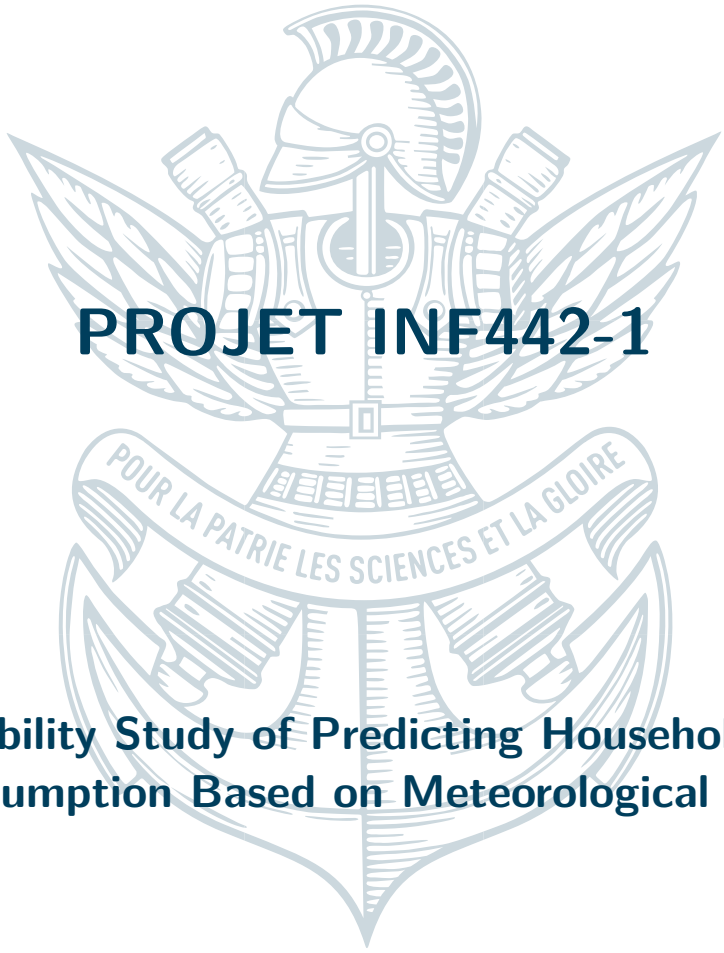

PROJET INF442-1

A Feasibility Study of Predicting Household Power Consumption Based on Meteorological Data

Mardi 21 Mai 2019

Skandère SAHLI
Alban ZAMMIT

TABLE DES MATIÈRES

1
RAPPEL DES ENJEUX DU PROJET

Ce projet s'inscrit dans la logique du cours de big data, et a pour objectif l'analyse de données et la recherche de corrélation entre elles. Nous avions à notre disposition des données de relevés météo de 2007 à 2011, ainsi que des données relatives à la consommation électrique sur la même période d'une maison située en banlieu parisienne. Intuitivement, il semble logique de supposer que la consommation électrique du foyer dépend des conditions climatiques, qui elles-même dépendent des saisons. Le but de ce projet était donc de vérifier cette intuition de plusieurs manières : classification, modélisation et apprentissage. Les relevés météos étaient effectués toutes le trois heures, tandis que les relevés de consommation électrique étaient actualisés toutes les minutes.

2
MISE EN PLACE ALGORITHMIQUE

2.1 Architecture algorithmique

Nous avons codé en `C++` pour réaliser ce projet. Cette section a pour but de présenter comment les classes que nous fabriquées fonctionnent entre elles.

Il y a d'abord trois classes qui permettent de lire les jeux de données : `DatasetConsumption`, `DatasetWeather`, toutes les deux héritant de l'interface `Dataset`. Ces trois classes sont testées à l'aide de deux autres fichiers `C++` : `test_dataset_consumption` et `test_dataset_weather`.

Ensuite, afin de réaliser le clusturing (question 1), en nous appuyant sur le TD3, nous avons créé trois classes. La première, `point`, sert à représenter un point dans l'espace où l'on réalise le clustering. Cette classe est ensuite importée dans `cloud`, qui représente le jeu de données, et dans lequel on réalise l'algorithme de k-means. L'exécution est réalisée par le fichier `test_k_means`.

Pour ce qui est de la détection des patterns de distribution de la consommation selon les heures, on a une classe `HourlyDistribution` qui essaye de l'approximer. Son exécution est ensuite réslisée par la classe `test_hourly_distribution`.

Enfin, pour la partie de régression sur les données météorologiques, deux classes, `LinearRegression` et `KnnRegression`, qui hérite de l'interface `Regression`, permettent d'obtenir deux prédictions différentes, qui sont éxécutées avec la classe `test_regression`

Le tout est résumé sur la figure 1. Par ailleurs, à l'exception des fichiers tests, toutes les classes sont accompagnées d'un fichier de header `.hpp`. Enfin, afin de faciliter la compilation, tous les lignes d'éxécution sont résumées dans un `Makefile`. De même, nous avons rédigé un fichier `Readme.txt` où toutes les procédures sont expliqués pour éxécuter les programmes.

2.2 Lecture et nettoyage des données

On lit les fichiers grace au package `ifstream`, et on place les données dans des `vector<vector<double> >`. Nous avons notamment utilisé le package `ctime` pour gérer les dates correctement. Nous représentons

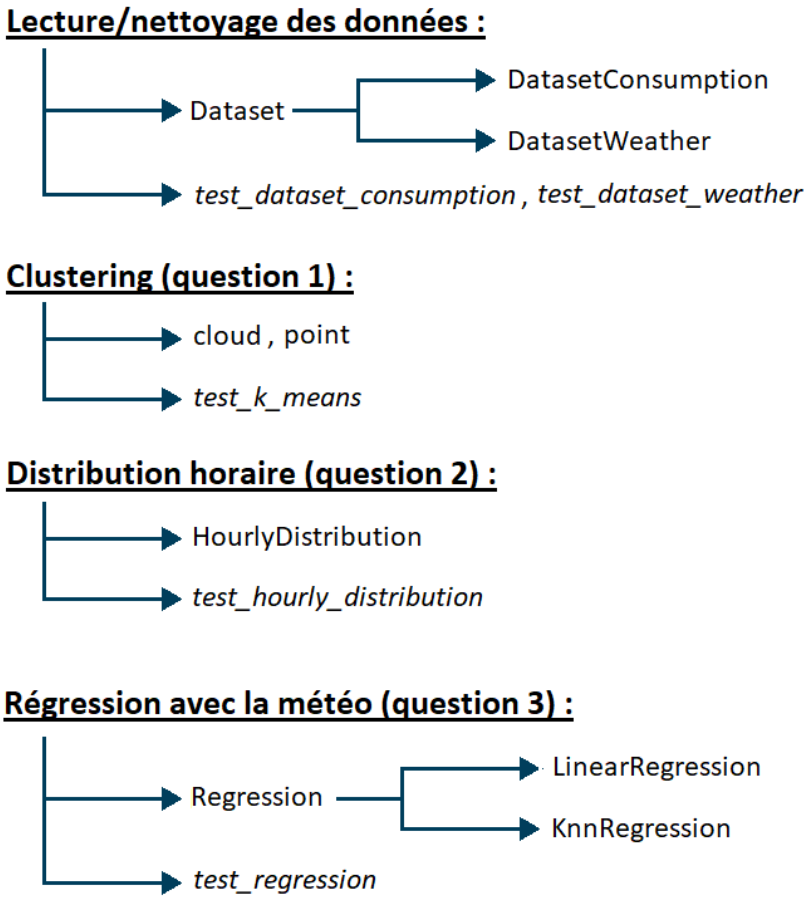

Figure 1 – Architecture des classes

en effet la date par le nombre entier de secondes écoulées depuis le 1er Janvier 2007, à 00 :00 :00.

Certaines données étant manquantes dans les fichiers, nous avons dû réaliser des interpolations, en faisant la moyenne des données antérieurs et postérieures. De même, afin de se prémunir contre les données abérrantes (*outliers*), nous avons utilisé le Z-score pour éliminer toutes les données qui dépassaient de plus ou moins trois fois l'écart-type de leur série. Par ailleurs, il manquait parfois des lignes entières dans les données météo (surement du à un oubli de mesure ou bien une défaillance du système à ce moment). Nous avons donc du rajouter ces données. Par soucis de simplification, et en considérant la rareté de cet évènement, nous avons décidé de simplement recopier la mesure précédente.

Enfin, nous avons ajouté deux paramètres qui furent très utiles dans la suite : un de période d'échantillonage (en moyennant les données si les fichiers sont trop rafinnés) et un de période globale pour sélectionner la période sur laquelle on désire les données.

En effet, un point très important pour la dernière partie du sujet était qu'il fallait que les données météos et les données de consommation correspondent point par point. Cette étape de récupération et de nettoyage des données était donc capitale pour le bon fonctionnement des algorithmes.

3
ANALYSE DES DONNÉES

3.1 Détection non-surpervisée de saisons par clustering

On peut intuitivement s'attendre à ce qu'il y ait des périodes dans l'année au cours desquelles les patterns de consommation électrique sont les mêmes, à l'instar des quatres saisons de l'année. Nous avons appliqué l'algorithme de k-means sur le jeux de données de consommations et essayé de repérer des clusters. Afin de rendre l'affichage plus digeste, nous avons moyenné les données de consommation jour par jour.

Comme **l'algorithme de k-means** renvoie une variance intracluster, dont la valeur intrinsèque n'est pas physiquement très parlante, nous avons décidé pour repérer les clusters d'afficher les points en 2 dimensions pour voir si de réels clusters se démarquaient. Il nous a donc fallu choisir les deux paramètres physiques les plus pertinents, i.e. les plus suceptibles à nos yeux de changer d'une saison à l'autre. Nous avons retenu deux combinaisons : consommation globale et voltage (pour la redondance qu'elle apporte puisque $P = U.I$), ainsi que consommation globale et la consommation dans la zone 3 (qui nous semblait la plus changeante en fonction des saisons avec l'air conditionné). Pour $k = 4$ on obtient les figures 2 et 3.

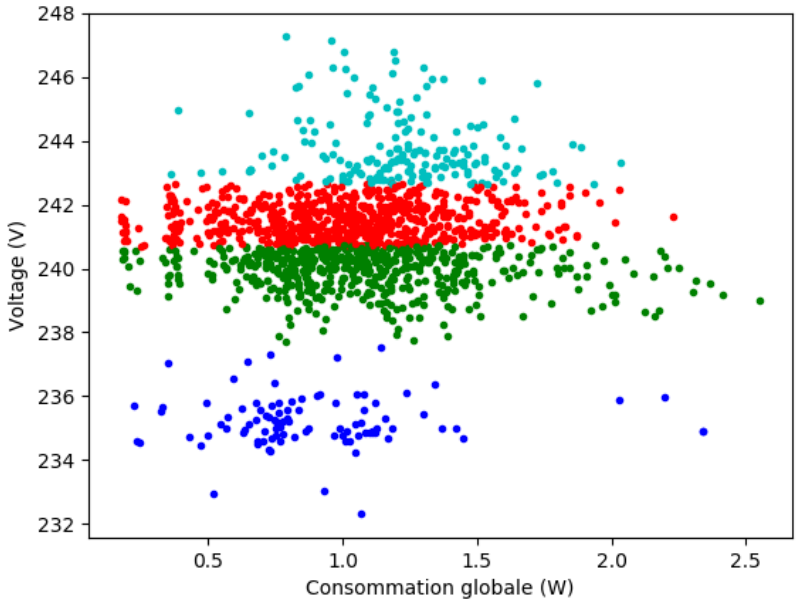

Figure 2 – Clustering pour k=4 des points formés par la consommation globale et le voltage

Le rendu n'est pas extraordinaire. **Il ne semble pas qu'on puisse facilement détecter des clusters qui identifient des saisons dans la consommation**. Même sur la figure 2, où certains points semblent se détacher en bas, un examen des dates associées montre en réalité qu'elles ap-

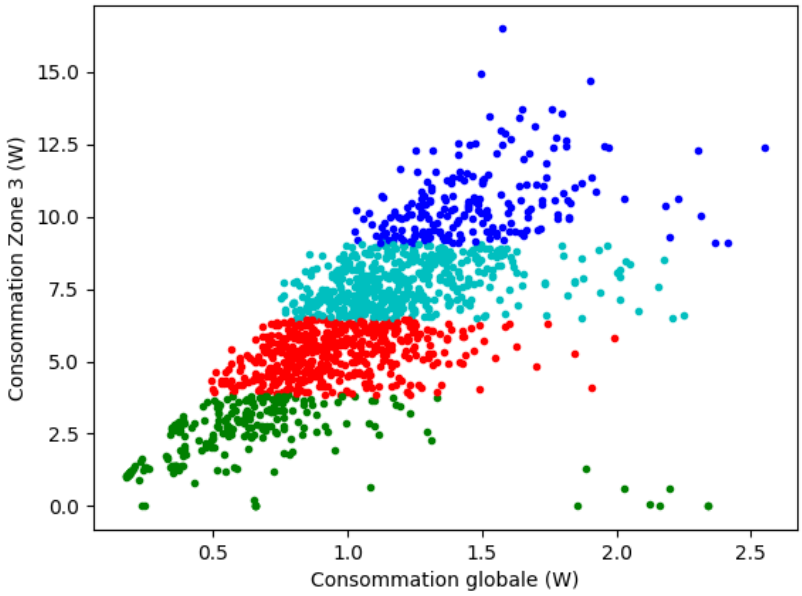

Figure 3 – Clustering pour k=4 des points formés par la consommation globale et la consommation en zone 3

partiennent toutes à un intervalle de temps compris entre le 1er Mars 2007 et le 1er Mai 2007. La consommation à cette periode a été très différente des autres certainement pour une raison exceptionnelle indépendante de la météo.

On peut toutefois essayer d'aller plus loin, en se disant que le k choisi n'est pas optimal. Nous avons alors tracé la variance intra-cluster en fonction de k (figure 4), qui confirme que le k optimal vaut 4 (méthode du coude.

Ci-dessous, quelques autres graphique de clustering pour les points formés par la consommation globale et le voltage.

Nous en avons conclu qu'il est difficile de repérer un pattern clair de saison au sein des données de consommation. Cela peut s'expliquer par différentes raisons. D'abord, la maison ne dispose pas de chauffage électrique, principale source de dépense énergétique dans un foyer, et quand on regarde les éléments énergivores dans les zones 1, 2 et 3 de la maison, on s'apperçoit que ce sont des appareils que l'on utilise assez indifférement à toute période de l'année. De plus, la maison est située en banlieue parisienne, dans un climat assez tempéré donc peu sujet à des variations fortes et brusques de température. Cette hypothèse de climat assez homogène à l'échelle de l'année est confirmé si l'on réalise un affichage des données météo des points formés par la température et la vitesse du vent (deux facteurs inluçant probablement le plus la consommation électrique), comme le montre la figure 8.

3.2 Mise en évidence d'une distribution horaire de consommation

Une autre question interessante à se poser est la distribution de la consommation sur une plage d'une heure, et plus précisément de savoir ci cette distribution est gaussienne. Par exemple, on

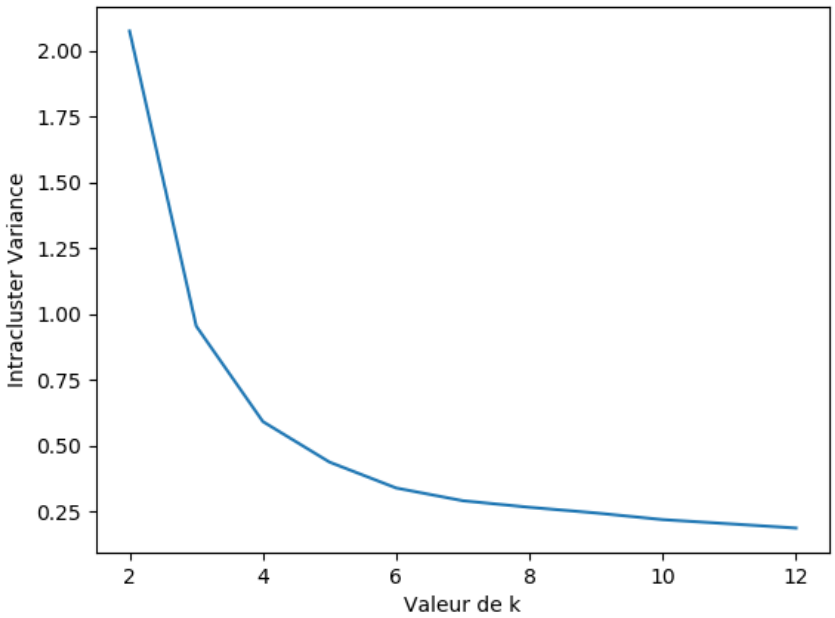

Figure 4 – Variance intra-cluster en fonction de k pour les points formés par la consommation gloabale et la consommation en zone 3 - Méthode du coude

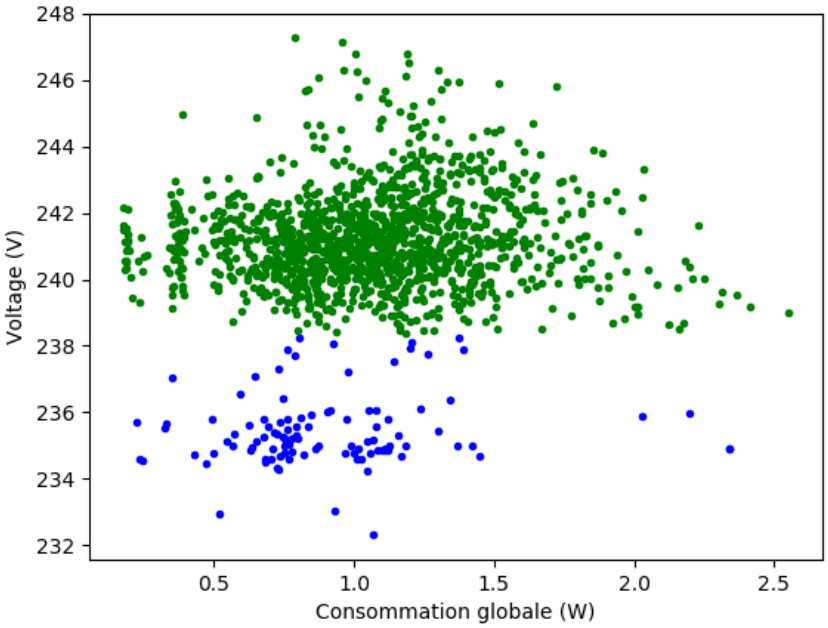

Figure 5 – Clustering pour k=2 des points formés par la consommation globale et le voltage

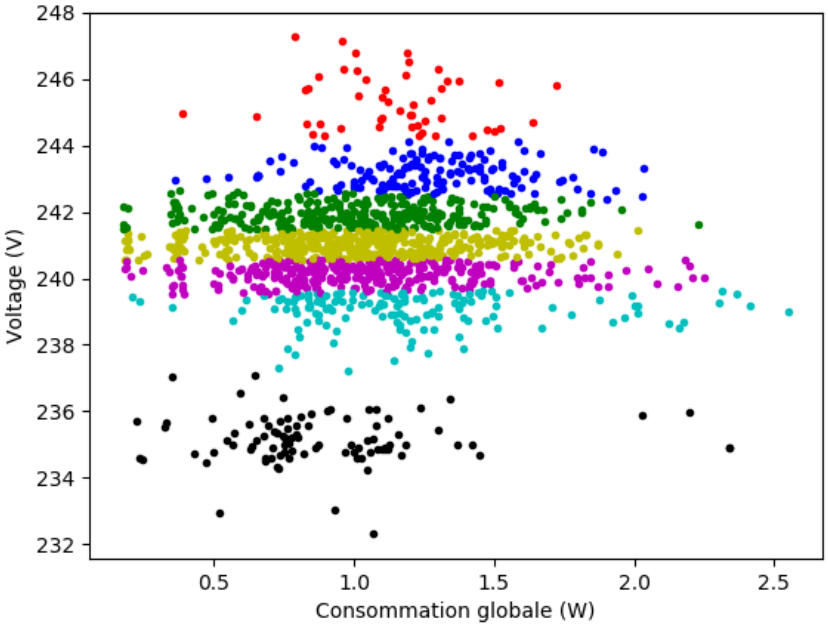

Figure 6 – Clustering pour k=7 des points formés par la consommation globale et le voltage

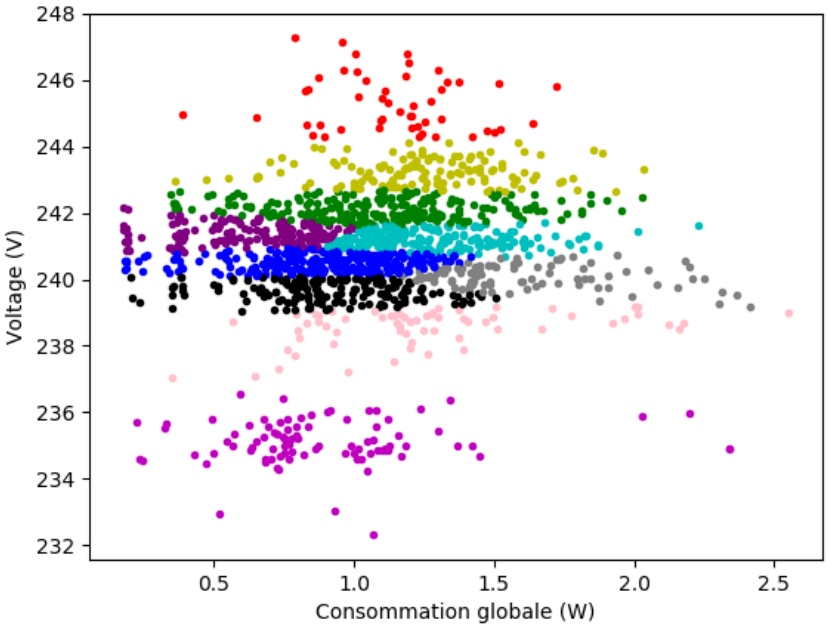

Figure 7 – Clustering pour k=10 des points formés par la consommation globale et le voltage

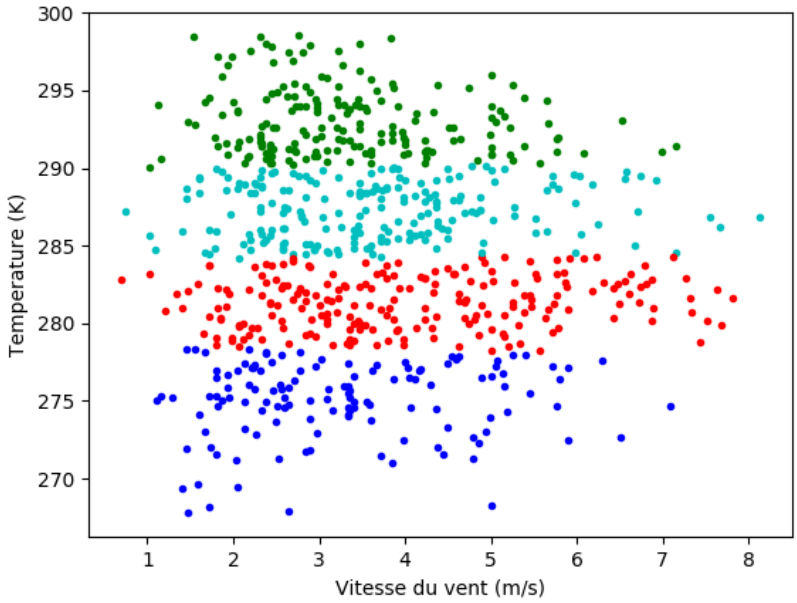

Figure 8 – Clustering pour k=4 des points formés par la température et la vitesse du vent

considère la plage horaire 8h-9h, et on prend, pour chacun des jours disponibles (il y en a environ $n = 365 \times 4 = 1460$) la consommation globale moyenne sur cette plage. On regarde alors la distribution de ce n-échantillon. Par exemple, sur pour la plage 8h-9h on a la distribution de la figure 9.

Cela ne ressemble pas vraiment à une Gaussienne..., et pour le montrer rigoureusement, nous allons utiliser **le test de Kolmogorov-Smirnov**. On commence par estimer la valeur en chaque point x (de l'espace de vie du n-échantillon $(x_i)_{i\in\{1,\dots,n\}}$ noté Ω) de la fonction de répartion empirique $F_n(x)$ du n-échantillon définie par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x\leq x_i}$$

Sans perte de généralité, on peut supposer que le n-échantillon est normalisé (il suffit de poser $(\widetilde{x_i})_{i\in\{1,\dots,n\}} = ((x_i - \hat{\mu_n})/\hat{\sigma_n}))_{i\in\{1,\dots,n\}}$ où $\hat{\mu_n}$ et $\hat{\sigma_n}$ sont respectivmement la moyenne et l'écart-type empiriques du n-échantillon). Alors, sous l'hypothèse que le n-échantillon suive une loi gaussienne (centrée réduite donc) de fonction de répartion notée F, on a par le théorème de Glivenko-Cantelli la convergence presque sûre suivante :

$$\sup_{x\in\Omega} \mid F_n(x) - F(x) \mid \overset{n\to+\infty}{\longrightarrow} 0$$

L'implémetation numérique de ce résultat (figure 10) consiste à calculer numériquement $D = \sup_{x\in\Omega} \mid F_n(x) - F(x) \mid$, et d'utiliser une table de correspondance avec la p-valeur associée à ce test pour confirmer ou infirmer l'hypothèse. Toujours avec l'exemple de la plage 8h-9h, on calcule $D = 0.0840$ avec trois chiffres significatifs. Dans une table de valeur, au seuil $\alpha = 0.05$, on a la valeur critique

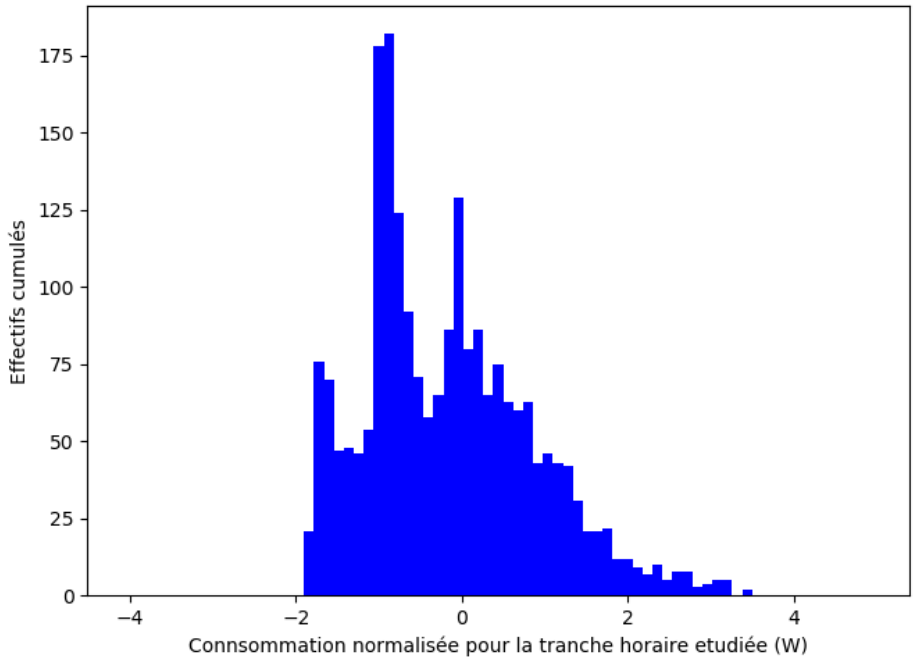

Figure 9 – Distribution empirique de la consommation moyenne horaire - Plage 8h-9h

$D_{max} = 0.0359$. Comme pressenti, comme $D > D_{max}$, **on rejète l'hypothèse de normalité de la distribution horaire**.

Pour les autres plages horaires, le problème reste le même. On peut chercher la plage horaire pour laquelle la valeur de D est minimale. On touve qu'il s'agit de la plage 21h-22h, avec un coefficient $D = 0.0608$ avec trois chiffres significatifs, toujours trop grand devant $D_{max} = 0.0359$ (cf figures 11, 12).

Malgré cet échec, on peut essayer d'interprêter les résultats. En effet, il semble que D explose pour des valeurs de x aux alentours de -2 à chaque fois. En fait, on remarque qu'il existe un seuil pour la consommation globale de la maison, qui peut s'interprêter comme les fonctions vitales de la maison qui fonctionnent toujours, même lorsque l'on ne fait rien dans le maison (ex : un réfrigérateur est branché en permanence), et qui accumule les toutes les valeurs d'une gaussienne obtenue pour des $x < -2$ sur -2. Hormis cela, la consommation, pour les $x > 0$ semble suivre une loi à peu près normale (cf. figures 9 et 11).

D'ailleurs, on pourrait s'imaginer qu'avec cette idée de seuil, la loi de consommation pour une plage horaire pourrait suivre une autre distribution. **Une loi exponantielle translatée convenabelement, qui possède également un seuil, semble par exemple très bien bien adaptée pour certaines plages horaires**, comme la plage 17h-18h (cf. figure 13).

Enfin, en ce qui concerne la distribution intracluster, il ne semble pas très pertinent de s'y attarder, puisque, comme nous l'avons vu plus tôt, **les clusters identifiées lors de la question 1 ne traduisent pas vraiment une réalité physique**.

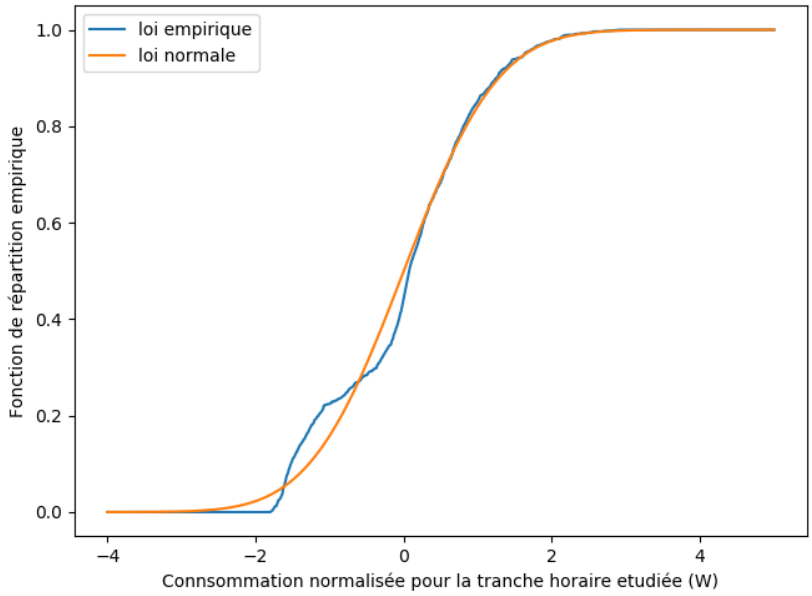

Figure 10 – Comparaison des fonctions de répartion empirique et cible - Plage 8h-9h

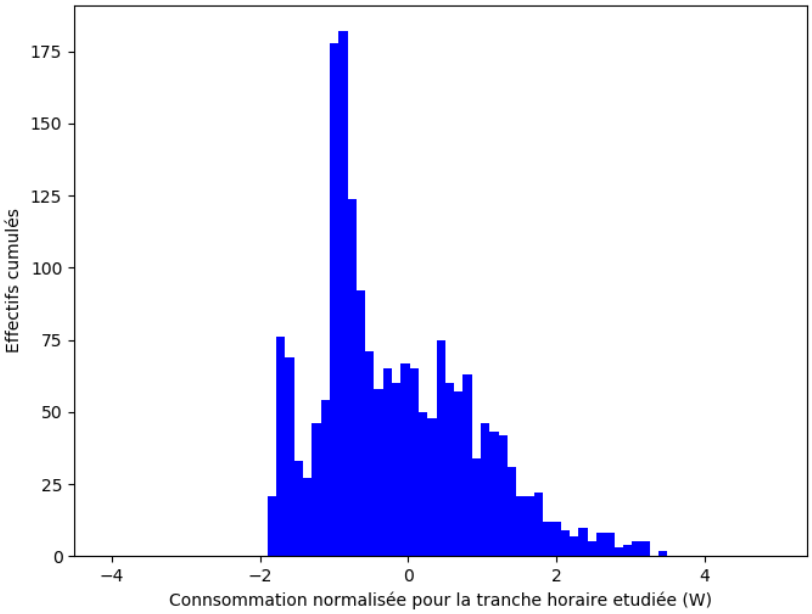

Figure 11 – Distribution empirique de la consommation moyenne horaire - Plage 21h-22h

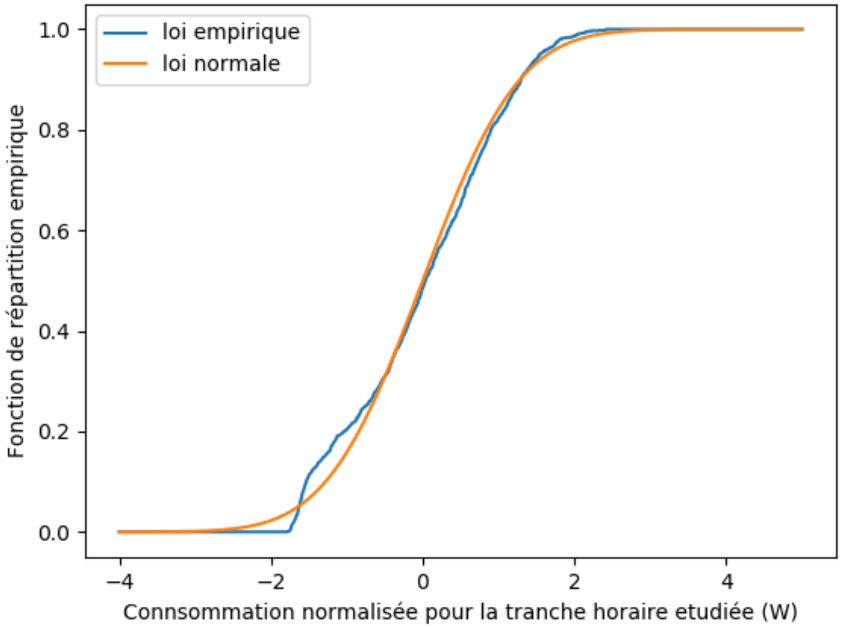

Figure 12 – Comparaison des fonctions de répartion empirique et cible - Plage 21h-22h

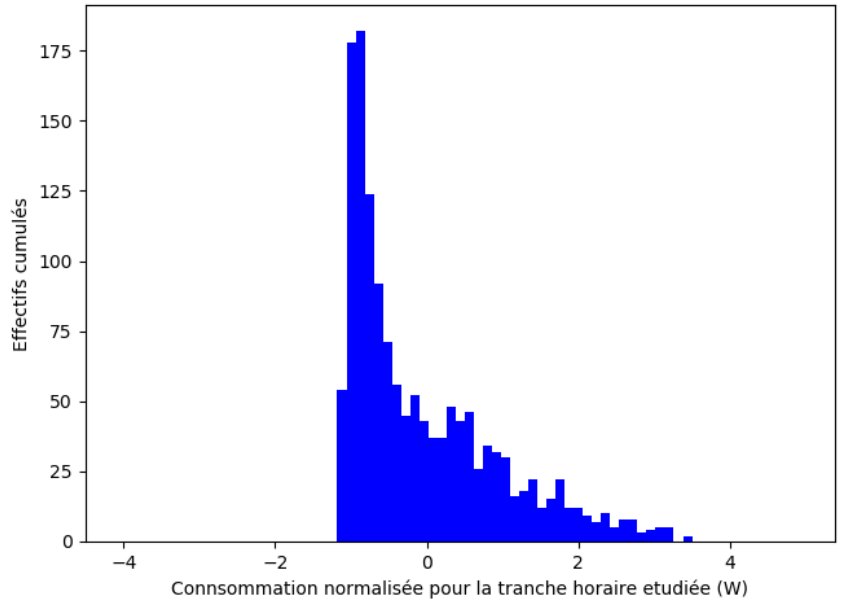

Figure 13 – Distribution empirique de la consommation moyenne horaire - Plage 17h-18h - On semlble reconnaitre une loi exponentielle

3.3 Prévision par régression de la consommation en fonction de la météo

Une dernière chose interessante à étudier est l'influence de la météo sur la consommation. Pour cela, on prend les données de météo et de consommation des années 2007 à 2009 comme jeu d'apprentissage et on essaye de prédire la consommation en 2010 en fonction de la météo.

Pour ce faire, il faut choisir judicieusement les données météo. En effet, on peut se douter que la température et la hauteur de la base des nuages inférieurs n'auront pas la même influence sur la consommation. D'autre part, nous nous somme particulièrement intéressés à la prédiction de la puissance active globale, paramètre le plus intéressant d'un point de vue pratique et économique.

Nous avons mis en oeuvre - en utilisant le TD7 - deux types de régression sur des données météorologiques moyennées à la journée : une régression k-NN et une regression linéaire. Pour choisir k, nous avons testé l'algorithme pour des valeurs allant de 1 à 40. Il s'avère que pour des $k > 8$ environ, la MSE variait très peu, comme le montre la figure 14. Nous avons donc choisi $k = 11$ pour toutes nos regressions

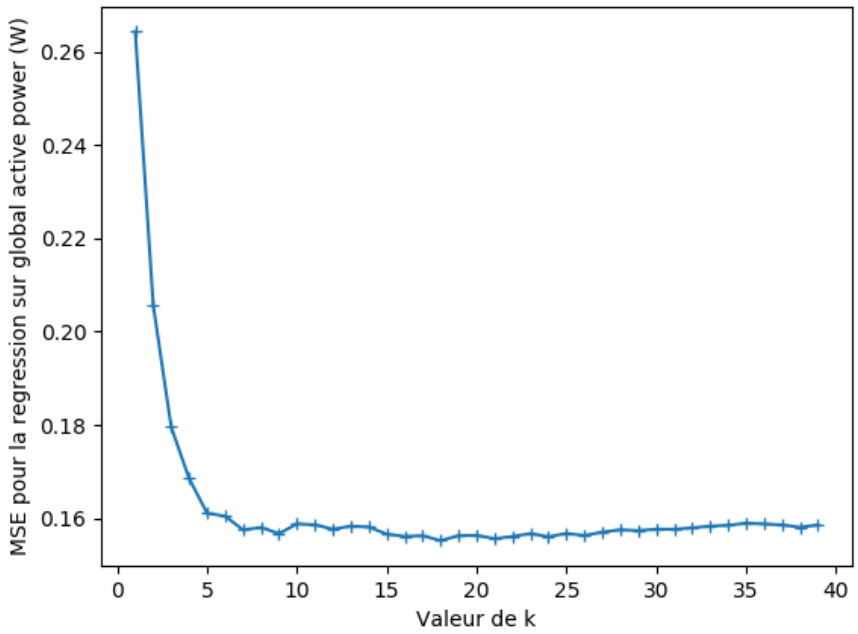

Figure 14 – Valeur de la MSE pour une régression k-NN sur la global active power, en fonction de k

Enfin, nous avons choisi de conserver seulement les colonnes Direction du vent, Vitesse du vent, Température et Humidité pour faire les regressions. En effet, physiquement ce sont celles qui ont le plus de sens, et qui intuitivement vont jouer un rôle sur la consommation de la maison. Par ailleurs, en testant avec toutes les colonnes non vide, nous obtenions des résultats un petit peu moins bon.

Techniquement, lors de l'exécution les résultats sont exportés dans deux fichiers un pour chaque type de regression : `output_linear_<num_col_regr>.txt` et `output_k_NN_<num_col_regr>.txt`. Les résultats sont résumés dans le tableau ci-dessous

Type de donnée	Moyenne	Linear (MSE)	k-NN (MSE)
Puissance active (W)	1.022	0.157	0.154
Puissance réactive (W)	0.114	0.0012	0.0011
Voltage (V)	239.2	8.107	7.595
Intensité (A)	4.362	2.689	2.668
Submetering 1 (W)	0.273	0.080	0.087
Submetering 2 (W)	0.918	0.662	0.716
Submetering 3 (W)	5.545	6.398	6.577

On peut visualiser le résultat sur la consommation globale (donnée la plus pertinente physiquement) sur la figure 15.

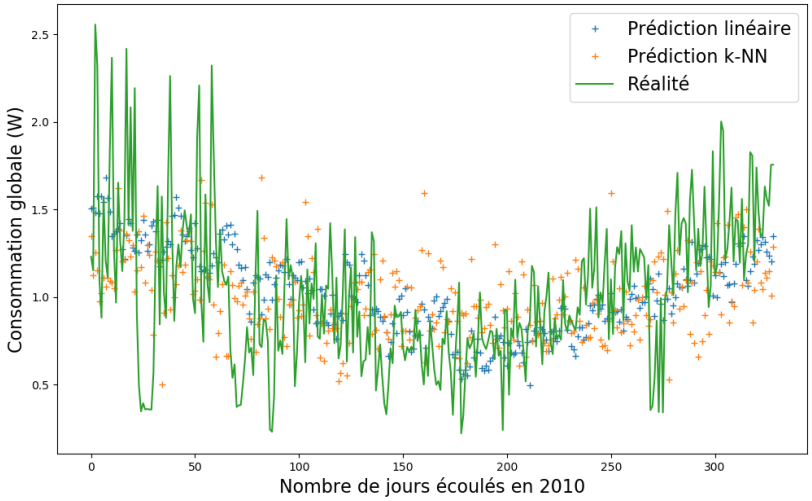

Figure 15 – Comparaison entre les prédictions et la réalité sur la consommation globale, pour un échantillonnage quotidien

Etonnament, en regardant dans le tableau, la régression la plus efficace, en terme de MSE sur, est la regression sur la puissance réactive. Cependant, il faut se méfier du système d'unité, car comme on le voit graphiquement avec les figures 16 et 17, cela n'est si net que cela...

Comme on peut le voir, sur les graphiques précédents, les prédictions (faites avec une regression linéraire ou k-NN) ont tendance à lisser les variations de temperature. Cela est certainement dû au fait que la taille du data set d'entrainement est trop petite (c'est comme si a partir de trois points on voulait anticiper le quatrième). Ainsi on observe un *underfitting* qui s'adapte très mal aux fortes variations. Pour essayer de tester les prédictions sur un jeu de donnée plus lisse, on peut changer de période d'échantillonnage, en passant à l'échelle de la semaine. Cela dimininue la taille du dataset, mais comme on peut le voir sur la figure 18, les prédictions semblent meilleures, et cela est confirmé par une meilleure MSE : 0.058 pour k-NN et 0.063 pour le modèle linéaire, soit 3 fois mieux que pour les regressions moyennées à la journée.

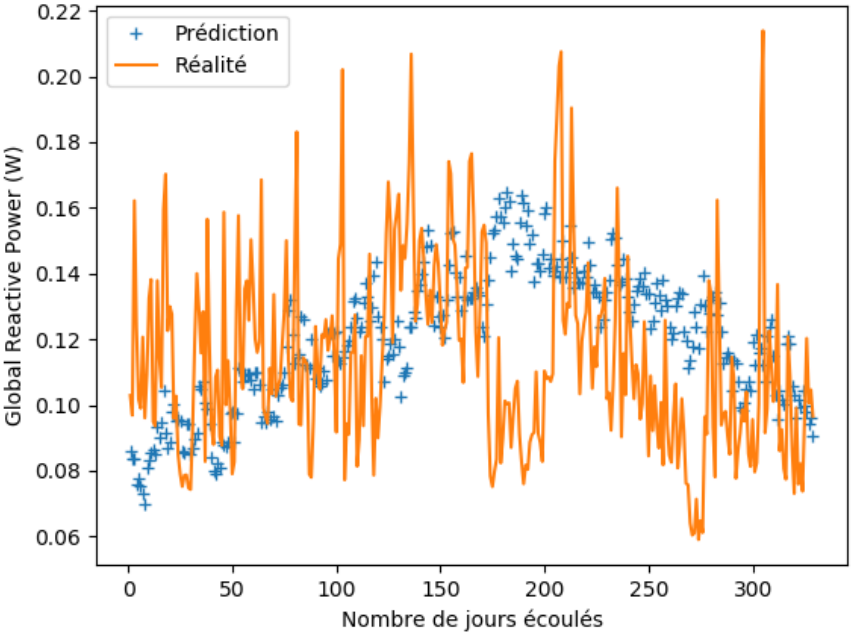

Figure 16 – Comparaison entre les prédictions et la réalité sur la consommation réactive, pour un échantillonnage quotidien, avec une regression linéraire

4 CONCLUSION

Ce projet, bien que prenant, était très intéressant et consistait pour nous en une première expérience en science des données. Nous nous sommes rendus compte de la difficulté du traitement des données, et du caractère chronophage de cette tâche. Par ailleurs, nous avons pu grandement améliorer nos compétences en `C++`, langage que nous avions trouvé difficile à appréhender au cours des TD.

Au-delà des algorithmes, nous avons dû utiliser notre sens physique et notre intuition pour analyser correctement les données et guider notre utilisation des différents algorithmes.

Enfin, ce projet peut avoir de réelles applications pour permettre aux ménages de mieux prédire leur consommation, en vue de réaliser des économies d'énergies.Ces résultats peuvent aussi être exploités par des fournisseurs d'électricité pour fixer le prix de leurs produits de manière optimale, ou encore anticiper les pics de consommation.

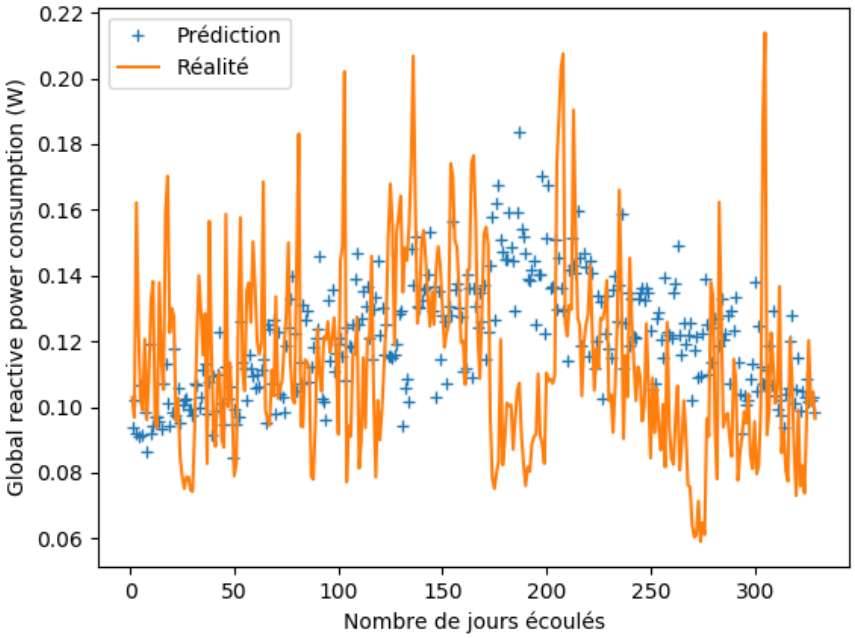

Figure 17 – Comparaison entre les prédictions et la réalité sur la consommation réactive, pour un échantillonnage quotidien, avec une regression k-NN

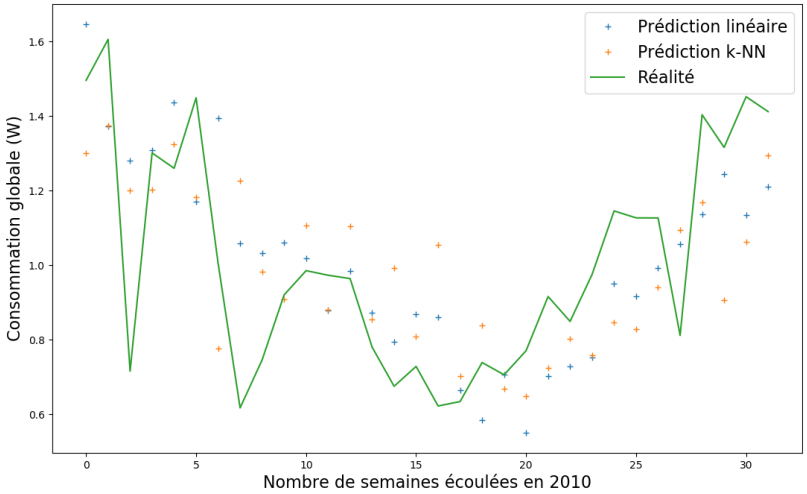

Figure 18 – Comparaison entre les prédictions et la réalité sur la consommation globale, pour un échantillonnage hebdomadaire