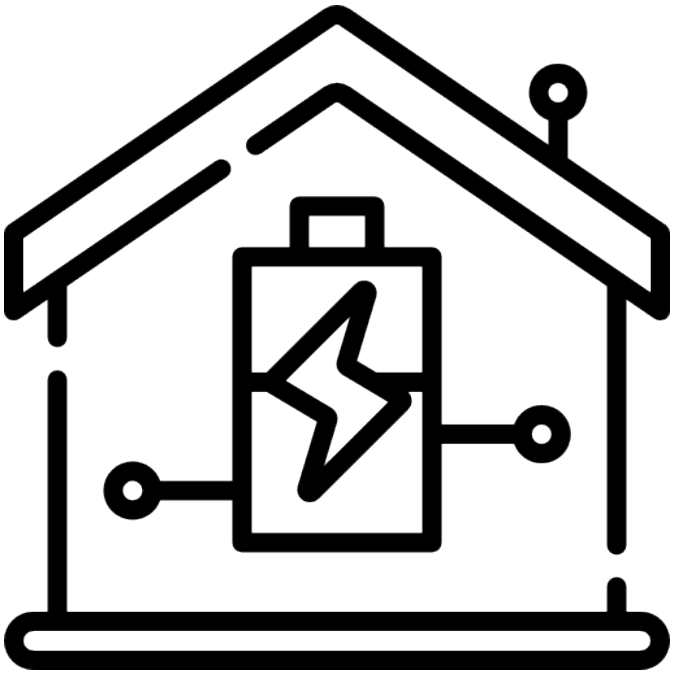
Prévision de la tension électrique d'un ménage

Courant Fanny & Massimi Camille
M2-S1 UE3 Séries temporelles

2021



Sommaire

1 Introduction

Pour notre projet, nous avons analysé un jeu de données d'EDF R&D mesurant la consommation d'énergie électrique d'un ménage. Les données ont été recueillies sur une période de près de 4 ans avec un taux d'échantillonnage par minute. Nous disposons de 2 075 259 mesures recueillies dans une maison située à Sceaux (7 km de Paris, France) entre décembre 2006 et novembre 2010 (47 mois). De plus, notre jeu de données contient 9 variables dont la date et l'heure du relevé.

Comme nous avons beaucoup de données, nous avons décidé de garder seulement les données de la dernière année et de s'intéresser à la variable **voltage**. Celle-ci représente la tension moyenne par minute (en volt). Nos données seront comprises entre le 26 novembre 2009 et le 26 novembre 2010.

L'objectif sera de prédire la tension électrique en Volt des 2 jours suivants.

2 Traitement des données

Les données ont un taux d'échantillonnage par minute. On a donc une très grande quantité de données avec beaucoup de bruit. Cela rend leur analyse difficile. Pour pallier à ce problème, nous avons décidé de nous ramener à un taux d'échantillonnage par heure en utilisant la moyenne.

Après avoir nettoyé et échantillonné les données, on observe une périodicité journalière avec une hausse de la tension électrique le midi et le soir (Figure 1).

Figure 1: Visualisation de la périodicité journalière

Les données comportent un certain nombre de valeurs manquantes. La représentation graphique (Figure 2) nous aide à situer les plages de données manquantes :

1. du 12 janvier 2010 à partir de 14h au 14 janvier 2010, 17h : il manque 52 mesures
2. du 20 mars 2010 à partir de 03h au 21 mars 2010, 11h : il manque 33 mesures
3. du 17 août 2010 à partir de 20h au 22 août 2010, 18h : il manque 119 mesures
4. du 25 septembre 2010 à partir de 02h au 28 septembre 2010, 16h : il manque 87 mesures

Une vérification plus approfondie nous indique la présence de 2 autres mesures manquantes :

1. le 31 octobre 2010 à 01h
2. le 11 novembre 2010 à 00h

Notre série ne présente pas de tendance particulière et les mesures semblent plus ou moins similaires d'une

année sur l'autre. Nous avons donc imputé les données de l'année précédente à nos mesures manquantes. Pour cela, nous avons appliqué la fonction *snaiv()* du package *forecast*. *snaiv()* est une méthode de prévision simple où chaque prévision est égale à sa valeur de l'année précédente. Nous avons maintenant un jeu de données complet, ne comportant plus de valeur manquante (Figure 3).

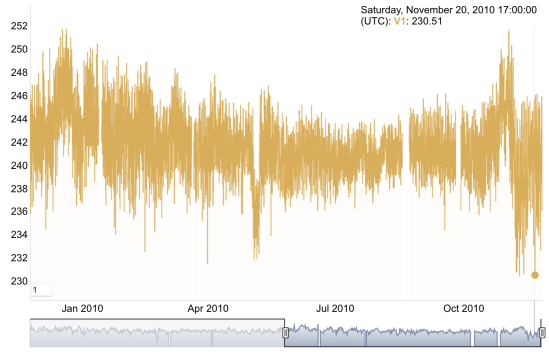


Figure 2: Avant l'imputation des données manquantes

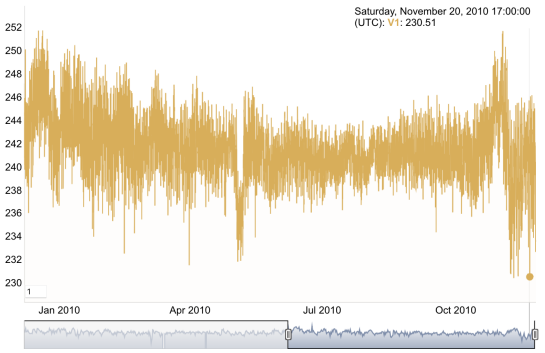


Figure 3: Après l'imputation des données manquantes

Le nombre important de données rend la périodicité difficile à visualiser lorsque l'on prend l'année entière. On constate que la tension électrique augmente légèrement en hiver et baisse à partir du mois de mai. Cette hausse en hiver peut être expliquée par la baisse des températures.

3 Analyses préliminaires

3.1 Stationnarité

Dans un premier temps, nous allons vérifier la stationnarité de la série. Visuellement, la série ne semble pas stationnaire, ceci est notamment dû à la hausse de la tension électrique en hiver. Nous allons effectuer les tests ADF (Dickey-Fuller Augmenté) et KPSS (Kwiatkowski-Philips-Schmidt-Shin) pour confirmer ou infirmer notre hypothèse.

```
Augmented Dickey-Fuller Test

data:  data$Voltage
Dickey-Fuller = -9.3332, Lag order = 20, p-value = 0.01
```

```
KPSS Test for Level Stationarity

data:  data$Voltage
KPSS Level = 14.137, Truncation lag parameter = 12, p-value = 0.01
```

Le test ADF rejette l'hypothèse H0 : "La trajectoire est issue d'un processus non stationnaire" au risque de 1 % et le test KPSS rejette l'hypothèse HO : "la trajectoire est issue d'un processus stationnaire" au risque de 1 %. Les tests se contredisent. La non stationnarité est confirmée. Nous devrons par la suite rendre notre processus stationnaire.

3.2 Les autocorrélations empiriques

Dans cette partie, nous allons analyser les autocorrélations de notre série. Nous pouvons commencer par une première approche graphique en utilisant un lag-plot (graphique de retardement). Celui-ci met successivement en relation chaque observation par rapport aux observations précédentes.

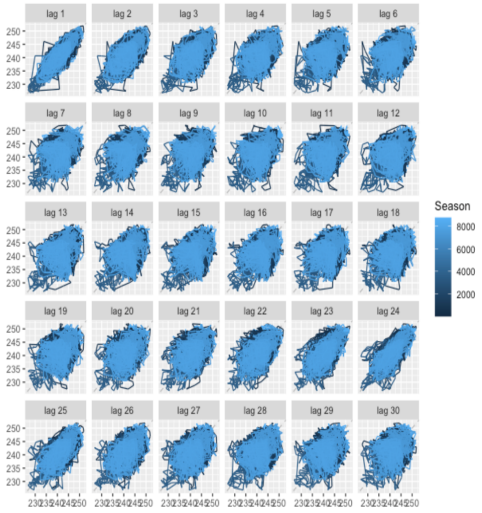


Figure 4: Lag-plot

Le lag-plot (Figure 4) met en évidence une corrélation élevée aux décalages 1 et 24.

Regardons maintenant l'ACF et la PACF.

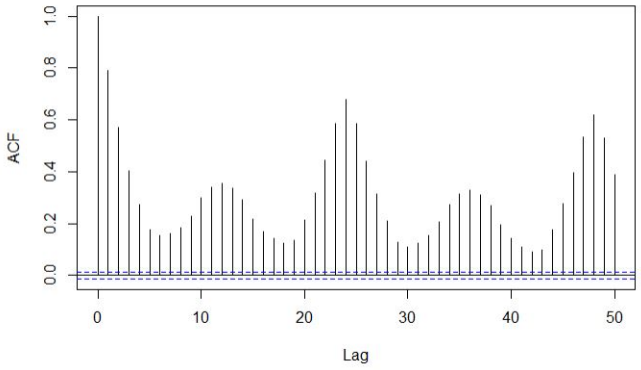


Figure 5: ACF

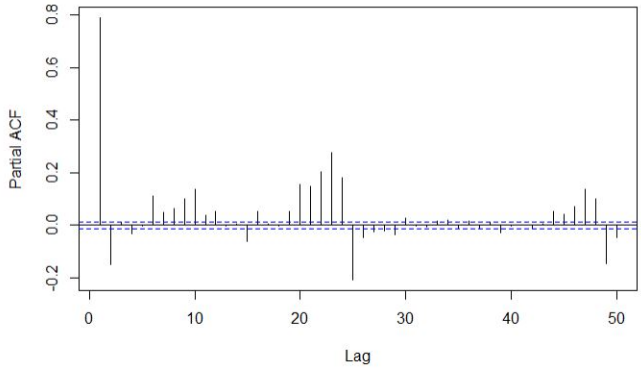


Figure 6: PACF

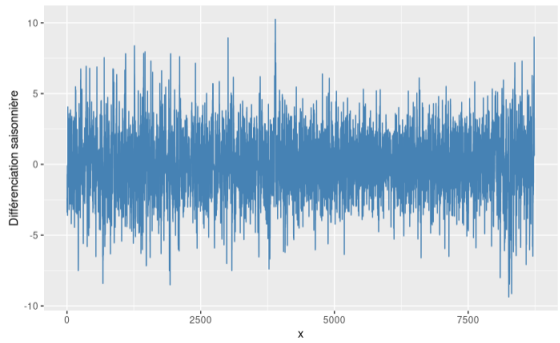
L'ACF montre qu'il existe des corrélations significatives pour les décalages 1 et 24 (Figure 5). Le décalage en 1 semble "évident", la tension mesurée au temps t est naturellement proche de la tension mesurée 1 heure plus tôt. De plus, on observe un motif de période 24, ce qui confirme la présence d'une périodicité et la non stationnarité de la série. La PACF (Figure 6) confirme cette saisonnalité.

Afin de se ramener à un processus stationnaire, nous devons gommer les effets saisonniers. Pour cela, nous pouvons appliquer une différenciation à notre série en utilisant un/des ordre(s) judicieux. C'est ce que nous allons faire dans les sections suivantes.

3.3 Analyse des incréments saisonniers

Le caractère saisonnier de la série nous amène à effectuer une première différenciation saisonnière d'ordre 24.

Visuellement (Figure 7), la stationnarité semble douteuse. Les tests ADF et KPSS semblent indiquer que la série est stationnaire. Regardons l'ACF et la PACF.




```
Augmented Dickey-Fuller Test

data:  VOLT_diff24
Dickey-Fuller = -19.733, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary


KPSS Test for Level Stationarity

data:  VOLT_diff24
KPSS Level=0.0096835, Truncation lag parameter=12, p-value=0.1
```

Figure 7: Série différenciée périodiquement

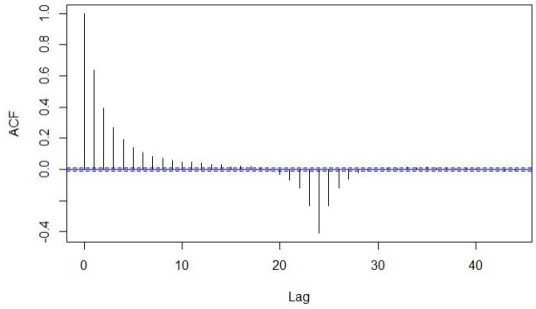


Figure 8: ACF

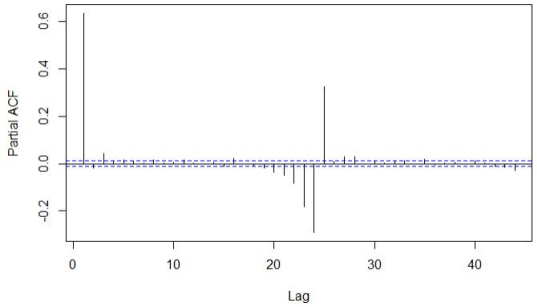
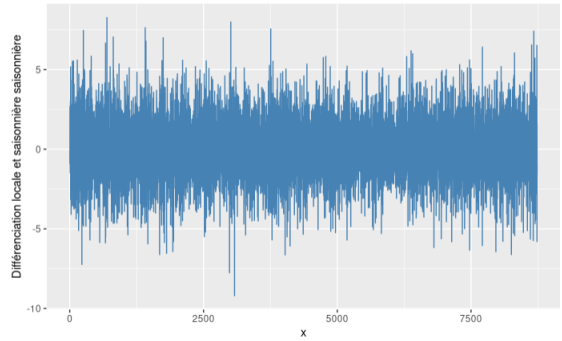


Figure 9: PACF

L'ACF et la PACF ne sont pas caractéristiques d'un ARMA. La stationnarité semble douteuse. Nous explorerons tout de même des modèles SARIMA où d=0 et D=1.

3.4 Analyse des incréments locaux et saisonniers

L'application de la différenciation saisonnière n'étant pas suffisante pour obtenir la stationnarité, nous pouvons effectuer en complément une différenciation locale. En effet, on observe une hausse de la tension électrique en hiver, que nous espérons éliminer avec cette nouvelle différenciation.




```
Augmented Dickey-Fuller Test

data:  VOLT_DIFF24D
Dickey-Fuller = -27.411, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary


KPSS Test for Level Stationarity

data:  VOLT_DIFF24D
KPSS Level=0.0010254, Truncation lag parameter=12, p-value=0.1
```

Figure 10: Série différenciée localement et périodiquement

Visuellement (Figure 10), la série semble stationnaire et les tests ADF et KPSS ne contredisent pas cette hypothèse.

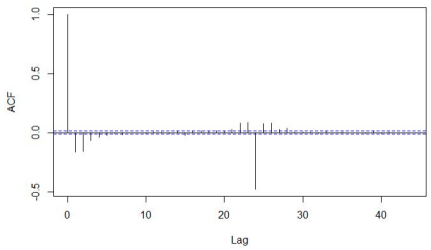


Figure 11: ACF

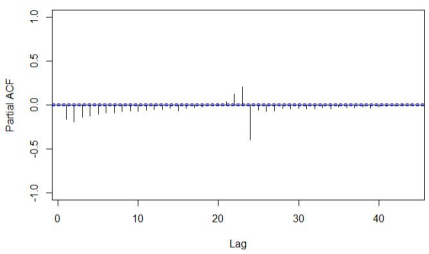


Figure 12: PACF

Les comportements de l'ACF et la PACF deviennent typiques d'un ARMA (décroissance rapide) même si on a des pics en 24. Ce qui nous amènera à explorer des modèles SARIMA avec d=1 et D=1.

4 Les modèles

L'étude des différenciations nous amène à explorer les modèles de type SARIMA où (d,D)=(0,1) et (d,D)=(1,1). Pour le modèle SARIMA où (d,D)=(0,1), nous nous sommes principalement appuyées sur la commande *auto.arima* afin de déterminer les ordres p, q, P et Q. Nous avons choisi de définir le meilleur modèle en minimisant le critère BIC (Bayesian Information Criterion). Pour le modèle où (d,D)=(1,1), nous avons également utilisé cette commande, mais nous nous sommes aussi basée sur l'ACF et la PACF. Ces corrélogrammes nous permettent de définir des combinaisons de modèle à tester où : $p = \{0$, ou plus $\}?$, $q = \{1, 2$ ou plus$\}?$, $P = \{0, 1\}?$ et $Q = 1?$.

En fonction des modèles proposés par *auto.arima* et de nos observations, nous avons étudié des modèles voisins afin d'obtenir le critère prédictif MSE (Mean Squared Error) le plus faible possible. Parmi ces modèles, nous en avons retenu 3 que nous allons détailler ci-dessous. Nous avons appliqué la même démarche d'analyse à tous les modèles testés.

4.1 Modèle 1, issu de la différenciation saisonnière

Le premier modèle est un SARIMA$(1,0,0) \times (1,1,0)_{24}$:

$$(I - B^{24})(I - \alpha_1 B^{24})(I - \phi_1 B)X_t = \epsilon_t$$

```
Series: X
ARIMA(1,0,0)(1,1,0)[24]

Coefficients:
         ar1     sar1
      0.6646  -0.4615
s.e.  0.0080   0.0095

sigma^2 estimated as 2.192:  log likelihood=-15828.53
AIC=31663.06   AICc=31663.07  BIC=31684.29
```

Les estimations de α_1 et ϕ_1 sont significatifs au risque de 5% d'après le test de Student. En effet, les quotients $|\frac{0.6646}{0.0080}| = 83,075$ et $|\frac{-0.4615}{0.0095}| \approx 48,58$ sont supérieurs à 1,96.

Regardons les résidus du modèle :

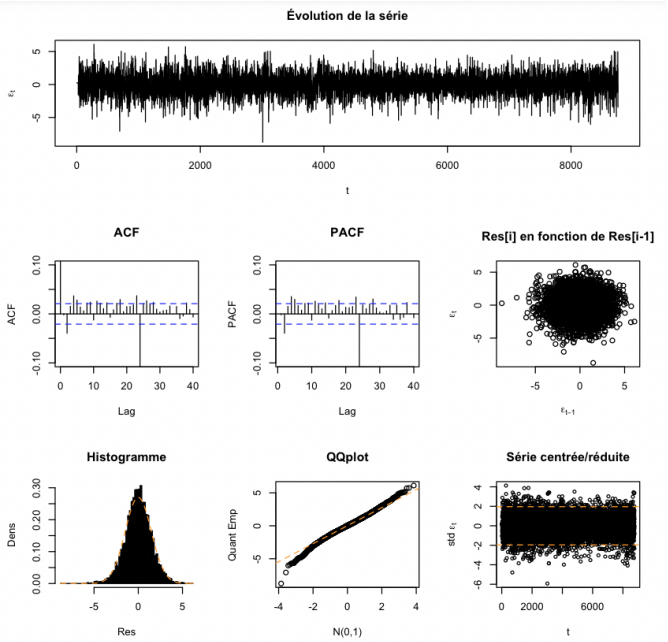


Figure 13: Analyse des résidus du modèle 1

Graphiquement, la normalité et la blancheur des résidus sont vérifiées. En effet, l'ACF et la PACF présentent des pics non significatifs, ce qui est caractéristique d'un bruit blanc. De plus, on peut voir qu'il n'y a pas de tendance dans les résidus (graphique 3 de la ligne 2). L'histogramme et le QQplot montrent que la densité des résidus suit une loi normale. Le dernier graphique en bas à droite indique que l'écart type des résidus se situe dans l'intervalle de confiance 0. Graphiquement, les résidus suivent une loi normale et sont semblables à un bruit blanc. Mais ces hypothèses ne sont pas confirmées par les tests suivants :

```
Shapiro-Wilk normality test

data:  Mod1$residuals[0:5000]
W = 0.99613, p-value = 3.55e-10
```

```
Box-Ljung test

data:  Mod1$residuals
X-squared = 248.44, df = 24, p-value < 2.2e-16
```

En effet, le test de Shapiro Wilk rejette l'hypothèse de normalité des résidus et la blancheur des résidus est rejetée par le test de test de Ljung-Box.

Afin de pouvoir comparer ce modèle aux autres, nous allons calculer la $MSE = \frac{1}{s}\sum_{k=1}^{s}(X_{n-k+1} - \hat{X}_{n-k+1})^2$, l'écart quadratique moyen entre les observations et les prédictions sur les 2 dernières périodes de notre modèle (dans notre cas, 1 période est égale à 1 jour). Pour cela, nous tronquons notre série en enlevant les 2 dernières périodes. Puis, nous entraînons notre modèle sur cette série tronquée afin de prédire les périodes tronquées (Figure 14). Nous obtenons une MSE de 4.656. Visuellement, les prédictions sont plutôt bonnes.

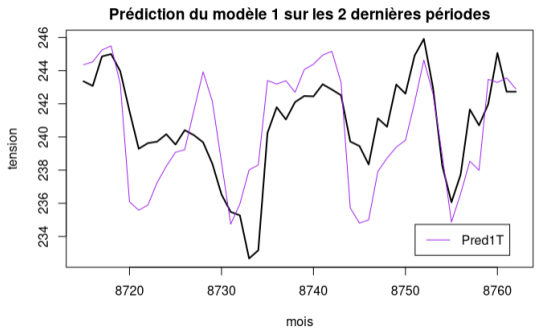


Figure 14: Prédictions du modèle 1 sur les 2 dernières périodes

4.2 Modèle 2 et 3, issus des différentiation saisonnière et locale

Le second modèle que nous allons étudier est un SARIMA$(0,1,1) \times (0,1,1)_{24}$:

$$(I - B^{24})(I - B)X_t = (I + \beta_1 B^{24})(I + \theta_1 B)\epsilon_t$$

```
Series: data$Voltage
ARIMA(0,1,1)(0,1,1)[24]

Coefficients:
          ma1     sma1
      -0.2687  -0.8928
s.e.   0.0147   0.0054

sigma^2 estimated as 1.897:  log likelihood=-15211.32
AIC=30428.64   AICc=30428.64   BIC=30449.87
```

Les estimations de β_1 et θ_1 sont significatifs au risque de 5%.

Regardons les résidus du modèle :

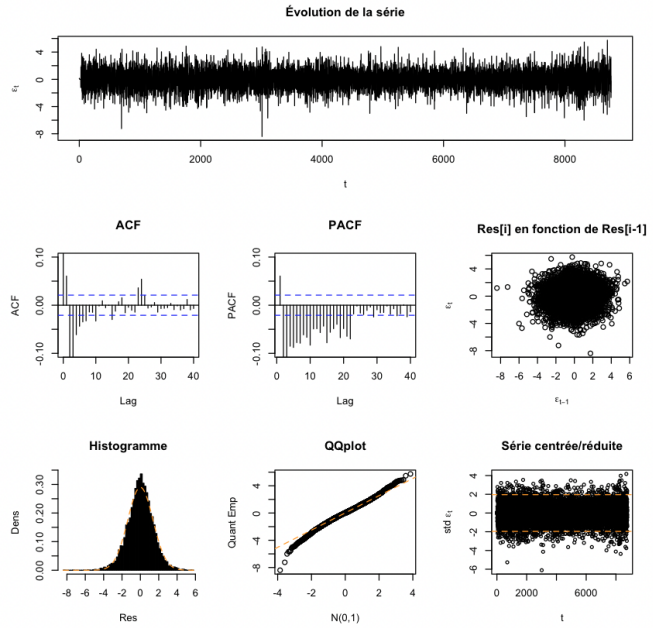


Figure 15: Analyse des résidus du modèle 2

Graphiquement, la normalité et la blancheur des résidus semblent vérifiées. Mais ces hypothèses ne sont pas confirmées par les tests suivants :

```
Shapiro-Wilk normality test

data:  Mod2$residuals[0:5000]
W = 0.99654, p-value = 2.127e-09
```

```
Box-Ljung test

data:  Mod2$residuals
X-squared = 625.55, df = 24, p-value < 2.2e-16
```

De la même façon que précédemment, nous calculons la MSE du modèle, qui est de 4.331. Les prédictions de ce modèle sont proches du modèle précédent.

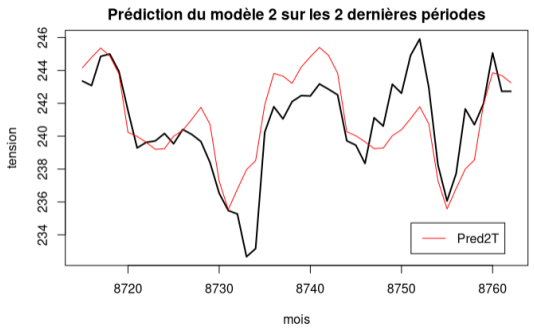


Figure 16: Prédictions du modèle 2 sur les 2 dernières périodes

Le troisième modèle est un SARIMA$(3,1,1)\times(1,1,1)_{24}$:

$$(I-B^{24})(I-B)(I-\alpha_1 B^{24})(I-\phi_1 B-\phi_2 B^2-\phi_3 B^3)X_t=(I+\beta_1 B^{24})(I+\theta_1 B)\epsilon_t$$

```
Series: data$Voltage
ARIMA(3,1,1)(1,1,1)[24]

Coefficients:
         ar1      ar2     ar3      ma1     sar1     sma1
      0.6596  -0.0647  0.0397  -0.9647  0.0700  -0.9098
s.e.  0.0113   0.0128  0.0111   0.0036  0.0122   0.0053

sigma^2 estimated as 1.658:  log likelihood=-14623.16
AIC=29260.31   AICc=29260.33   BIC=29309.84
```

Les estimations de α_1, ϕ_1, ϕ_2, ϕ_3, β_1, θ_1 sont significatifs au risque de 5%.

Regardons les résidus du modèle :

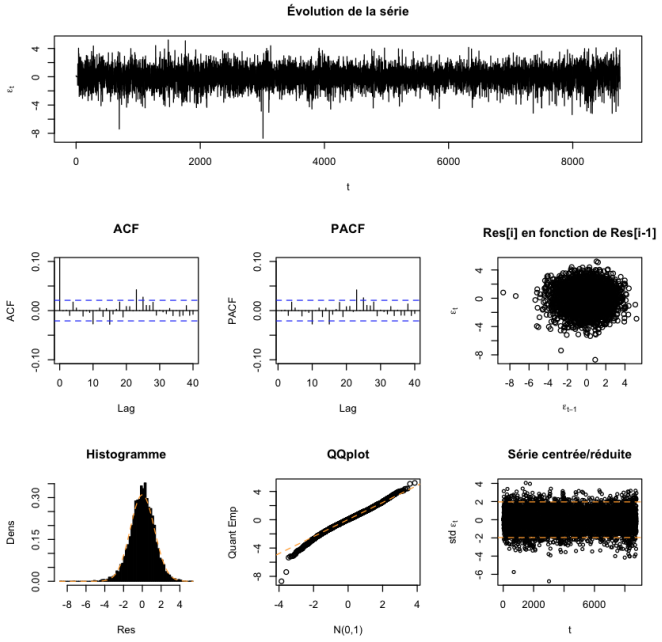


Figure 17: Analyse des résidus du modèle 3

Graphiquement la normalité et la blancheur des résidus sont vérifiées. Mais encore une fois, ces hypothèses ne sont pas confirmées par les tests suivants :

```
Shapiro-Wilk normality test

data:  Mod3$residuals[0:5000]
W = 0.99536, p-value = 1.564e-11
```

```
Box-Ljung test

data:  Mod3$residuals
X-squared = 41.326, df = 24, p-value = 0.01534
```

La MSE de ce modèle est de 4.622. Les prédictions de ce modèle sont également semblables aux autres modèles.

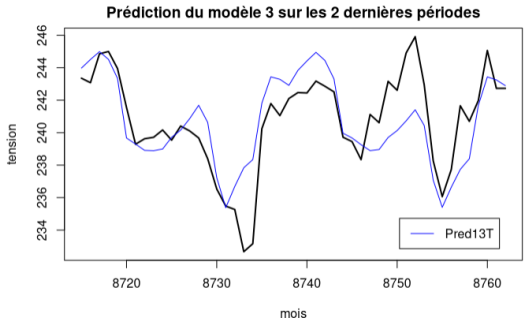


Figure 18: Prédictions du modèle 3 sur sur les 2 dernières périodes

4.3 Récapitulatif, et choix du modèle

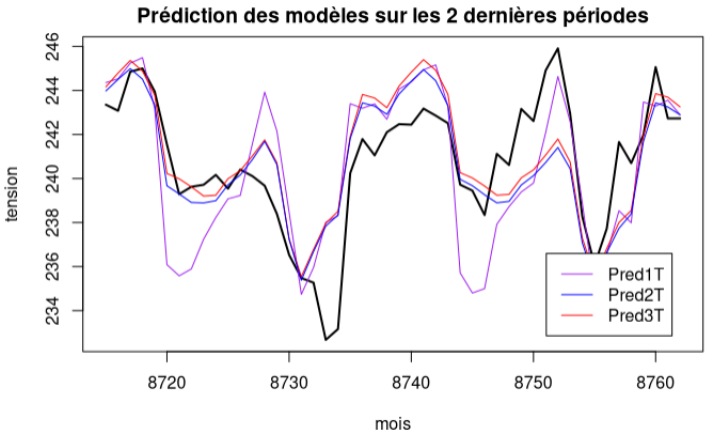


Figure 19: Prédictions des modèles sur sur les 2 dernières périodes

Mod1.bic	Mod2.bic	Mod3.bic
31684.29	30449.87	29309.84

MSE1	MSE2	MSE3
6.69112	4.032953	4.069899

Figure 20: BIC et MSE

Les modèles présentés ci-dessus donnent des prédictions assez semblables. Nous choisirons comme modèle celui avec la MSE la plus faible, même si elles sont très proches. Nous allons donc estimer les prédictions des 2 prochaines périodes avec le modèle 2, qui est le SARIMA$(0,1,1) \times (0,1,1)_{24}$.

5 Prédiction de la tension électrique

Nous obtenons les prédictions suivantes, pour les 2 prochaines périodes, avec le modèle 2 :

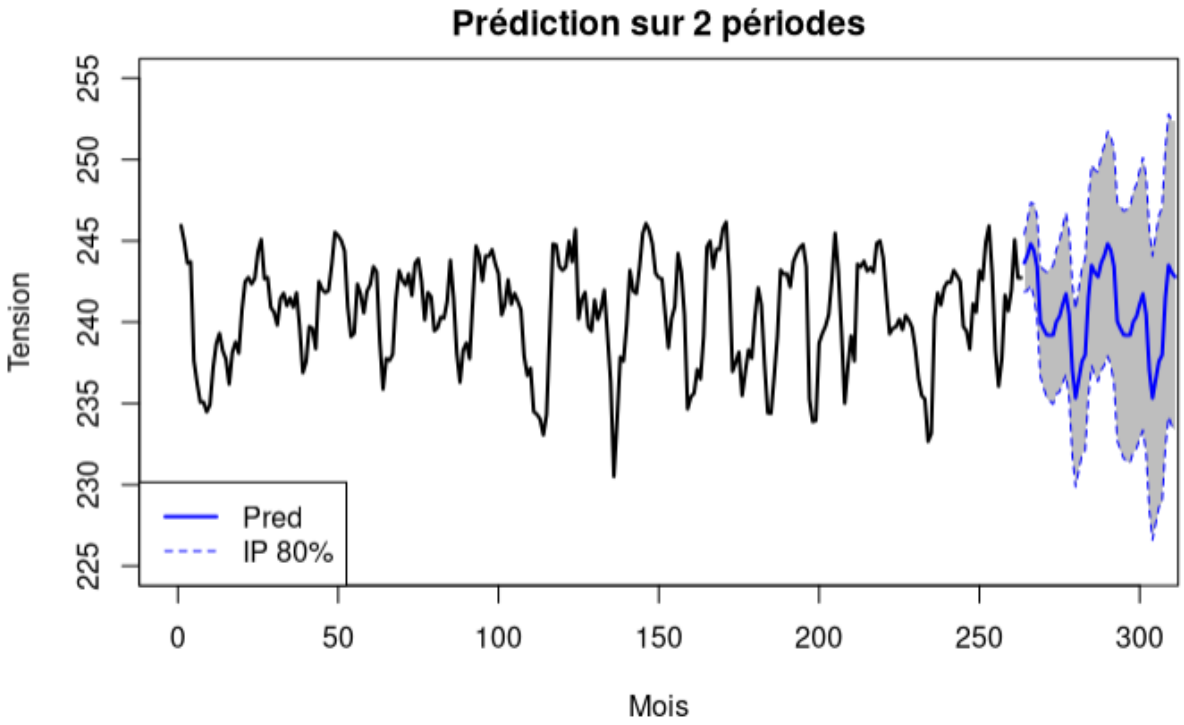


Figure 21: Prédiction sur les 2 prochaines périodes

Notre série présente beaucoup de bruit au niveau du motif saisonnier. Ce qui peut expliquer une amplitude plutôt élevée de l'intervalle de confiance. On remarque également que celle-ci augmente lors de la prédiction de la 2^{de} période par rapport à la 1^{re}. Ce qui semble cohérent.