Série Chronologique - Projet

Contents

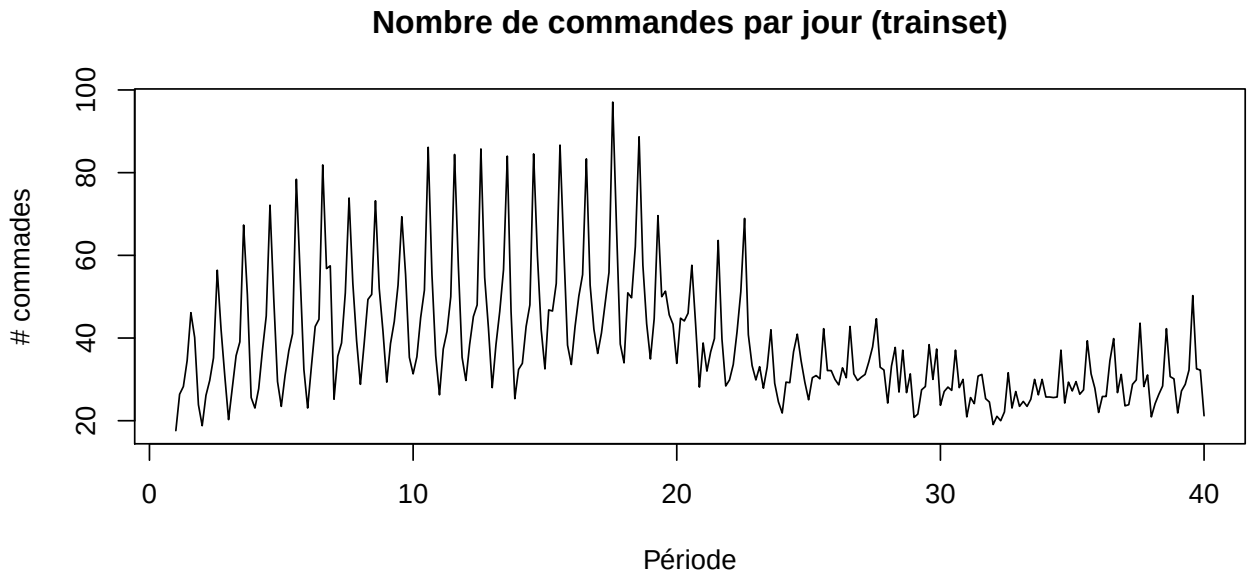

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

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

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

1 Introduction

L'objet d'étude de ce projet est une série temporelle décrivant le nombre de commandes par jour d'une 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 de la famille des modèles $ARMA$ afin de prédire les valeurs du jeu de test soit 21 jours.

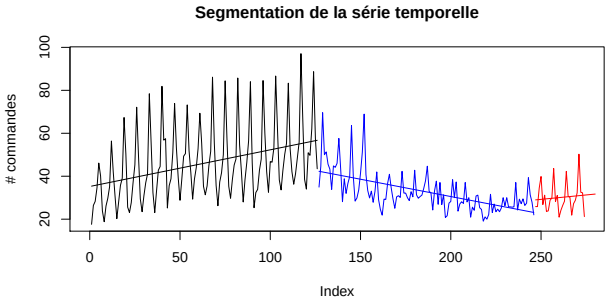


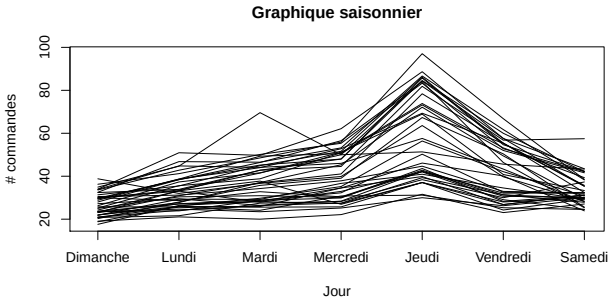
2 Analyse visuelle

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.



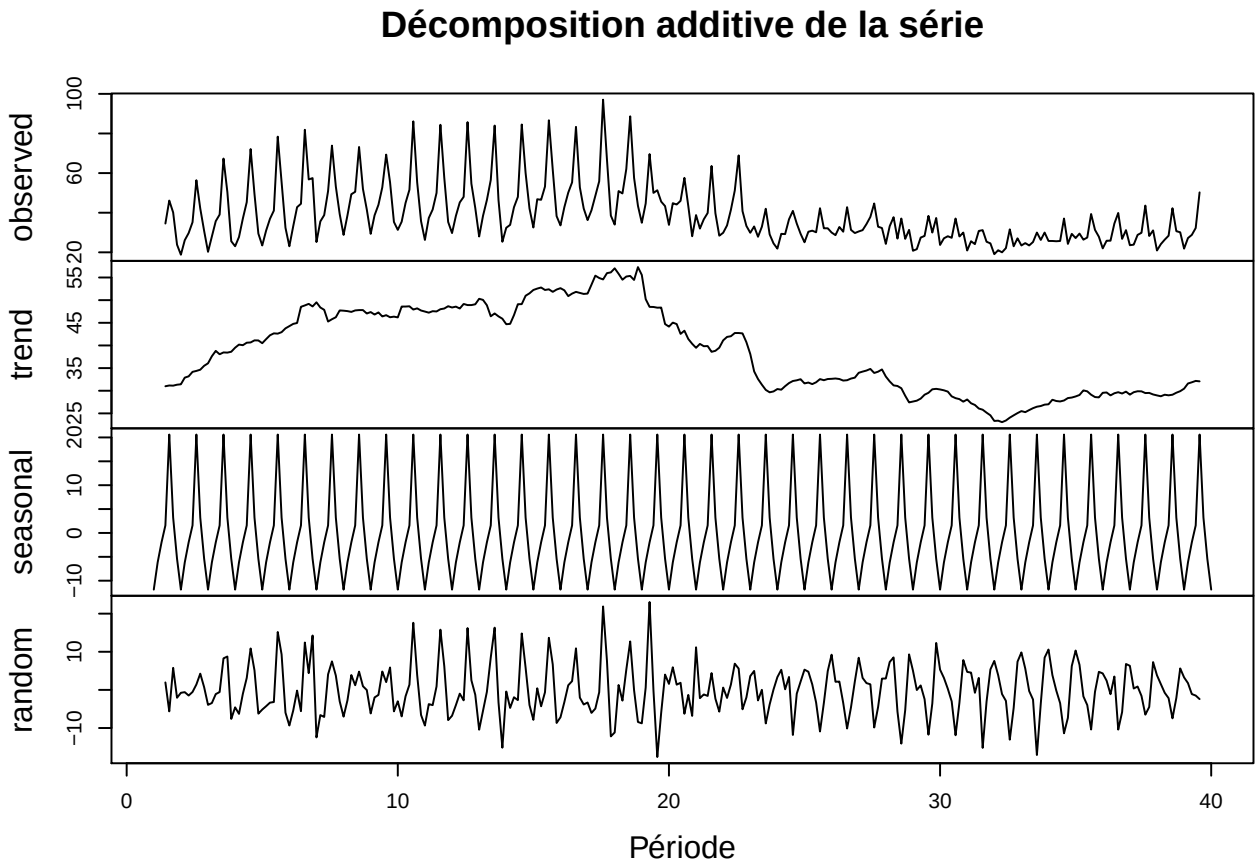


Les modèles $ARMA$ font l'hypothèse que la série temporelle est stationnaire (du second ordre), c'est à dire que sa moyenne et sa covariance sont invariantes par translation dans le temps. Pour rendre cette série stationnaire, nous allons utiliser plusieurs approches telles que la décomposition (partie 2) et la différentiation (partie 3). Puis nous établirons des modèles de la famille $ARIMA$ à l'aide de la fonction `auto.arima` et `checkup_res` (partie 3) avant d'évaluer nos modèles sur le jeu de donnée test avec des critères prédictifs (partie 4).

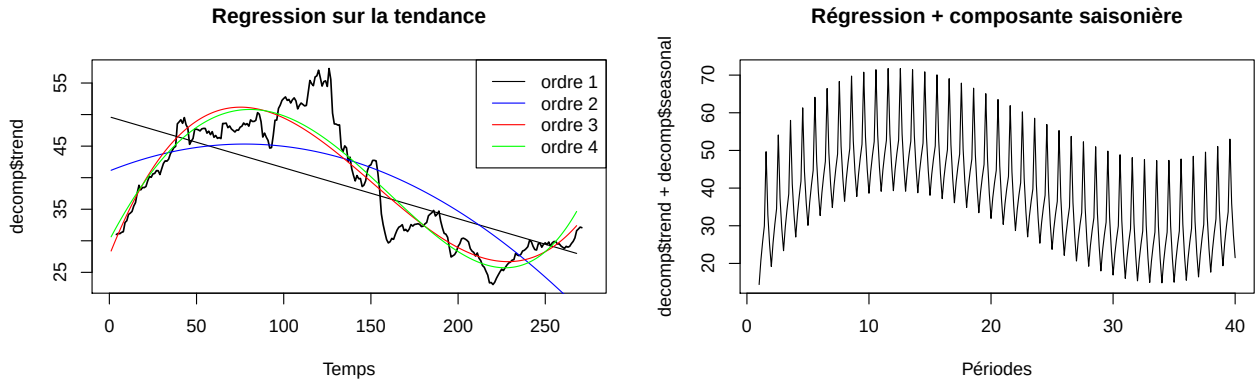
3 Décomposition

Afin de valider cette hypothèse, il est possible d'appliquer la procédure suivante. Dans un premier temps, on applique sur la série un opérateur moyenne mobile $M_m(B) = \frac{1}{2m+1} \sum_{n=1}^{m} B^{-k}$ qui élimine les tendances τ-périodiques avec B l'opérateur de retard tel que $B^k X_t = X_{t-k}$. Puis, on estime par régression la tendance

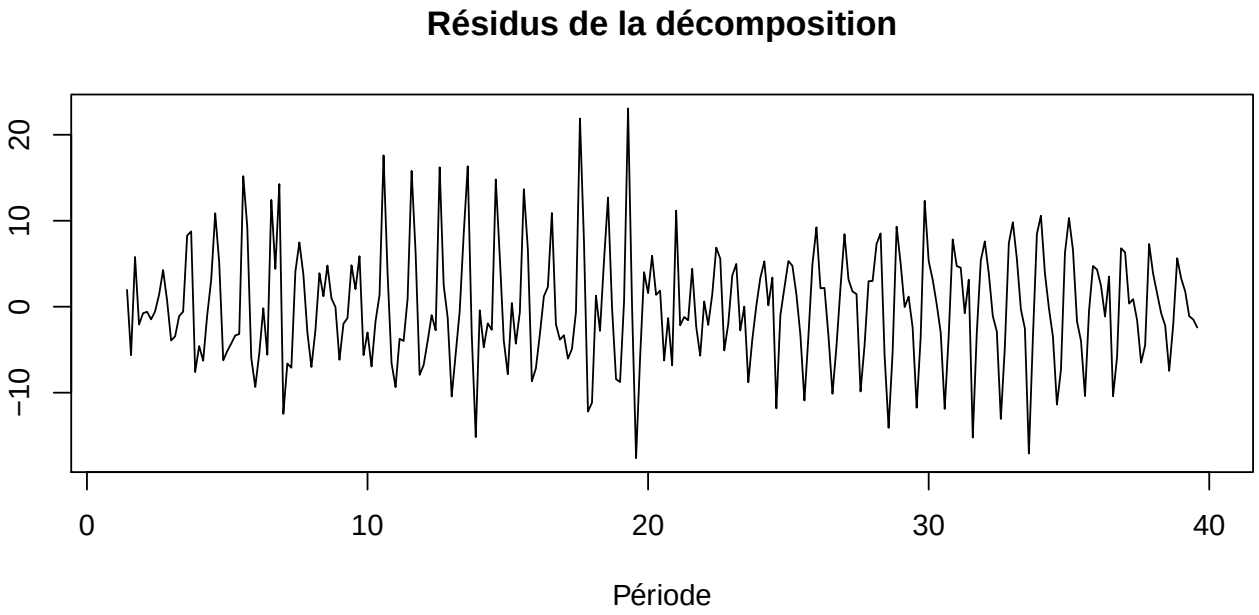
sur la série filtrée. Enfin, on estime la composante saisonnière en effectuant la moyenne des périodes et en dupliquant ce résultat autant de fois que nécéssaire. On obtient la décomposition ci-dessous.



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 avec un R^2 ajusté de 0.88. Le principe de parcimonie nous pousse à garder la régression linéaire d'ordre 3. Analysons à présent les résidus de la décomposition.



Il semble y avoir une légère sésonalité dans les résidus.

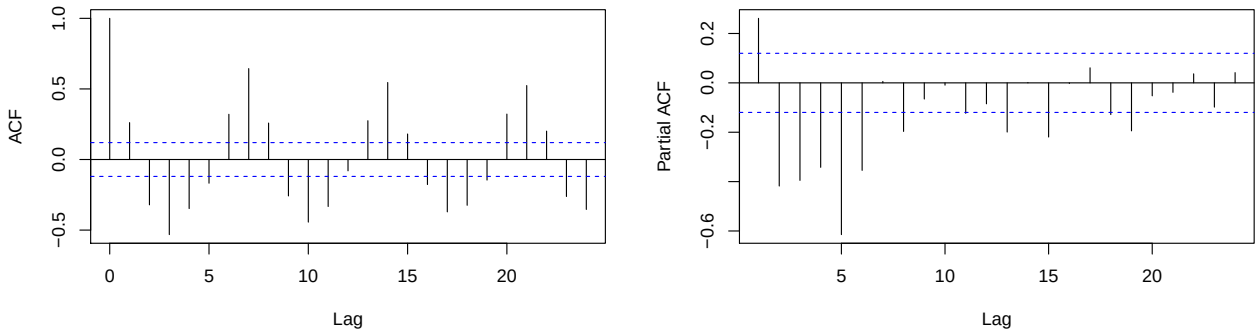
En plus de l'aspect visuel, les tests suivants permettent d'évaluer l'hypothèse de stationarité : - ADF avec $\mathcal{H}_0$: "la trajectoire est issue d'un processus non stationnaire" contre $\mathcal{H}_1 = \bar{\mathcal{H}}_0$. - KPSS avec $\mathcal{H}_0$: "la trajectoire est issue d'un processus stationnaire" contre $\mathcal{H}_1 = \bar{\mathcal{H}}_0$.

Test	Statistique	p-value
ADF	-12.135	<0.01
KPSS	0.020884	>0.1

Les tests s'accordent. Le test ADF rejette l'hypothèse de non stationarité (p-value«0.05) alors que le test KPSS ne rejette pas l'hypothèse de stationarité (p-value»0.05). Nous avons donc obtenu des résidus stionnaires d'après ces tests.

Présentons maintenant deux outils indispensable à l'étude des processus ARMA: 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 $AR(p)$.

De plus, on justifie une tentative de modélisation par un processus $ARMA$ lorsque la série est considérée comme stationnaire et que ses ACF et PACF empiriques montrent une décroissance rapide.



L'ACF présente des pics périodiques majoritairement significatifs avec une faible décroissance. On observe

une corrélation élevée pour les lag multiple de 7. La PACF possède des corrélations partielles significatives aux lags 1 à 6. Ainsi, il est clair que nous encore avons une saisonalité d'ordre 7.

En conclusion, les résidus du modèle de decomposition ne sont pas stationnaires et présentent une saisonalité d'ordre 7.

4 Différenciation

4.1 Différentiation simple

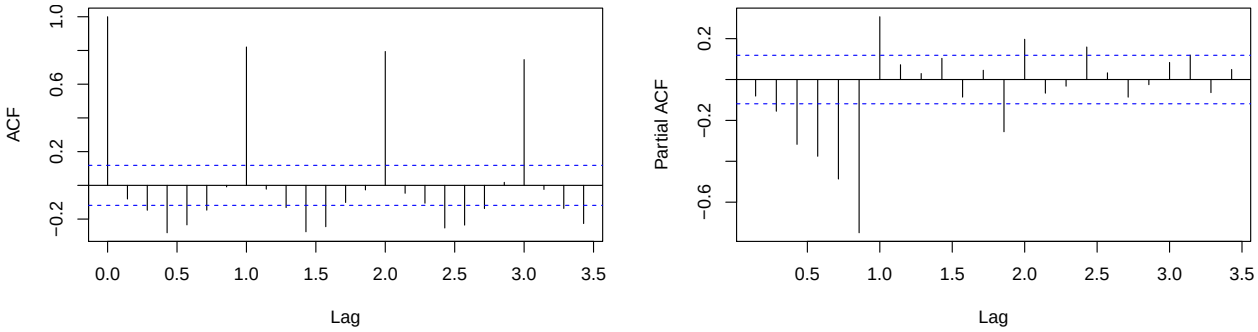
Afin de rendre une série stationnaire, il est possible d'intégrer cette dernière en lui appliquant un filtre de la forme $(I - B)$. On regarde alors si $\Delta X_t = (I - B)X_t$ est stationnaire.

Cette méthodes ne donne pas de résultats satisfaisant sur les résidus de la décomposition (cf. Annexe 1). Nous allons alors appliquer cette méthode sur la série originale.

Les tests de stationarité nous incitent à penser que la série différanciée (ΔX_t) est pas stationnaire.

Test	Statistique	p-value
ADF	-11.875	<0.01
KPSS	0.06673	>0.1

Néanmoins, l'ACF et la PACF présentent repectivement des autocorrélationS très significatives sur les lags 7, 14, 21, 28 et 1 à 7.



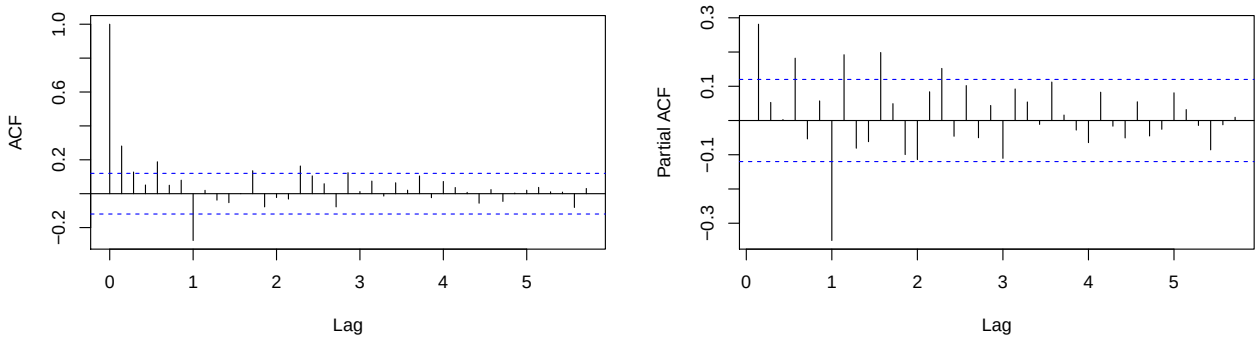
Nous devons conclure que les incréments ne sont pas stationnaires car ils contients encore une composante saisonnière.

4.2 Différentiation saisonnière

La différentiation saisonnière consiste à appliquer un filtre de la forme $(I - B^s)$ à notre série. On obtient alors les incréments saisonniers $(\Delta_s X_t = (I - B^s)X_t)$. L'application de ce filtre sur la série semble tout indiqué car il élimine les tendances s-périodique. Les tests ADF et KPPSS s'accordent sur l'hypothèse de stationarité.

Test	Statistique	p-value
ADF	-11.875	<0.01
KPSS	0.06673	>0.1

L'ACF montre une décroissance rapide avec des pics significatifs en 1, 4 et 7. La PACF possède aussi une décroissance rapide et des pics significatifs en 1 et 7. On pourra alors tenter une modélisation $ARMA$ sur la série $Y_t = \Delta_7 X_t = (I - B^7)X_t$.

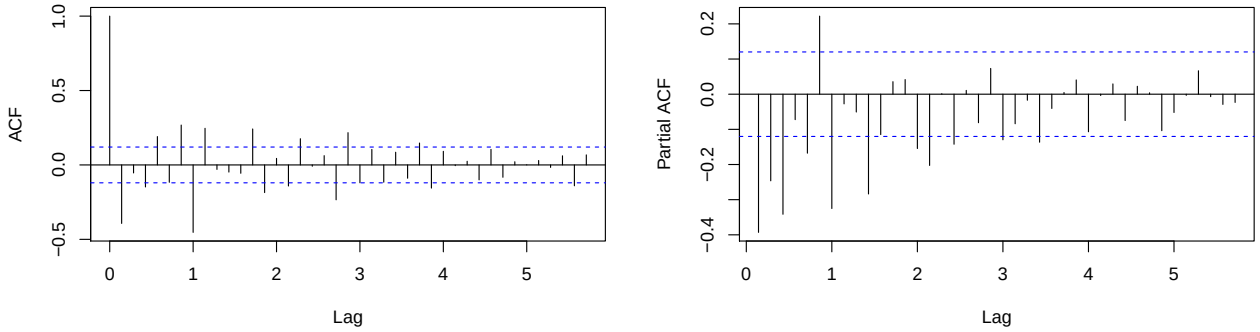


Essayons de différencier à nouveau la série, c'est à dire d'appliquer le filtre $(I - B)(I - B^7)$ sur la série.

Les tests vont encore dans le sens de la stationnarité.

Test	Statistique	p-value
ADF	-11.875	<0.01
KPSS	0.06673	>0.1

L'ACF et la PACF donnent de moins bon résultat dans le sens où l'ACF donne des résultats similaires mais la PACF décroit beaucoup moins vite et possède d'avantage de pics significatifs.



Conclusion, la série différenciée avec le filtre $(I - B^7)$ remplie le critère de stationarité. Il est donc possible d'appliquer un modèle de type $SARIMA(p, d, q)(P, D, Q)_7$ avec $(d, D) = (0, 1)$.

5 Modélisation

Nous avons vu dans la partie précédante qu'une différanciation d'ordre 1 avec le filtre $(I - B^7)$ donne une série stationnaire. Cela motive l'idée d'appliquer un modèle $SARIMA(p, 0, q)(Q, 1, P)_7$. Nous déterminons les coefficients p, q, Q et P avec la méthode `auto.arima`. Il est important d'effectuer la recherche du meilleur modèle avec les paramètres `allowdrift=TRUE` et `include.mean=TRUE` car la différanciation effectuée élimine totalement les tendance d'ordre 1. Le résultat de la recherche est présenté ci-dessous:

```
## Series: trainset
## ARIMA(3,0,4)(1,1,1)[7] with drift
##
## Coefficients:
##           ar1     ar2     ar3     ma1      ma2      ma3      ma4     sar1
##       -0.4062  0.7424  0.5789  0.8136  -0.4264  -0.7571  -0.0860  0.0946
## s.e.   0.2302  0.0761  0.2071  0.2412   0.1341   0.1803   0.1273  0.1039
##          sma1   drift
##       -0.6364  0.0233
## s.e.   0.0811  0.1187
##
```

```
## sigma^2 estimated as 29.69:  log likelihood=-828.21
## AIC=1678.42   AICc=1679.46   BIC=1717.88
```

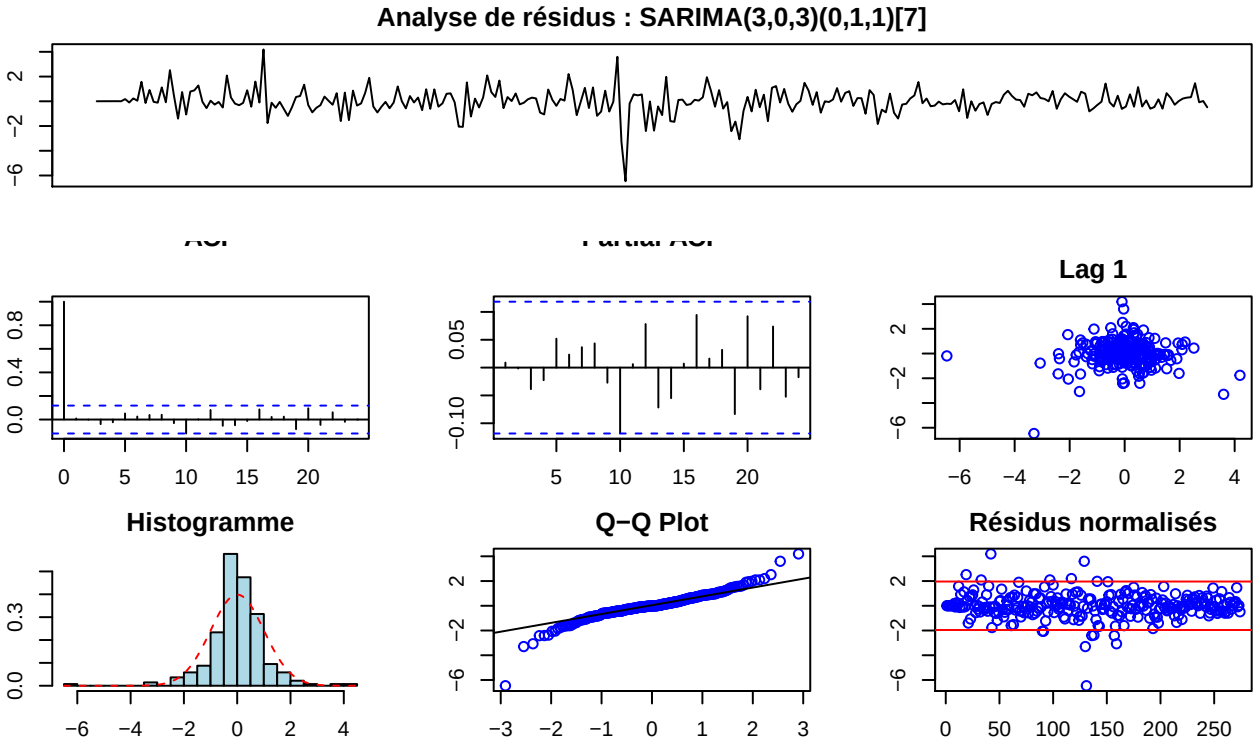
Pour estimer la pertinance du modèle et en particulier la significativité d'un coefficient, on divise la valeurs de ce derniers par son écart type. Le quotient doit se trouver hors de l'intervalle $[-1.96, 196]$ afin que le paramètre soit considéré comme significatif. Les coefficients ma4, sar1 et drift non sont pas significatifs. Testons alors le modèle $SARIMA(3,0,3)(0,1,1)_7$ sans tendance linéaire.

```
## Series: trainset
## ARIMA(3,0,3)(0,1,1)[7]
##
## Coefficients:
##           ar1     ar2     ar3     ma1      ma2      ma3     sma1
##       -0.2260  0.7077  0.4283  0.6253  -0.4771  -0.6145  -0.5843
## s.e.   0.1795  0.0799  0.1416  0.1571   0.0976   0.0923   0.0649
##
## sigma^2 estimated as 29.49:  log likelihood=-828.87
## AIC=1673.75   AICc=1674.31   BIC=1702.45
```

Les coefficients sont tous significatifs. Le score BIC obtenu pour ce modèle est de 1702,45.

Pour considérer que le modèle est bon, les résidus doivent former si possible un bruit blanc. Autrement dit, les résidus réduits doivent être stationnaires, centrés, indépendants et suivre une loi normale. Pour ce faire nous utilisons différants outils statistiques tels que les tests ADF et KPSS présentés précédamment ainsi que les tests de Ljung-Box d'hypothèse nulle $\mathcal{H}_\prime$: "La série temporelle ne possède pas d'autocorrélation." et de Shapiro d'hypothèse nulle $\mathcal{H}_\prime$: "L'échantillon suit une loi normale."

Test	Statistique	p-value
ADF	-5.7081	<0.01
KPSS	0.18528	>0.1
Ljung-Box	1.9463	0.9628
Shapiro	0.91315	1.657e-11



Les résidus de ce modèle 1 semblent tout à fait correspondre à un bruit blanc. En effet, l'ACF et la PACF n'ont aucun pics hors du couloir de non-significativité, les tests ADF et KPSS estiment les résidus stationnaires et le test de Ljung-Box avec une p-value relativement élevée de 0.9628 ne rejette pas l'hypothèse d'indépendance. Néanmoins, leur distribution est légèrement trop concentrée pour coller parfaitement à une distribustion normal. La p-value du test de Shapiro en témoigne.

Ce modèle est relativement satisfaisant. D'autres modèles ont été évalués en Annexe 3, notamment en appliquant des transformations de Box-Cox et logarithmique. L'annexe 4 contient les résultats des modélisation explotant le paramètre `xreg` de la fonction `auto.arima`. Les modèles retenus sont les suivants :

Modèle	Transformation	BIC
	Décompositon	
SARIMA(3,0,3)(0,1,1)[7]	Aucune	1702.45
modèle 2	Log	
modèle 3	Box-Cox	

6 Prédiction

Précédement, la comparaison des modèles était basées sur des critères intégrant la notion de parciemonie avec le score BIC et l'étude des coefficients significatif. Nous portions aussi une grande importance à ce que les résidus du modèle soit un bruit blanc gaussien.

Il est temps de comparer nous modèles sur des critères prédictifs. La procédure d'évaluation consiste en une validations simple. Les scores utilisés pour comparer les modèles sont la RMSE et la métrique MAPE pour Mean Absolute Percentage Error.

7 Annexe 1

8 Annexe 2