MINI PROJET :

Analyse des séries temporelles de la consommation d'électricité, de gaz et d'eau de ESIEE Paris

DSIA-5102B
2022

Adam **GOUJA**

Valentin **TAILLANDY**

Table des matières

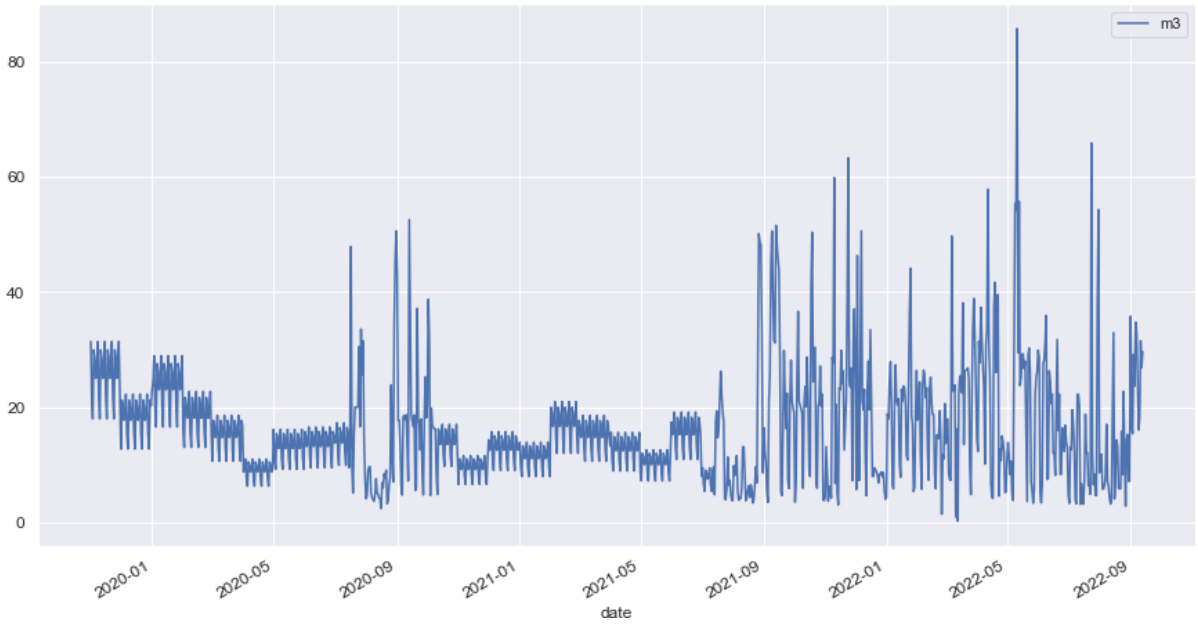
Introduction

Exploration des données

Eau

Nous avons pour le jeu de données représentant la consommation d'eau de ESIEE Paris en m3 :

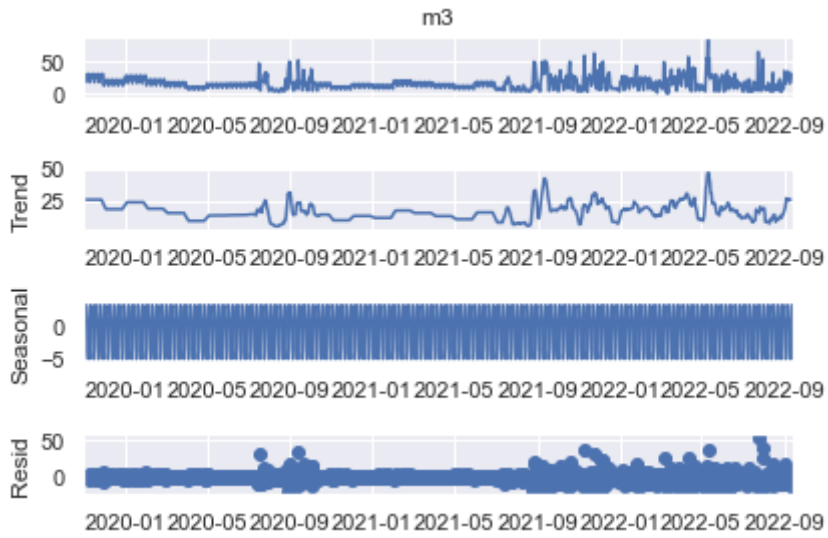
- Une range de date du 01 novembre 2019 au 13 septembre 2022
- Quantité d'eau utilisée en m3
- Données journalières
- Aucune valeur 'encodée' manquante
- Une absence de données pour le 21 et 22 mars 2022



Quantité d'eau consommée en rapport à la date

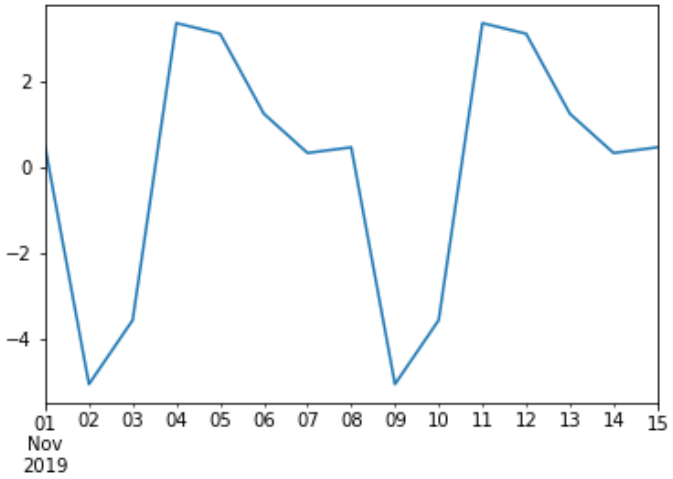
Pour ce qui est des données absentes du jeu de données, la range étant limité à 2 jours, nous pouvons nous contenter d'appliquer un SMA (Simple Moving Average) sur une période de 4 jours et ainsi de combler le trou avec un lissage des données intermédiaires.

Nous pouvons percevoir une certaine saisonnalité dans le signal, nous allons donc le décomposer :



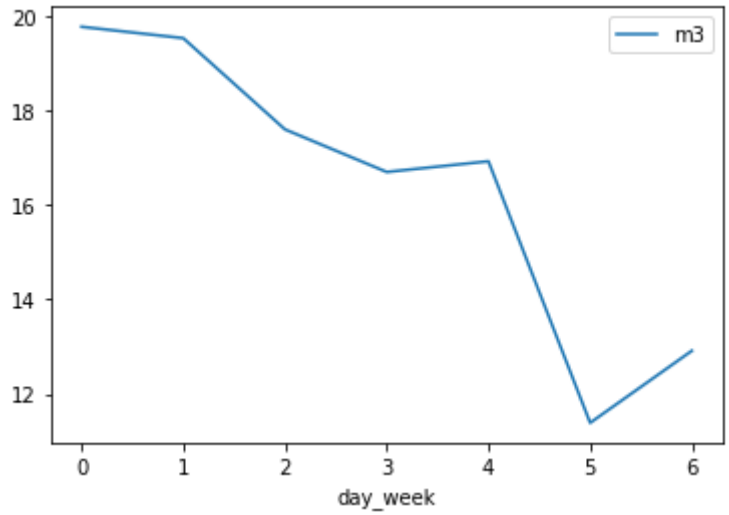
Décomposition du signal m3

Un motif semble se répéter, mais en se concentrant sur la saisonnalité obtenue observons :



Saisonnalité du signal m3

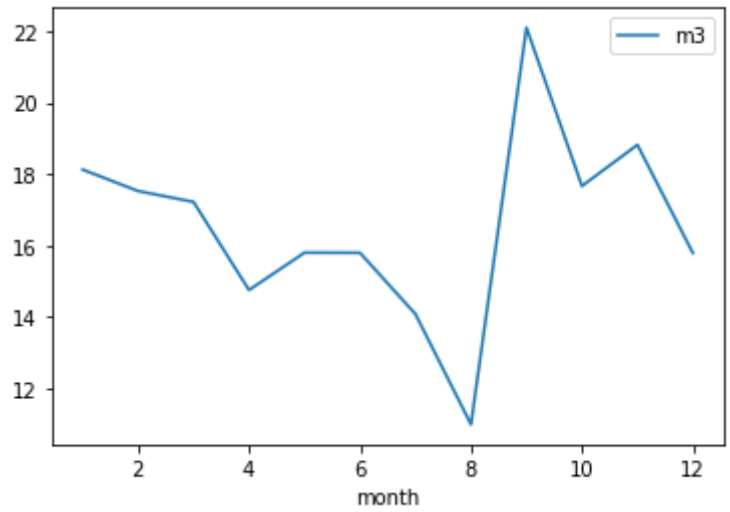
Le signal se répète tous les 7 jours, ce qui implique une saisonnalité hebdomadaire et une période de 7 jours. Nous pouvons imaginer une consommation accrue due à la présence d'étudiant (chasse d'eau, fontaine à eau, station de lavage pour la cantine) du lundi au vendredi. Cette hypothèse est renforcée par le fait que le 04 Nov. 2019 était un lundi (jour d'affluence) avec un décroissement sur la semaine et un creux pour le week-end. Nous observons également ceci en faisant la moyenne en groupant par jour :



Quantité d'eau moyenne en m3 consommée groupée par jour

Du lundi (0) au dimanche (6) nous voyons que la consommation est en moyenne à son maximale le lundi, décroît le long de la semaine avec une petite augmentation le vendredi avant de chuter pour le week-end, l'influence étant moindre.

En suivant cette hypothèse, nous pouvons donc nous questionner sur une saisonnalité au travers des mois, les étudiants étant absents pendant les vacances scolaires :



Quantité d'eau moyenne en m3 consommée groupée par mois

En effet, nous voyons un creux lors des vacances universitaires, les étudiants et personnels sont absents pour le mois d'août. Nous voyons également un pique au retour des étudiants en septembre et un décroissement jusqu'en avril, avec une augmentation pour le mois de mai et un plateau jusqu'à juin. Nous pouvons donc nous demander si l'arrivée des beaux jours et de température plus élevée, d'un taux d'humidité moindre ou encore moins de précipitation

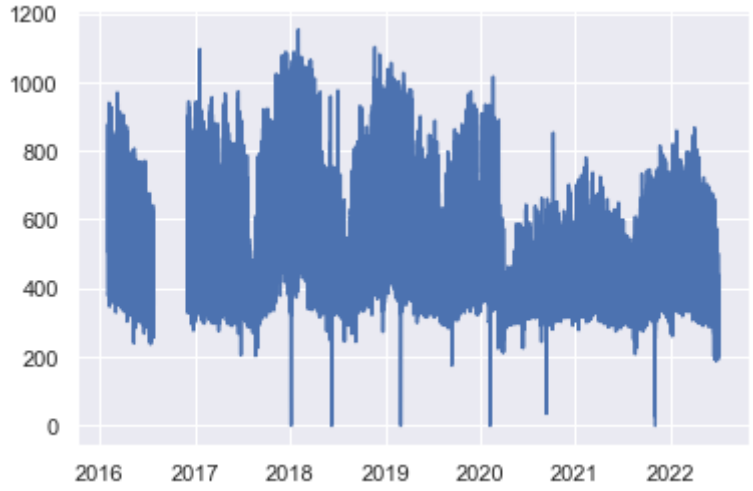
comparée à février ou à mars implique l'utilisation de plus d'eau en moyenne pour entretenir la verdure au sein de ESIEE Paris, d'où l'augmentation pour mai et avril.

Nous pouvons d'ores et déjà ajouter month et day_week qui sont présents implicitement dans les données, représentant respectivement le mois et le jour de la semaine.

Électricité

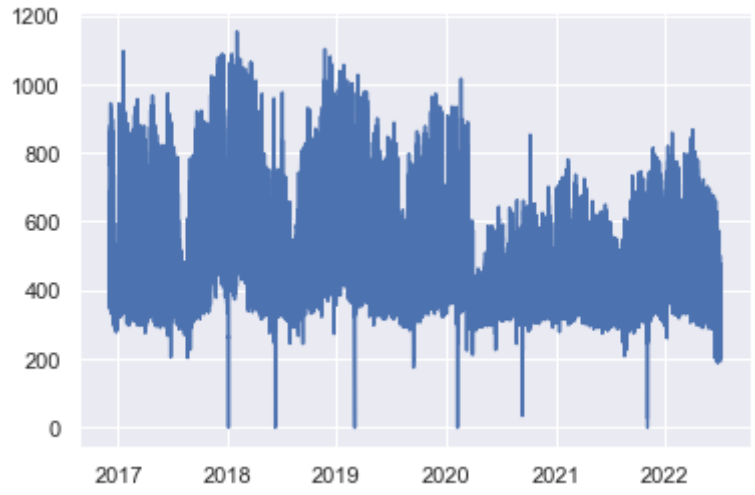
Nous avons pour le jeu de données représentant la consommation d'électricité de ESIEE Paris en kW :

- Une range de date du 01 février 2016 au 05 juin 2022
- Quantité d'électricité consommée en kW
- Données toutes les 10 minutes
- 18422 valeurs 'encodées' manquantes pour kW
- Une absence de données pour le 04 novembre 2021 de 8h20 à 11h40



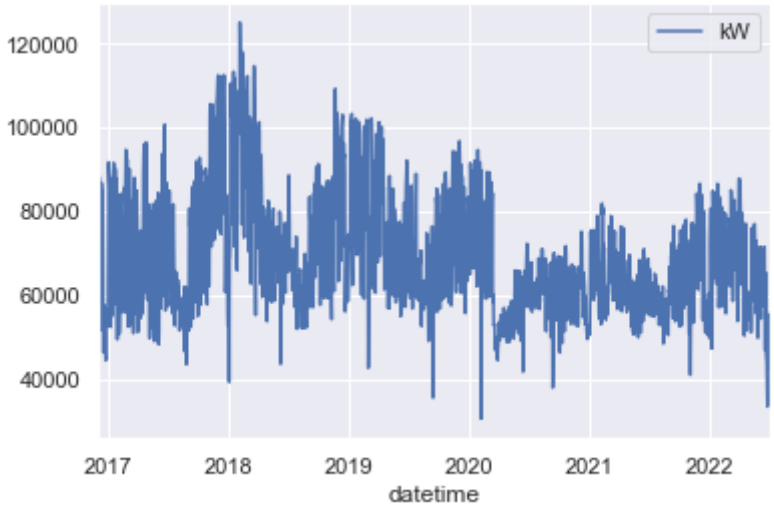
Quantité d'électricité en kW consommée en rapport à la date

Nous pouvons voir un trou juste avant 2017, nous avons suffisamment de données pour ce dataset, là où eau et gaz s'étendent de 2018 ou 2019 à 2022, nous pouvons couper le dataset et ne pas combler la partie vide. Dans le cas où nous voudrions combler ce vide, nous pouvons imaginer créer un modèle et prédire à reculons les données et ainsi combler avec des approximations, mais possédant suffisamment de données nous avons décidés de simplement le couper.



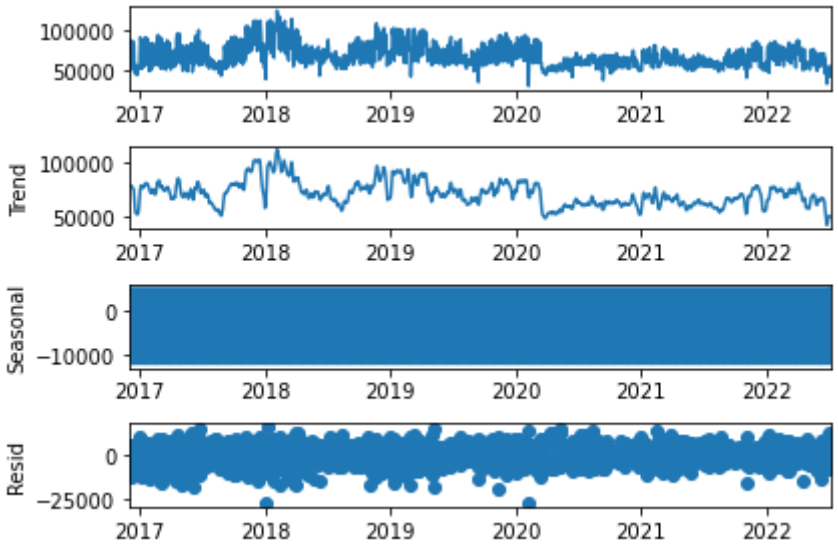
Quantité d'électricité en kW consommée en rapport à la date (coupé)

Cependant, nous avons toujours des données manquantes pour le 04 novembre. A l'instar de l'eau nous allons appliquer un SMA et ainsi combler les dates manquantes en lissant. Une fois fait, nous avons des observations toutes les 10 minutes, nous allons donc grouper les données en faisant une somme pour avoir la consommation sur la journée et avoir un référentiel commun.



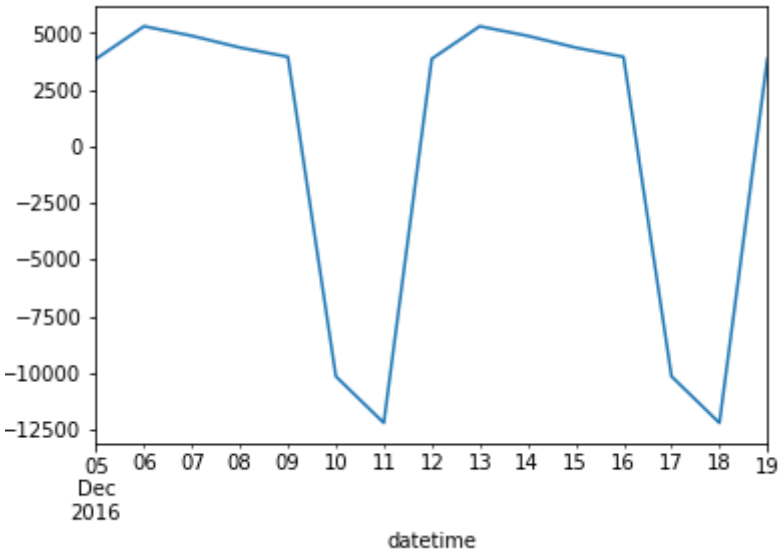
Quantité d'électricité en kW consommée groupé par jour en rapport à la date

Nous pouvons percevoir une certaine saisonnalité dans le signal, nous allons donc le décomposer :



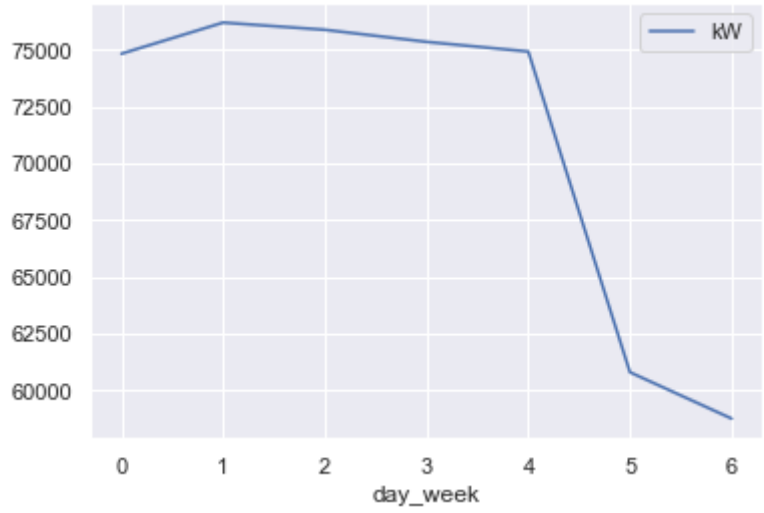
Décomposition du signal kW

Tout comme l'eau, nous avons une tendance et motif qui semble se répéter. En se concentrant sur la saisonnalité, nous pouvons observer :



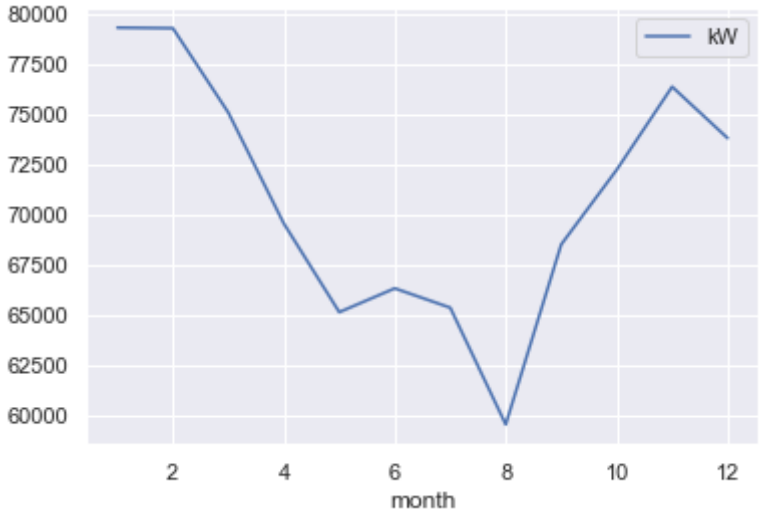
Saisonnalité du signal kW

De même, nous observons un creux pendant 2 jours, l'hypothèse appliquée au signal de l'eau, que la présence des élèves influe sur la consommation s'applique. La période est également de 7 jours.



Quantité de kW moyenne consommé groupé par jour

En faisant un groupement par jour, nous voyons bien le creux de consommation (non nulle) entre le week-end et les jours de la semaine, qui sont en moyenne au même niveau.



Quantité de kW moyenne consommé groupé par mois

De même, ici, nous voyons un creux pour le mois d'août, une augmentation pour les mois de septembre à novembre (temps de cours « plein ») avec une légère descente pour les fêtes de Noël et une décroissance vers le mois de mai et un pic en juin (avec potentiellement la remise des diplômes) avant de diminuer vers le mois d'août.

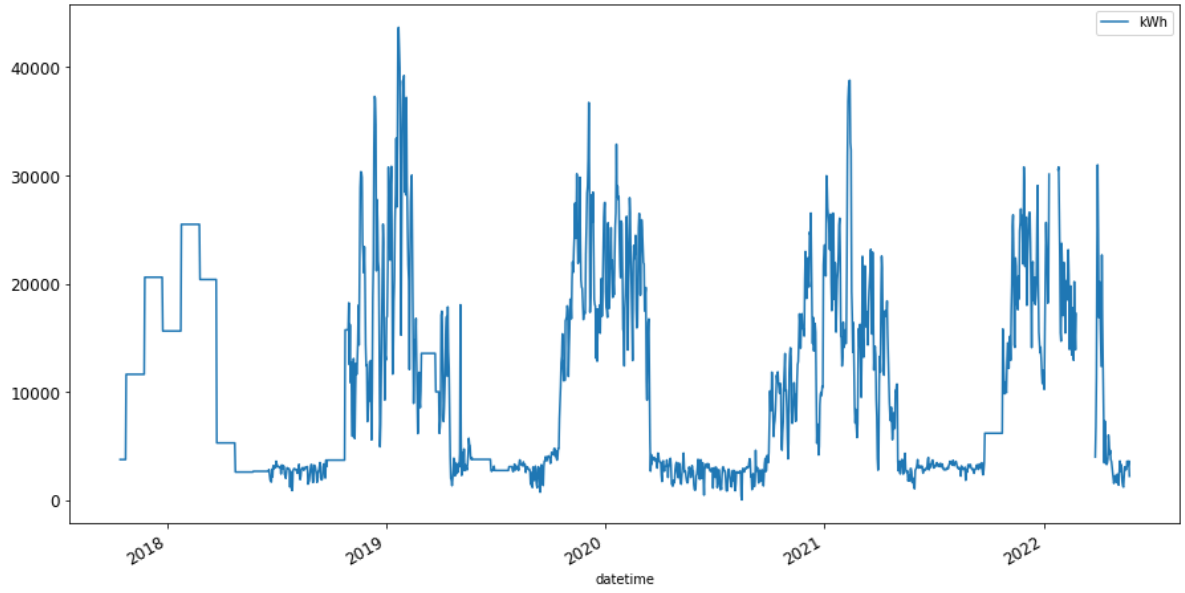
Nous pouvons d'ores et déjà aussi ajouter month et day_week qui sont présents implicitement dans les données, représentant respectivement le mois et le jour de la semaine.

Gaz

Nous avons pour le jeu de données représentant la consommation de gaz de ESIEE Paris en kWh, m3 et Nm3 :

- Une range de date du 15 octobre 2017 au 24 mai 2022
- Rapport de 14,59375 pour passer de m3 à kWh
- Données tous les jours
- 42 valeurs 'encodées' manquantes pour les trois
- Une absence de données du 25 au 27 mars et avril 2022

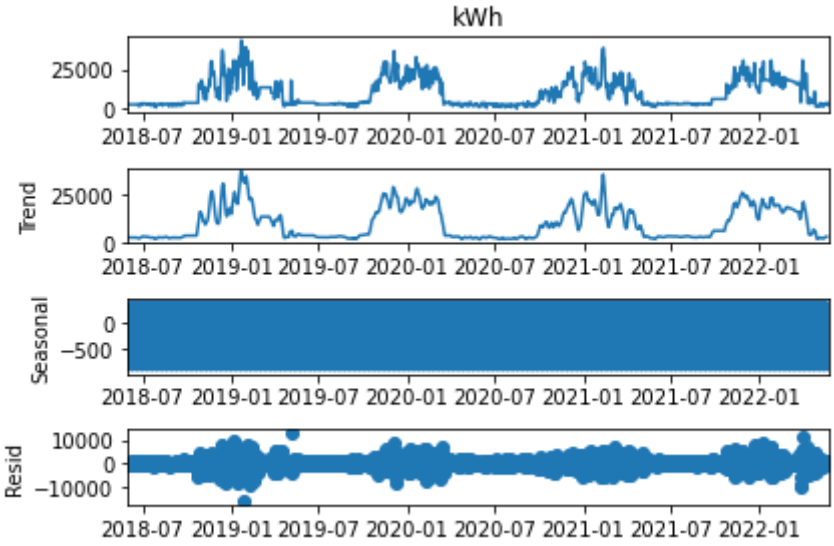
Pour passer de m3 au kWh ou encore au Nm3, nous utilisons une constante, en scalant nos données, nous obtiendrons le même signal pour les 3 données, nous avons donc décidé de ne prendre que le kWh.



Quantité de gaz en kWh consommée en rapport à la date

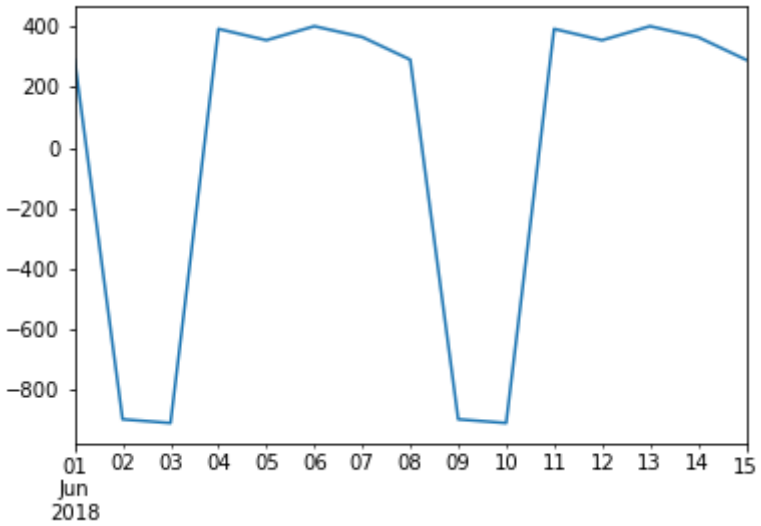
Pour ajouter les dates manquantes, nous appliquons un SMA sur la période et de même pour les valeurs manquantes aux dates présentes. Nous remarquons également un lissage du signal de 2018, où des observations sont rentrées pour une période plus grande, une observation par mois. Nous allons nous séparer de cette partie afin de préserver le signal.

Nous décomposons donc le signal :



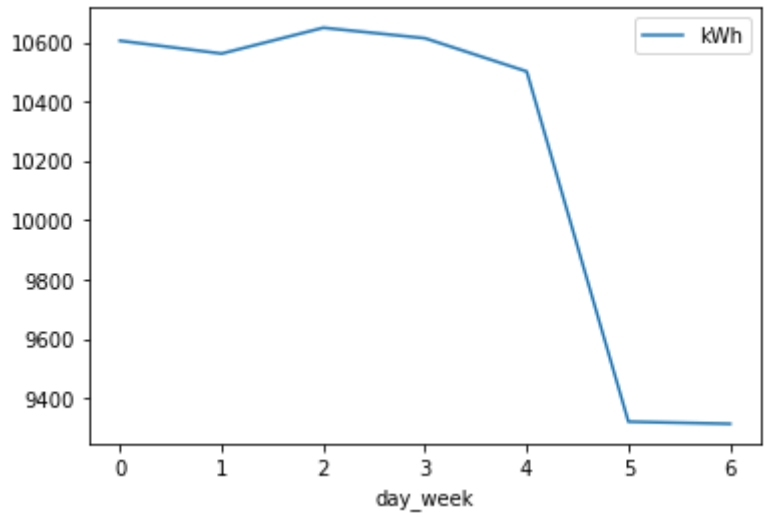
Décomposition du signal kWh

Nous pouvons observer que la « trend » se répète au fil des mois, en regardant la saisonnalité de plus près :



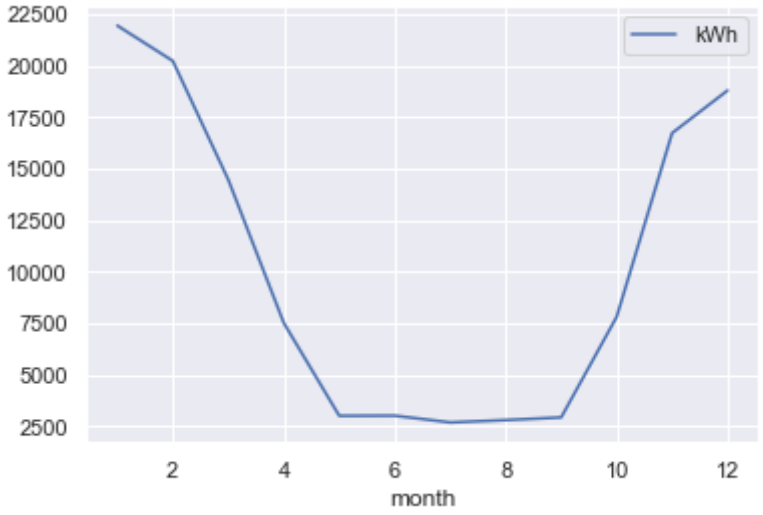
Saisonnalité du signal kWh

Le week-end du 02 et 03 juin 2018 marque comme les signaux précédant un creux lors des week-ends et quantité stable consommée durant la semaine.



Quantité de kWh moyenne consommé groupé par jour

Nous pouvons l'observer ici, avec une consommation réduite le week-end, le personnel étant réduit et les étudiants absents.



Quantité de kWh moyenne consommé groupé par mois

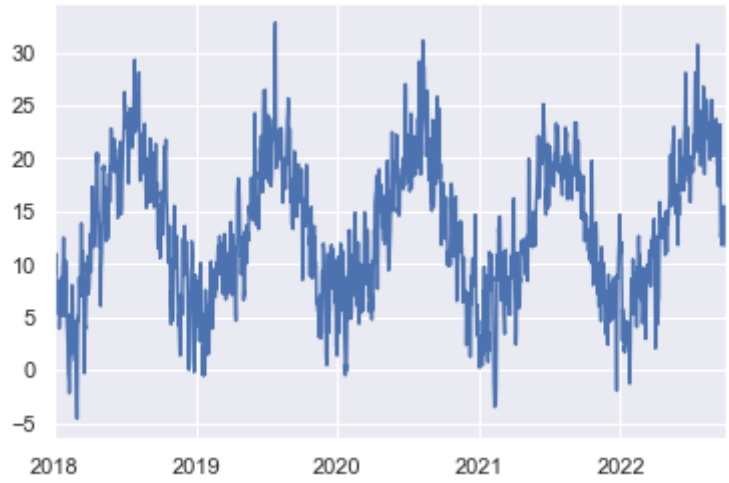
Nous pouvons voir ici un grand creux pour les mois de mai et début septembre, ceci peut se référer à la présence d'étudiant ou encore l'augmentation des températures. ESIEE Paris doit utiliser le gaz pour chauffer ses locaux et ainsi, à l'approche de l'hiver la consommation est à son maximale pour décembre et janvier et diminue. En émettant cette hypothèse, nous pouvons nous orienter vers l'utilisation de la température au cours de l'année, quand il fait plus froid, une plus grande quantité de gaz est utilisé pour chauffer les locaux.

Feature engineering

Nous avons donc déjà ajouté pour les trois jeux de données, le mois et le jour de la semaine. En suivant les hypothèses que nous avons émises précédemment, nous pouvons ajouter plusieurs features et les évaluer.

Température extérieure

Nous allons ajouter la température moyenne sur la journée à nos données, cette information peut être corrélée à la consommation de gaz, des températures plus faibles demandant une plus grande quantité de gaz pour chauffer le bâtiment à la même température. Nous prenons les températures moyennes dans le département du 93.



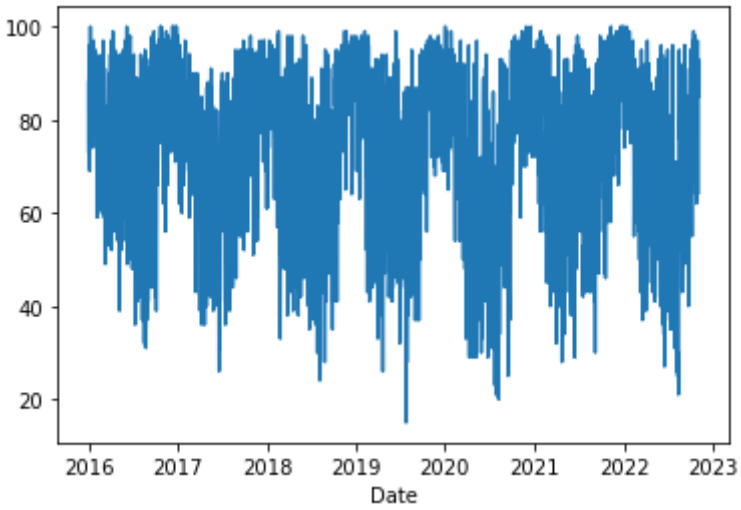
Température en °C moyenne par rapport à la date

Données provenant de :

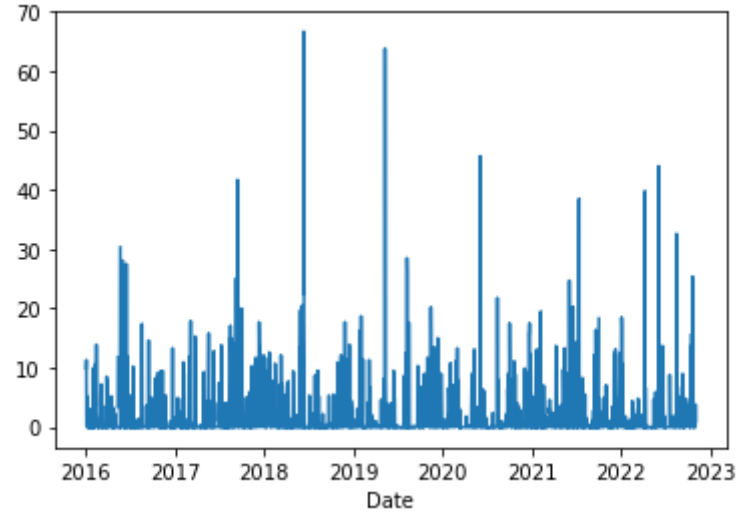
https://www.data.gouv.fr/fr/datasets/temperature-quotidienne-departementale-depuis-janvier-2018/

Humidité extérieure et Précipitations

En revenant sur notre hypothèse de consommation d'eau, nous pouvons également prendre l'humidité extérieure, il n'y a pas la nécessité d'arroser si le taux d'humidité est élevé ou encore s'il a plu. De même, une atmosphère humide est plus difficile à chauffer et la consommation de gaz ou d'électricité peut se voir augmenter.



Humidité en % par rapport à la date



Précipitation en mm par rapport à la date

Données provenant de :

https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm

Weekend, Vacances et Distanciel

C'est en suivant notre hypothèse sur les jours de la semaine et les mois que nous avons construits 2 nouvelles features, week-end qui représente la valeur binaire pour si c'est un samedi ou un dimanche et vacances qui englobent les grandes vacances scolaires en juillet jusqu'à fin août.

En plus de ces deux features, nous avons essayé d'implémenter les jours de distanciel :

- Les vendredis pour l'année scolaire 2022/2023
- Les lundis pour l'année scolaire 2021/2022
- Les deux jours distanciels lundi et vendredi de février 2021 à juillet 2021
- La période distanciel complet de novembre 2020 à février 2021
- Les deux jours distanciels lundi et vendredi de septembre 2020 à novembre 2020

Affluence

Cependant, nous devons noter que week-end, vacances, distanciel ou même jours de la semaine ne sont que des valeurs exogènes aux modèles et servent à simuler une feature que nous ne possédons pas, mais qui pourraient bénéficier le modèle : l'affluence de ESIEE Paris.

En effet, il y a une corrélation entre les vacances et l'affluence à ESIEE Paris, il y a plus de personnes quand ce ne sont pas les vacances, mais ce n'est pas car il y a moins de personnes que c'est nécessairement les vacances, d'où la présence du jour de la semaine, il peut y avoir moins de monde hors vacances si c'est un week-end ou encore un jour férié.

Ainsi, prenons le cas d'un lundi du même mois, avec les mêmes températures, mais une affluence plus grande sur le deuxième lundi (plus de cours que le précédent) :

1. Sans l'information sur l'affluence, le modèle serait tenté de prédire la même valeur
2. Avec l'information sur l'affluence, le modèle pourrait ainsi prédire une valeur différente

En effet, avec plus d'affluence, une consommation électrique ou d'eau plus importe est à envisager du fait de l'affluence, mais nous pouvons également émettre l'hypothèse qu'avec plus de monde, il y aurait moins à chauffer, car plus de monde, les salles se chaufferaient en partie grâce à la chaleur humaine. De plus, si tenté que la pièce est ventilée entre les cours, nous obtiendrons un cycle de chauffage et de refroidissement.

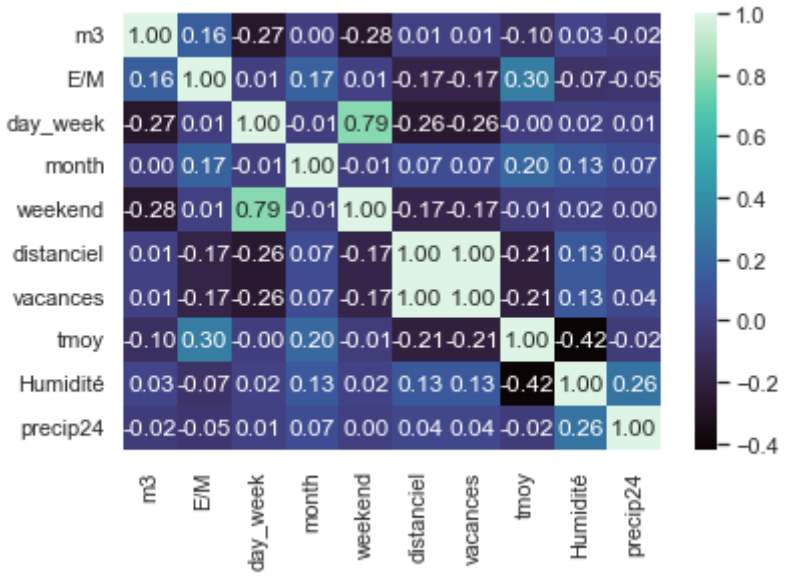
C'est avec ces considérations, que l'ajout d'une feature d'affluence qui, avec les outils que dispose l'ESIEE sur l'ADE, prédise l'affluence sur tel jour et que cette information soit donnée au modèle pour prédire avec nous l'espérons plus de précision.

Étude des features

Nous allons maintenant étudier les features et leurs corrélations avec la consommation d'eau, d'électricité et de gaz. Les trois sont partiellement corrélés entre eux, mais en raison du point que vous avons soulevé plus haut sur l'affluence, nous n'utiliserons pas la consommation d'eau pour prédire l'électricité ou encore le gaz et vice-versa.

Eau

Nous affichons en premier la matrice de corrélation entre les features :

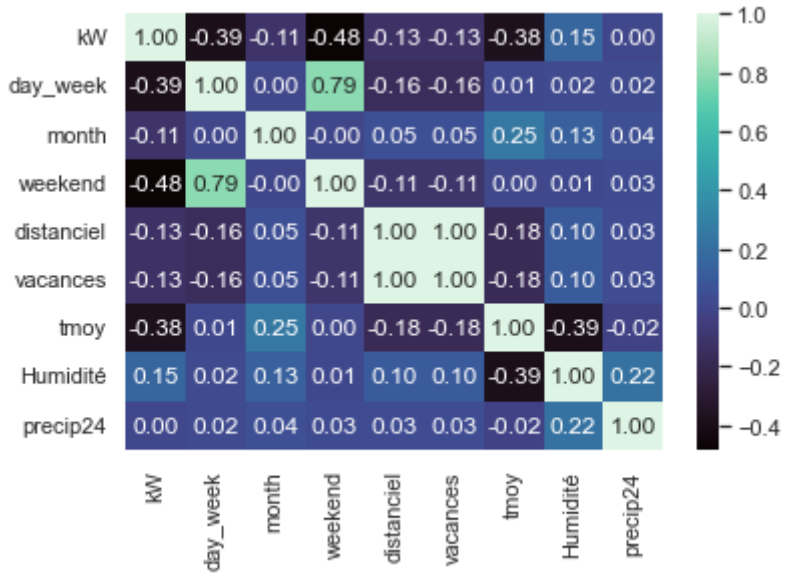


Matrice de corrélation Eau

Nous pouvons voir que ce qui est le plus corrélé à notre target est le jour de la semaine et la valeur binaire du week-end. Ils sont corrélés négativement, ce qui signifie que quand c'est un week-end, la consommation d'eau est diminuée. Ce sont des valeurs de classifications, les nuages de points ne transmettent donc peu d'informations, mais ils sont disponibles dans le notebook.

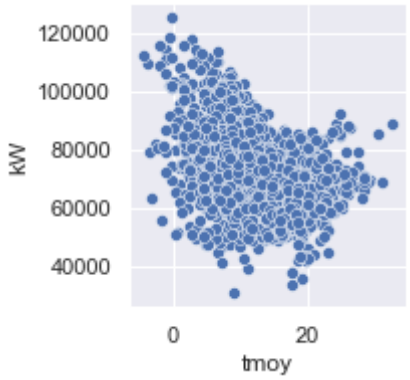
Électricité

Nous affichons en premier la matrice de corrélation entre les features :



Matrice de corrélation Électricité

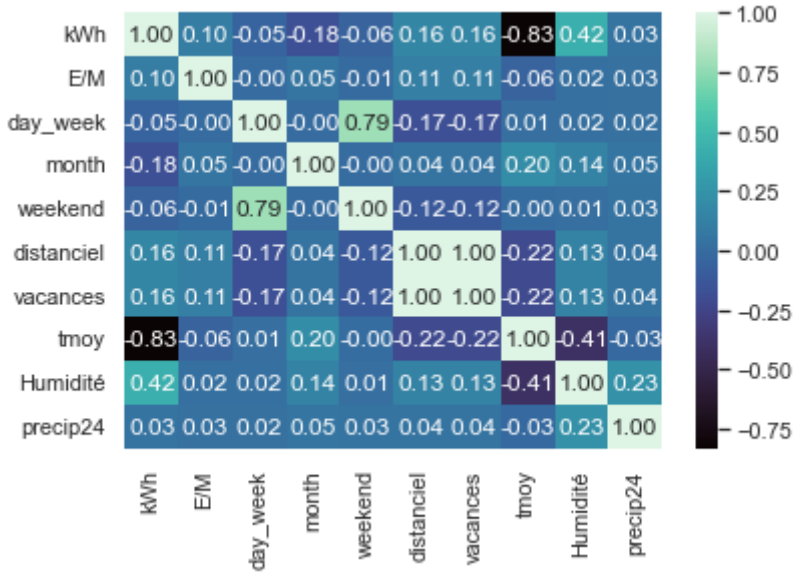
Nous pouvons voir que ce qui est le plus corrélé à notre target est le jour de la semaine, la valeur binaire du week-end, mais également la température moyenne. Les trois sont négativement corrélés, ce qui implique une consommation d'électricité réduite quand c'est le week-end ou alors quand les températures augmentent. Ce sont des valeurs de classifications, les nuages de points ne transmettent donc peu d'informations, mais ils sont disponibles dans le notebook, pour ce qui est de la température moyenne :



Consommation électrique en kW par rapport à la température moyenne

Gaz

Nous affichons en premier la matrice de corrélation entre les features :



Matrice de corrélation Gaz

Nous pouvons voir que ce qui est le plus corrélé à notre target est la température moyenne, mais également le taux d'humidité. La température est négativement corrélée, quand les températures descendent la quantité de gaz consommée augmente, à l'inverse quand le taux d'humidité augmente, la quantité de gaz consommée augmente, ce qui rejoint le fait que plus l'air est humide plus il est difficile à chauffer.

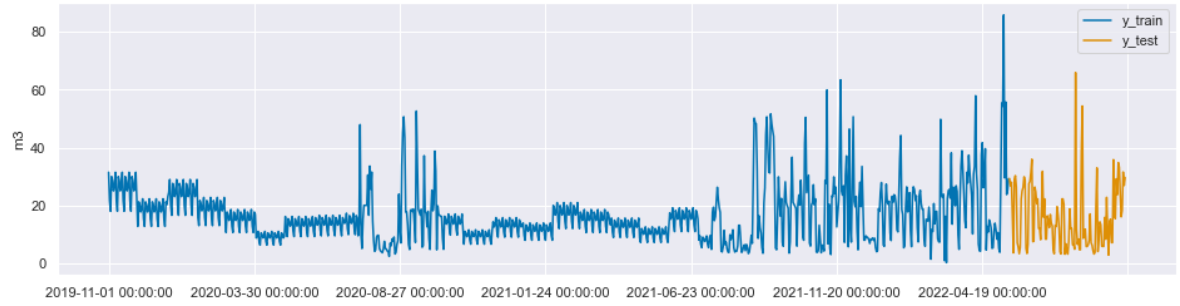
Mise en place des modèles

Nous allons maintenant mettre en œuvre trois modèles pour chacun des datasets, un suivant une méthode naïve, un suivant une méthode de deep learning et le dernier suivant une méthode statistiques.

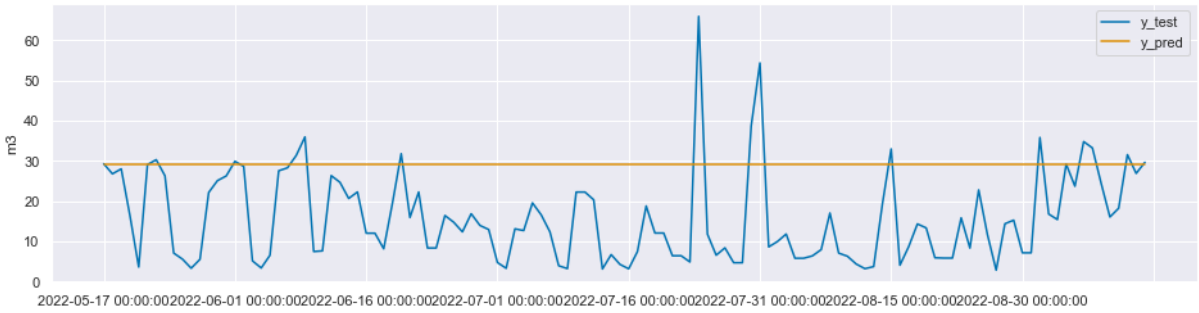
Eau

Naïve

Pour le signal suivant, avec un test size de 120 :

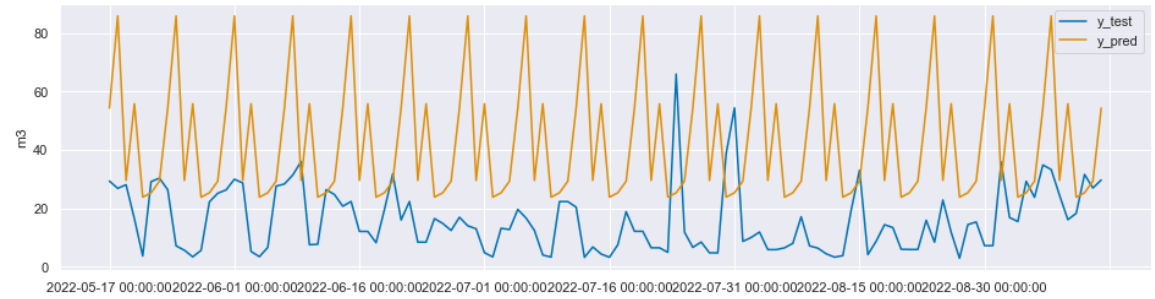


En utilisant prédicteur naïf nous obtenons le résultat suivant :



Avec un SMAPE (Symmetric mean absolute percentage error) de 0.5183.

En ajoutant la périodicité de 7 jours nous obtenons :

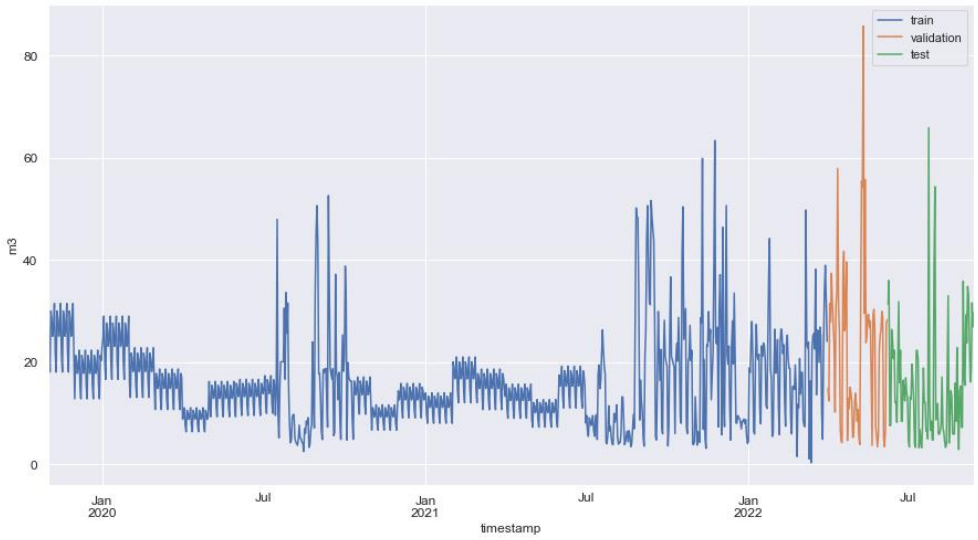


Avec un SMAPE de 0.6222

Ajouter la périodicité permet de suivre l'allure du signal, mais un offset est présent exacerbant ainsi l'écart et résultant en un taux d'erreur plus élevé.

Deep learning

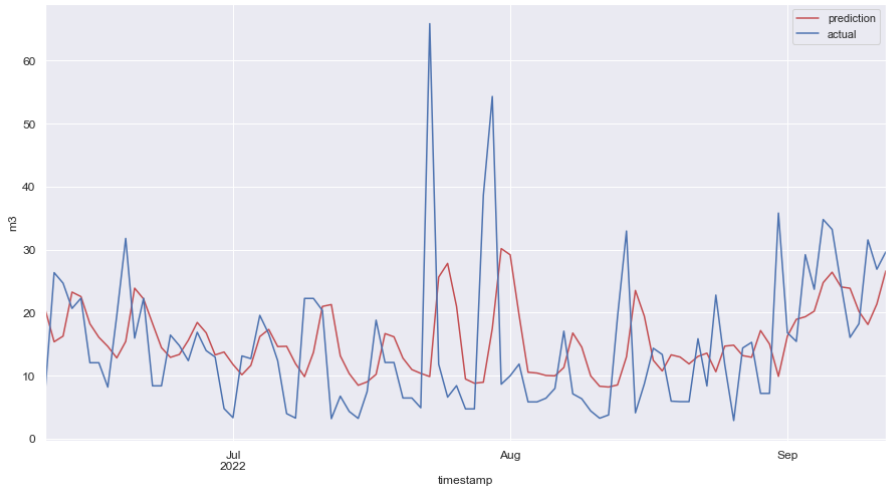
Avec un T = 3 et Horizon = 1 pour le signal suivant :



Pour le model RNN suivant :

```
Model: "sequential_20"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 gru_59 (GRU)                (None, 3, 15)             855

 gru_60 (GRU)                (None, 10)                810

 dense_20 (Dense)            (None, 1)                 11

=================================================================
Total params: 1,676
Trainable params: 1,676
Non-trainable params: 0
_________________________________________________________________
```

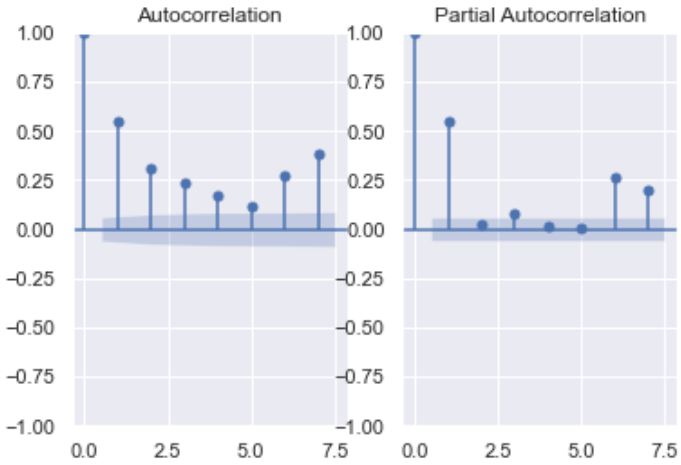
En utilisant le signal m3 d'origine et le jour de la semaine, nous obtenons les prédictions suivantes :



Avec un SMAPE de 0.5722.

Statistiques

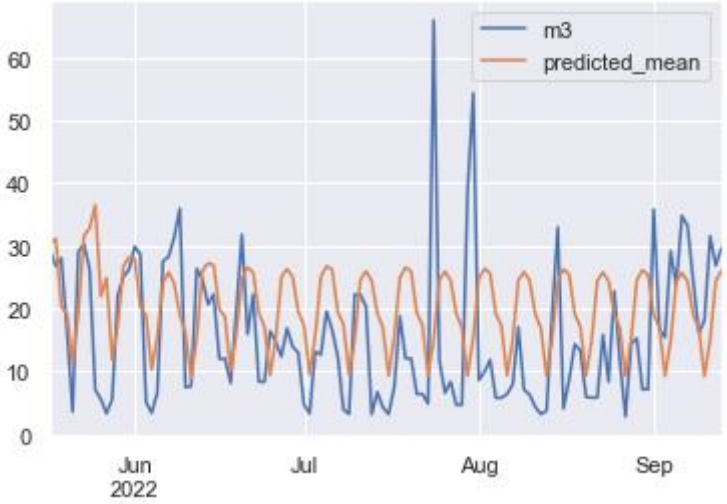
Nous allons implémenter un SARIMAX en régression dynamique, nous considérons le jour de la semaine comme valeur exogène et la consommation en m3 d'eau comme endogène.



Nous mettons donc en place un ARIMA (1,0,2), AR(1) avec le pic sur le acf, I(0) nous n'avons pas fait de transformation et MA(2) grâce au pic sur le pacf. On ajoute le seasonal order avec les valeurs :

- P = 7
- D = 0
- Q = 2
- M = 7 pour représenter la périodicité du signal

Nous obtenons la prédiction suivante :



Avec un SMAPE de 1.1542.

Comparaison

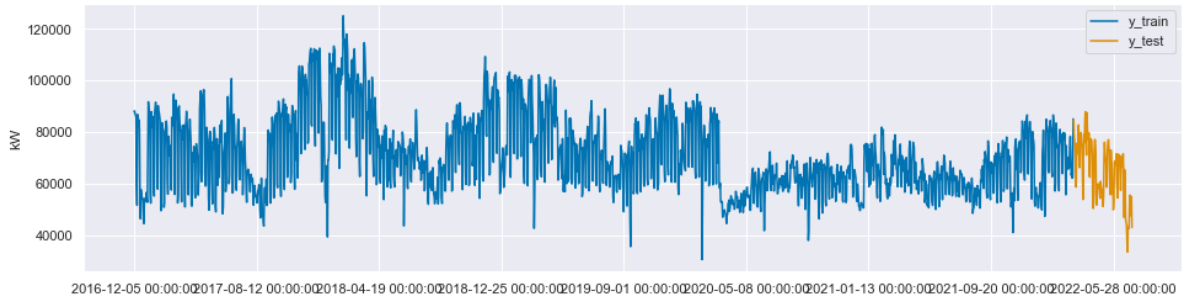
Nous pouvons voir que la méthode naïve sans périodicité n'est pas adaptée, en ajoutant la périodicité nous obtenons un meilleur résultat. Comparé au modèle avec réseau de neurones, il n'est pas meilleur, ni plus robuste, car il ne prend pas en compte les piques soudains et la tendance comme le fait le réseau de neurones ou encore le SARIMAX.

Pour ce qui est du modèle statistique, nous avons la tendance qui est prise en compte, mais comparé au réseau de neurones, il ne peut pas suivre les piques soudains ou du moins tenter de les prédire comme le fait le réseau de neurones. Nous notons cependant que le réseau de neurones est en décalage, il permet une bonne approximation générale, mais malgré qu'il soit le meilleur modèle pour détecter les piques, il peine quand même à le faire. Le SARIMAX prend plus en considération les fluctuations, mais le modèle en réseau de neurones est plus robuste et prédit en moyenne mieux.

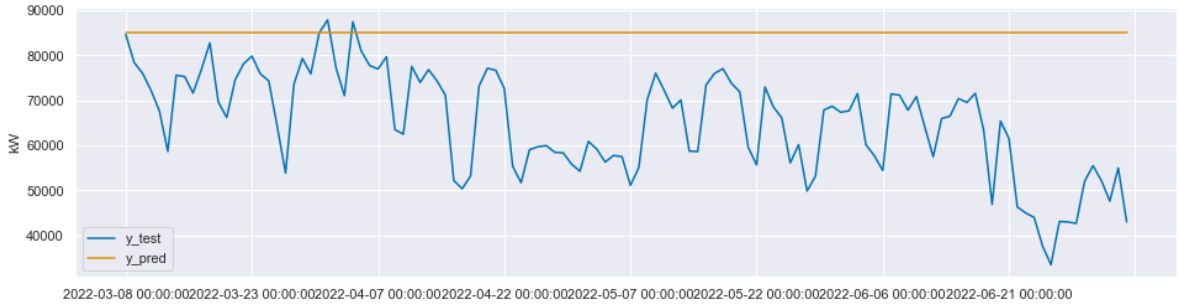
Électricité

Naïve

Pour le signal suivant, avec un test size de 120 :

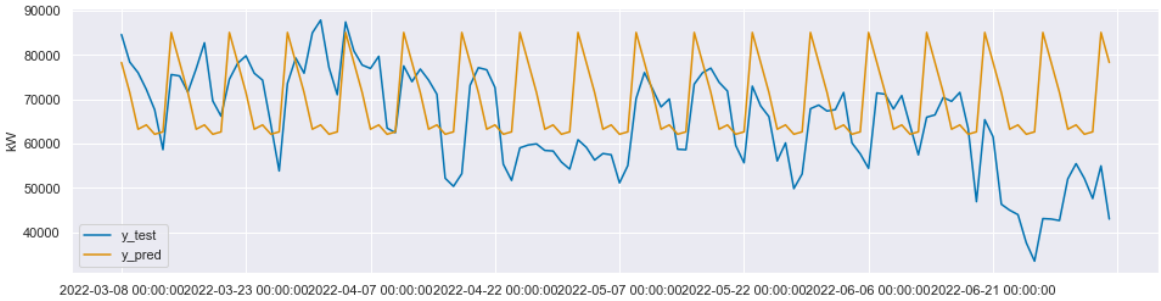


En utilisant prédicteur naïf nous obtenons le résultat suivant :



Avec un SMAPE de 0.2365

En ajoutant la périodicité de 7 jours nous obtenons :

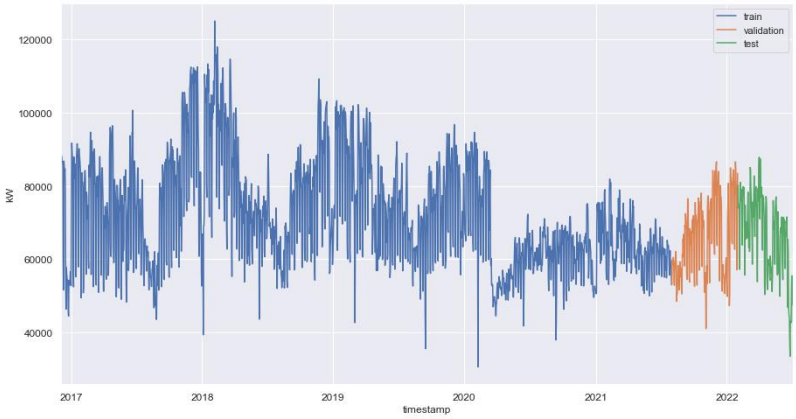


Avec un SMAPE de 0.15.

Ajouter la périodicité a permit d'obtenir un meilleur résultat.

Deep learning

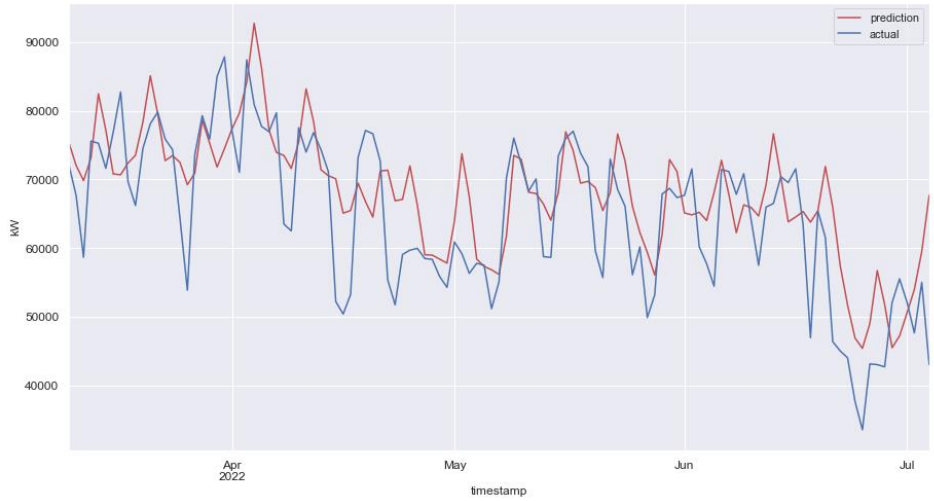
Avec un T = 3 et Horizon = 1 pour le signal suivant :



Pour le model RNN suivant :

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 gru (GRU)                   (None, 3, 25)             2250

 gru_1 (GRU)                 (None, 3, 15)             1890

 gru_2 (GRU)                 (None, 10)                810

 dense (Dense)               (None, 1)                 11

=================================================================
Total params: 4,961
Trainable params: 4,961
Non-trainable params: 0
_________________________________________________________________
```

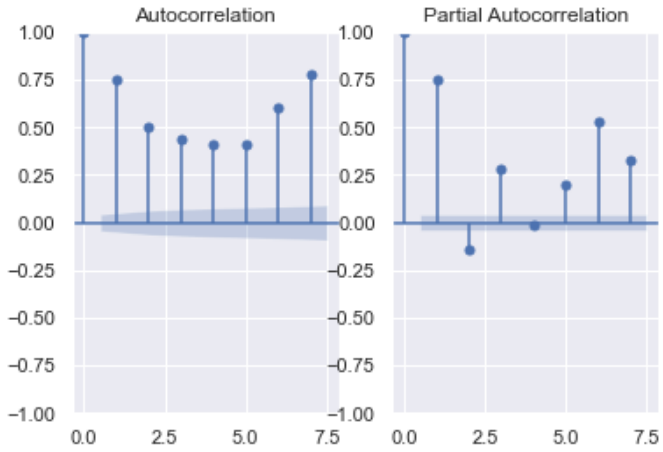
En utilisant, le signal kW, la température moyenne et si c'était un weekend, nous obtenons la prédiction suivante :



Avec un SMAPE de 0.0971.

Statistiques

Nous allons implémenter un SARIMAX en régression dynamique, nous considérons la température moyenne, le jour de la semaine et le weekend comme des valeurs exogènes et la consommation en kW comme endogène.

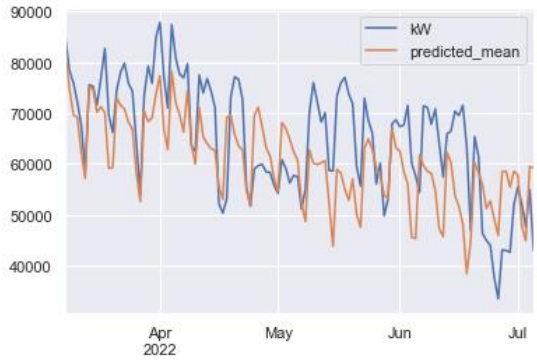


Nous avons trouvé un pic sur le acf à 7, ce qui est logique car nous avons des données journalières. Nous avons donc appliqué un lag de 7 pour visualiser l'acf et le pacf.

Nous mettons donc en place un ARIMA(1,0,4), AR(1) avec le pic sur le acf, I(0) nous n'avons pas fait de transformation et MA(4) grâce au pic sur le pacf. Nous ajoutons le seasonal order avec les valeurs :

- P = 7
- D = 0
- Q = 0
- M = 7 pour représenter la périodicité du signal

Nous obtenons la prédiction suivante :



Avec un SMAPE de 0.1287.

Comparaison

Le naïf, même en prenant en compte la périodicité du signal ne peut pas capturer la tendance et sera ainsi constant ce qui en fait déjà une mauvaise méthode comparée aux deux autres.

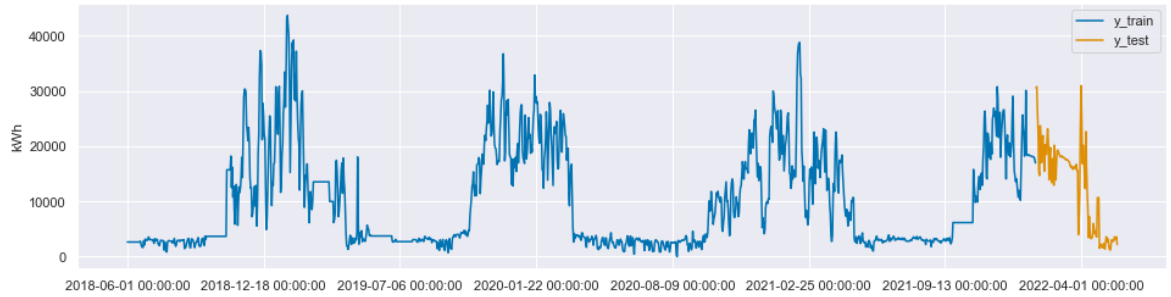
Pour ce qui est du SARIMAX et du modèle en réseau de neurones, là où le SARIMAX parvient à suivre les pics et à mieux épouser la tendance générale du signal en suivant les pics là où le model à réseau de neurones est légèrement en retard. Cependant, même si le SARIMAX suit la tendance, le modèle en réseau de neurones peut prédire les pics arrivant d'un coup comme pour le mois de mai à juin, où le réseau de neurones réussit à le prédire.

Le modèle en réseau de neurones est ainsi plus robuste au changement brusque et à la fluctuation sur la tendance, mais le SARIMAX capte mieux la périodicité.

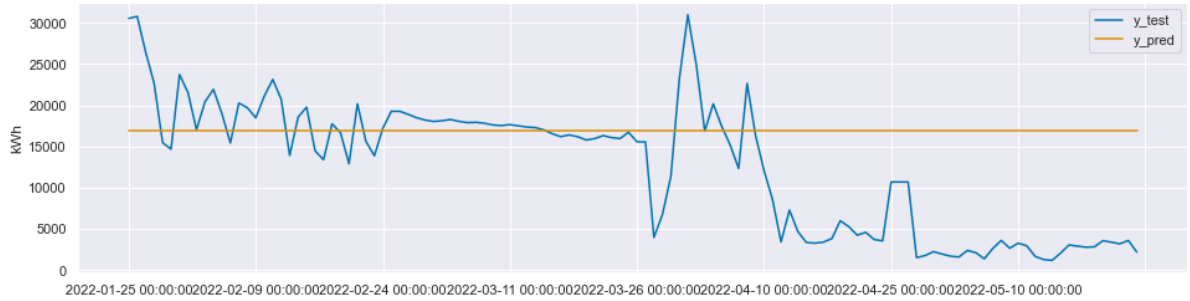
Gaz

Naïve

Pour le signal suivant, avec un test size de 120 :



En utilisant prédicteur naïf nous obtenons le résultat suivant :



Avec un SMAPE de 0.3977.

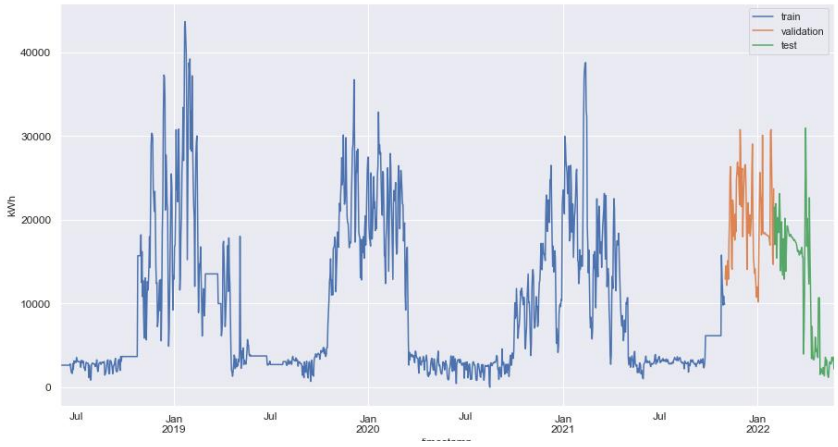
En ajoutant la périodicité de 7 jours nous obtenons :

Avec un SMAPE de 0.3959.

Ajouter la périodicité n'a que légèrement amélioré le modèle.

Deep learning

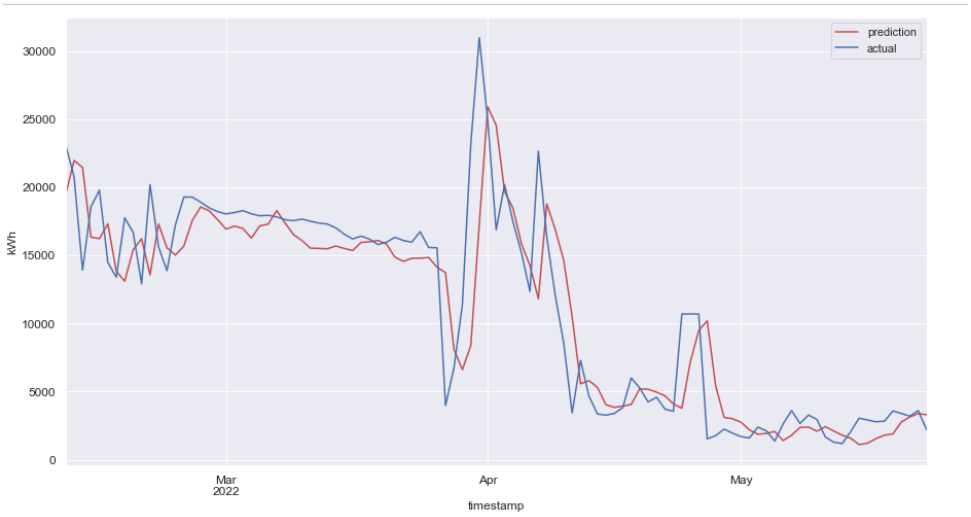
Avec un T = 10 et Horizon = 1 pour le signal suivant :



Pour le model RNN suivant :

```
Model: "sequential_25"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 gru_73 (GRU)                (None, 10, 25)            2250

 gru_74 (GRU)                (None, 10, 20)            2820

 gru_75 (GRU)                (None, 15)                1665

 dense_25 (Dense)            (None, 1)                 16

=================================================================
Total params: 6,751
Trainable params: 6,751
Non-trainable params: 0
_________________________________________________________________
```

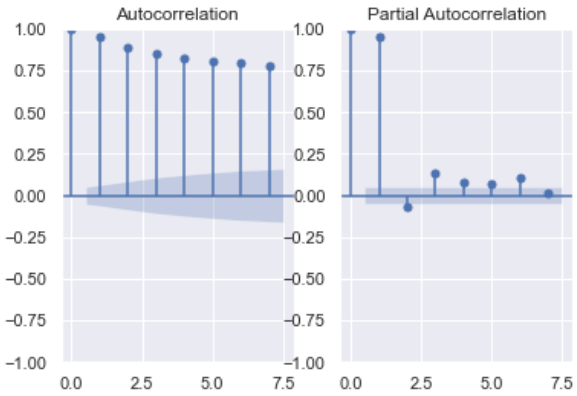
En utilisant, le signal kWh, la température moyenne et l'humidité extérieure, nous obtenons la prédiction suivante :



Avec un SMAPE de 0.3312.

Statistiques

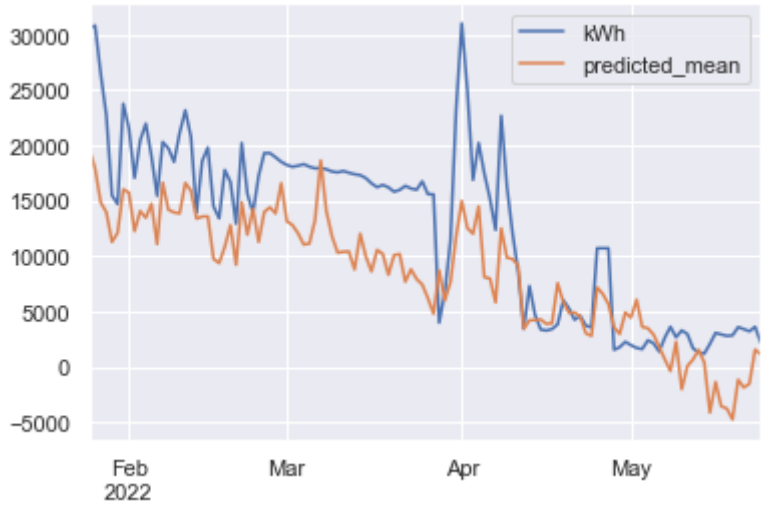
Nous allons implémenter un SARIMAX en régression dynamique, nous considérons la température moyenne, le jour de la semaine et le weekend comme des valeurs exogènes et la consommation en kWh comme endogène.



Nous mettons donc en place un ARIMA(1,0,7), AR(1) avec le pic sur le acf, I(0) nous n'avons pas fait de modification et MA(7) grâce au pic sur le pacf. Nous ajoutons le seasonal order avec les valeurs :

- P = 5
- D = 0
- Q = 6
- M = 7 pour représenter la périodicité du signal

Nous obtenons la prédiction suivante :



Avec un SMAPE de 0.5271.

Comparaison

Le model en réseau de neurones reste le meilleur, malgré un léger retard il suit le signal et parvient à relativement bien le prédire en suivant la tendance. Le modèle naïf lui ne suit pas la tendance et même en appliquant la périodicité, ses performances sont moindres comparés au SARIMAX, qui malgré quelques écarts réussit à suivre la tendance du signal et les fluctuations de façon assez précise, surement grâce à la température et l'humidité qui sont fortement corrélées à la consommation de gaz. A pars un pique sur le mois de mars, nous voyons que la prédiction est une projection sous-estimée de la consommation réelle. Ainsi, le SARIMAX est une bonne alternative, même si le réseau de neurones reste le plus robuste au changement brusque et à la tendance.

Conclusion

Nous avons vu que dans la majorité des cas, un réseau de neurones est plus général et permet de prédire les fluctuations inattendues au détriment de d'un léger retard de prédiction, là où le SARIMAX suit mieux la saisonnalité au détriment de la possibilité de prédire et prendre en compte les piques ou creux inattendues.