# Projet Séries Chronologiques

CORDIER Mathis - MADELAINE Céline

Nous disposons dans cette étude de 2 vecteurs composés chacun de 43824 données :

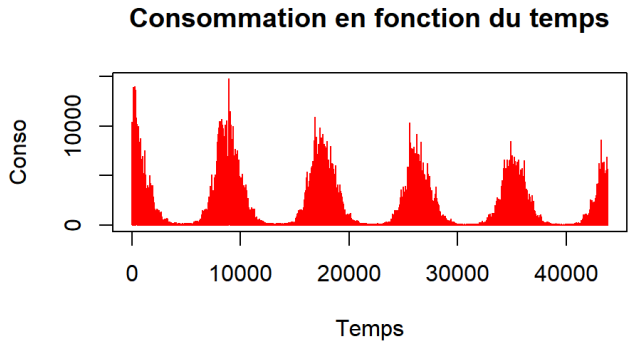
- Conso qui contient la consommation électrique horaire d'un foyer
- Temp qui contient la température extérieure horaire de la station la plus proche de ce foyer

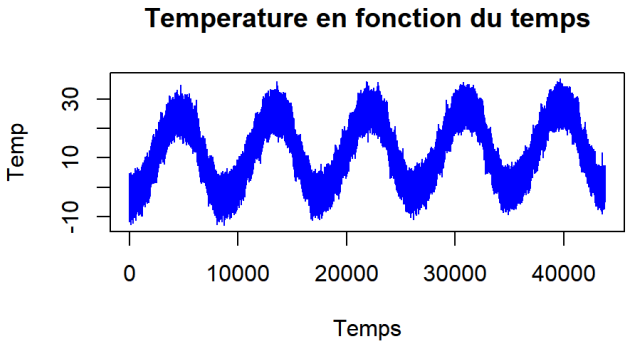
# I - Etude des données horaires

Nous allons dans cette partie chercher à prédire la consommation horaire du foyer pour la première semaine de janvier 2018.

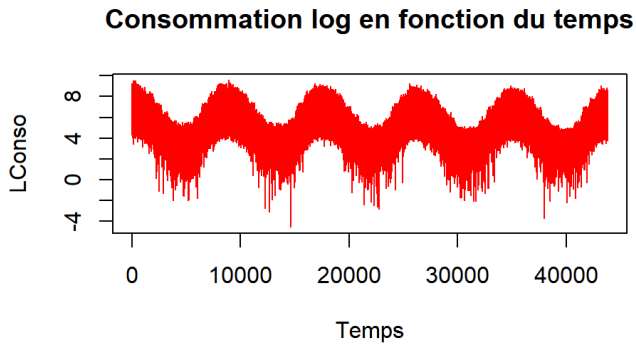
## 1) Premières observations - Statistiques descriptives

En étudiant les statistiques descriptives de nos séries et les graphiques ci-dessous, nous avons fait le choix d'étudier LConso obtenue par transformation de Box-Cox de Conso (aux valeurs positives) plutôt que cette dernière.

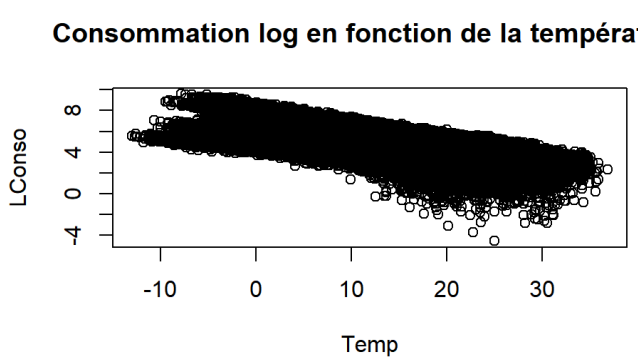


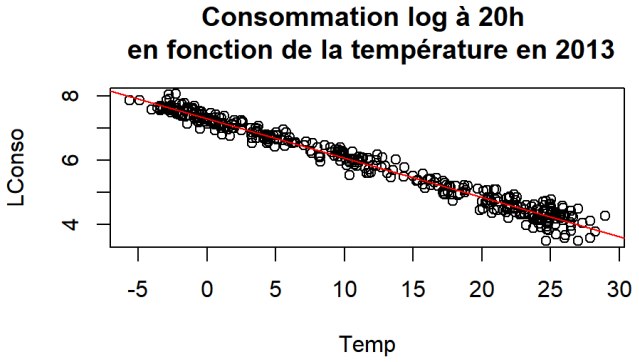




La représentation ci-dessous de notre série LConso nous montre qu'elle semble adaptée à notre problème.

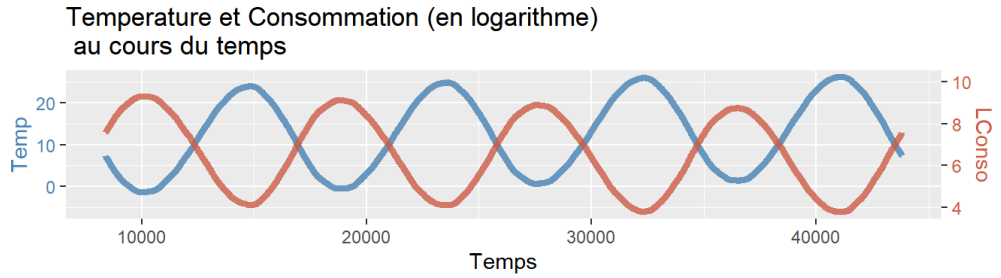



Comme nous pouvons le constater sur le graphique ci-dessous, il y a une relation linéaire entre la température et la consommation (en logarithme). Cela est confirmé par la corrélation fortement négative de -0.66 entre nos 2 séries. Cette relation est d'autant plus mise en avant si nous observons nos données à un instant précis de la journée comme c'est le cas sur le graphique de droite ci-dessous.

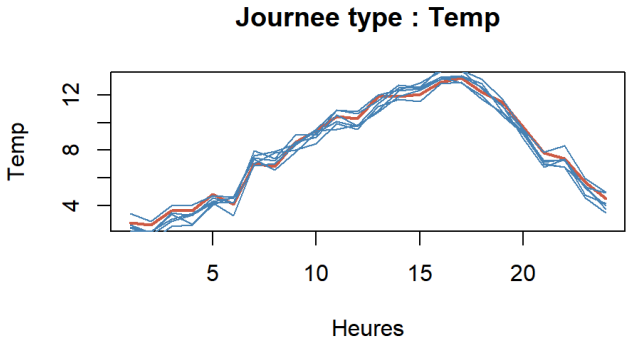


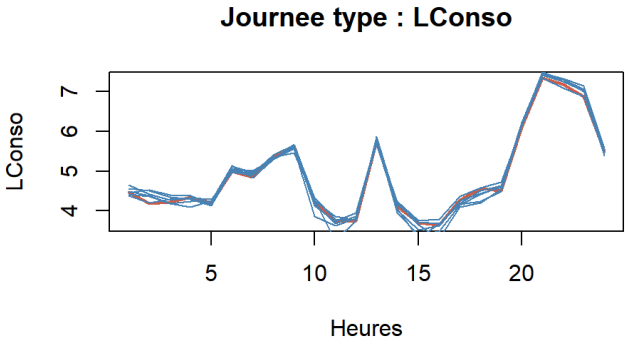



En superposant nos 2 séries, nous constatons à nouveau qu'il y a un lien entre elles.

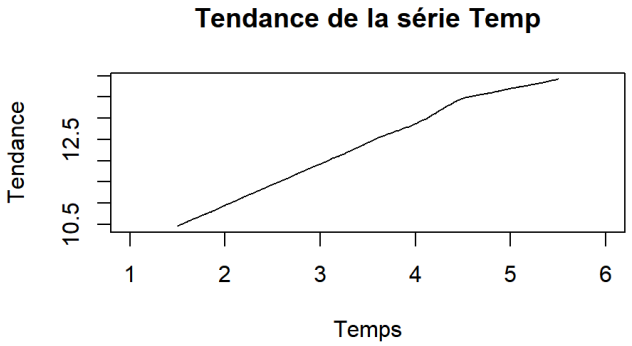


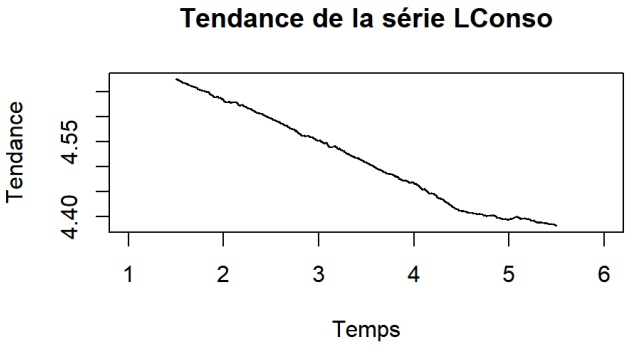

De plus, nous remarquons une forte périodicité au sein de Temp ainsi qu'au sein de LConso. Les graphiques ci-dessous représentent une superposition pour chaque jour de la semaine de la moyenne par heure de la consommation (en logarithme) et de la température. Nous constatons que notre hypothèse selon laquelle la consommation du dimanche (en rouge sur nos graphiques) serait différente des autres jours n'est pas validée. En effet, la périodicité semble de 24h. De plus, il y a des pics de consommation le matin, le midi, et le soir.




Nous remarquons également qu'entre 2013 et 2017, la température a augmenté et la consommation (en logarithme) a diminué :



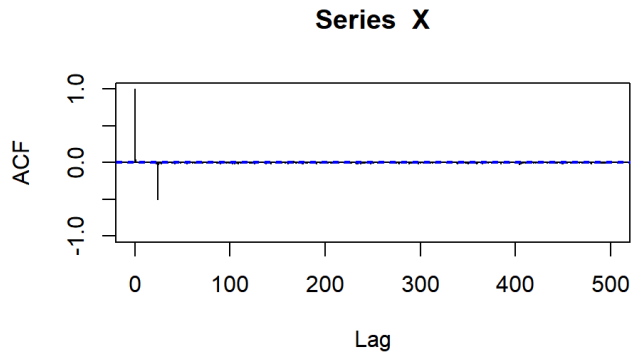

## 2) Etude de la stationnarité de LConso

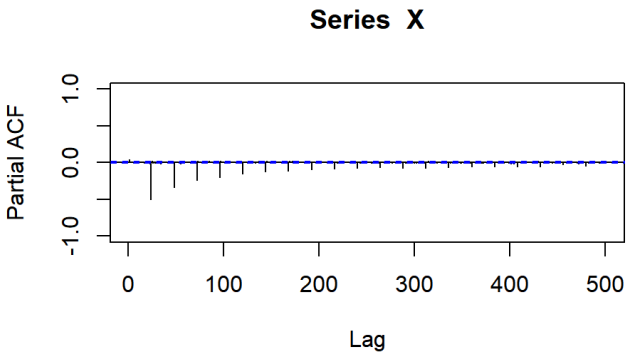
Afin d'obtenir la stationnarité dans notre modèle, nous avons différencié notre série selon les 3 combinaisons suivantes :

- d=1, D=0
- d=1, D=1
- d=0, D=1

Nous en avons conclu que la série sera stationnaire en prenant D=1. En effet, les tests ADF et KPSS ainsi que les ACF et PACF nous orientent vers ce choix.

Par exemple, dans le cas d=0 et D=1, l'ACF et la PACF ci-dessous montrent que le modèle est bien stationnaire. De plus, il semble judicieux de s'orienter vers P=0 et Q=1 : l'ACF présente un unique pic en 24 (hormis celui en 0) et la PACF s'éteint à partir d'un certain rang.




# 3) Etude de la stationnarité de Temp

Nous avons fait de même pour la série Temp et les 3 combinaisons ci-dessus valideront la stationnarité de notre série dans notre modèle. En effet, les 3 mènent à la stationnarité de la série.

# 4) Modèle sur données connues

Dans un but d'économie de calculs, nous avons conservé uniquement les données des 52 dernières semaines.

## A - Choix du modèle pour LConso

Dans un premier temps, nous avons cherché le modèle le plus approprié pour prédire LConso pour les 2 dernières semaines de décembre en prenant en compte la série Temp. Pour cela, nous avons conservé d'une part les données de la semaine 1 à la semaine 50 et d'autre part les données des semaines 51 et 52, pour les séries LConso et Temp réduites à la dernière année.

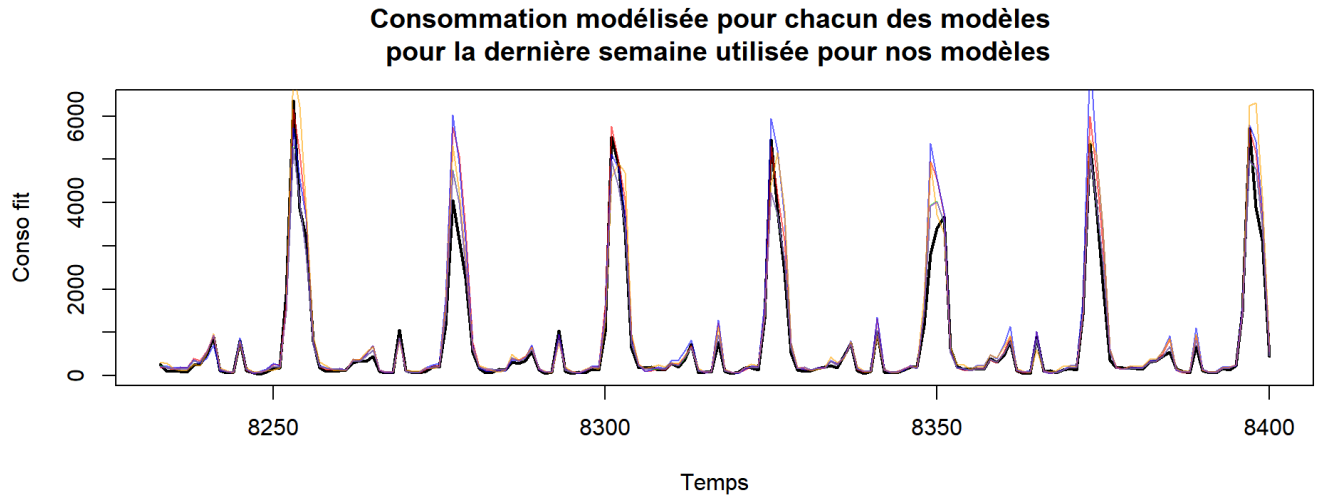
Nous sélectionnerons le modèle le plus pertinent selon les critères suivants : la normalité des résidus, la blancheur des résidus, la significativité des paramètres, la longueur de la mémoire, l'AIC, le BIC, la logvraisemblance, la MSE, la MAPE, la représentation graphique de notre prédiction. Voici les 5 modèles (avec régression linéaire sur la température) que nous mettons en comparaison :

- SARIMA(0,0,0)(2,1,0)[24] : proposé par la fonction auto.arima en imposant d=0, D=1
- SARIMA(5,1,0)(2,1,0)[24] : proposé par la fonction auto.arima en imposant d=1, D=1
- SARIMA(3,0,0)(0,1,1)[24] : proposé par tatonnement avec d=0, D=1, P=0, Q=1
- SARIMA(1,1,0)(0,1,1)[24] : proposé par tatonnement avec d=1, D=1, P=0, Q=1
- SARIMA(2,0,2)(0,1,1)[24] : proposé par tatonnement avec d=0, D=1, P=0, Q=1

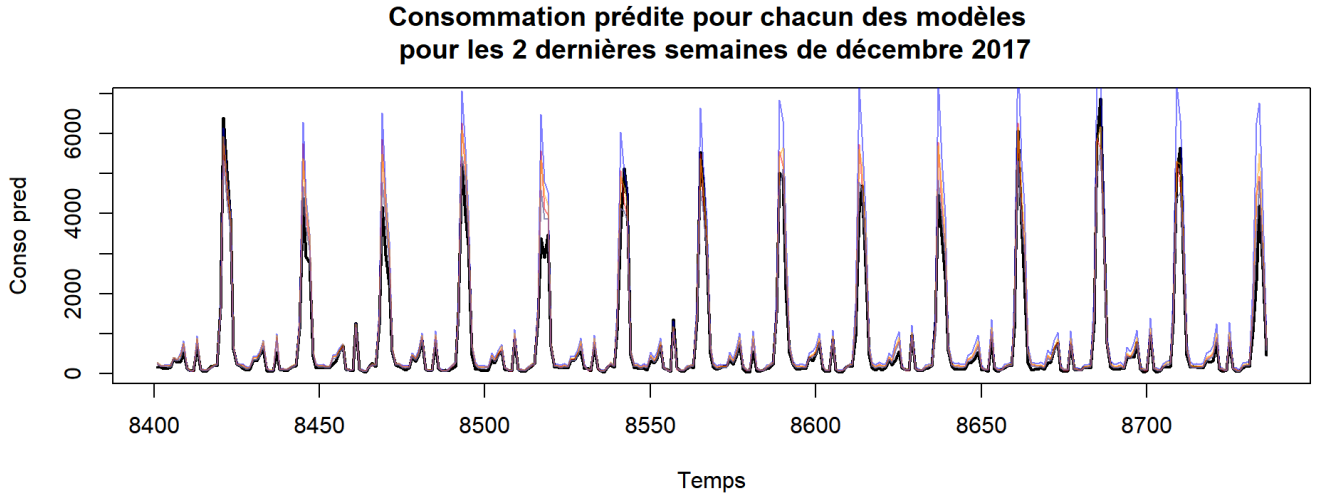
Nous constatons qu'aucun des modèles ne semble avoir des résidus gaussiens au seuil de 5%. Nous allons donc nous baser sur les autres critères pour faire notre choix de modèle dont nous présenterons ultérieurement un diagnostic de la normalité des résidus.

Les modèles 3 et 5 semblent les plus intéressants au sens de la blancheur des résidus.

En reprédisant les 50 semaines nous ayant servi dans notre modèle, nous obtenons ce graphique avec les vraies valeurs (en noir), les valeurs obtenues avec le modèle 1 (en rouge), le modèle 2 (en bleu), le modèle 3 (en vert), le modèle 4 (en orange) et le modèle 5 (en violet) :




Nous allons maintenant prédire les valeurs de Conso pour les 2 dernières semaines de décembre afin de voir si l'un des modèles est mis en avant pour sa bonne prédiction. Voici le graphique associé avec les vraies valeurs (en noir), les valeurs obtenues avec le modèle 1 (en rouge), le modèle 2 (en bleu), le modèle 3 (en vert), le modèle 4 (en orange) et le modèle 5 (en violet):




Les modèles 1 et 2 ne semblent pas les mieux adaptés pour la prédiction des données inconnues.

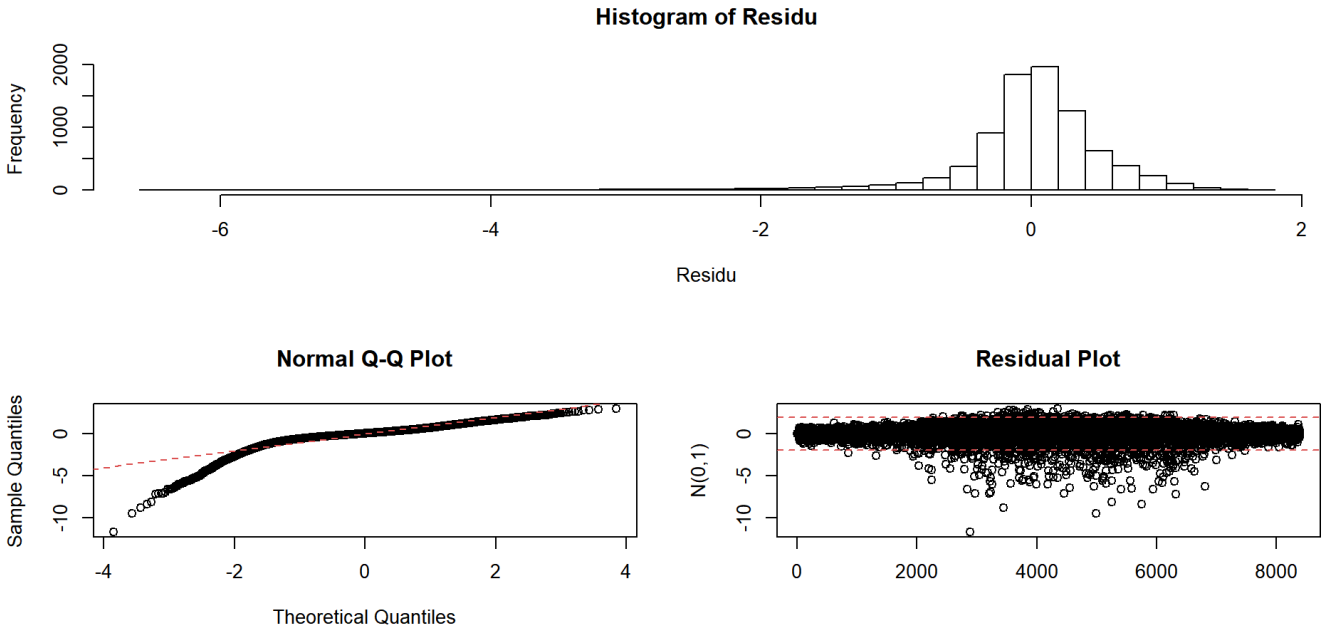
Voici le tableau récapitulatif de nos critères, précisant l'AIC, le BIC, la log-vraisemblance, la MSE, la MAPE, le nombre de paramètres significatifs et le nombre de paramètres non significatifs :

```
##        AIC      BIC    LogLik       MSE     MAPE N signif N non signif
## 1 15878.96 15907.10 -7935.482  39442410 30.60113        3            0
## 2 17061.89 17125.18 -8521.943 139057565 55.97357        8            0
## 3 14033.04 14075.24 -7010.519  21110994 21.85311        4            1
## 4 17264.77 17292.90 -8628.384  44096392 38.57641        3            0
## 5 14037.38 14086.61 -7011.690  21119137 21.85194        4            2
```

Les MSE et MAPE sont calculées sur les prédictions de Conso en prenant en compte le fait que l'hypothèse de normalité soit peu crédible. Cela explique le fait que notre erreur soit assez importante. En effet, la MAPE pour les meilleurs modèles était de moins de 2% avec LConso.

En prenant en compte tous ces critères, le modèle qui semble le plus adapté est le modèle numéro 3 : SARIMA(3,0,0)(0,1,1)[24] avec régression linéaire sur la température. En voici les diagnostics de normalité et de blancheur :

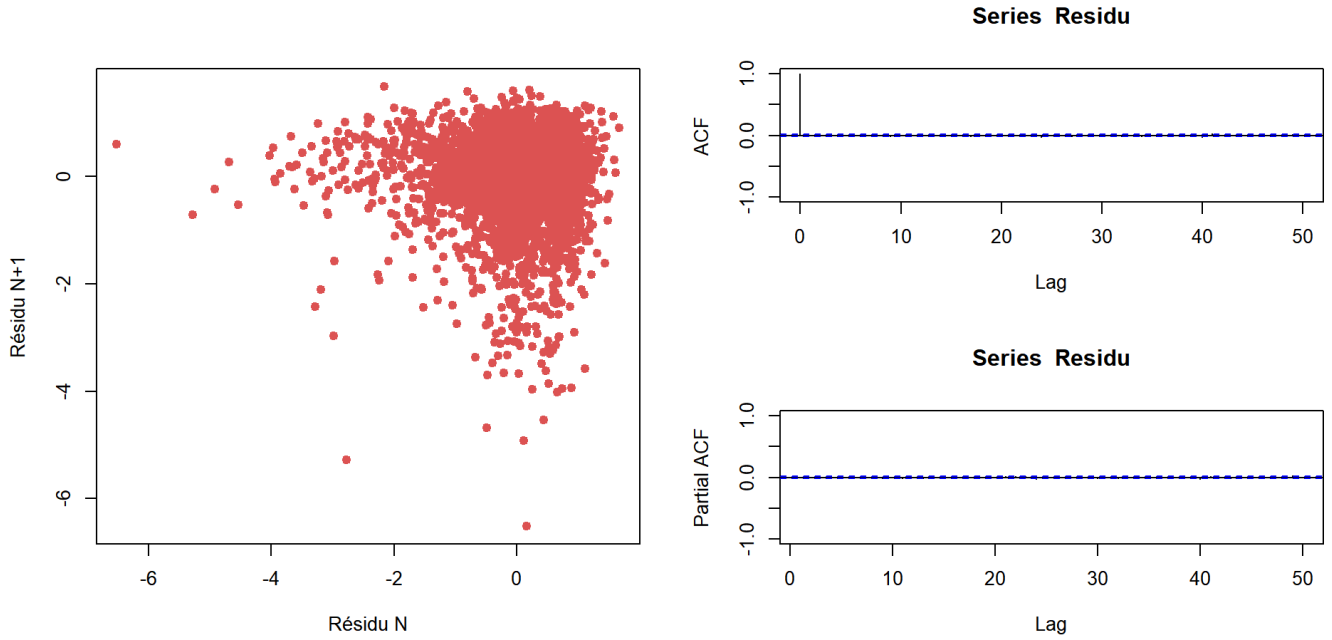
Le diagnostic de normalité des résidus :







Le diagnostic de blancheur des résidus :




Comme nous l'avions constaté et comme c'était le cas pour tous les modèles, le test de Shapiro nous indique au seuil de 5% de rejeter H0, l'hypothèse de normalité.

Cependant, le modèle 3 nous propose un histogramme des résidus ne contredisant pas l'hypothèse de normalité des résidus. De plus, 98.7% des résidus centrés réduits sont compris entre -1.96 et 1.96. Nous constatons tout de même que le QQ-Plot est cohérent avec le test de Shapiro.

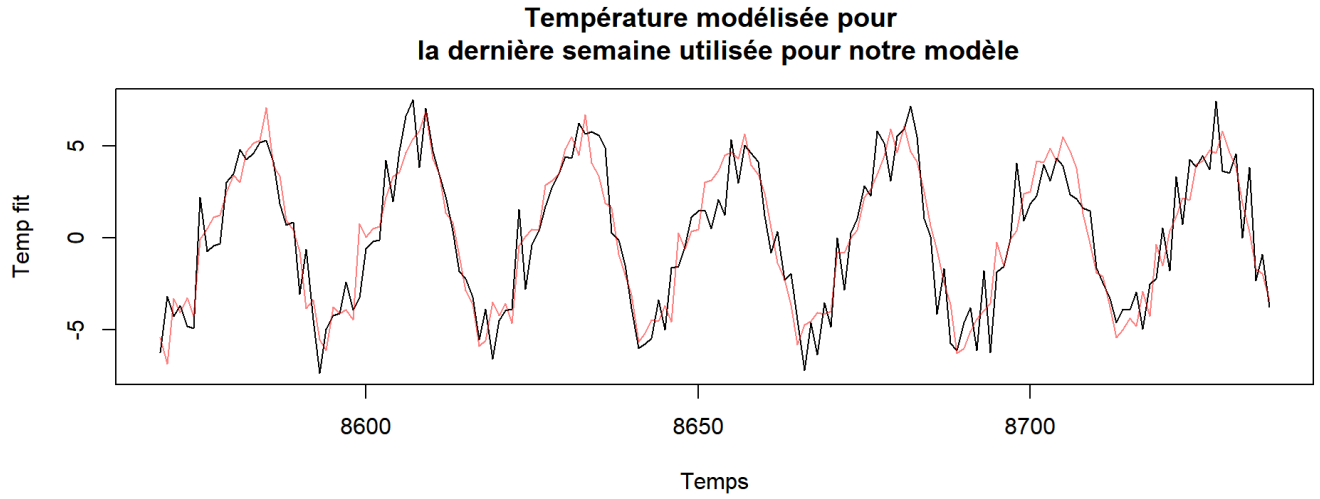
Nous avons envisagé de moyenniser les modèles 3 et 5 qui étaient en concurrence, cependant nous n'avons pas fait ce choix. En effet, nous avons calculé en valeur absolue la différence entre nos consommations prédites pour les deux dernières semaines de décembre 2017 pour les modèles 3 et 4. La somme de ces différences ne vaut que 32.65, ce qui est négligeable au vu des valeurs prises par la vraie consommation sur cette période (valeur moyenne de 755.49 et valeur médiane de 177.04). De plus, 32.65 est obtenu sur 336 (= 24x7x2) valeurs, ce qui confirme que l'écart est très faible entre ces 2 modèles.

## B - Choix du modèle pour Temp

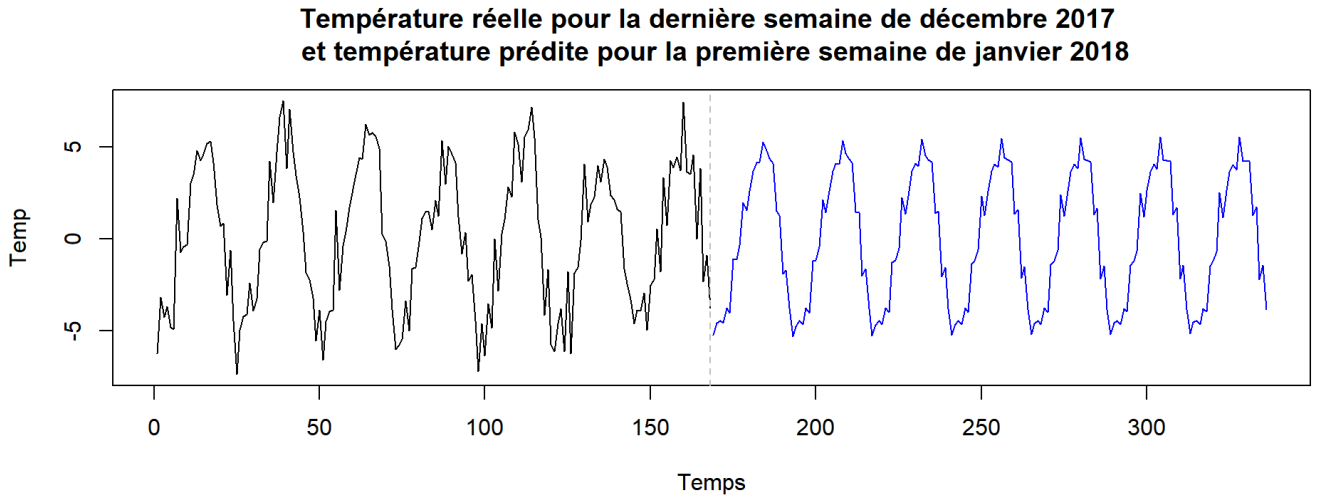
Afin de prédire la consommation de la première semaine de janvier 2018 en prenant en compte la température sur la même période, il nous faut la prédire. Pour cela, nous avons choisi comme nous le suggère la fonction auto.arima, le modèle SARIMA(0,0,5)(1,1,2)[24].

Nous confirmons par les tests ADF et KPSS ainsi que sur l'ACF et la PACF que la série obtenue par cette différenciation est stationnaire.

De plus, ce modèle ne semble pas être le meilleur pour modéliser la série comme le confirme le graphique ci-dessous présentant en noir les vraies données de la dernière semaine de décembre 2017 et en rouge les données modélisées. Cependant, par le même procédé que pour LConso, en comparant les différents modèles que nous avions envisagés, il s'agissait du meilleur prédicteur.

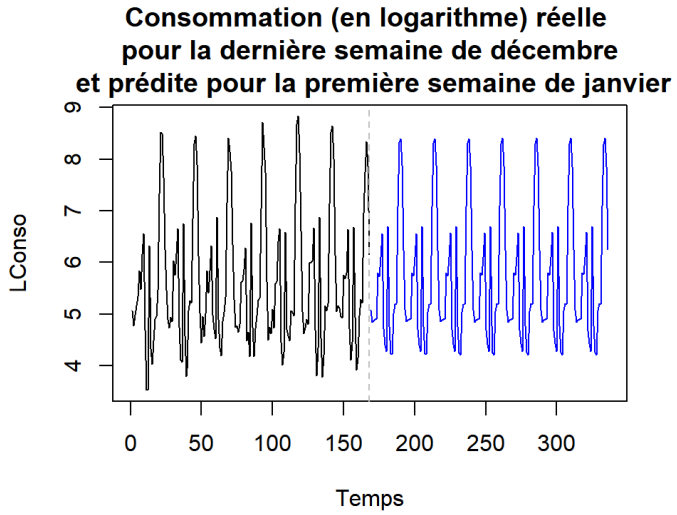


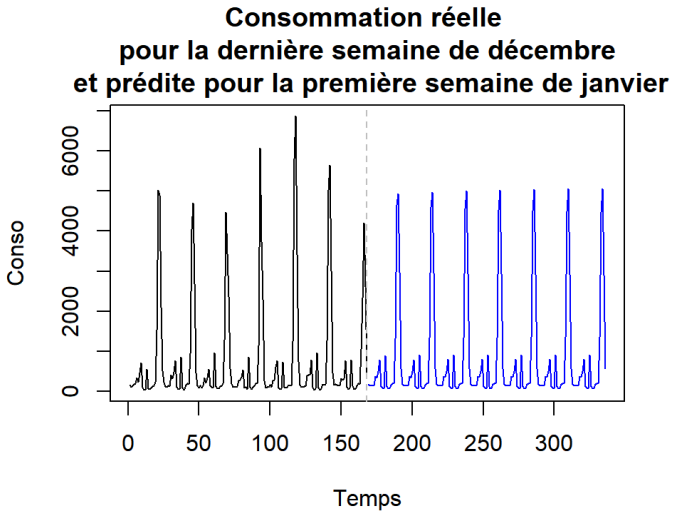
Nous obtenons la prédiction de la température de la première semaine de janvier 2018. Ci-dessous est représentée la température de la dernière semaine de décembre 2017 ainsi que la prédiction de la première semaine de janvier 2018.




## 5) Prédiction de la consommation horaire de la première semaine de janvier 2018

En prenant le modèle 3 pour la prédiction de LConso, et en utilisant la prédiction de la température pour la première semaine de janvier, nous avons pu obtenir la prédiction de LConso et par conséquent la prédiction de Conso, pour la première semaine de janvier 2018. Voici ci-dessous les représentations graphiques pour LConso (à gauche) et Conso (à droite)




Afin de mieux prédire les pics, il serait envisageable de mettre en place un autre modèle sur ces pics.

# II - Etude des données journalières

Nous allons dans cette partie chercher à prédire la consommation journalière du foyer pour janvier 2018.

## 1) Travail préliminaire

Nous avons agrégé nos données horaires de la façon suivante : pour chaque jour nous avons calculé la moyenne des consommations horaires et des températures. Nous obtiendrons la prédiction de la consommation totale journalière de janvier 2018 une fois les consommations moyennes horaires prédites.

Nous avons obtenu les même statistiques descriptives que dans l'étude des données horaires soit :

- la nécessité de faire la transformation de Box-Cox sur nos données journalières de consommation
- la dépendance linéaire entre la consommation journalière en logarithme et la température journalière moyenne
- une augmentation de la température journalière et une diminution de la consommation journalière sur les 5 dernières années

De plus, nous avons constaté que nos 2 séries (consommation en logarithme et température) ne sont pas stationnaires, et pour les rendre stationnaires, il faudra effectuer l'une des différenciations suivantes :

- pour la consommation en logarithme : d=0 et D=1, d=1 et D=1, d=1 et D=0
- pour la température : d=1 et D=1

Nous notons tout de même qu'en appliquant D=1 avec une saisonnalité de 365, nous incluons dans nos calculs un biais lié à l'année bissextile.

## 2) Choix des modèles

En utilisant le même procédé que lors de l'étude des données horaires, nous avons séparé nos séries en 2 : d'une part les jours jusqu'au 30 novembre 2017 et d'autre part les jours du mois de décembre 2017. Nous avons ensuite entrainé plusieurs modèles, respectant les hypothèses de II-1), sur les jours jusqu'au 30 novembre 2017, puis nous avons selectionné ceux modélisant le mieux notre série. Pour sélectionner notre modèle final, nous avons prédit les données des jours de décembre 2017 et retenu le modèle ayant le meilleur profil pour la prédiction et les critères de qualité des modèles vus précédemment.

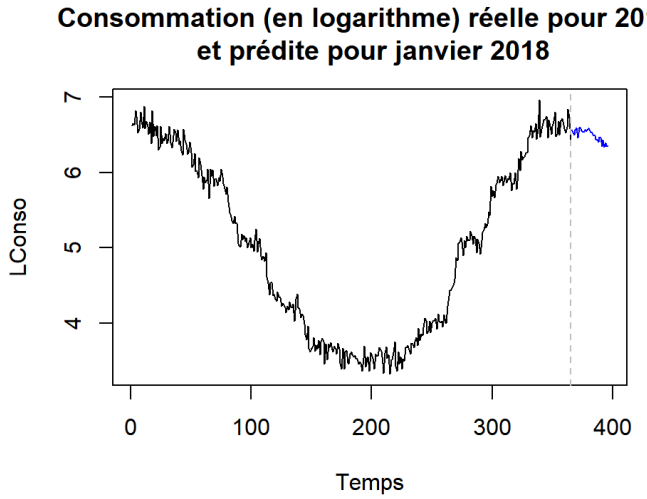
Nous constatons que la blancheur et la normalité des résidus semblent vérifiées.

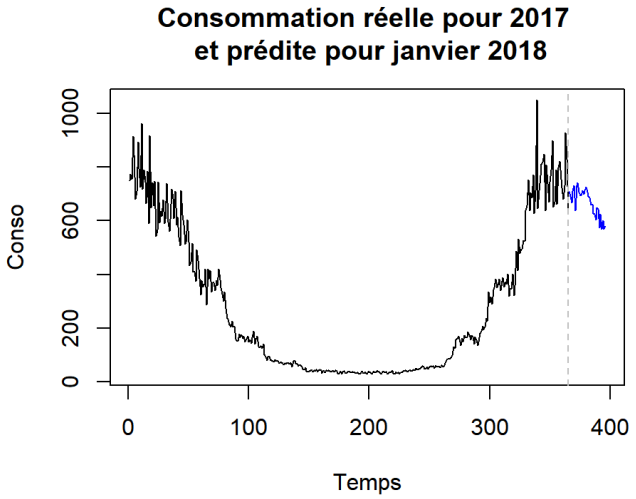
Voici les modèles que nous avons retenu :

- pour la consommation en logarithme : ARIMA(0,1,1) avec régression linéaire sur la température
- pour la température : SARIMA(0,1,1)x(0,1,0)[365]

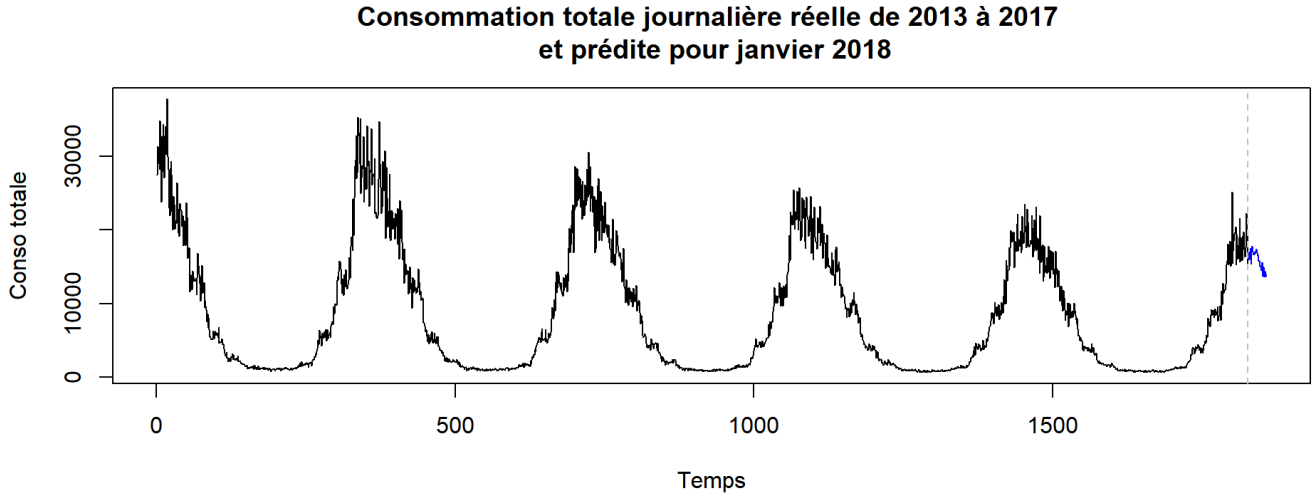
## 3) Prédiction

Nous avons ci-dessous la prédiction de la consommation moyenne horaire par jour connue en 2017 en noire et celle prédite pour janvier 2018 en bleue.




Dans le but d'obtenir la consommation journalière totale du foyer en janvier 2018, nous avons multiplié les valeurs prédites par 24. En voici la représentation graphique :




# III - Etude des données mensuelles

Nous allons dans cette partie chercher à prédire la consommation mensuelle du foyer pour janvier, février et mars 2018.

## 1) Travail préliminaire

Nous avons agrégé nos données horaires de la façon suivante : pour chaque jour nous avons calculé la moyenne des consommations horaires et des températures. Puis, nous avons calculé la moyenne de ces valeurs par mois, en tenant compte du nombre de jours par mois non constant (28,29,30 ou 31).

Nous obtiendrons la prédiction de la consommation totale mensuelle de janvier, février et mars 2018 une fois les moyennes horaires par mois prédites.

Nous avons obtenu les même statistiques descriptives que dans l'étude des données horaires et journalières. Par conséquent, nous avons réalisé la transformation de Box-Cox sur nos données mensuelles de consommation.

De plus, nous avons constaté que nos 2 séries (consommation en logarithme et température) ne sont pas stationnaires, et pour les rendre stationnaires, il faudra effectuer l'une des différenciations suivantes :

- pour la consommation en logarithme : d=0 et D=1, d=1 et D=1
- pour la température : d=1 et D=1 avec une saisonnalité de 12 liée au nombre de mois par an.

Nous ne disposons que de 5 périodes, le choix D=1 implique par conséquent un faible nombre de paramètres dans le choix de nos modèles.

# 2) Choix des modèles

En utilisant le même procédé que lors de l'étude des données horaires et journalières, nous avons séparé nos séries en 2 : d'une part les mois jusqu'à septembre 2017 inclus et d'autre part les jours du mois d'octobre, novembre et décembre 2017. Nous avons ensuite entrainé plusieurs modèles, respectant les hypothèses de III-1), sur les mois jusqu'à septembre 2017 inclus, puis nous avons selectionné ceux modélisant le mieux notre série. Pour sélectionner notre modèle final, nous avons prédit les données des mois d'octobre, novembre et décembre 2017 et retenu le modèle ayant le meilleur profil pour la prédiction et les critères de qualité des modèles vus précédemment.

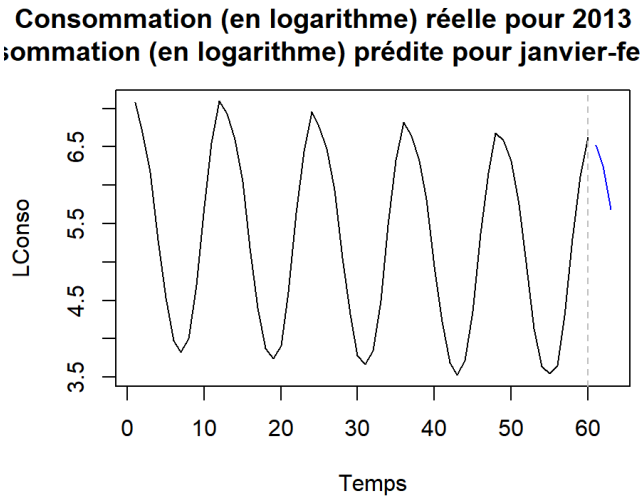
Nous constatons que la blancheur et la normalité des résidus semblent vérifiées.

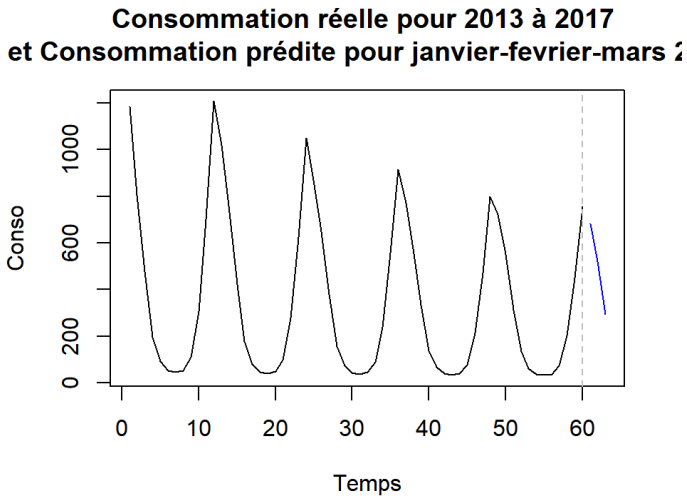
Voici les modèles que nous avons retenu :

- pour la consommation en logarithme : SARIMA(0,0,0)x(0,1,0)[12] avec régression linéaire sur la température
- pour la température : SARIMA(0,1,0)x(0,1,0)[12]

# 3) Prédiction

Nous avons ci-dessous la prédiction de la consommation moyenne horaire par mois connue entre 2013 et 2017 inclus en noire et celle prédite pour janvier, février et mars 2018 en bleue.




Dans le but d'obtenir la consommation mensuelle totale du foyer en janvier, février et mars 2018, nous avons multiplié les valeurs prédites par le nombre de jours dans le mois puis par 24. En voici la représentation graphique :