POLYTECHNIQUE MONTRÉAL

Équipe 19

Cours : IND6212 Exploration de données industrielles
Présenté au professeur Bruno Agard

Réalisé par : Félix Blanchard (2285987) et Leila Rouga (1570533) Avec *Python*

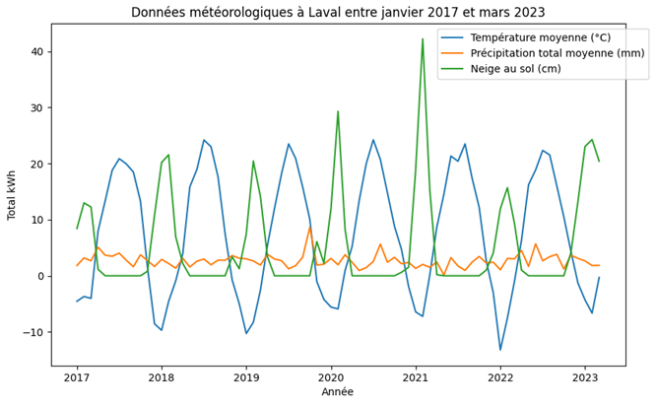
Description du contexte :
Notre étude se propose d'analyser la consommation d'électricité mensuelle en kWh dans la municipalité de Laval selon les différents secteurs suivant : (1) Le **secteur agricole** concernant les exploitations agricoles, (2) le **secteur commercial** concerne les entreprises de vente et de services, (3) le **secteur industriel** regroupe les installations de production et de fabrication, (4) le **secteur institutionnel** englobe les établissements publics et privés comme les hôpitaux et les écoles, (5) et le **secteur résidentiel** fait référence aux logements individuels et collectifs.
Ces secteurs présentent des profils de consommation énergétique distincts, susceptibles d'être affectés différemment par les variations météorologiques. Parmi les variables climatiques considérées, (A) la température mensuelle moyenne (en °C), (B) la précipitation quotidienne moyenne par mois (en mm) et (C) la quantité de neige au sol quotidienne moyenne par mois (en cm).

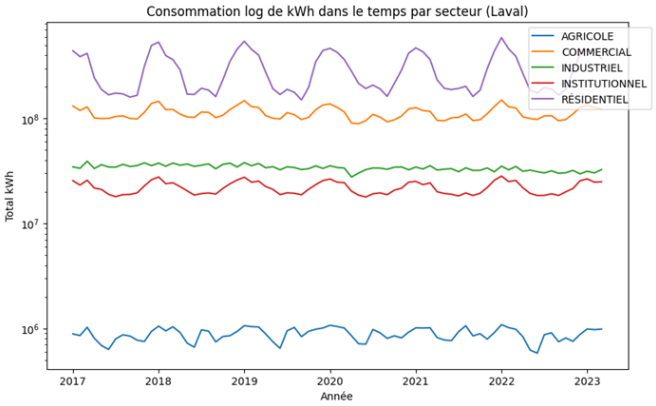
Problématique :
Quelle est l'influence des variables météorologiques (énumérées précédemment) sur la consommation d'électricité mensuelle (en kWh) dans les différents secteurs d'activité à Laval ?

Préparation, description et sources des données :
Nos données proviennent de deux sources différentes, les données énergétiques proviennent d'un fichier Excel (.xlsx) fournit par Hydro-Québec comportant toutes les consommations en kWh par secteur, par mois et par municipalité du Québec entre 2016 et mars 2023. Ce fichier comporte un peu moins de 500 000 lignes et beaucoup de données manquantes. Après avoir exploré ce fichier, nous avons établi quelques municipalités proches de Montréal avec des intervalles de dates sans données manquantes. Nous avons finalement sélectionné Laval qui avait un historique de données complet sur la période janvier 2016 à mars 2023.
Notre deuxième source de données provient d'Environnement Canada et est disponible à travers une librairie *Python*. Ces données sont fournies de façons journalières et comportent une trentaine de colonnes. Pour avoir accès aux données de Laval, nous avons entré les coordonnées géographiques de la municipalité et le programme nous a présenté avec le centre d'étude météorologique le plus proche (dans notre cas, il se situe à l'aéroport Pierre Eliott Trudeau). Comme Environnement Canada n'a pas de centre à Laval, nous avons pris ce jeu de données qui se situe à environ 10 km du lieu d'intérêt. Les enregistrements d'un certain nombre de variables n'ont commencé qu'en janvier 2017, alors nous avons décidé de démarrer notre étude à ce moment-là. Nous avons commencé par enlever l'ensemble des colonnes sans données (15 colonnes), puis nous avons enlevé 2 colonnes dont la corrélation était de 1 avec la variable « température » puisqu'elles étaient directement calculées à partir de celle-ci. Puis nous avons enlevé toutes les colonnes jugées non intéressantes à l'étude pour diminuer la taille de notre DataFrame (comme longitude, latitude, nom de la station météo…). Nous avons donc retenu les variables : "date", "température moyenne", "précipitation totale en mm" et "neige au sol en cm". Après quoi, nous avons commencé à transformer les données pour avoir des valeurs mensuelles et non journalières, nous avons calculé une moyenne mensuelle pour chacune des trois variables explicatives et nous avons fait un nouveau DataFrame. Avec ces valeurs, on a plus que 75 lignes, une pour chaque mois, de janvier 2017 à mars 2023. Ensuite, nous avons importé notre fichier Excel nettoyé ne comprenant que la consommation mensuelle en kWh à Laval par secteur entre 2017 et mars 2023. Nous avons fait un DataFrame pour chaque secteur (5) puis nous avons concaténé chacun de ces DataFrame à notre DataFrame traité des données météo, obtenant ainsi 5 DataFrame exploitable pour notre analyse, pour chaque industrie.

Description statistique des données :

Nous avons 4 variables indépendantes et 1 variable dépendante. 4 de ces variables sont continues et une est catégorielle. Dans la construction de cette étude, on a analysé la consommation d'électricité par secteur (celui-ci étant notre variable catégorique). Ainsi, nous n'analyserons pas la consommation totale de l'électricité à Laval. Toutes nos données continues (température moyenne, précipitation moyenne, neige au sol moyenne et consommation d'électricité) sont temporelles. Ci-dessous, on peut observer la distribution des variables météorologiques à gauche et de la variable consommation par secteur à droite (Le graphique de droite utilise une échelle logarithmique pour faciliter la visualisation des différences d'échelle entre les secteurs).





La première chose qu'on peut constater est une forte saisonnalité dans les données, nous reviendrons plus tard sur ce point pour expliquer comme cela impacte notre étude et plus précisément la préparation des données.

kWh\ Secteur par mois	AGRICOLE	COMMERCIAL	INDUSTRIEL	INSTITUTIONNEL	RÉSIDENTIEL
Moyenne	881 114.84	112 601 202.93	33 832 875.46	21 908 118.68	292 371 911.83
Médiane	885 538.00	109 263 501.50	33 820 692.25	21 479 273.10	243 863 710.00
Écart-type	123 580.33	14 867 834.03	2 257 694.83	2 968 642.55	121 878 897.25
Min	580 011	89 128 162.2	27 627 894.16	17 837 244.8	149 920 722
Max	1 085 066	149 819 164.1	39 113 501.82	28 249 812.3	588 696 319

Nous remarquons que la consommation mensuelle d'électricité dans les secteurs agricole, commercial, institutionnel et industriel ont une médiane proche de leur moyenne. Certains secteurs consomment énormément d'énergie (Résidentiel, Commercial) tandis que d'autres sont vraiment minime (Agricole). Le secteur résidentiel révèle une fluctuation très marquée (écart-type de 121 878 897.25 kWh, soit environ 50% de la médiane). De plus, le résidentiel possède une moyenne bien supérieure à la médiane, avec des pics saisonniers dans la consommation comme présenté dans le graphique.

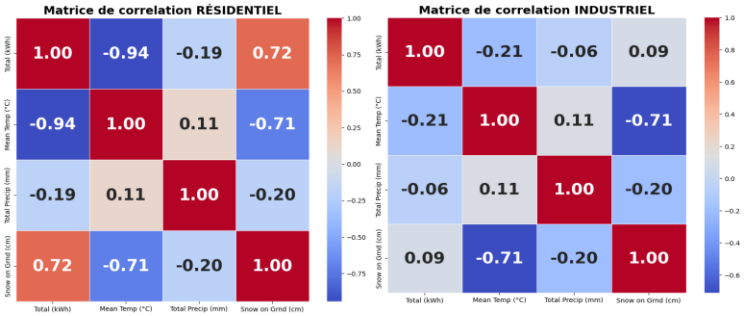
Données\ Par mois	Mean Temp (°C)	Total Precip (mm)	Snow on Grnd (cm)
Moyenne	7.31	2.70	5.50
Médiane	7.50	2.64	0.18
Écart-type	11.01	1.25	8.69
Min	-13.23	0.13	0
Max	24.24	8.65	42.25

Les variables météorologiques possèdent chacune des caractéristiques distinctes : la Température Moyenne (°C) affiche une grande variabilité saisonnière, avec une moyenne de 7.31°C et des fluctuations entre -13.23°C et 24.24°C. Les Précipitations Journalières Moyenne (mm) se répartissent régulièrement autour d'une moyenne de 2.70 mm, indiquant une relative constance mensuelle. En revanche, la Neige au Sol journalière moyenne (cm) montre une disparité notable, avec une médiane de 0.18 cm et des épisodes de plus de 40 cm, suggérant des variations significatives dans la quantité de neige

présente. Ces différences peuvent impacter divers aspects, comme la consommation d'énergie, particulièrement pour le chauffage, soulignant ainsi l'importance de comprendre les variations météorologiques pour potentiellement prédire la consommation électrique.

Nouvelle modification des données en prenant en compte la saisonnalité :
Dans la suite de ce rapport, nous utiliserons les sigles AS pour « avec prise en compte de la saisonnalité » et SS pour « sans prise en compte de la saisonnalité ».
Les données soumisea à de la saisonnalité peuvent être décomposées de la façon suivante : $Y_{data} = Saisonnalité + Tendance + résidu$, pour essayer d'améliorer les performances des modèles prédictifs que nous développeront par la suite, nous avons choisi de considérer cette saisonnalité pour la retirer lorsque nous développons nos modèles et pour la réintroduire ensuite, lors de l'évaluation ($Y_{kWh} = modèle(Tendance + résidu) + Saisonnalité$). Nous souhaitons aussi savoir si cette prise en compte a un impact important sur l'amélioration de nos prédictions, c'est pourquoi nous développerons nos modèles à la fois sur des données désaisonnalisées (AS) et d'autres non-désaisonnalisées (SS). Pour ce faire, nous avons rajouté des colonnes à chacun de nos 5 DataFrames comprenant chaque variables AS et SS, puis une colonne de plus qui correspond à la saisonnalité retirée de la variable à expliquer pour la réintroduire dans l'évaluation, comme présenté dans la dernière équation. La saisonnalité a été retirée par la fonction *seasonal_decompose* de la librairie *statsmodels.tsa.seasonal* disponible sur *Python*.

Matrices de corrélations :



Les matrices de corrélations révèlent ce qui suit dans chacun des secteurs : **Résidentiel :** la température et la neige ont des effets très forts sur la consommation électrique, avec la température ayant un impact particulièrement significatif, ce qui reflète la haute sensibilité de ce secteur aux variations climatiques, probablement en raison des besoins de chauffage. **Agricole :** l'impact de la température sur la consommation d'électricité est modéré, contrairement au secteur résidentiel où l'impact est très fort. La neige, bien que positivement corrélée, semble avoir un impact plus limité en agriculture comparé à sa forte influence dans le résidentiel. **Commercial :** la température a une corrélation négative modérée avec la consommation d'électricité, et la neige montre une association positive modérée, indiquant une sensibilité aux conditions hivernales moins marquée que dans le résidentiel mais plus significative qu'en agriculture. **Industriel :** les variations météorologiques ont le moins d'impact sur la consommation électrique dans ce secteur, avec des corrélations faibles dans les deux cas, suggérant une indépendance relative de la consommation d'électricité par rapport aux conditions météorologiques. **Institutionnel :** la consommation est très fortement influencée par la température, un peu comme dans le résidentiel, et la neige a également une corrélation positive modérée. Dans tous les secteurs, les précipitations ont une corrélation plutôt faible sur la consommation d'électricité.

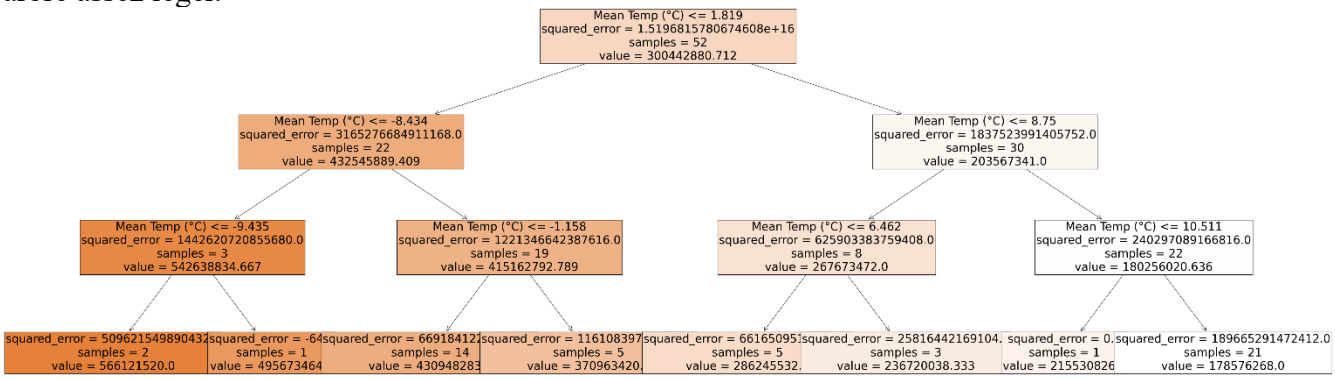
Explication du choix de type de modèle :
Notre étude comporte 4 variables explicatives et une variable à expliquer, de plus nous n'avons que 75 lignes de données par DataFrame de secteur, ce qui représente un relativement faible volume de données. De ce fait, nous avons évalué qu'il serait plus pertinent de se concentrer sur des méthodes de prédictions

interprétables. En mettant l'emphase sur « l'explicativité » de nos modèles, nous écartons d'emblée les réseaux de neurones, les forêts aléatoires et le gradient boosting comme méthode d'analyse. De plus, nous devons prendre en compte que nous avons une variable à prédire dont la nature est « continue ». De même, nos variables explicatives pour chaque secteur sont aussi « continues ». Ainsi, nous avons décidé d'utiliser les arbres de régressions et de comparer leurs performances à des régressions linéaires multiples.

Développement des modèles :

1. **Arbres de régressions**

Nous avons fait deux arbres de régressions pour chaque secteur, un AS et un SS, de manière à pouvoir comparer l'impact de la prise en compte de celle-ci dans le contexte de notre étude. Pour ce faire, nous avons utilisé la librairie *scikitlearn* et appelé la fonction : *DecisionTreeRegressor*. Pour paramétrer nos arbres, nous avons séparé l'ensemble d'entraînement et de test en deux sous-ensembles de respectivement 70% et 30% des données. Après avoir fait plusieurs ajustements sur la profondeur de l'arbre, nous avons trouvé que les performances étaient les plus intéressantes avec une profondeur de 3. Cela concorde avec notre volonté « d'interprétabilité » puisque nous n'avons que 3 variables explicatives et cela donne un arbre assez léger.



Ci-dessus, se trouve l'arbre de régression SS pour le secteur de consommation "Résidentiel" (notre meilleur secteur en termes de performance). Chaque mois est classé dans une des huit catégories, la première variable discriminante est la « température (°C) », pour l'arbre AS et SS. L'ensemble des autres nœuds de l'arbre ont aussi une discrimination faite sur la température pour le modèle SS. Pour celui AS, la variable neige au sol est aussi présente. On peut ainsi déduire que pour le secteur résidentiel, la consommation d'électricité est expliquée en grande majorité par la température, dans une moindre mesure par la présence de neige, mais pas en fonction de la pluie. Ce constat n'est pas partagé dans tous les secteurs, par exemple, pour le secteur agricole le modèle intègre les 3 variables. Il en va de même pour l'ensemble des autres arbres, ou les 3 variables sont présentes, sauf dans le cas du secteur institutionnel et commercial SS ou il n'y en a que 2. Dans l'ensemble, la variable « température » constitue le nœud le plus populaire, mais sa popularité diminue légèrement pour les modèles AS.

2. **Régressions linéaires multiples**

Pour notre modèle de régression, on a aussi fait un modèle AS et un modèle SS, cela pour chacun des 5 secteurs. La librairie est *scikitlearn* et la fonction appelée est *LinearRegression*. Avec 30% de test et 70% réservé à l'entraînement. Pour chacun des modèles, l'équation des modèles est la suivante :

$$\boldsymbol{Y_{kWh/secteur} = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + e}$$

Avec :
$\beta_0, \ldots, \beta_3$ *: Les coefficients*
X_1 *: La température moyenne par mois*
X_2 *: La précipitation quotidienne moyenne par mois*
X_3 *: La quantité de neige au sol quotidienne moyenne par mois*
e *: L'erreur résiduelle*

Dans le cas du secteur résidentiel SS on obtient :

$$Y_{kWh/résidentiel} = 383\,196\,244 - 9\,942\,255 \times X_1 - 7\,427\,259 \times X_2 + 903\,265 \times X_3$$

Dans cette équation, l'augmentation de 1°C de température mensuel moyenne diminue la prédiction de consommation d'électricité de presque 10 millions de kWh pour le mois à Laval, -7,4 millions pour l'augmentation de 1 mm de pluie quotidienne moyenne sur le mois et l'augmentation de 1cm de neige sur le mois entraîne une augmentation de consommation de 0,9 million de kWh. Cela avec une consommation de base de plus de 383 millions de kWh par mois si toutes variables sont à 0.
Les résultats des métriques seront présentés à la prochaine partie, pour analyser plus en détail les coefficients nous avons standardisés les β_i de manière à les comparer entre eux. Pour ce faire, nous avons utilisé la fonction *zscore* de la librairie *statsmodels.api*. Celle-ci nous indique par exemple que dans le cas du secteur résidentiel SS : $X_1 = -0.8671$, $X_2 = -0.0717$, et $X_3 = 0.0933$ et AS : $X_1 = -0.8202$, $X_2 = -0.1109$, et $X_3 = 0.0438$. On constate ainsi que X_1 est la variable la plus influente comme nous l'avons déduit dans l'analyse des arbres. Cela se reflète dans les autres secteurs avec plus ou moins d'écart, l'écart le plus marqué est dans le secteur institutionnel SS avec $X_1 = -0.9394$, $X_2 = -0.0024$, et $X_3 = -0.0003$. Le moins marqué est dans le cas du secteur agricole SS avec $X_1 = -0.2824$, $X_2 = -0.0874$, et $X_3 = 0.3459$.

Choix des métriques d'évaluations :
Pour analyser les résultats, nous avons sélectionné trois métriques. D'abord le R^2, car c'est la mesure qui fait référence dans la comparaison de modèles de régression. Ensuite, c'est une mesure très facilement interprétable, puisqu'elle représente la part de Y expliqué par le modèle ce qui permet de comparer les différentes industries entre elles assez facilement. Finalement, le R^2 plutôt que le R^2 ajusté, car nous avons le même nombre de données et de variables à chaque industrie, donc l'ajustement est inutile dans notre situation. Les deux autres métriques RMSE/MAE normalisées (en fonction de la moyenne de chaque secteur) sont pertinentes pour comparer les mêmes industries entre elles, car chaque même industrie est présente 4 fois pour chaque option : avec ou sans prise en compte de la saisonnalité, et l'arbre de régression contre la régression linéaire.

Présentation des résultats :

Secteur/Métrique		Avec prise en compte de la saisonnalité			Sans prise en compte de la saisonnalité		
		RMSE normalisé	MAE normalisé	R2	RMSE normalisé	MAE normalisé	R2
Arbre de décision	Agricole	0.083	0.065	0.666	0.102	0.085	0.487
	Commercial	0.049	0.041	0.859	0.05	0.037	0.851
	Industriel	0.071	0.061	0.18	0.087	0.07	-0.257
	Institutionnel	0.034	0.028	0.918	0.039	0.054	0.792
	Résidentiel	0.033	0.048	0.985	0.077	0.064	0.961
Moyenne des résultats :		0.054	0.0486	0.7216	0.071	0.062	0.5668
Régression linéaire multiple	Agricole	0.09	0.071	0.607	0.14	0.116	0.041
	Commercial	0.046	0.037	0.873	0.095	0.074	0.466
	Industriel	0.075	0.06	0.082	0.081	0.069	-0.103
	Institutionnel	0.035	0.026	0.93	0.051	0.041	0.814
	Résidentiel	0.052	0.042	0.982	0.151	0.132	0.849
Moyenne des résultats :		0.0596	0.0472	0.6948	0.1036	0.0864	0.4134

Analyse des résultats :
Analyse macro : Nous avons de biens meilleurs résultats avec la prise en compte de la saisonnalité (AS). Le R^2 est de 0,72 et 0,69 en moyenne AS contre 0,57 et 0,41 SS. Les arbres sont un peu mieux que la régression, AS et beaucoup mieux SS. En effet, on passe de 0,69 à 0,72 pour l'un contre 0,41 à 0,57 pour l'autre en moyenne de score de R^2. Soit une augmentation de 4% de performance moyenne contre une

augmentation de presque 40%. Ce qui témoigne d'une meilleure robustesse des arbres sur les données SS. Il n'y a que pour le secteur industriel SS ou la performance diminue en utilisant les arbres.
Le RMSE et le MAE normalisé indiquent dans chaque cas la même conclusion, sauf dans le cas du MAE normalisé AS où celui est plus performant en moyenne dans le cas de la régression. Cela vient nuancer l'intérêt de choisir un arbre vis-à-vis d'une régression dans notre situation avec des données désaisonnalisé (AS).

Analyse micro : Les secteurs avec les meilleures performances sont le secteur résidentiel, institutionnel et commercial. Ils ont tous un R^2 presque toujours supérieur à 80% et de quasiment 100% dans le cas AS pour le secteur résidentiel, ce qui est une performance extrêmement prometteuse. Le secteur agricole a une performance moyenne (entre 0% et 67% dépendamment du modèle) et le secteur industriel à une mauvaise performance avec parfois un R^2 négatif, ce qui indique qu'en prédisant toujours la moyenne, on aura de meilleures performances.
Le meilleur modèle pour le secteur agricole est l'arbre AS et le pire est la régression SS. Pour le secteur commercial le meilleur modèle est la régression AS et le pire est la régression SS. Pour le secteur industriel le pire est l'arbre SS et le meilleur est l'arbre AS. Le meilleur modèle pour le secteur institutionnel est la régression AS et le moins bon est l'arbre SS. Finalement, le meilleur modèle associé au secteur résidentiel est l'arbre AS et le pire est la régression SS. Ainsi, il existe une homogénéité dans la forte performance des modèles AS, en revanche, il ne semble pas avoir de consensus clair entre l'arbre et la régression.

Critique de l'étude et des résultats :
Ainsi, les types de modèles que nous avons sélectionnés se sont révélés pertinents dans le cadre de notre étude puisqu'ils ont démontré des résultats très satisfaisant (surtout pour 3 secteurs), et ont exploités avec pertinence les variables explicatives de notre modèle puisque les secteurs avec de grandes performances sont aussi ceux dont les variables explicatives étaient le plus corrélé avec la consommation d'électricité. En revanche, les données météorologiques ne semblent pas pertinentes lorsqu'il s'agit d'expliquer la consommation électrique du secteur industriel et il faudrait analyser plus en détail des caractéristiques de ce secteur pour identifier des variables explicatives avec un plus fort potentiel.
Concernant le paramétrage de nos modèles, dans le cas de l'arbre, une profondeur de 3 s'est révélée adéquate et souvent, l'augmentation de celle-ci ne contribuait pas à l'augmentation de la performance. Dans le cas de la régression, nous sommes restés sur une régression linéaire simple, mais nous aurions pu ajouter des effets d'interactions et des effets quadratiques, cela aurait apporté une nouvelle dimension à notre analyse tout en augmentant potentiellement la performance de prédiction.
Finalement, nos résultats ont démontré de solides bases pour répondre notre problématique comme nous l'avons fait dans la partie « analyse des résultats ».

Conclusion :
Ainsi, la consommation de l'électricité mensuelle à Laval peut être prédite de façon assez précise pour 3 secteurs (Résidentiel, Institutionnel et Commercial), elle peut être vaguement prédite dans le secteur agricole et très mal prédite dans le secteur industriel. Cela, en fonction de 3 variables météorologiques présentées précédemment. Parmi ces variables, on voit que leur influence individuelle diffère d'un modèle/secteur à l'autre avec toutefois un consensus sur l'importance de la variable « température » vis-à-vis des 2 autres. Dans le choix de la construction des modèles, la prise en compte de la saisonnalité (AS) s'est révélée déterminante, avec des performances moyennes généralement un peu supérieur dans le cas des arbres plutôt que dans le cas des régressions, mais pas suffisamment importantes pour imposer un type de modèle plutôt que l'autre.
Finalement, la surprise de cette étude est la mise en lumière de la robustesse des arbres de régressions vis-à-vis des régressions linéaires sur l'analyse de données non désaisonnalisées (SS).