Projet Final FDEC

Le projet FDEC de cette année est inspiré d'un défi de la conférence **EGC 2018** (**Un défi sous le soleil de l'Île de La Réunion**).

Le Laboratoire d'Énergétique, d'Électronique et Procédés (LE2P) et le Laboratoire d'Informatique et de Mathématiques (LIM) de l'Université de La Réunion ont proposé d'analyser des données de flux/rayonnements solaires à l'Île de La Réunion. Ce projet s'inscrit dans le cadre de la politique de développement vers l'autonomie énergétique de cette île.

Un historique de données de capteurs multi-sources sur plusieurs années est mis à votre disposition sous forme de séries temporelles multivariées.

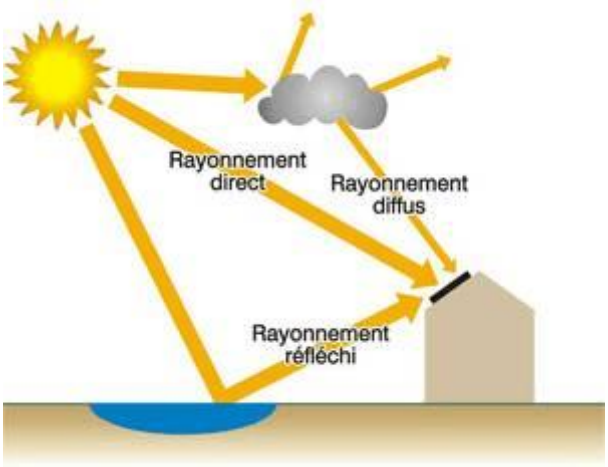
1. Données

Pour mesurer le rayonnement solaire, quinze stations équipées de capteurs SPN1 (Sunshine Pyranometer) sont réparties sur l'Île de La Réunion. Le rayonnement solaire peut être décomposé en trois flux :

- Le flux global F_{Global}
- Le flux diffus (ou réfléchi) F_{Diffus}
- Le flux direct F_{Direct} :

$$F_{Direct} = F_{Global} - F_{Diffus}$$

Dans le domaine de la recherche sur l'énergie solaire, on s'intéresse aussi à l'indice de fraction directe $\boldsymbol{k_b}$**,** défini comme le rapport du flux direct et du flux global, afin de représenter le rayonnement solaire journalier. Intuitivement, lorsque cet indice est proche de 1, le flux direct est proche du flux global et on est en présence d'une journée ensoleillée ; inversement, lorsque l'indice est proche de 0, la journée est nuageuse.



Les capteurs permettent d'obtenir les composantes diffuses et globales du flux solaire toutes les minutes. Ces capteurs sont associés à des capteurs météorologiques qui permettent d'obtenir – au pas de la minute aussi – la température, la pression atmosphérique, le taux d'humidité dans l'air ainsi que la force et la direction du vent.

Ainsi, vous disposerez de 2 ans d'historique de données de flux solaire et de données météorologiques (locales aux capteurs de flux solaire) sous forme de séries temporelles numériques au pas de la minute et par station. Les données ont été mises à disposition de la communauté sous la forme de 5 fichiers disponibles dans un dossier compressé au format ZIP. Chaque fichier contient les données collectées par une des stations pendant deux ans.

Plus précisément, chaque station SPN1 fournit les sept mesures suivantes :

- FG_avg (en W/m^2) : le flux global
- FD_avg (en W/m^2) : le flux diffus
- Patm_avg (en *hPa*) : la pression atmosphérique
- RH_avg (en %) : le taux d'humidité dans l'air
- Text_avg (en °*C*) : la température extérieure
- WD_MeanUnitVector (en degré) : la direction du vent
- WS_avg (en *m*/s) : la vitesse du vent

Ci-dessous un exemple de table de données des sept mesures pour une station SPN1 :

moufia_2014_2015

Timestamp	FD_Avg	FG_Avg	Patm_Avg	RH_Avg	Text_Avg	WD_MeanUnitVector	WS_Mean
2014-01-01 00:00:00	7.999	1.759	973.1667	68.56667	24.6	100.998	4.616667
2014-01-01 00:01:00	1.361	4.084	973.1667	68.71667	24.58333	94.43468	3.75
2014-01-01 00:02:00	3.574	6.297	973.15	68.83333	24.51666	97.8131	5.583333
2014-01-01 00:03:00	5.673	2.212	973.1166	69.15	24.5	92.55701	5.3
2014-01-01 00:04:00	1.588	6.524	973.0999	69.33334	24.5	98.83556	4.65
2014-01-01 00:05:00	3.574	4.311	973.0667	69.23333	24.5	101.1774	6.366667
2014-01-01 00:06:00	3.801	4.425	973.0833	69.46667	24.5	88.854	4.716667

Les données peuvent être téléchargées à partir de la page Arche du module.

2. Objectifs et tâches du projet

Bien que ce projet soit ouvert, nous suggérons quelques pistes de travail (non-exhaustives et donc non-restrictives) :

- Clustering de séries temporelles journalières (par exemple, la classification de journées type en fonction des données de flux solaires et/ou météorologiques)
- Analyse des corrélations entre données de flux solaires et données météorologiques ;
- Analyses liées à la détection d'anomalies, d'évènements extrêmes
- Visualisations de masses de données de séries temporelles

- Une tâche importante portera sur la prédiction du flux global F_{Global} ou de l'indice k_b de fraction directe à l'horizon *H*+1, *H*+2, …, *J*+1, J+2…

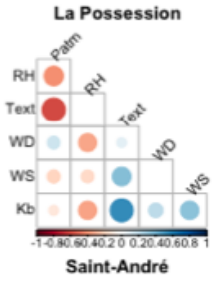
L'utilisation de données externes (*open data*) est autorisée tant qu'elles sont publiquement disponibles.

3. Points à considérer

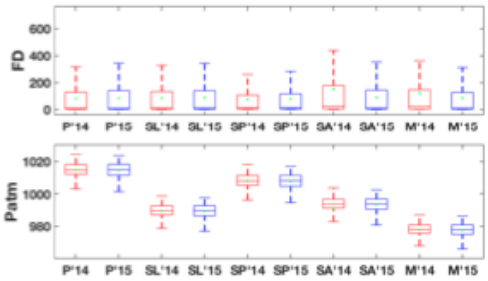
Certains points à prendre en compte pour réaliser cette analyse sont liés à la manière d'explorer les données et, plus tard, à la manière d'évaluer les résultats de la modélisation.

Certains des outils qui peuvent être utilisés pour explorer les données sont :

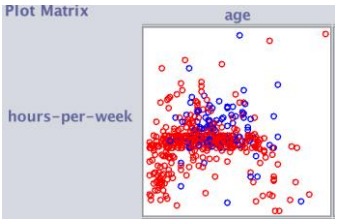
- Indices de corrélation ;



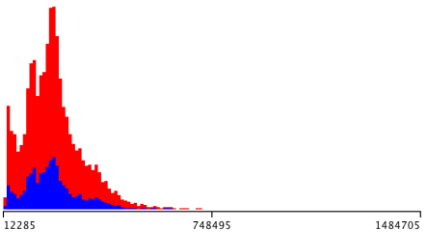
- Box plot ;



- Nuages de points ;



- Histogrames ;



Certains des métriqies qui peuvent être utilisés pour évaluer le modèle sont :

- R Squared / Adjusted R Squared:

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- Mean Squared Error (MSE) (et Root Mean Squared Error (RMSE))

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Certains des modules / packages qui peuvent être utilisés pour analyser et évaluer les données sont :

- Python

sklearn.linear_model.**LinearRegression**

sklearn.metrics.**r2_score**

sklearn.metrics.**mean_squared_error**

sklearn.metrics.**mean_absolute_error**

- WEKA

Weka.classifiers.functions.**LinearRegression**

4. Critères à examiner

Le rapport à élaborer comportera au moins 10 pages (en tenant compte du fait que les figures et les tableaux seront inclus). La quantité de figures et de tableaux doit être adéquate par rapport au texte et à la discussion inclus, il n'est pas conseillé de ne pas utiliser ces éléments de manière exagérée. Également, pour chaque tableau et figure inclus, il est nécessaire qu'ils soient mentionnés dans le texte avec leur description respective.

Un des critères les plus importants à considérer est l'analyse et la discussion des résultats, et pas seulement la description des résultats. Les critères qui seront pris en compte pour l'évaluation du projet sont mentionnés ci-dessous ;

Définition du problème à modéliser + pré-traitement des données + ingénierie des données (5 pts)

- Le problème à modéliser est techniquement correct (*Médiocre / Partiellement mal orienté / Approprié*)
- Utilisation d'outils de visualisation pour l'exploration de données (*Pas d'utilisation / Partiellement / Oui*)
- Qualité technique (*Médiocre / Partiellement mal orienté / Approprié*)
- Complétude (*Pauvres / Bonnes idées mais auraient pu en faire plus / Complet)*

Discussion de la performance sur les données (5 pts)

- La validation a été correctement configurée (*Oui / Non*)
- Plusieurs algorithmes ont été testés (*0, 1, 2, +2*)
- Des arguments appropriés ont été évalués (*Pas de test / Partiellement approprié / Oui*)
- Les résultats sont correctement présentés (*Non / Partiellement / Oui*)

Discussion du (des) modèle (s) final (s) (5 pts)

- La discussion est techniquement correcte (*Aucun / partiel ou incorrect / complet*)
- La discussion est terminée et les points évidents ont été abordés (*Aucun / partiel ou incorrect / complet*)

Clarté du rapport et code Python / flux de travail Weka (5 pts)

- Texte (*Très peu clair / Peu clair dans certaines parties / Clair*)
- Figures (*Difficile à lire / Certaines ne sont pas claires / Claires*)
- Tables (*Difficile à lire / Clair*)
- Code / flux de travail (*Difficile à lire / Clair*)