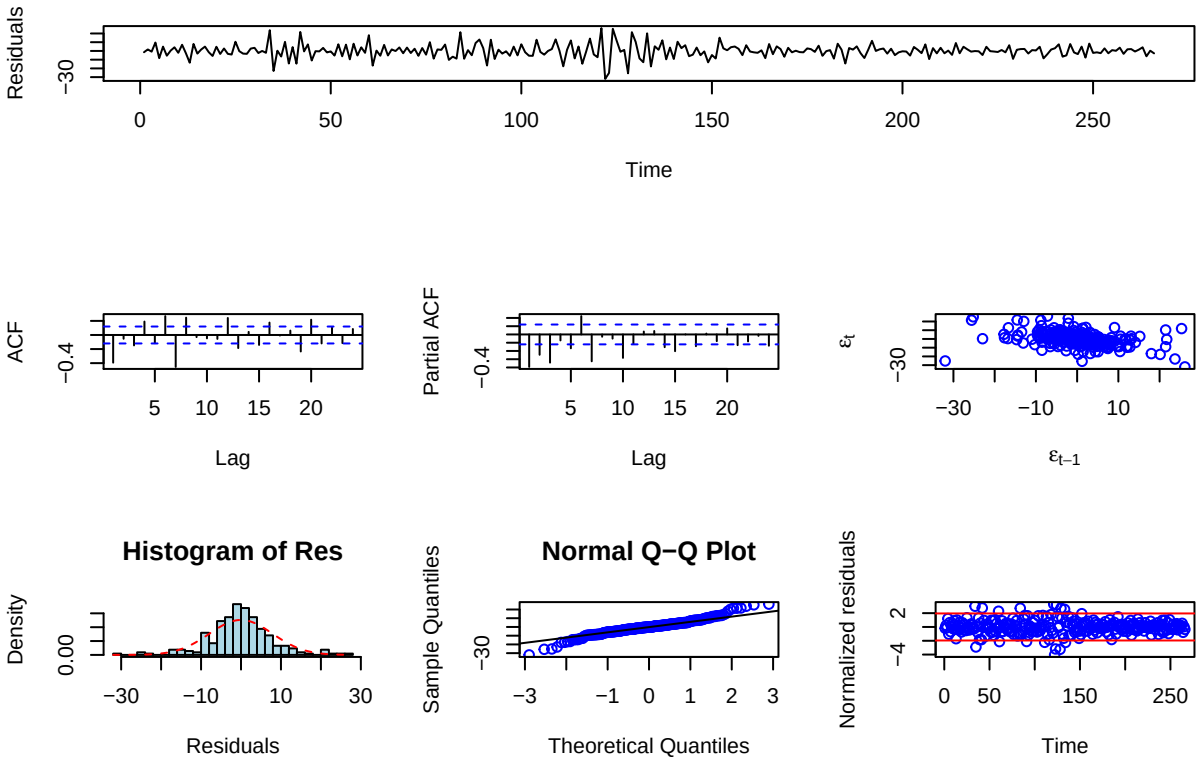
```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 138.08, df = 7, p-value < 2.2e-16

## Warning in kpss.test(res): p-value greater than printed p-value

##
```

```
##  KPSS Test for Level Stationarity
##
## data:  res
## KPSS Level = 0.012151, Truncation lag parameter = 5, p-value = 0.1
```

```
## Warning in adf.test(res[which(!is.na(res))]): p-value smaller than printed p-
## value
```




```
##
##  Augmented Dickey-Fuller Test
##
## data:  res[which(!is.na(res))]
## Dickey-Fuller = -9.4455, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.95969, p-value = 9.213e-07
```

Les tests vont dans le sens de la stationnarité. L'ACF et la PACF donnent de moins bon résultat dans le sens où il y a respectivement au moins 6 pics significatifs dont certains aù delà de 10. La série ne semble toujours suivre une loi gaussienne.

Conclusion, les séries différenciées avec les filtres $(I-B^7)$ et $(I-B)(I-B^7)$ remplissent le critère de stationarité. Il est donc possible d'appliquer un modèle de type $SARIMA(p,d,q)\times(P,D,Q)_7$ avec $(d,D)=(0,1)$ ou $(d,D)=(1,1)$.

Modélisation

=> Définition du modèle à développper avec un exemple.

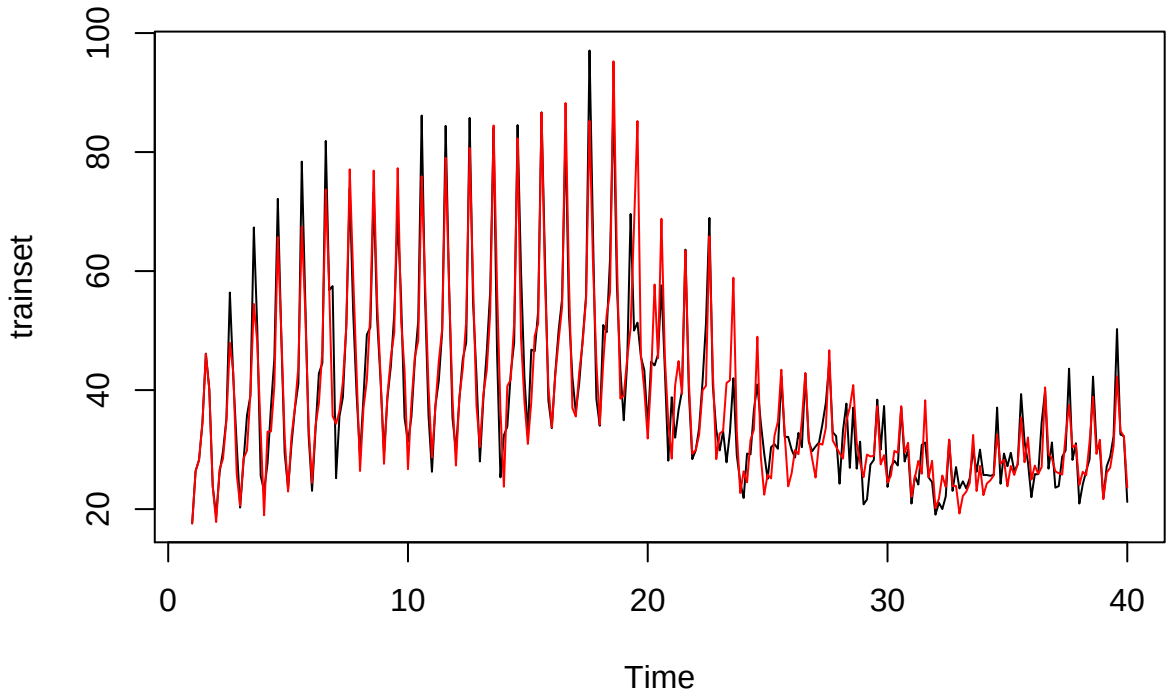
En pratique, on détermine les coefficients avec la méthode auto.arima.

```
model = auto.arima(trainset, allowmean=TRUE, allowdrift=TRUE)
model
```

```
## Series: trainset
## ARIMA(3,0,4)(1,1,1)[7]
##
## Coefficients:
##           ar1     ar2     ar3     ma1      ma2      ma3      ma4    sar1
##       -0.4059  0.7405  0.5775  0.8137  -0.4239  -0.7542  -0.0849  0.0947
## s.e.   0.2322  0.0760  0.2081  0.2432   0.1350   0.1807   0.1276  0.1040
##          sma1
##       -0.6365
## s.e.   0.0811
##
## sigma^2 estimated as 29.58:  log likelihood=-828.23
## AIC=1676.46   AICc=1677.32   BIC=1712.33
```

=> Détails : Tous les termes sont significatifs. On obtient un score BIC de 1796.31.

```
plot(trainset)
lines(model$fitted, col='red')
```



Visuellement, le modèle s'adapte relativement bien aux données. Analysons les résidus de ce premier modèle.

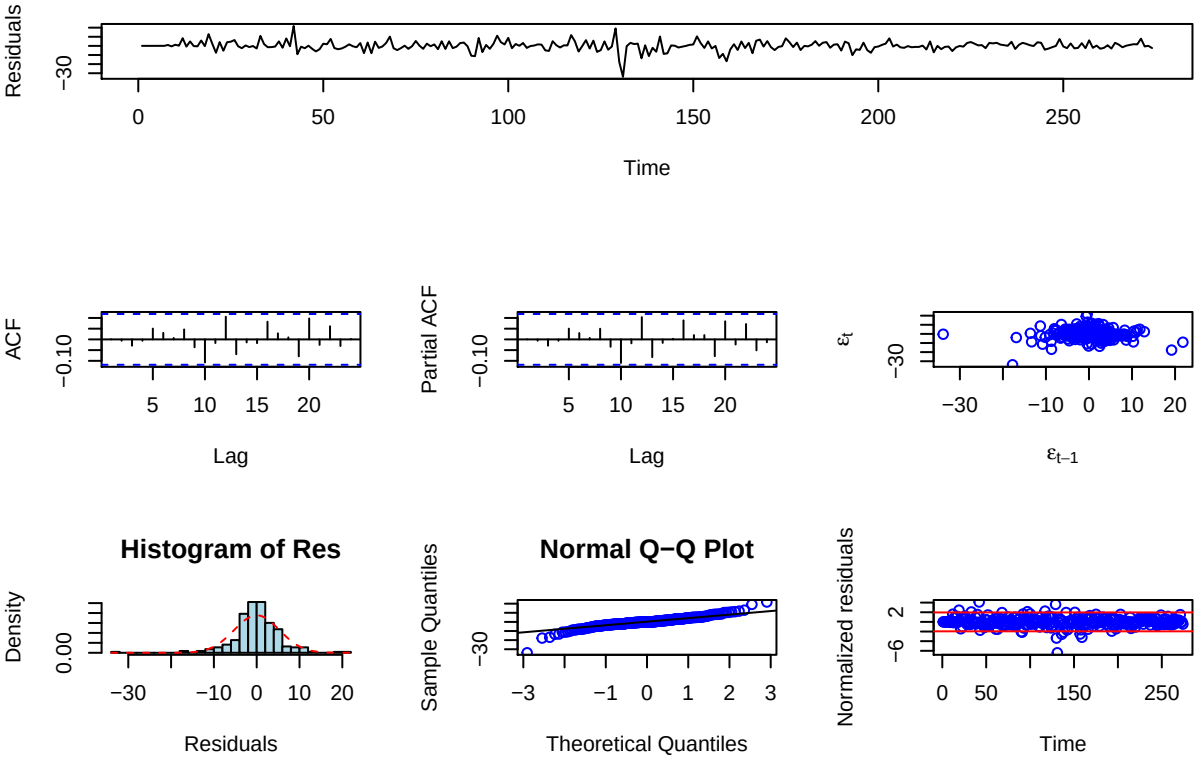
```
check_stat(model$residuals)
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 1.2492, df = 7, p-value = 0.9897
```

```
## Warning in kpss.test(res): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  res
## KPSS Level = 0.18513, Truncation lag parameter = 5, p-value = 0.1
```

```
## Warning in adf.test(res[which(!is.na(res))]): p-value smaller than printed p-
## value
```




```
##
##  Augmented Dickey-Fuller Test
##
## data:  res[which(!is.na(res))]
## Dickey-Fuller = -5.7924, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.91268, p-value = 1.521e-11
```

Pour considérer que le modèle est bon, les résidus doivent former un bruit blanc. Cela se décline en en deux partie.

Stationarité des résidus Box-Ljung : bb au risque de 5% KPSS : pas non-stationnaire ADF : stationnaire Plot : ras ACF : pas de pics significatifs PACF : pas de pics significatifs Lag 1 : indpt

Normalité des résidus Shapiro : pas gaussien Histogramme : pas gaussien Q-Q plot : pas gaussien