Étude d'un système de vélos partagés

Samuel TAULEIGNE

Lycée La Martinière Monplaisir, Lyon

2018

Introduction

- Système de vélos partagés (Vélo'v) : utilisateurs occasionnels ou réguliers
- Stations fixes permettant la mobilité
- But de ce travail : étudier la disponibilité des vélos aux bornes des stations Vélo'v, tenter d'en dégager des tendances d'évolutions voire de prévoir cette évolution
- Ce que nous allons faire : établir plusieurs méthodes de modélisation et de prévision, puis faire intervenir (et varier) des paramètres extérieurs.

Table des matières

Base de données

Un grand jeu de données à manipuler

- 348 stations Vélo'v (338 sont présentes dans notre base de données) comportant 4000 vélos
- Entre 10 et 40 bornes à chaque station
- Environ 3 mesures par heure pour chaque station chaque jour
- 6 mois d'études (maintenant 8)
- Base de données de taille totale : 5.000.000 d'entrées
- Traitement préalable nécessaire (données anarchiques)

Base de données

Voyez vous-mêmes ...

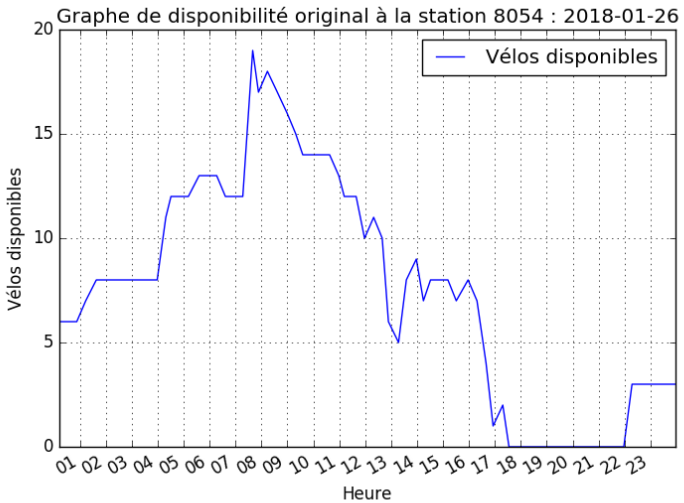
*data_all_Lyon_Septembre_2017

Chargement de **data_all_Lyon_Septembre_2017** depuis **Volume de 4,0 GB /media/samuel/44D5-77FD/TIPE/Données**


```
[{"status": "OPEN", "contract_name": "Lyon", "download_date": 1504240812, "bike_stands": 16
"download_date": 1504240812, "bike_stands": 22, "number": 1002, "last_update": 150424054800
15, "number": 1003, "last_update": 1504240541000, "available_bike_stands": 15, "available_b
1504240355000, "available_bike_stands": 10, "available_bikes": 0}, {"status": "OPEN", "cont
"available_bikes": 7}, {"status": "OPEN", "contract_name": "Lyon", "download_date": 1504240
"contract_name": "Lyon", "download_date": 1504240812, "bike_stands": 11, "number": 1013, "l
1504240812, "bike_stands": 17, "number": 1016, "last_update": 1504240462000, "available_bik
"last_update": 1504240459000, "available_bike_stands": 18, "available_bikes": 0}, {"status"
"available_bike_stands": 9, "available_bikes": 1}, {"status": "OPEN", "contract_name": "Lyo
3}, {"status": "OPEN", "contract_name": "Lyon", "download_date": 1504240812, "bike_stands":
"Lyon", "download_date": 1504240812, "bike_stands": 17, "number": 1024, "last_update": 1504
"bike_stands": 20, "number": 1031, "last_update": 1504240638000, "available_bike_stands": 2
"last_update": 1504240503000, "available_bike_stands": 21, "available_bikes": 1}, {"status"
"available_bike_stands": 16, "available_bikes": 0}, {"status": "OPEN", "contract_name": "Ly
13}, {"status": "OPEN", "contract_name": "Lyon", "download_date": 1504240812, "bike_stands"
```

Base de données
Choix de la représentation

Mon intérêt s'est porté plus spécialement sur l'occupation d'une station Vélo'v, ainsi on trace le nombre de vélos disponibles en fonction du temps :



Le nombre total de bornes est 20.
Il n'y a aucun vélo disponible de 17h30 à 22h.
Ce jour-là, on peut déposer un vélo à tout moment de la journée.

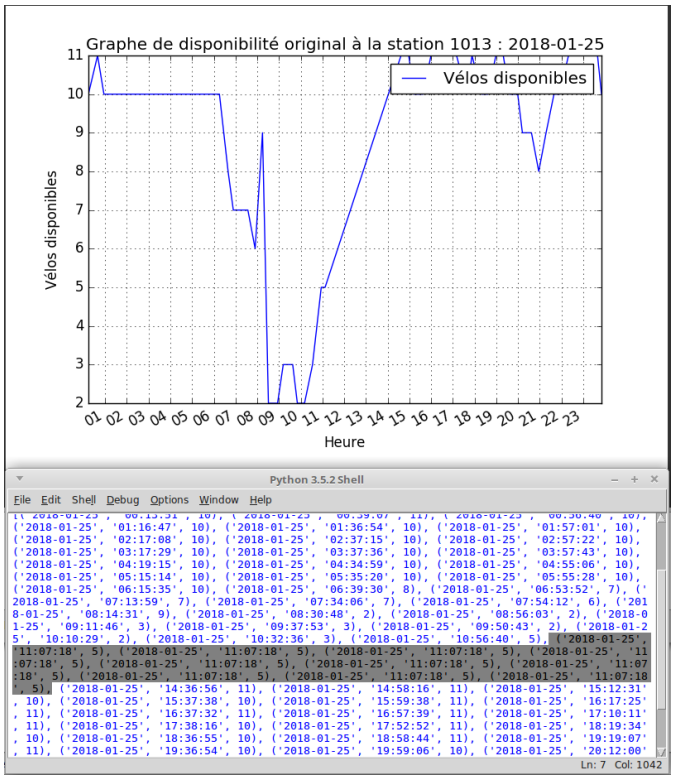
Base de données
Ma base de données

J'ai collecté les données publiques mensuelles du système Vélo'v, puis je les ai traitées pour ne conserver que les données utiles.

Ainsi, j'ai construit une base de données regroupant les mesures de disponibilité de vélos aux stations Vélo'v.

Base de données

Mais la réalité a rattrapé cet idéal de données ...



Voici un exemple de donnée erronée :

- Une plage de données constantes douteuse
- Mais surtout une plage de données absente !

Notre base de données est donc incomplète, erronée.

Base de données

Comment gérer le manque de données ?

Plusieurs solutions sont envisageables :

1. Retirer la journée comportant des données erronées
2. Ne pas considérer la tranche horaire concernée par les données erronées

La première ne convenant pas pour effectuer des prévisions, nous choisissons la deuxième méthode en adaptant les algorithmes.

Algorithme naïf
Principe de l'algorithme

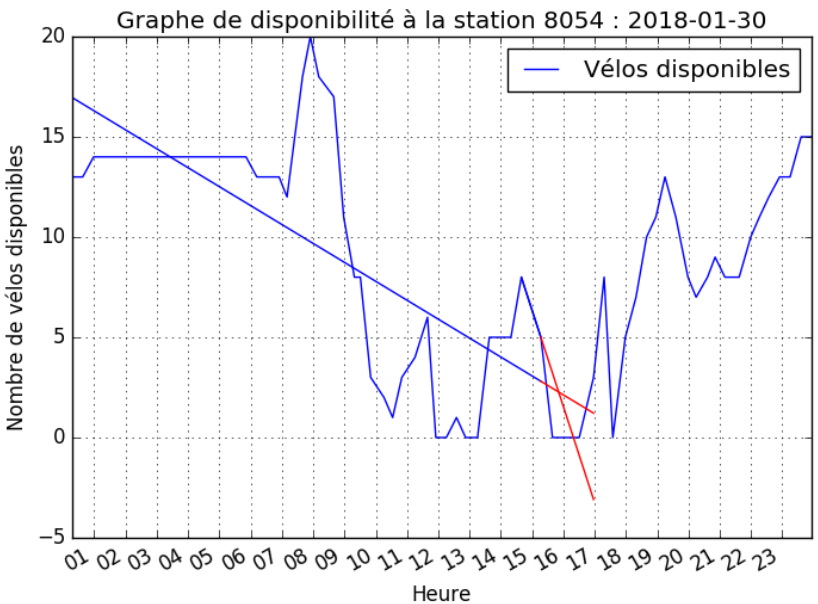
Cet algorithme ne nécessite que les dernières valeurs :

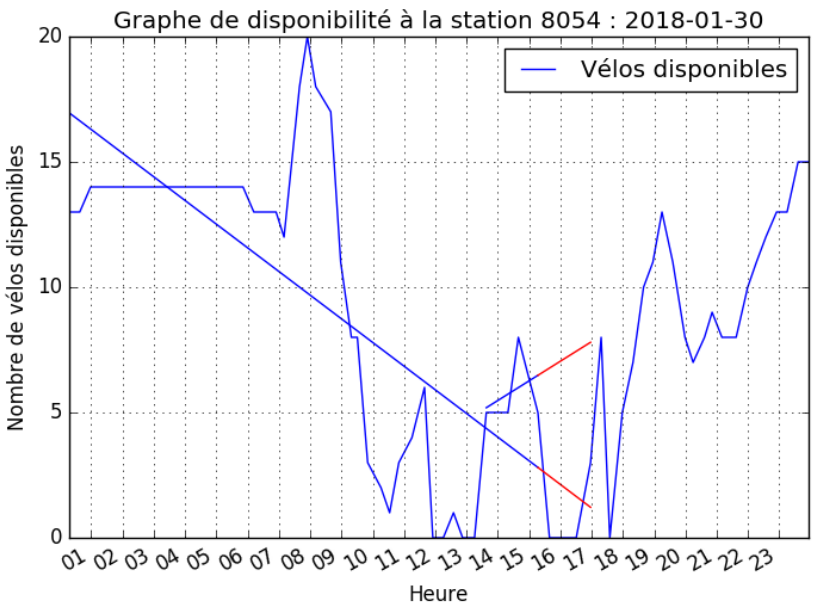
1. On fixe le nombre de données que l'on veut prendre en compte.
2. Il est ensuite possible d'effectuer une régression linéaire aux précédentes données.
3. On trace donc cette droite afin de visualiser la modélisation ainsi que la prévision.

Cela revient à considérer que la consommation de vélos est constante.

Algorithme naïf
Naïf mais efficace

En effet, pour une première approximation, la modélisation et la prévision semblent efficaces si l'on peut faire varier efficacement les paramètres de régression linéaire (en cours):





Algorithme par saisonnalité
Principe de l'algorithme

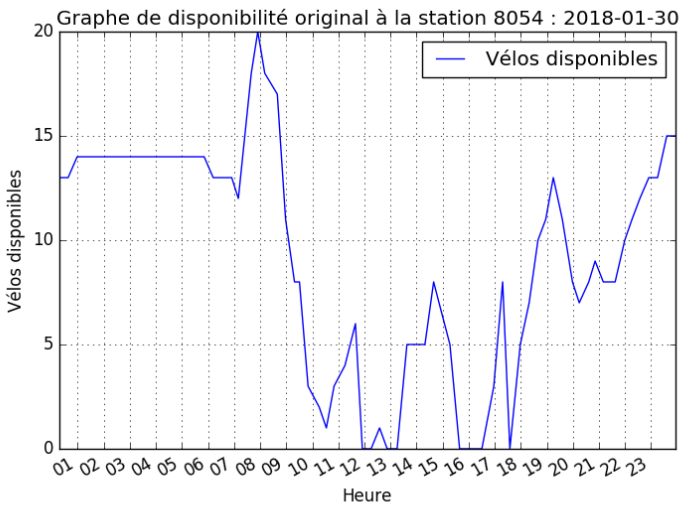
Cet algorithme vise à dégager des tendances globales d'évolution des disponibilités :

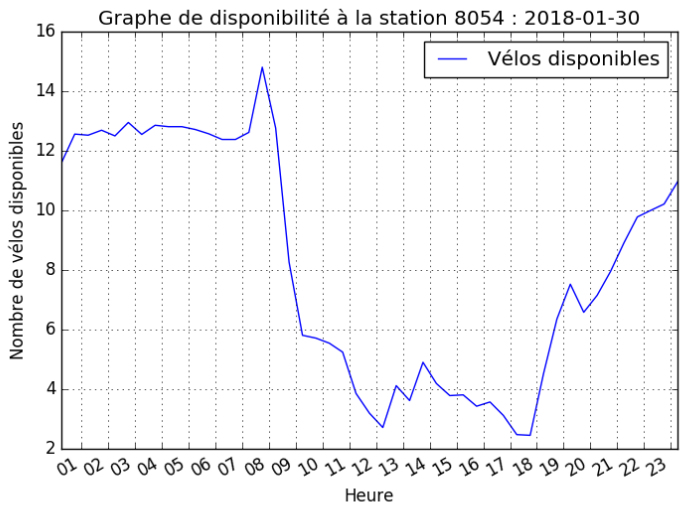
1. Identifier le jour de la semaine correspondant au jour pour lequel on effectue la prévision
2. Création d'un ensemble de vecteurs de données correspondant
3. Calcul de la moyenne globale sur chaque tranche horaire définie

Cette modélisation consiste à considérer que les semaines se répètent à l'identique.

Algorithme par saisonnalité
Une efficacité discutable

En effet, des événements liés à la vie de la ville peuvent influer sur les valeurs et donc perturber les calculs si l'on n'utilise pas assez de valeurs.





Algorithme par recherche de semblables

Principe de l'algorithme

Cet algorithme vise à comparer les données déjà collectées pour cette journée à toutes les autres journées de la base de données :

1. Calculer la somme des carrés des écarts entre les mesures
2. Déterminer le minimum de ces écarts
3. Tracer la courbe du jour semblable

Cette modélisation consiste à penser que le début de la journée détermine l'évolution de la disponibilité ce jour-là.

Algorithme par recherche de semblables

Une modélisation fidèle ...

La méthode de comparaison est fidèle puisqu'on cherche à minimiser les écarts.

Ainsi, on reproduit la journée qui a donné des mesures similaires.

La prévision à court terme est ainsi assez efficace.

Algorithme par recherche de semblables

Qui pose tout de même problème pour les prévisions

Mais il y a tout de même discontinuité au niveau des dernières données collectées.

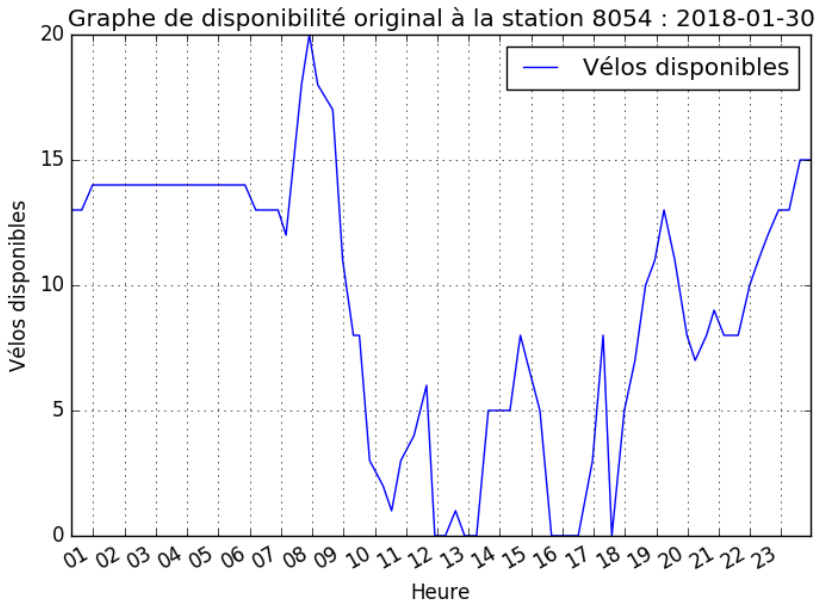
La méthode peut poser problème si on ne prend pas en compte assez de valeurs.

Mais la prévision à plus long terme est difficile.

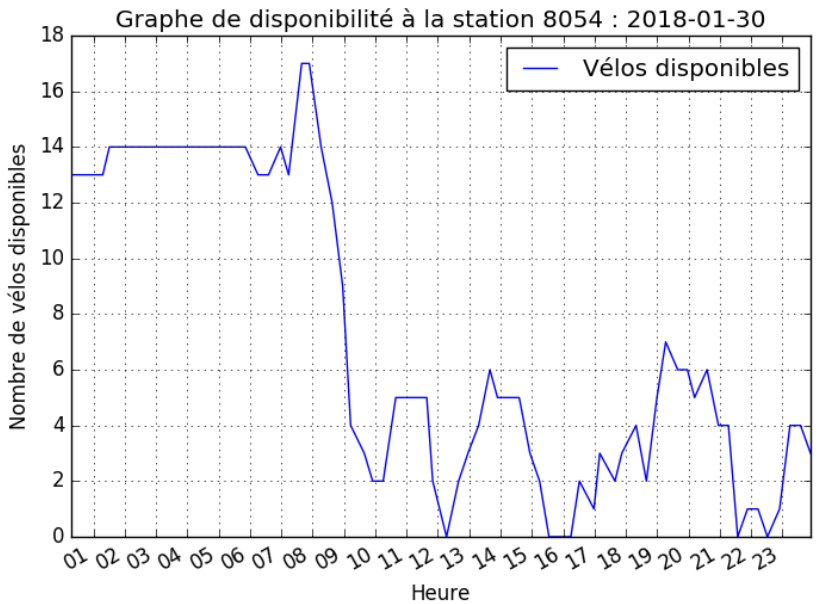
Algorithme par recherche de semblables

Voyez plutôt :

La situation originale :



La prévision (après 15^h) :



Analyse des résultats
Méthode de calcul de fiabilité RMSE

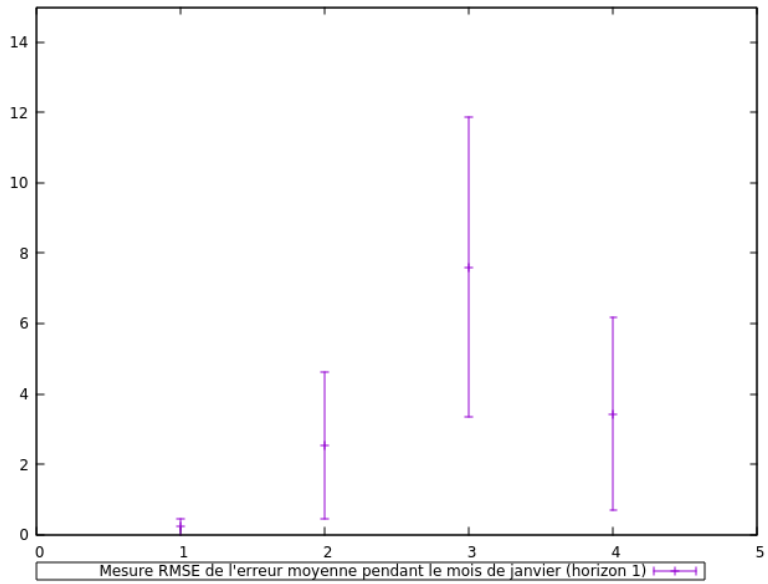
Soit n $\in \mathbb{N}$, i $\in [\![0, n[\![$, X_i les données réelles et Y_i les prévisions sur les données X_i.

On évalue l'erreur par la méthode Root Mean Square Error :

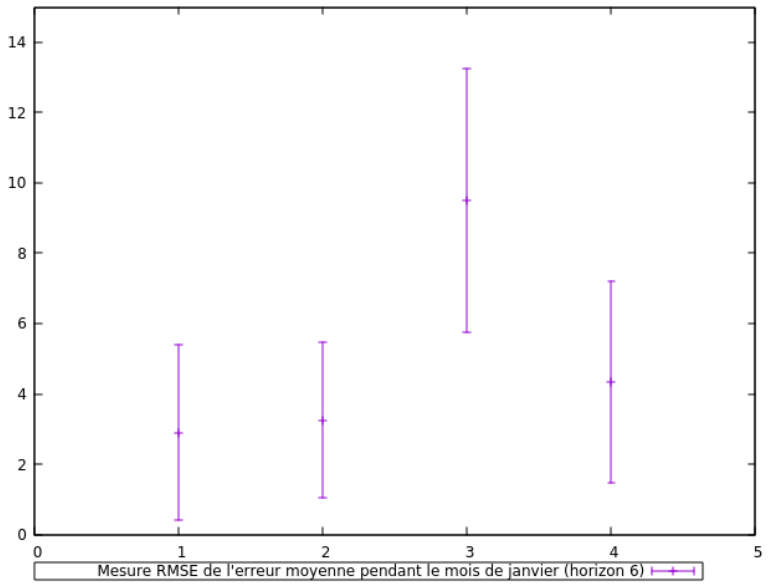
$$\text{RMSE(X, Y)} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(Y_i - X_i)^2}$$

Analyse des résultats
Comparaison des méthodes à différents horizons

Prévisions à horizon 1 sur le mois de Janvier (station 8054):

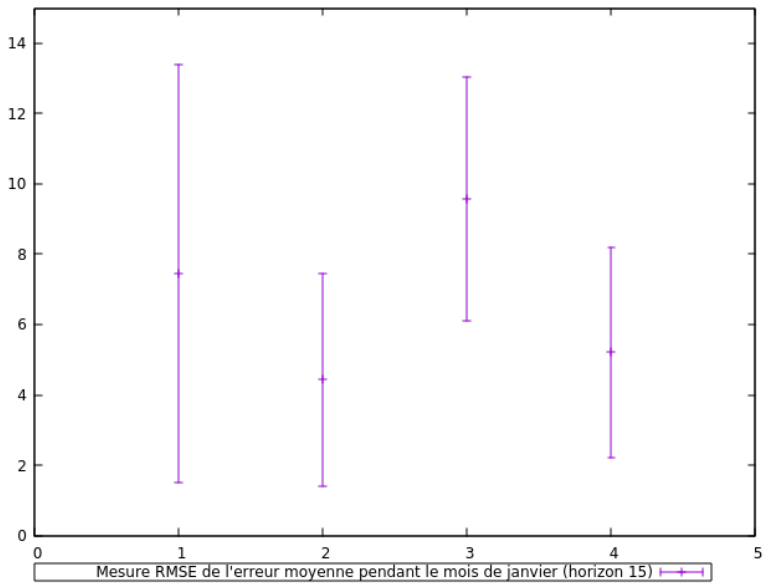


Prévisions à horizon 6 sur le mois de Janvier (station 8054):



Analyse des résultats
Comparaison des méthodes à différents horizons

Prévisions à horizon 15 sur le mois de Janvier (station 8054):



Prévisions sur le mois de Janvier (station 8054):

- Méthode1 : très efficace à court terme, mais de moins en moins à plus long terme
- Méthode 2 par saisonnalité : la moins efficace
- Méthode 3 par recherche de semblables : assez efficace, stable

Analyse des résultats
Interprétation des résultats

CERTES

- Prévisions à court terme très efficaces par une simple régression linéaire
- Prévisions à plus long terme assez efficace par une recherche de semblables

MAIS

- Saisonnalité non-vérifiée : mauvais choix ?

Merci de votre attention

La méthode naïve est, presque paradoxalement, la méthode la plus efficace. Cela n'est pas très étonnant et apparemment, c'est assez fréquent en statistique, d'après Jairo Cugliari, statisticien, au Laboratoire ERIC à Bron, rencontré le 31 Janvier 2018.

Nous avons encore quelques études en cours non-présentées : influence des conditions météorologiques, de la position géographique, du paramètre de régression linéaire, etc.

Je suis maintenant disponible pour des questions supplémentaires.

Prise en compte de données supplémentaires

- Données météorologiques
- Position géographique
- Événements particuliers