Valentin Braux-Guillin
Quentin Assémat

June 2021

PROJECT INF442-1

A feasibility study of predicting household power consumption based on meteorological data

INTRODUCTION

L'objectif de ce projet est d'analyser des données de consommation relevées sur une habitation pendant quatre ans [1], afin d'identifier des structures dans les modes de consommation et d'établir des corrélations avec les données météorologiques correspondant à ces années [2].

Nous avons d'abord procédé au nettoyage des jeux de données, puis au pre-processing nécessaire pour les utiliser dans le cadre des différents objectifs. Nous avons ensuite tenté de détecter des périodes de consommation dans l'année sous forme de "saisons" par du clustering avec l'algorithme des k-moyennes. De plus, nous avons recherché la présence de distributions connues dans la façon dont l'énergie était consommée selon les heures de la journée, en les visualisant et en utilisant le test de Kolmorogov-Smirnov. Enfin, nous avons tenté de prédire la consommation d'énergie en 2010 à partir de celle de 2007 à 2009, en utilisant les données météorologiques rendues disponibles par Météo France. Nous avons évalué les performances de nos prédicteurs, basés sur des algorithmes de plus proches voisins ou de régression linéaire.

PRE-PROCESSING DES DONNÉES

Afin de réaliser le nettoyage et le pre-processing des données, nous avons utilisé Python et plus particulièrement la librairie Pandas très utile pour manipuler des dataframes. Nous avons d'abord chargé l'ensemble des fichiers disponibles pour chaque type de données en un seul dataframe. Pour les données météorologiques, nous avons avant toute chose éliminé toutes les lignes correspondant à des stations différentes de celle d'Orly.

Nous avons commencé par nous occuper des données manquantes dans les jeux de données. Pour les données de consommation, nous avons remarqué qu'il n'y avait jamais uniquement quelques données manquantes pour une entrée, mais soit toutes les colonnes soit aucune. Nous avons donc décidé de directement supprimer toutes les entrées inutilisables.

Pour les données météorologiques, la situation était plus délicate et nous avons procédé de plusieurs façons : tout d'abord suppression des colonnes redondantes ou ne correspondant pas à des mesures (par exemple celles qui indiquent la méthode de mesure), et élimination des colonnes dont le nombre de données non manquantes était inférieur à 20% du nombre d'entrées. Nous avons fait une exception pour la colonne correspondant à la hauteur de neige, puisque nous avons remarqué que l'absence de données correspondait en général à une absence de neige. Nous avons donc rempli les vides par des 0, après une interpolation linéaire visant à combler les petits trous. Enfin pour toutes les autres colonnes, nous avons également procédé par interpolation linéaire.

Nous avons ensuite terminé le nettoyage par la détection et l'élimination des outliers en utilisant le Z-score. Nous avons considéré comme outlier toute entrée dont le Z-score dépassait un threshold fixé pour au moins une colonne. Afin de choisir le threshold à utiliser, nous avons tracé le graphe du nombre d'entrées conservées en fonction du threshold dans la figure 1.

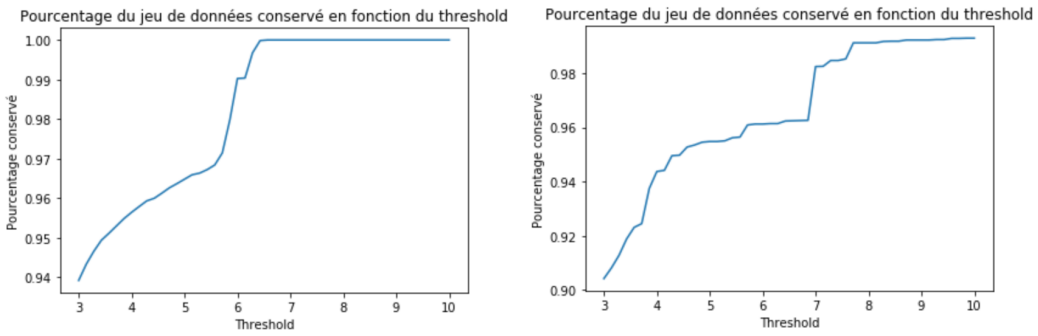

FIG. 1 – *Pourcentage de données conservées pour la consommation (gauche) et la météo (droite).*

Pour des données normalement distribuées, un threshold de 3 donne environ 0.3% d'outliers. Nous sommes ici très loin de cette valeur avec un pourcentage d'outliers plus de 20 fois supérieur dans les deux cas. Cela s'explique d'abord par le fait que nous travaillons en grande dimension (7 pour la consommation, 34 pour la météo) et que nous avons choisi de supprimer toutes les données pour lesquelles au moins une des métriques pouvait être considérée comme un outlier. Cela augmente donc fortement le nombre d'outliers détectés.

Nous observons un fort changement dans la courbe qui se produit aux environs d'une valeur de threshold de 6.5 pour la consommation et de 8 pour la météo. Après cette valeur, le nombre d'outliers stagne, et correspond probablement au nombre de "vrais" outliers. C'est donc cette valeur du threshold que l'on choisit pour éliminer les outliers. Nous élimons ainsi 0.56% des données de consommation et 0.88% des données météorologiques.

Nous avons enfin réalisé le pre-processing nécessaire aux utilisations des données. Nous avons tout d'abord réalisé la normalisation des jeux de données afin de donner le même poids à toutes les métriques dans les analyses. En effet les écarts d'échelle étaient très importants et sans connaissance de l'importance relative des données, il était raisonnable de les normaliser.

Pour les données de consommation, nous les avons ensuite moyennées par jour et par heure, et nous avons de plus séparé les données en 24 fichiers contenant chacun l'évolution de la consommation horaires sur une heure précise de la journée.

Pour les données météorologiques, nous les avons également moyennées par jour, puis nous les avons séparées en données d'entraînement (2007 à 2009) et de test (2010). Avant de les concaténer avec les données de consommation séparées de la même façon, nous avons dû éliminer toutes les dates qui n'étaient présentes que dans l'un ou l'autre des jeux de données. Nous avons ensuite pu créer nos jeux de données d'entraînement et de test, en éliminant les colonnes de date, heure et station.

Une dernière étape de pre-processing utile pour le dernier objectif a été de transformer les données météorologiques par des polynômes de degré 2 afin de pouvoir réaliser une régression polynomiale avec un simple algorithme de régression linéaire.

DÉTECTION DE SAISONS

À des fins de mise en pratique de nos nouvelles connaissances en programmation et d'efficacité algorithmique, nous avons choisi de réaliser tous nos algorithmes de traitement de données en C++, notamment avec les librairies Eigen, ANN... Toutes nos visualisations ont elles été réalisées sur Python à l'aide de la librairie Matplotlib.

Pour déterminer la présence de saisons dans la distribution de la consommation au cours de l'année, nous avons utilisé l'heuristique de Lloyd pour la résolution du problème des k-moyennes. Nous avons d'abord utilisé les méthodes Elbow et Silouhette afin de déterminer le bon hyperparamètre k à choisir pour le clustering.

Nous les avons calculées en partant de trois initialisations différentes (Random, Forgy, k-means++) afin d'obtenir une certitude maximale sur l'hyperparamètre. Pour toutes les initialisations et les deux méthodes, nous avons trouvé que le paramètre $k = 4$ était optimal. Les graphes étant tous similaires, nous présentons uniquement les résultats obtenus pour l'initialisation Forgy dans la figure 2.

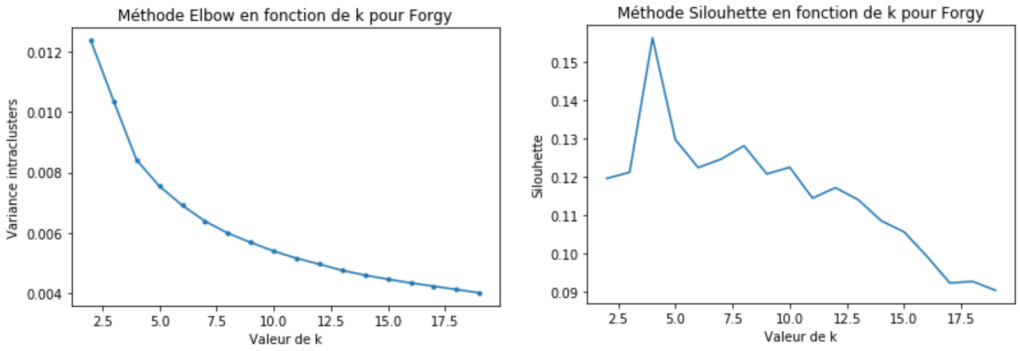

FIG. 2 – *Méthodes Elbow et Silouhette pour la détermination du nombre de clusters optimal.*

Nous avons ensuite réalisé le clustering sur les données de consommation normalisées pour cette valeur de k, encore une fois avec les différentes initialisations. Nous avons trouvé la même variance intraclusters finale dans tous les cas. Cela nous conduit à penser qu'elle correspond bien au minimum global. Nous utilisons donc pour la suite les résultats du clustering donné par l'initialisation de k-means++. Afin de mieux comprendre la façon dont le clustering a eu lieu, nous visualisons la répartition dans les quatre "saisons" dans la figure 3.

Nous pouvons alors caractériser les quatres saisons. La saison 0 est assez étonnante car n'est représentée presque que lors de l'année 2007. La saison 1 correspond vraisemblablement à la période d'été, puisqu'elle est concentrée sur les mois de juin, juillet et août sur les quatre années. La saison 2 semble correspondre à l'hiver, avec une concentration sur les mois de décembre, janvier, février et mars. Enfin la saison 3 est assez répartie et semble correspondre à la consommation modérée qui a notamment lieu au printemps et à l'automne (mais pas uniquement).

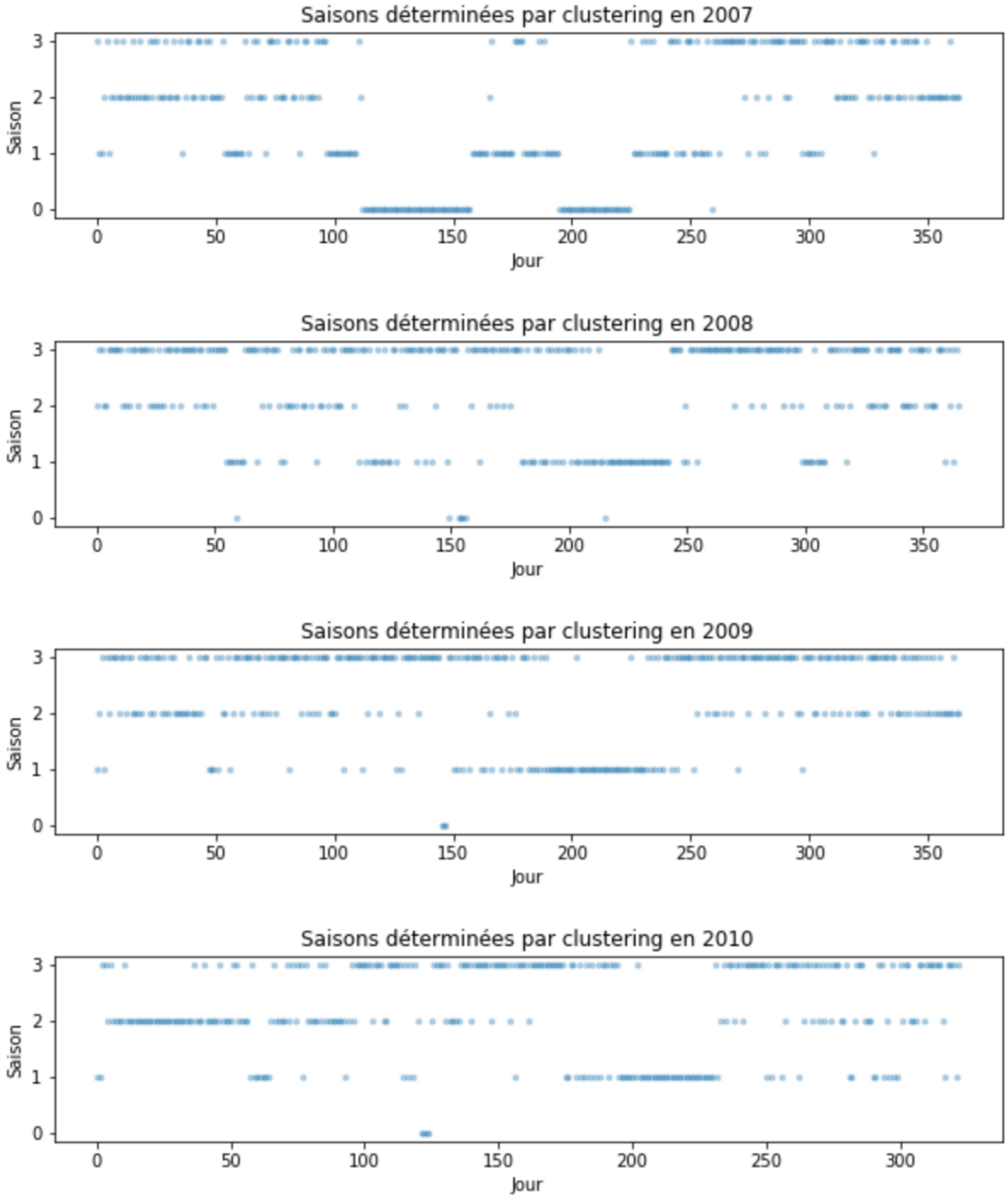

FIG. 3 – *Répartition des saisons au cours des années.*

Nous pouvons ainsi conclure qu'il y a entre 3 et 4 saisons par an (la quatrième saison n'étant justifiée que lors de l'année 2007), qui correspondent grossièrement aux saisons réelles avec une période froide, une période chaude et une période plus modérée.

DISTRIBUTION DE LA CONSOMMATION HORAIRE

Nous avons ensuite étudié la distribution de la consommation horaire selon le critère “global active power” au fil du temps sur des heures précises. Nous nous attendions à observer des distributions proches de gaussiennes, mais cela n’a pas été le cas. Nous visualisons les résultats pour quelques heures dans la figure 4.

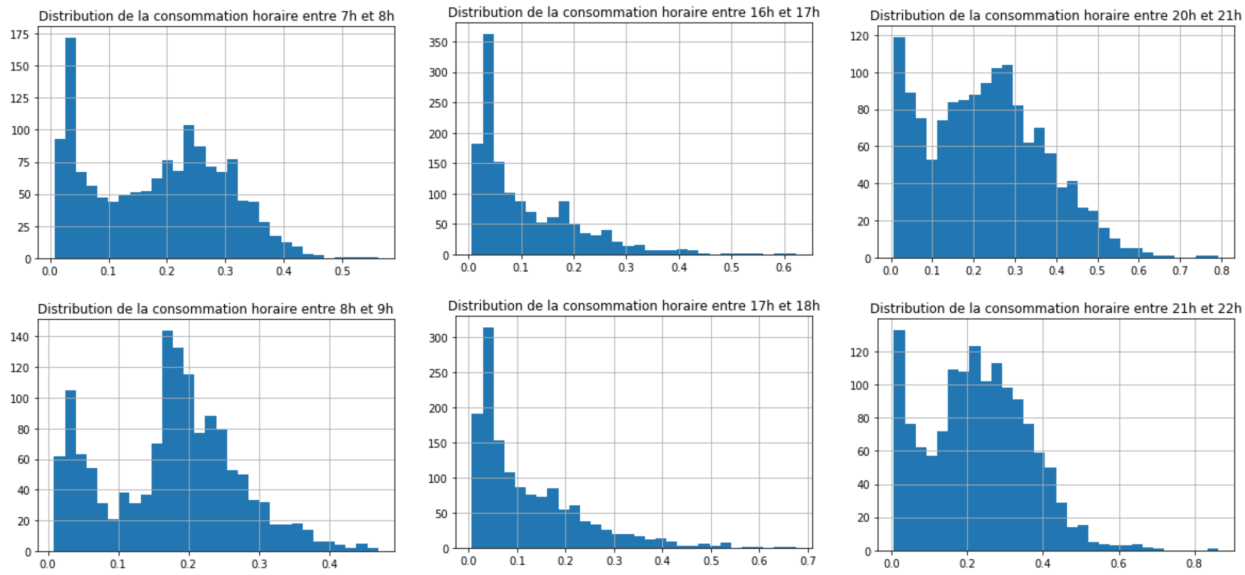

FIG. 4 – *Distribution de la consommation horaire pour différentes heures.*

Nous observons alors plusieurs comportements différents. Aux heures 7 et 8, nous observons ce qui ressemble à une superposition de deux distributions, d’une gaussienne étroite et d’une autre un peu plus large. Aux heures 16 et 17, il n’y a plus qu’un pic de distribution, qui ressemble à une exponentielle. Enfin aux heures 20 et 21, la distribution semble à nouveau contenir deux pics, mais ils sont plus aplatis.

Afin d’obtenir des informations plus quantitatives sur la distribution de la consommation, nous avons déterminé les résultats du test de Kolmorogov-Smirnov et la p-valeur associée (à 10^{-15} près) pour chaque heure. Nous les représentons dans le tableau 1.

Heure	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Kolmorogov	9.3	10.2	10.7	11.0	10.7	9.8	3.9	3.6	3.1	4.6	4.1	3.4
p-valeur	0	0	0	0	0	0	10^{-13}	10^{-11}	10^{-8}	0	10^{-15}	10^{-10}
Heure	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Kolmorogov	4.0	4.5	5.6	6.3	6.1	5.6	4.7	2.9	2.1	1.8	2.7	5.6
p-valeur	10^{-10}	0	0	0	0	0	0	10^{-7}	10^{-4}	10^{-3}	10^{-6}	0

TAB. 1 – *Résultats du test de Kolmorogov-Smirnov et p-valeur pour les différentes heures.*

Ainsi, l'hypothèse selon laquelle est la distribution est gaussienne peut largement être rejetée (la p-valeur est très inférieure à 0.05 pour toutes les heures). On voit cependant qu'elle varie beaucoup, puisqu'elle est inférieure à 10^{-15} pour la plupart des heures et qu'elle atteint 10^{-3} entre 21h et 22h.

Nous avons ensuite cherché à déterminer si des distributions différentes apparaissaient en utilisant les saisons déterminées précédemment. Comme nous pouvons le voir dans la figure 5, pour certaines heures et certaines saisons la distribution semble bien plus proche d'une gaussienne, par exemple pour les heures 20 et 21 aux saisons 2 et 3.

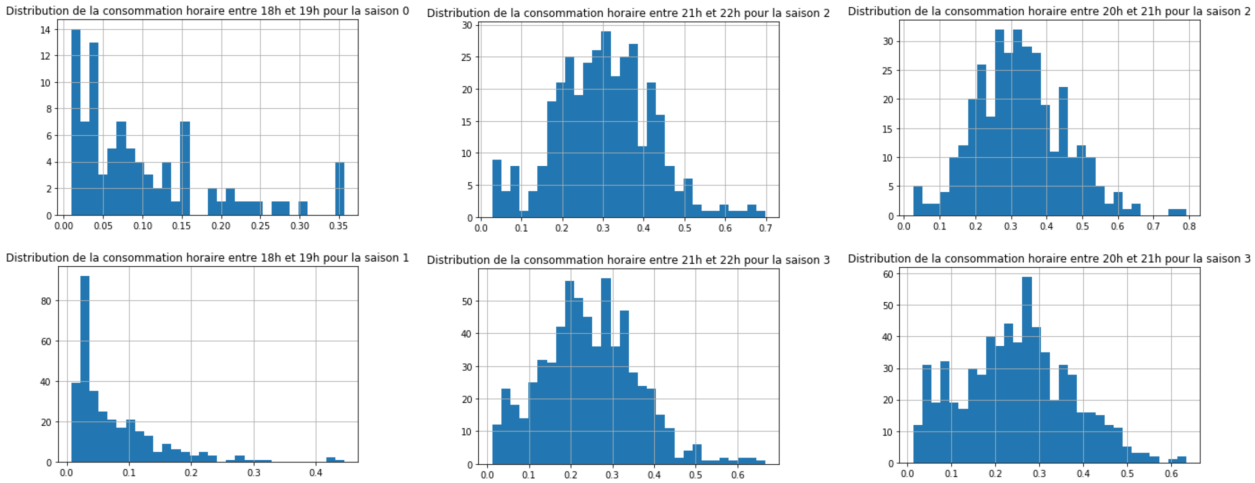

FIG. 5 – *Distribution de la consommation horaire par saisons pour quelques heures.*

Nous pouvons à nouveau déterminer les valeurs du test de Kolmogorov et la p-valeur associée pour les saisons intéressantes et quelques heures de la journée. Certaines de ces valeurs sont recensées dans le tableau 2. Nous trouvons alors des résultats très intéressants, puisque durant les saisons 2 et 3, l'hypothèse selon laquelle la distribution de la consommation horaire entre 20h et 21h ou entre 21h et 22h est une gaussienne ne peut plus être rejetée (la p-valeur est très supérieure à 0.05).

Saison	**2**				**3**			
Heure	**18**	**19**	**20**	**21**	**18**	**19**	**20**	**21**
Kolmorogov	1.6	1.0	0.7	0.6	2.9	1.5	0.9	0.8
p-valeur	10^{-2}	0.2	0.6	0.8	10^{-7}	10^{-2}	0.3	0.5

TAB. 2 – *Résultats du test de Kolmorogov-Smirnov et p-valeur par saisons pour quelques heures.*

La détection de saison nous permet donc de conclure qu'il existe pour certaines heures une distribution proche d'une gaussienne pour la consommation horaire, conditionnellement au choix d'une bonne saison. Cependant, pour la plupart des heures même la séparation en saison ne permet pas de s'approcher d'une distribution gaussienne et beaucoup d'histogrammes restent proches d'une distribution exponentielle. Cela signifie que sur ces heures, la consommation est en général assez faible et que les consommations supérieures à la moyennes sont rares. À l'inverse durant les heures de la soirée, la consommation semble répartie autour d'une moyenne toujours assez élevée.

INFLUENCE DE LA MÉTÉO

Nous avons enfin tenté de prédire la consommation en 2010 en utilisant les données de consommation des années antérieures ainsi que les données météorologiques correspondantes. Nous avons d'abord effectué une simple régression linéaire sur le jeu de données d'entraînement pour prédire la métrique "global active power". Nous avons tenté de régulariser la régression linéaire avec plusieurs coefficients différents en suivant la méthode ridge, mais tous nos essais nous ont donné une valeur de R^2 inférieure à celle obtenue sans régularisation. Nous avons donc utilisé les résultats de la régression linéaire classique pour prédire et pour en déduire des corrélations entre les paramètres.

Nous avons alors mesuré la performance de notre prédicteur sur le jeu de données de test et nous avons trouvé un coefficient $R^2 = 0.623$ pour la régression linéaire, soit un coefficient $R = 0{,}79$. Ce n'est pas une performance très satisfaisante mais elle montre bien l'existence d'une corrélation entre les données. Nous pouvons alors faire des déductions sur la façon dont les données météorologiques influencent la consommation, à partir des coordonnées du vecteur de coefficients β représentées dans le tableau 3. Nous avons mis en gras les coefficients les plus importants en termes d'intensité.

Métrique	**Coefficient**
biais	0.0798568
pmer	0.00609899
tend	0.0535041
cod_tend	0.014226
dd	-0.00812423
ff	-0.012525
t	-0.0299947
td	-0.040781
u	**0.0754571**
vv	0.0170034
ww	-0.00670193
w1	0.0436859

Métrique	**Coefficient**
w2	-0.0681074
n	0.00168108
nbas	-0.0108261
hbas	0.0153587
cl	-0.00918462
cm	0.015158
ch	0.00161848
pres	0.00383123
tn12	0.00321222
tx12	**-0.110239**
rafper	0.05881
ht_neige	**-0.0809063**

Métrique	Coefficient
ssfrai	-0.0192155
rr1	0.0384838
rr3	-0.0355849
phenspe1	**0.0775889**
phenspe2	-0.000469906
phenspe3	**0.0712553**
nnuage1	-0.0197563
ctype	-0.00250493
hnuage1	-0.000148738
nnuage2	-0.0104875
ctype2	-0.0227675
hnuage2	-0.0178437

TAB. 3 – *Coefficients de la régression linéaire pour déterminer la consommation active.*

Ainsi nous pouvons observer que le paramètre influençant le plus la consommation active est "tx12", la température maximale sur les 12 dernières heures. De façon très naturelle, plus la température maximale est élevée, moins la consommation est élevée. De façon plus surprenante, on observe que la hauteur de neige influence également négativement la consommation, alors qu'on s'attend à ce qu'une hauteur de neige élevée implique une consommation élevée pour combattre le froid. Enfin l'humidité joue également un rôle important, et les phénomènes spéciaux 1 et 3 dont nous n'avons pas réussi à obtenir les descriptions sur le site de Météo France.

Nous avons également procédé à la prédiction des mesures “sub metering 1”, “sub metering 2” et “sub metering 3”. Cependant les résultats ont été bien moins concluants puisque nous avons obtenu des coefficients R^2 respectivement égaux à 0.126, 0.097 et 0.295. Nous en déduisons que ces mesures sont assez faiblement corrélées aux données météorologiques, ce qui est assez logique puisque l'utilisation des appareils de la cuisine, de la buanderie ou du chauffe-eau ne dépendent a priori pas vraiment de la météo.

De plus, afin de vérifier que la corrélation entre les données météorologiques et les données de consommation était suffisamment importante pour que les considérations qualitatives précédentes soient bien fondées, nous avons réalisé une régression polynomiale de degré 2 avec le jeu de données préalablement transformé. Nous obtenons alors de meilleurs résultats sur le jeu de données de test avec un coefficient $R^2 = 0.792$, soit un coefficient $R = 0.890$.

Nous avons enfin voulu tester une autre méthode de prédiction par l'algorithme de régression par plus proches voisins. Cependant celle-ci ne s'est pas révélée concluante puisque nous avons obtenu des erreurs quadratiques moyennes environ 5 fois supérieures à celles obtenues par régression linéaire, quelles que soient les colonnes à prédire.

CONCLUSION

Finalement, nous avons mis en évidence l'existence de périodes semblables à des saisons dans la répartition de la consommation d'énergie au cours de l'année. Nous avons également étudié la distribution de la consommation horaire, et conclu contre l'intuition qu'elle n'était pas gaussienne en général, mais qu'elle pouvait s'en approcher lorsqu'elle était conditionnée par les saisons mises en évidence précédemment. Enfin nous avons tenté d'utiliser les données météorologiques pour prédire la consommation d'énergie, ce qui se révèle moyennement efficace en pratique mais permet de mettre en avant des corrélations.

Ce projet présente toutefois plusieurs limites, notamment le fait que le jeu de données utilisé pour les données de consommation est basé sur une unique maison et donc largement biaisé par les habitudes personnelles de ses habitants. De plus, les métriques météorologiques utilisées sont également corrélées entre elles, et il pourrait être efficace d'utiliser des techniques de réduction de dimension préalablement à la prédiction pour obtenir de meilleurs résultats.

RÉFÉRENCES

[1] Georges Hébrail, Alice Bérard. *Individual Household Electric Power Consumption Data Set.* UC Irvine Machine Learning Repository

[2] Météo France. *Données SYNOP Essentielles OMM.* Données Publiques de Météo France