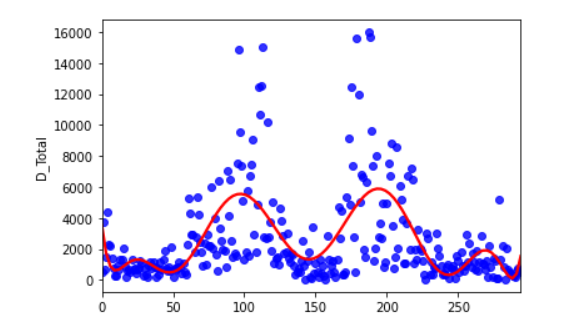
|  |
| --- |
| Université de Montpellier |
| Prédiction de fréquentation des cyclistes dans Montpellier |
| [Sous-titre du document] |

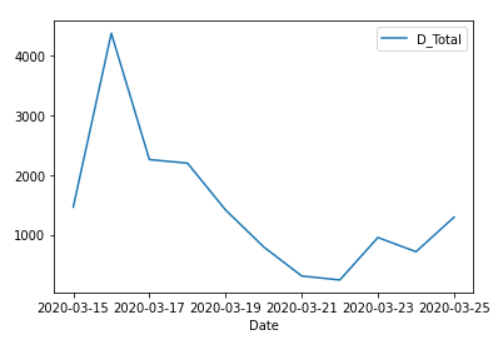
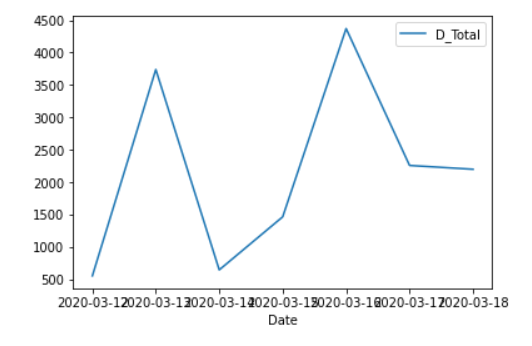
|  |
| --- |
| FRANC Pierre-Louis  21/03/2021 |

# Introduction

 Il est nécessaire de commencer par étudier la distribution des données.

On observe ici qu’il y a un socle de base de 0 à 2000 cyclistes par jour toute l’année. De plus, on repère 2 pics pour lesquels on soumet l’hypothèse de l’augmentation des températures et l’arrivé d’une météo plus clémente. Cependant il y a un gap entre ces deux pics qui arrive au moment de la période des vacances scolaires (commence début juillet et termine fin aout).

Ces deux hypothèses mettent en avant le principe de saisonnalité, c’est-à-dire que si elles s’avèrent être vrai ce phénomène se répètera d’année en année.

D’un point de vue plus microscopique, une répétition est identifiable par semaine. Cependant cette répétition n’est pas stable (si l’hypothèse de la météo est vérifiée cette non-stabilité pourrait être expliqué) et ne permet pas d’identifier une saisonnalité précise.

## Les données :

Les données fournies sont celle de la fréquentation de cycliste dans Montpellier pour différent créneau horaire dans une journée sur un an (de mars 2020 à mars 2021).

|  |  |
| --- | --- |
| La date (format date time) | La fréquentation (Integer) |
| La date | Date time |
| La fréquentation | Un entier naturel |

D’après nos hypothèses posées en introduction, il a été nécessaire de récupérer les données météo à Montpellier sur la même période.

Données météo

|  |  |
| --- | --- |
| Type de donnée | Format |
| Température maximale | Entier |
| Température minimal | Entier |
| Vitesse max du vent | Entier naturel |
| Précipitation sur la journée | Décimal |
| Humidité | Pourcentage |
| Couverture nuageuse | Pourcentage |
| Indice de chaleur | Entier |
| Température ressentie du vent | Entier |
| Temps de soleil | Heure |

Données des vacances.

En science des données lorsqu’un problème se base sur une série de données en fonction du temps il faut se pencher sur les techniques des séries temporelles. Trois grands choix s’offrent à nous.

## Choix du modèle :

En étude de série temporelle 3 types de technique sont utiliser, un modèle appeler ARIMA qui va décomposer le signal en 3 signaux : la tendance la saisonnalité et le bruit

Le modèle Fast fourrier transformation (FFT) qui représente la composition de plusieurs signaux sinuosidaux. La répétition de ces signaux permettra d’identifier la saisonnalité et d’émettre une prévision.

Ces deux modèles ont l’inconvénient de supposé qu’il y ait une saisonnalité et ne prends pas en compte les événement extérieurs comme la météo ils sont bien pour dégager une tendance mais ont plus de mal pour une prévision a un instant T

La dernière solution est d’établir un modèle de deep learning. Ce modèle nécessite une grande quantité de donnée en supplément de la valeur de fréquentation des cyclistes. D’après notre première observation et sachant que toutes les valeurs sont indépendantes, pour mettre en place ce modèle des données de météorologique serait nécessaire ainsi que des données sur les vacances scolaires. Cependant ce modèle à l’inconvénient qu’il serait biaisé en se basant uniquement sur la météo et les vacances car d’autres facteurs peuvent rentrer en jeu

Pour pouvoir comparer nos différents modèles nous utiliserons la Root mean square error (RMSE) qui permet de mesurer la différence entre les données réelle et leur prévision par le modèle.

Dans notre cas nous avons commencé par un modèle de Deep learning pour avoir une première idée du résultat

# Modèle de Deep Learning

Le modèle de deep learning est une méthode qui consiste à obtenir, grâce à un grand ensemble de données, différentes une corrélation qui mène à un résultat.

Dans notre cas, nous avons mis en place un modèle de Dense Neural Network à l’aide de la librairie Keras . Le but étant d’inséré les données météo jour par jour et d’obtenir la fréquentation des cyclistes en sorti.

L’architecture optimal trouver se composait de 6 couches de neurones chacune activé par une unité linéaire rectifié (ReLu) :

Une couche d’entrer composé de 10 neurones correspondant au 10 données météorologique et la données des vacances.

Couche 2 : 15 neurones

Couche 3 : 15 neurones

Couche 4 : 8 neurones

Couche 5 : 4 neurones

Couche 6 : un neurone de sorti

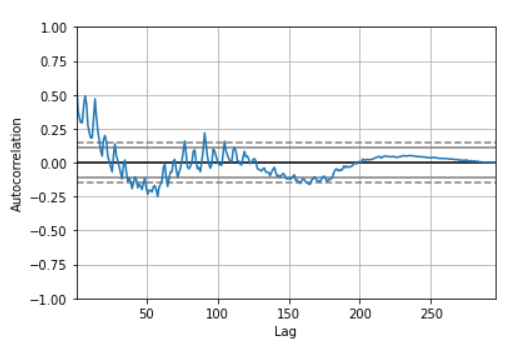
Ce modèle a obtenu un RMSE de 2915 sachant que la fréquentation moyenne dans Montpellier est de 2392.

Ce résultat étant loin de nos attentes D’autre méthode se sont offerte a nous.

# Modèles ARIMA

Cette méthode est propre aux prévisions sur séries temporelle.

## Méthodologie suivie :

Nous avons vérifié dans un premier temps l’autocorrélation d’un grand nombre de lag. On identifie une sinusoïde qui tends vers 0. Nous utiliserons un AR d’ordre de 2 donc p=2

Nous avons ensuite étudié si les données suivaient une tendance. Deux approches s’offraient à nous, l’approche macro en considérant que les données formeront une saisonnalité par année. La seconde (approche micro) étant de considéré une saisonnalité par semaine.

### Approche Micro :

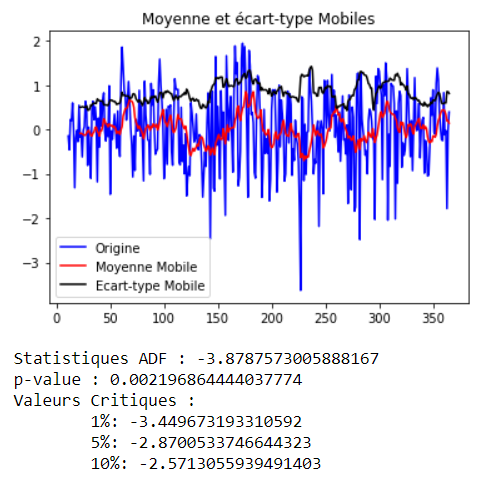
Les données par semaine ont une répétition mais trop chaotique. Beaucoup de variation et de bruit s’y insère. Pour cette approche la tendance est complexe à modéliser.

### Approche Macro :

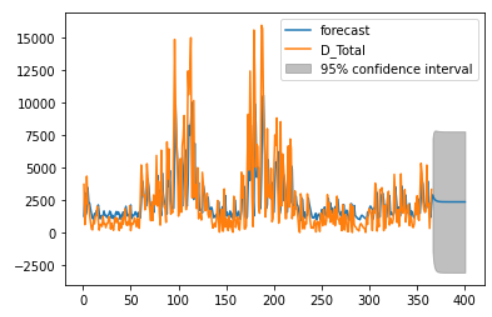
Ayant des données sur qu’une année c’est une hypothèse de considéré que les données suivent une saisonnalité par année. Cette approche permet de dire que les données sont stationnaires.

Sachant que les données ont une distribution stationnaire le facteur I est nul

Nous avons suivi l’approche Macroscopique.

Nous avons dans un premier temps loguarisé les données ce qui nous a permis de nous affranchir des problèmes de tendance. Voici les résultats :

En utilisant le modèle ARIMA de la librairie statsmodèle nous avons obtenu un RMSE de 2806

Ce score étant encore loin de nos attentes une autre méthode s’offre à nous en termes de séries temporelle.

Les transformations de fourrier rapide (FFT) se compose par la répétition périodique d’un schéma de données. Sachant que l’arima repose sur le principe de saisonnalité et qu’il n’a pas réussi à en détecter une, nous avons pris la décision de ne pas tester de modèle FFT.

Ayant des données météo nous nous sommes rabattus sur une régression linéaire

# Modèle de régression linéaire

La régression linéaire est une méthode qui nécessite plusieurs types de donnée. En se basant uniquement sur les données de fréquentations des cyclistes nous ne pouvons pas utiliser un modèle de régression linéaire. Ayant les données météorologiques nous pouvons utiliser ce type de modèle cependant cela signifie que nous prenons en hypothèse que la météo à un impact sur l’utilisation du vélo à Montpellier.

En Appliquant ce modèle nous avons obtenu un RMSE de 2180 ce qui est le score le plus faible que nous ayons obtenus.

Sachant que R2= 0.39 c’est aussi le R2 le plus grand obtenu. Cela signifie qu’environ 39% des variations observé peuvent être expliqué par les données d’entré du modèle.

# Conclusion

D’après l’étude de chacun de nos modèles sur un jeu de données d’un an, aucune saisonnalité n’est observable. En revanche l’hypothèse sur la météo est en partie vérifié par les résultats de la régression linéaire. Elle impacte donc la fréquentation des cyclistes dans Montpellier d’environ 39%.

Le but de ce projet était d’émettre une prévision du nombre de cycliste le 2 avril. Ce modèle a pour données d’entrés des données météo. Les prévisions actuelles pour le 2 avril étant les suivantes :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MAX\_TEMPERATURE\_C, | MIN\_TEMPERATURE\_C | WINDSPEED\_MAX\_KMH | PRECIP\_TOTAL\_DAY\_MM | HUMIDITY\_MAX\_PERCENT | CLOUDCOVER\_AVG\_PERCENT | HEATINDEX\_MAX\_C | WINDTEMP\_MAX | SUNHOUR | Hollyday |
| 18 | 11 | 45 | 0 | 31 | 12 | 14 | 13 | 11.5 | 0 |

Ce qui donne une prévision de 12071 personnes utilisant leur vélo dans Montpellier à cette date.

Plusieurs aspect ressorte de cette étude, dans un premier temps que la météo et les vacances d’été ont un rôle dans la pratique du vélo à Montpellier. Dans un second temps, qu’aucune saisonnalité n’a pu être déterminé avec ce jeu de donnée mais qu’avec un ensemble sur plusieurs années, les modèles que nous avons étudiés pourraient avoir une efficacité bien supérieure.