Projet de prédictions de séries temporelles

COLLIN Hugo

ℹ️ Pour exécuter le code du projet, assurez-vous de spécifier le chemin du dossier racine lors du chargement des données, ainsi que le chemin du dossier d'exportation lors de la génération du fichier Excel contenant les prédictions.

I. Chargement des données et préparation

Au niveau du chargement et de la mise en forme des données, la colonne temporelle a été transformée en format date/heure (`POSIXct`) pour une gestion temporelle précise, puis les données ont été séparées en trois ensembles :

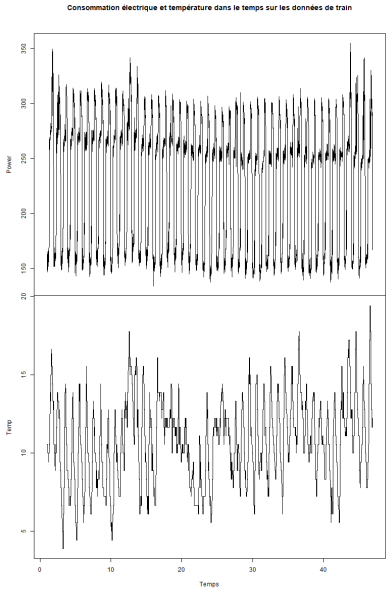
- **Entraînement (train)** : pour ajuster les modèles (données du 1 janvier jusqu'au 15 février 2010).
- **Validation (validate)** : pour évaluer les modèles (données du 16 février 2010).
- **Test (test)** : pour tester les performances finales (données du 17 février 2010).

Enfin, les données d'entraînement, de validation, et de test ont été transformées en séries temporelles multivariées, avec une fréquence de 96 observations par jour (correspondant à des intervalles de 15 minutes, car il y a 96 intervalles de 15 minutes dans une journée). Pour les modèles à réseaux de neurones, les données ont été spécifiquement enrichies, en extrayant des caractéristiques temporelles supplémentaires, telles que l'heure et le jour de l'année, afin de mieux capturer les variations cycliques et contextuelles liées au temps.

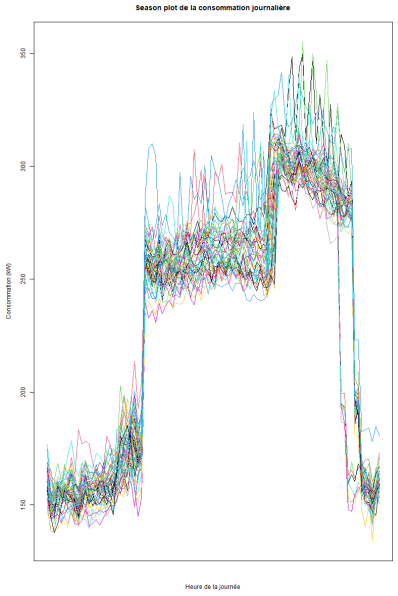
II. Analyse exploratoire des données

Pour comprendre les caractéristiques des données des séries temporelles, une analyse exploratoire a été réalisée :

1. Des graphiques des séries temporelles pour la consommation électrique et la température ont été générés. Ceux-ci montrent des tendances et des structures saisonnières évidentes, notamment des cycles journaliers.

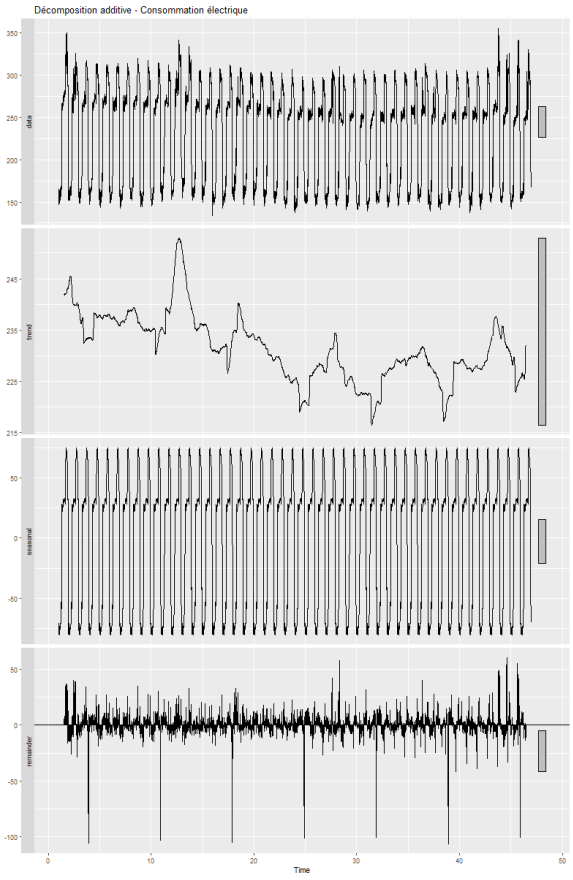


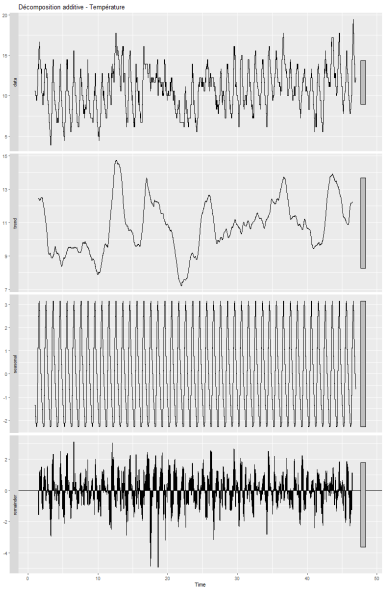
2. La création d'un season plot de la consommation représentant la consommation électrique au cours d'une journée a été créée, mettant en évidence :
   - **Une répétition quotidienne** : nous observons une structure répétitive dans la consommation électrique au fil de la journée. Les pics de consommation apparaissent systématiquement le matin vers 7h-8h pour se stabiliser jusqu'à 15h-16h. Puis un second pic survient vers 16h-17h jusqu'à 21h-22h, ce qui reflète les habitudes humaines, comme le début de la journée de travail et les activités domestiques en soirée.
   - **Des creux durant la nuit :** la consommation est minimale pendant la nuit, entre 23h et 5h, période où l'activité est généralement plus faible.
   - **Une variabilité inter-journalière** : bien que la forme globale reste constante d'un jour à l'autre, il existe une variabilité notable d'une journée à l'autre. Cela pourrait être lié à des facteurs externes comme la température, les jours de semaine (travail vs week-end).



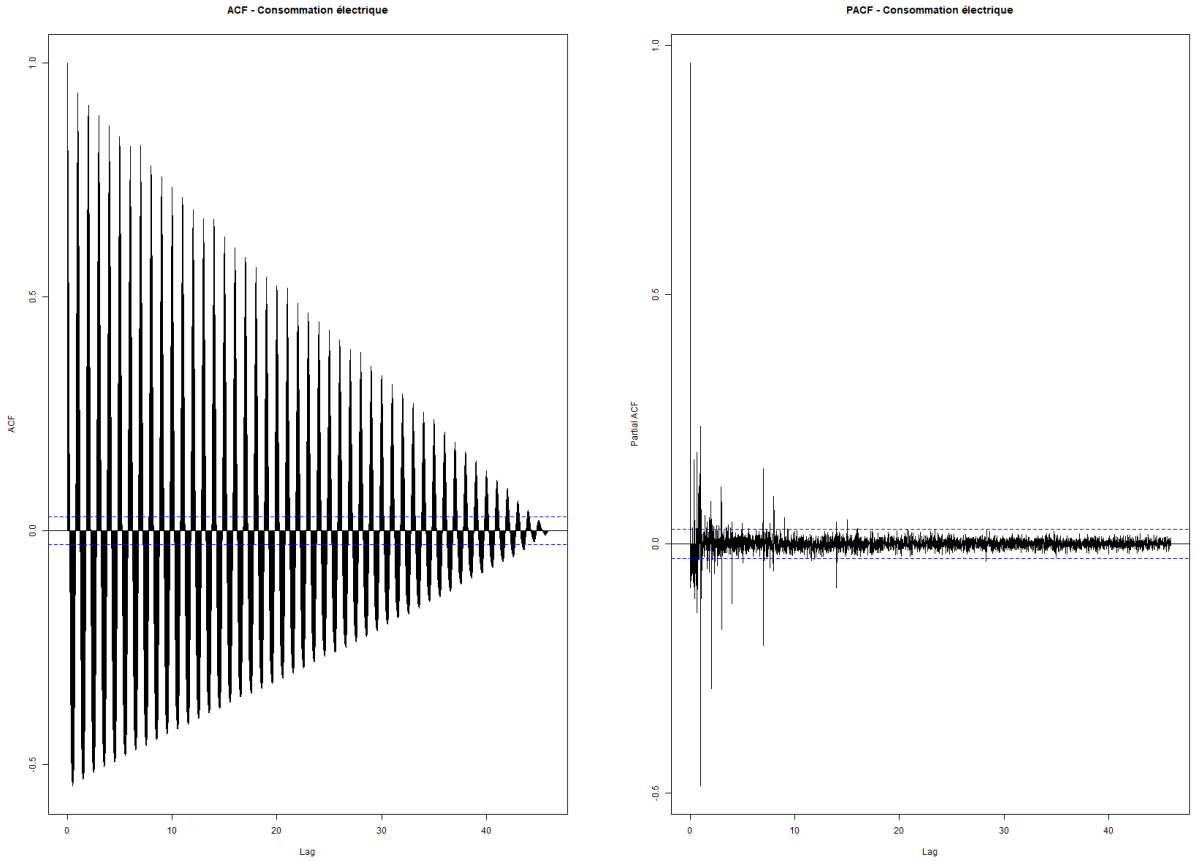
3. Une décomposition additive des séries temporelles a été effectuée, ce qui a révélé, concernant les données de la consommation électrique :
   - **Tendance :** nous observons une tendance globale avec des fluctuations significatives à la baisse au début puis une stabilisation. Cela indique une dynamique de diminution générale de la consommation électrique, et nous pouvons voir qu'à l'inverse la tendance de la température augmente, ce qui est logique, car plus les températures augmentent moins le besoin de chauffer est nécessaire.

- **Saisonnalité :** la composante saisonnière montrent des cycles réguliers très marqués que ce soit pour la consommation ou pour la température. Cela suggère une forte dépendance temporelle, probablement sur des cycles quotidiens (par exemple, variation jour/nuit ou jours ouvrés vs week-end), ce qui rejoint les conclusions effectuées avec le season plot.
- **Résidus :** les résidus semblent être distribués de manière aléatoire, ce qui est souhaitable, cependant bien que globalement aléatoires, les résidus montrent une amplitude variable avec des pics plus marqués à certains moments. Cela pourrait refléter des anomalies ou des événements spécifiques influençant la consommation d'électricité.





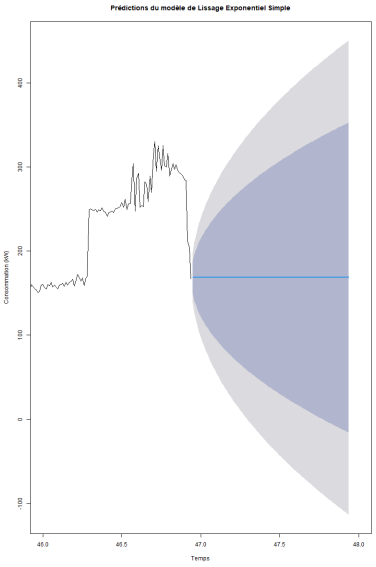
4. Les analyses des autocorrélations ont montré :
   - **ACF :** révèle un décalage significatif, un déclin progressif et des pics périodiques, témoignant d'une forte dépendance temporelle. Cela est cohérent avec la saisonnalité observée dans le season plot. Le graphique ACF présente une configuration sinusoïdale marquée et périodique, confirmant un comportement saisonnier prononcé. De plus, la lente décroissance des autocorrélations suggère que la série est non stationnaire et caractérisée par des fluctuations périodiques.
   - **PACF :** les premiers lags présentent des coefficients significatifs au-delà des intervalles de confiance, indiquant une corrélation forte pour les lags initiaux. Après les premiers lags, les valeurs de la PACF se situent rapidement dans l'intervalle de confiance.

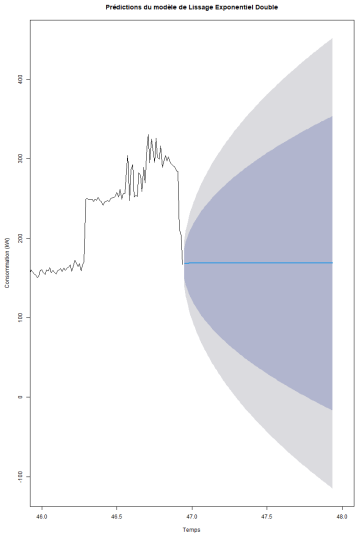


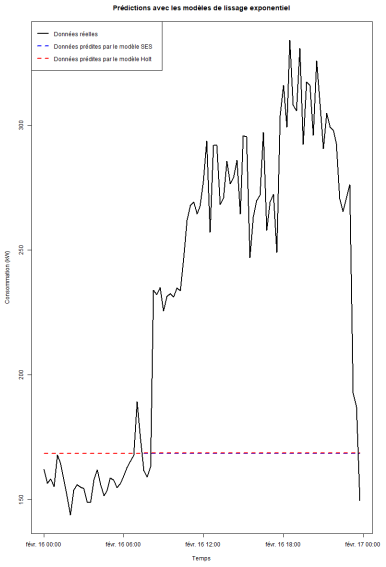
III. Utilisation des modèles de prédictions

1. Modèles de lissage exponentiel

Les modèles prédisent une valeur constante après la dernière observation (ligne bleue horizontale). Nous obtenons de mauvais résultats car les modèles ne prennent pas en compte les tendances ni la saisonnalité, ce qui explique l'absence de variation dans la prévision. Il n'est donc pas adapté pour faire une prédiction sur notre jeu de données. Concernant les lissages exponentiels triples : Holt-Winters Method - Additive Seasonal et Multiplicative seasonal il est impossible de les utiliser car la fréquence des données est trop élevée. De manière générale, les modèles de lissage exponentiel ne sont pas adaptés pour prédire ces données, car elles contiennent une tendance et une saisonnalité significatives.



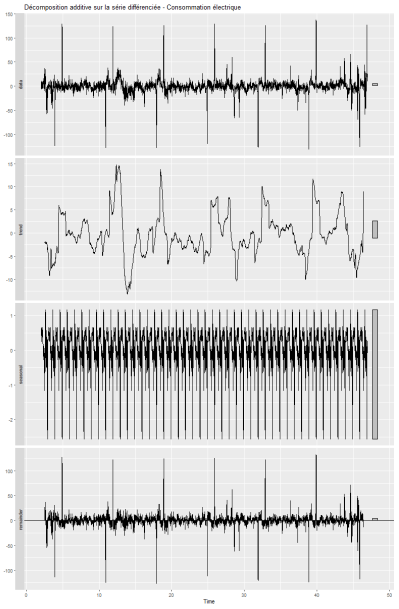




2. Modèles ARIMA

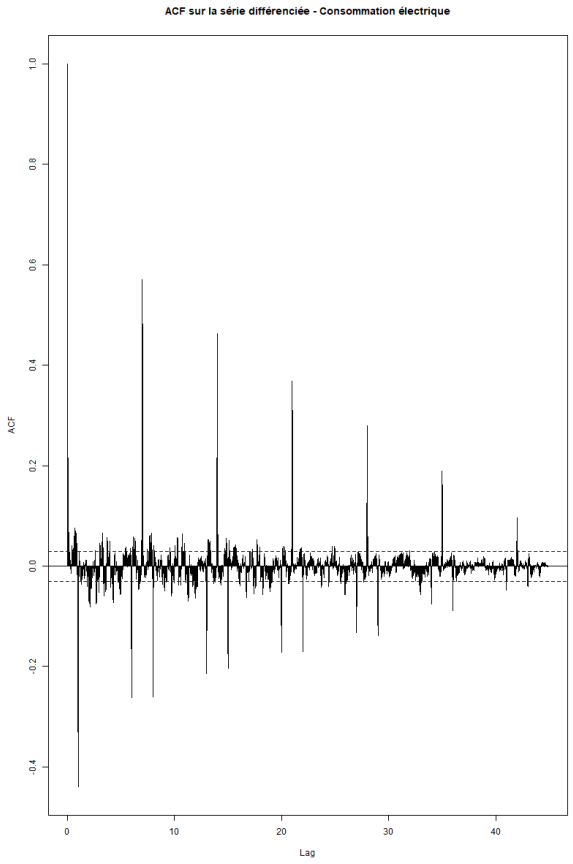
Avant de pouvoir utiliser les modèles ARIMA suite aux conclusions obtenues lors de la partie , nous allons effectuer une différenciation pour enlever la tendance et la saisonnalité. Analysons ce que nous obtenons :

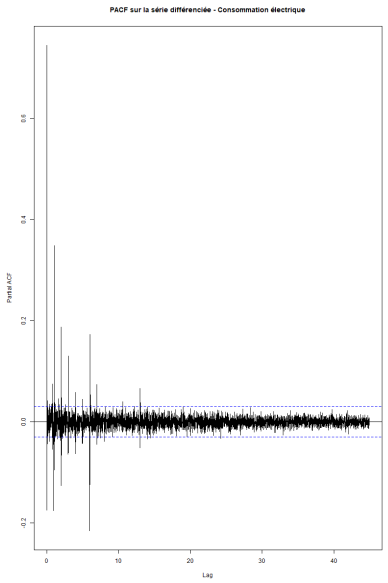
- **Données :** nous observons une série plus stationnaire en apparence, avec des fluctuations autour de zéro. Les pics et les creux importants qui reflétaient la saisonnalité journalière ont pour la plupart disparu.
- **Tendance :** nous observons des fluctuations, mais sans tendance globale. Cela suggère que la différenciation a permis de supprimer la tendance linéaire.
- **Saisonnalité :** après la différenciation saisonnière avec un lag de 96, la composante saisonnière devrait être proche de zéro et c'est ce que l'on observe sur le graphique. Cependant une saisonnalité est toujours présente et montre un comportement cyclique marqué. En ayant essayé d'appliquer une différenciation saisonnière supplémentaire, cela surdifférenciait trop la série temporelle, c'est donc pour cela que j'ai décidé de ne pas en appliquer une deuxième. Donc les modèles capturant directement cette composante auront sûrement de meilleurs résultats.
- **Résidus :** les résidus présentent des fluctuations aléatoires avec des pics occasionnels. Ces pics pourraient indiquer des anomalies ou des données bruitées qui ne sont pas expliquées par le modèle de décomposition.



Puis les analyses des autocorrélations ont montré :

- **ACF :** montre encore des pics significatifs (beaucoup moins nombreux qu'avant la différenciation) notamment aux premiers lags, qui décroissent progressivement vers zéro. Cela suggère la présence d'une composante MA (Moyenne Mobile).
- **PACF :** possède un pic important au début, puis des pics plus faibles et une décroissance vers zéro. Cela suggère la présence d'une composante AR (Autorégressive).





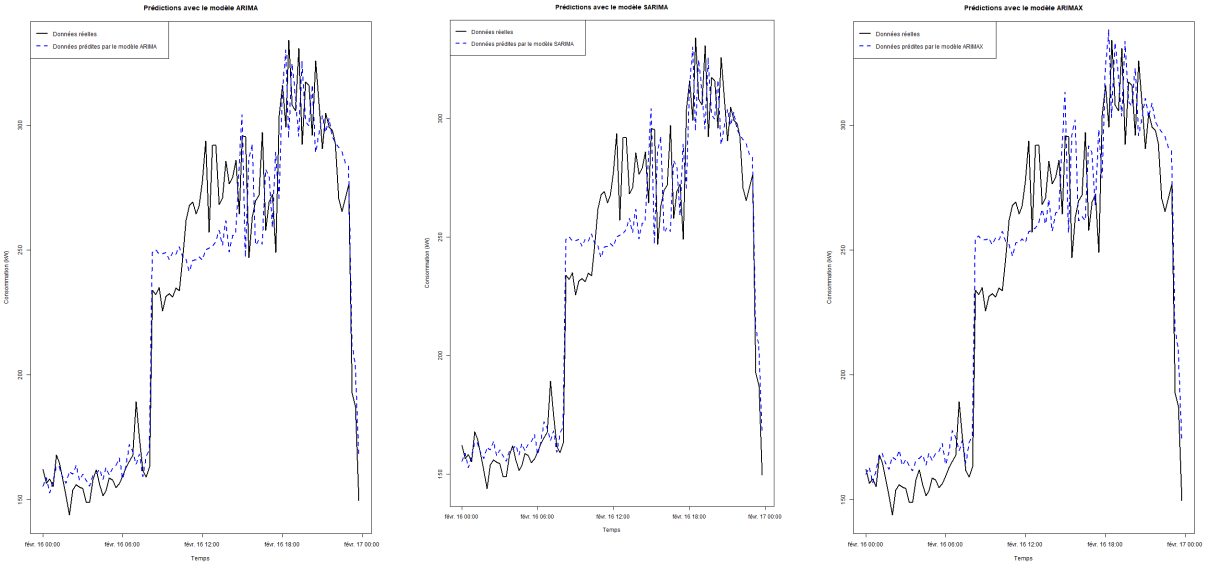
Enfin, concernant le test de Ljung-Box, nous obtenons une p-value extrêmement faible (< 0.05) qui indique que l'hypothèse nulle d'absence d'autocorrélation dans les résidus est rejetée. Cela signifie qu'il existe encore une autocorrélation significative dans la série différenciée. Pour l'Augmented Dickey-Fuller test la p-value est faible (< 0.05), ce qui indique que l'hypothèse nulle de non-stationnarité est rejetée. Cela confirme que la série différenciée est considérée comme stationnaire. Cependant, bien que le test ADF confirme la stationnarité globale, le résultat du test de Ljung-Box montre qu'il reste une structure d'autocorrélation dans les résidus qui peut être associée à la présence d'une structure saisonnière que nous n'avons pas réussit à totalement supprimer.

Pour le choix des modèles ARIMA, j'ai décidé de tester un modèle ARIMA classique (Modèle Auto-Régressif Intégré Moyenne Mobile), un modèle SARIMA (Modèle Auto-Régressif Intégré Moyenne Mobile avec Saisonnalité) et un modèle ARIMAX (Modèle Auto-Régressif Intégré Moyenne Mobile avec Exogène).

Voici les résultats sur le jeu de validation :

Modèle	RMSE	MAPE
ARIMA	19.78361	6.303264
SARIMA	19.78361	6.303264
ARIMAX	20.32967	7.239171

Globalement, ces trois modèles obtiennent les mêmes performances que l'on rajoute la composante de saisonnalité ou la variable de la température. La statistique du test de Ljung-Box a une valeur très élevée et la p-valeur est extrêmement faible, inférieure au seuil de signification (< 0.05) pour ces trois modèles. Rejeter l'hypothèse nulle avec une p-valeur aussi basse signifie qu'il existe une autocorrélation significative dans les résidus des modèles. En d'autres termes, les modèles ne capturent pas toute l'information présente dans les données, et il reste des motifs non expliqués dans les résidus. Ces modèles ne sont donc pas les plus adéquats pour ces données.

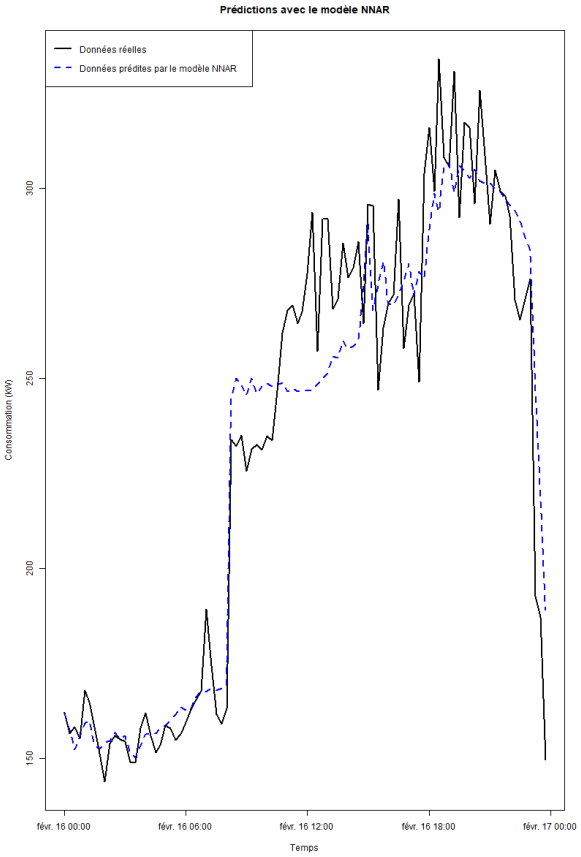


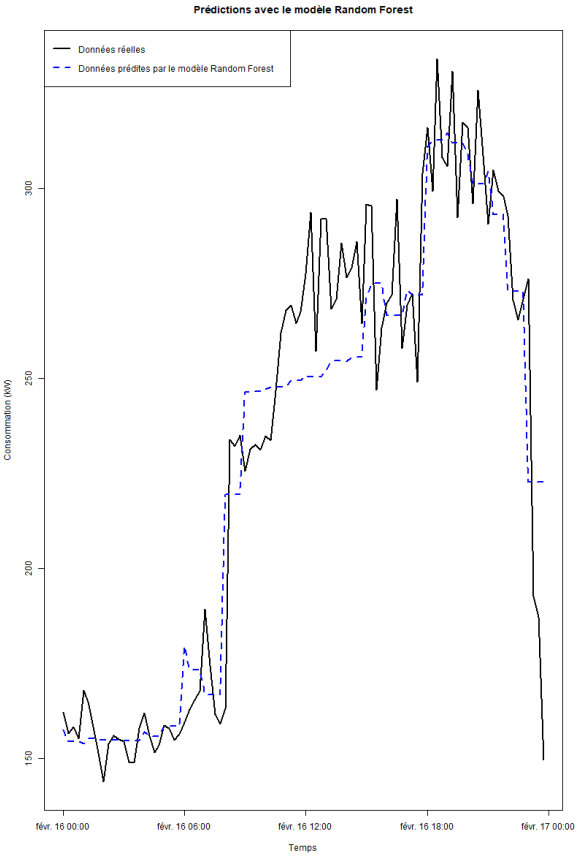
3. Modèles à réseaux de neurones

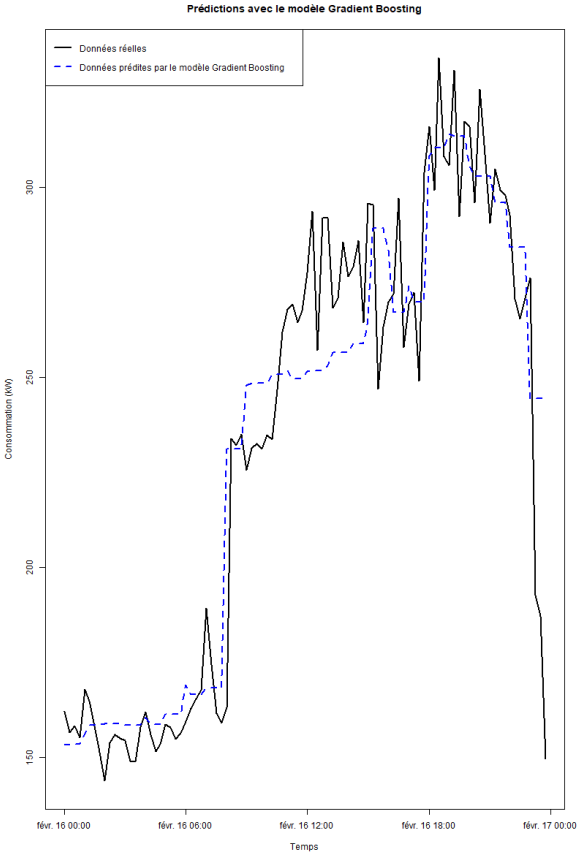
- **NNAR :** dans l'implémentation, j'ai utilisé la fonction `nnetar()` du package `forecast`. Ce modèle ajuste automatiquement ses paramètres internes lors de l'entraînement, il n'y a donc pas d'hyperparamètres à optimiser directement via une grille de recherche.
- **Random Forest :** l'optimisation des hyperparamètres est cruciale pour ses performances. J'ai utilisé la fonction `train()` du package `caret` pour réaliser une validation croisée (5-fold) et rechercher la meilleure combinaison de l'hyperparamètre `mtry`.
- **Gradient Boosting :** j'ai utilisé l'implémentation XGBoost via le package `xgboost`. L'optimisation des hyperparamètres a été effectuée avec `train()` et une validation croisée (5-fold). La grille de recherche `gb_grid` explorait plusieurs combinaisons de `nrounds` (nombre d'arbres), `max_depth` (profondeur maximale des arbres), `eta` (taux d'apprentissage), `gamma` (paramètre de régularisation).
- **SVM :** j'ai utilisé un kernel radial (`svmRadial` dans `caret`). L'optimisation des hyperparamètres `C` (coût de pénalité) et `sigma` (paramètre du kernel) a été réalisée avec `train()`, une validation croisée (5-fold).
- Pour le modèle linéaire, il sert de base de comparaison pour les modèles à réseaux de neurones (il n'y a pas d'hyperparamètres à optimiser).

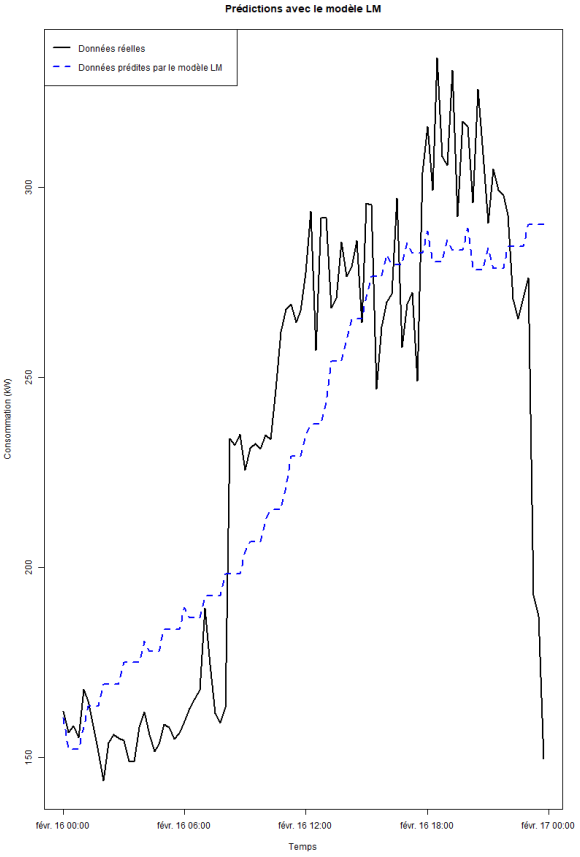
Modèle	RMSE	MAPE
NNAR	16.97597	5.122007
Random Forest	19.49814	6.069644
Gradient Boosting	19.14427	5.864172
SVM	24.36839	8.667843
LM	32.99828	11.988642

Le modèle NNAR présente les meilleures performances avec le RMSE et MAPE les plus bas, ce qui indique une meilleure précision des prévisions. Les modèles Random Forest et Gradient Boosting montrent des performances comparables, bien que légèrement inférieures au NNAR. Le SVM a des performances moins bonnes, et le modèle linéaire (LM) est le moins performant, ce qui souligne l'importance d'utiliser des modèles plus complexes pour cette tâche. L'optimisation des hyperparamètres a permis d'améliorer les performances des modèles Random Forest, Gradient Boosting et SVM. Concernant le modèle NNAR, bien qu'il n'ait pas d'hyperparamètres à optimiser manuellement, s'est avéré être le plus performant pour cette tâche de prévision de la consommation électrique.


Prédictions avec le modèle NNAR
Données réelles
Données prédites par le modèle NNAR
Consommation (kW)
150
200
250
300
févr. 16 00:00
févr. 16 06:00
févr. 16 12:00
févr. 16 18:00
févr. 17 00:00
Temps


Prédictions avec le modèle Random Forest
Données réelles
Données prédites par le modèle Random Forest
Consommation (kW)
150
200
250
300
févr. 16 00:00
févr. 16 06:00
févr. 16 12:00
févr. 16 18:00
févr. 17 00:00
Temps


Prédictions avec le modèle Gradient Boosting
Données réelles
Données prédites par le modèle Gradient Boosting
Consommation (kW)
150
200
250
300
févr. 16 00:00
févr. 16 06:00
févr. 16 12:00
févr. 16 18:00
févr. 17 00:00
Temps


Prédictions avec le modèle LM
Données réelles
Données prédites par le modèle LM
Consommation (kW)
150
200
250
300
févr. 16 00:00
févr. 16 06:00
févr. 16 12:00
févr. 16 18:00
févr. 17 00:00
Temps

IV. Conclusion

L'analyse des performances sur le jeu de validation a révélé des différences significatives entre les modèles. Le modèle NNAR s'est distingué par ses excellentes performances, surpassant les autres modèles en termes de précision globale. Les modèles Random Forest et Gradient Boosting ont également montré des résultats prometteurs, bien que légèrement inférieurs au NNAR. Le SVM et le modèle linéaire se sont avérés moins performants, soulignant l'importance d'utiliser des approches plus adaptées à la complexité des données. Concernant les modèles ARIMA, ils sont moins bon que les modèles à réseaux de neurones.

Une distinction cruciale a été faite concernant l'utilisation de la température comme variable prédictive. L'architecture du modèle NNAR, telle qu'implémentée dans le package `forecast`, est intrinsèquement conçue pour les séries temporelles univariées. Elle ne permet pas d'intégrer directement des variables exogènes comme la température. Par conséquent, pour la prévision sans information sur la température, le modèle NNAR a été retenu comme le plus approprié en raison de ses performances supérieures observées lors de la phase de validation.

En revanche, les modèles Random Forest, Gradient Boosting et SVM peuvent intégrer des variables exogènes. Parmi ceux-ci, le Gradient Boosting a démontré des performances compétitives lors de la validation. Ainsi, pour la prévision intégrant la température, le modèle Gradient Boosting a été choisi.

Voici les résultats globaux des modèles sur les métriques RMSE et MAPE calculée sur le jeu de donnée de validation :

Modèle	RMSE	MAPE
ARIMA	19.78361	6.303264
SARIMA	19.78361	6.303264
ARIMAX	20.32967	7.239171
NNAR	16.97597	5.122007
Random Forest	19.49814	6.069644
Gradient Boosting	19.14427	5.864172
SVM	24.36839	8.667843
LM	32.99828	11.988642

Enfin, voici l'affichage graphique des prédictions pour la journée du 17 février 2010 :