Prédiction de la consommation électrique d'une Marina par les méthodes CART et forêts aléatoires.

Sullivan ROYER, Olivier FRUCHIER, Dorian GACHON, Thierry TALBERT
Laboratoire PROMES-CNRS, Université de Perpignan Via Domitia

RÉSUMÉ – De nos jours, nous assistons à l'émergence des smarts grids et microgrids dont l'objectif est de perfectionner la gestion énergétique du réseau électrique en améliorant l'efficacité de la production, du transport, de la distribution et de la consommation de l'électricité. Cet article traite de la partie consommation. Le but est de modéliser la consommation électrique d'une marina afin d'en prédire les valeurs et permettre à la régie de la maîtriser. Pour cela, ce sont les méthodes CART et forêts aléatoires qui ont été employées. Les résultats obtenus à l'échelle mensuelle sont satisfaisants et encouragent à poursuivre le travail jusqu'à l'échelle de la minute.

Mots-clés – Microgrid, consommations, modélisation, prédictions, régression, machine learning.

1. INTRODUCTION

De nos jours, on peut constater l'émergence et l'amélioration des technologies exploitant les énergies renouvelables et apparaître de nouvelles technologies qui améliorent l'efficacité énergétique [5]. C'est ainsi qu'évolue le réseau électrique conventionnel en réseau intelligent, dit Smart Grid [6][7] et Microgrid [8][9]. Cette évolution doit permettre de prendre en compte les actions des acteurs du système électrique, tout en assurant une livraison d'électricité plus efficace, économiquement viable et sûre. C'est dans ce contexte que se place notre projet. Il s'agit d'une marina qui fait face à une consommation énergétique croissante chaque année de la part de ses plaisanciers et qui souhaite d'avantage maîtriser la consommation des fluides (eau et électricité). Premier port de plaisance d'Europe, la consommation d'énergie électrique de la marina est équivalente à une petite ville.

Pour atteindre ses objectifs, le port a équipé son réseau électrique d'un réseau de communication utilisant le Courant Porteur de Ligne (CPL) afin de faire transiter des données de comptage, mesures de consommation d'énergie et de qualité du réseau. Alors que la majorité des travaux de recherche actuels sur les microgrids se focalisent sur la production [10][11][12], notre travail porte sur la partie consommation, plus précisément sur sa modélisation et sa prédiction comme elle a déjà pu être traitée dans [13][21][22] par exemple ou encore dans le travail de thèse de Benjamin Goehry soutenue en décembre 2019 [14].

Le plan de l'article est de présenter, dans un premier temps, les méthodes de modélisation employées, l'ensemble des entrées/sortie du modèle et les paramètres utilisés ; puis dans un deuxième temps de discuter des résultats. Enfin conclure sur le travail effectué et exposer les suites envisagées.

2. MODÉLISATION

Nous cherchons à modéliser la consommation électrique de la marina pour améliorer le système en place ou bien en concevoir un nouveau, plus performant. Le but est de prédire les valeurs de consommation à partir d'un ensemble de données d'observation explicatives. Les modèles que nous allons utiliser sont issus d'une agrégation d'arbres de régression : les méthodes CART et forêts aléatoires. La méthode CART nous vient de Breiman en 1984 (réédition en 2017)[15] et la notion de forêts aléatoires a été introduite par ce même Breiman en 2001 [16].

2.1. Méthode CART

L'acronyme CART signifie Classification And Regression Tree. Les algorithmes d'arbres de décision consistent à déterminer un ensemble de conditions logiques de partition (division) du type Si...-Alors... afin de prévoir les valeurs ou classifications prévues d'une variable. Cette variable à modéliser ou prévoir est qualitative (classification) ou quantitative (régression). Dans le cas de la consommation d'électricité, il s'agit d'un problème de régression puisque l'on veut déterminer une consommation de la marina facturée en kWh.

Un arbre est construit à partir de données d'observations explicatives que l'on appelle prédicteurs. La méthode CART vérifie récursivement pour chaque nœud si une séparation est possible sur la base du prédicteur choisi au hasard pour ce nœud.

2.2. Forêts aléatoires

Aussi connues sous le nom anglophone « Random Forests », les forêts aléatoires sont une combinaison ou une agrégation d'un grand nombre de modèles issus des arbres de régression. Les forêts utilisent le hasard pour améliorer les performances de l'algorithme CART décrit dans la sous-section précédente. Une forêts est composée de plusieurs arbres construits indépendants entre eux afin de rendre le modèle plus efficace. [17]

Principe : on tire au hasard dans la base d'apprentissage B échantillons avec remise. Pour chaque échantillon i on construit un arbre CART Gi(x) selon un algorithme légèrement modifié. À chaque fois qu'un nœud doit être coupé, on tire au hasard une partie des attributs (q parmi les p attributs) et on choisit le meilleur découpage dans ce sous-ensemble. Les arbres sont moins corrélés car ils sont sont construits sur des échantillons différents et ont appris sur un ensemble différent d'attributs.

Chaque arbre est petit et moins performant mais l'agrégation compense ce manquement parce que chaque attribut se retrouve typiquement dans plusieurs arbres. On utilise l'erreur Out Of Bag (OOB) pour prévenir le sur-apprentissage (on choisit B là où l'erreur se stabilise et ne descend plus) [18].

2.3. Prédicteurs

La modélisation par forêts aléatoires et méthode CART présente l'avantage d'utiliser un très grand nombre de variables explicatives qui peuvent même être de type varié. Appelées prédicteurs, les variables que nous utilisons sont de types météorologique et électrique. Les prédicteurs sont un regroupement de données historiques météorologiques et de données de consommation du passé telles que la hauteur des précipitations, la température extérieure, l'humidité, la direction et la vitesse du vent ; ainsi que les données historiques de consommation (cf. Fig. 1).

2.4. Données d'entraînement et données de test

Le jeu de données historiques est découpé en deux jeux : un jeu d'entraînement et un jeu de test qui incluent les données

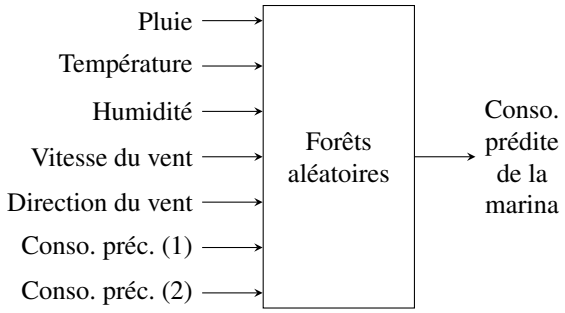

Fig. 1. Prédiction de la consommation électrique de la marina par un modèle de forêts aléatoires avec la liste des prédicteurs en entrée du modèle selon le pas de temps. (1) mois ou jour ou heure précédent. (2) même mois l'année précédente ou même jour la semaine précédente ou même heure le jour précédent.

météorologiques et de consommation électrique. L'idéal étant d'entraîner les modèles les mêmes mois qui seront testés, donc de disposer de plusieurs années d'historiques.

Plusieurs pas de temps ont été utilisés pour la modélisation des consommations mais ont été retenus les modèles mensuels et horaires. Mensuels parce que ces modèles sont construits à partir de données de consommations facturées de plusieurs années, et horaires parce que les modèles ont un horizon de prédiction à court terme utile à la régie de la marina.

2.4.1. *Modèles mensuels*

Pour construire un modèle mensuel nous utilisons les données provenant des factures mensuelles à notre disposition. Ensuite, il convient de travailler avec des données de moyennes météorologiques mensuelles fournies par MétéoFrance et d'utiliser comme mesures passées la consommation du mois précédent et la consommation du même mois l'année précédente.

Le jeu de données mensuelles complet comprend des données de janvier 2016 à octobre 2019. Nous allons couper ce jeu en deux : les données du jeu d'entraînement sont les données mensuelles de janvier 2016 à décembre 2018 ; et les données du jeu de test comprend les consommations mensuelles de l'année 2019, du mois de janvier au mois d'octobre. Ainsi on entraîne le modèle sur trois ans, soit 36 mois et on le teste sur dix mois.

2.4.2. *Modèles horaires*

Pour construire un modèle horaire nous utilisons les données de consommation provenant d'une centrale d'agrégation et d'un superviseur (SCADA). Les données enregistrées sont les courants de phase ; les tensions simples et composées ; les puissances actives, réactives et apparentes de chaque phase et totale.

Nous nous sommes servi des puissances actives mesurées en kW comme sortie à prédire. Ensuite, les données météorologiques nous sont fournies par la station météo de Port Camargue au format horaire. Ces données incluent les grandeurs de température, pluie, vitesse et direction du vent. Notons l'absence des mesures d'humidité par rapport au modèle mensuel. Comme mesures passées, nous ajoutons les consommations du jour précédent à la même heure et du même jour à la même heure la semaine précédente.

Le jeu de données horaires complet comprend des données de septembre 2018 à mai 2019, puis de juillet à novembre 2019. Cependant, ces données sont incomplètes jusqu'en janvier. En effet, il y a un trou (absences de données) systématique entre 2 et 10 h tous les matins, dû à un mauvais paramétrage mais rétabli en début d'année. Nous coupons ce jeu en deux : les données du jeu d'entraînement sont la première partie des données (septembre 2018 à mai 2019) ; et les données du jeu de test sont la deuxième partie (août à novembre 2019). Ainsi on entraîne le modèle sur neuf mois et on le teste sur quatre. À l'exception d'août, ces mois sont présents dans la partie entraînement l'année précédente.

2.5. *Paramètres*

Les forêts aléatoires sont une méthode statistique paramétrique et les paramètres modifiables par l'utilisateur sont : le nombre d'arbres de la forêt, le nombre de variables à partitionner à chaque nœud et le nombre minimum de feuilles d'un arbre. Plusieurs paramètres ont été testées afin d'améliorer le modèle mais le nombre de variables à partitionner à chaque décision est laissé par défaut et correspond à un tiers du nombre total.

Plusieurs tests ont été lancés et on a tracé à chaque fois l'erreur de régression en fonction du nombre d'arbres et de feuilles. Seulement un graphique par pas de temps est affiché mais il est représentatif de la majorité des différents cas rencontrés.

2.5.1. *Modèles mensuels*

Pour le modèles mensuels, sur la Fig. 2, si on se focalise sur le nombre de feuilles et l'erreur de régression, on peut classer les performances de la meilleure à la pire comme suit : 5 feuilles, 10 feuilles et 20 feuilles. Les courbes d'erreur pour 5 et 10 feuilles ont des évolutions plus ou moins proches et convergent vers une valeur située entre 1 et 2.10^8 alors que la courbe de 20 feuilles se stabilise entre $5,5$ et 6.10^8. À propos du nombre d'arbres de la forêt, on peut voir que l'erreur se stabilise à partir de 60 arbres environ.

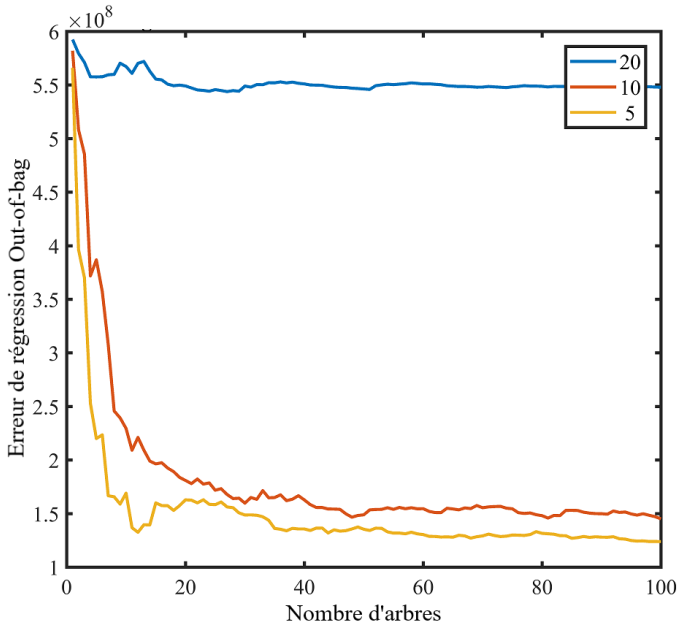

Fig. 2. Évolution de l'erreur de régression en fonction des nombres d'arbres et de feuilles pour des modèles mensuels.

Nous avons réalisé une vingtaine de tests pour déterminer quels sont les meilleurs choix de nombre d'arbres à comparer, et les résultats de l'erreur de régression sont les suivants :

— elle est 9 fois sur 20 à peu près égale pour 20 et 40 arbres ;
— elle est 3 fois sur 20 plus basse pour 20 arbres ;
— elle est 8 fois sur 20 plus basse pour 40 arbres.
— On remarque une « stabilité » à partir de 60 arbres.

2.5.2. *Modèles horaires*

Pour les modèles horaires, la Fig. 3 montre que les courbes d'erreur ont la même allure quelque soit le nombre de feuilles.

Cependant, pour d'avantage de précision, en observant les variations au-delà de 60 arbres où les courbes convergent, on peut établir un ordre de performance décroissant par courbe avec entre parenthèses le nombre de feuilles : violet (5), jaune (10), bleu (20), rouge (40).

Nous constatons que les erreurs de régression des courbes convergent vers une valeur proche de 300. Nous constatons également qu'un nombre de 40 arbres semble suffisant pour obtenir une modèle satisfaisant. Finalement, pour ces raisons et pour rester homogène avec le modèles mensuels, nous allons tester des forêts de 40, plus et moins 20 arbres constitués de 10 et 5 feuilles.

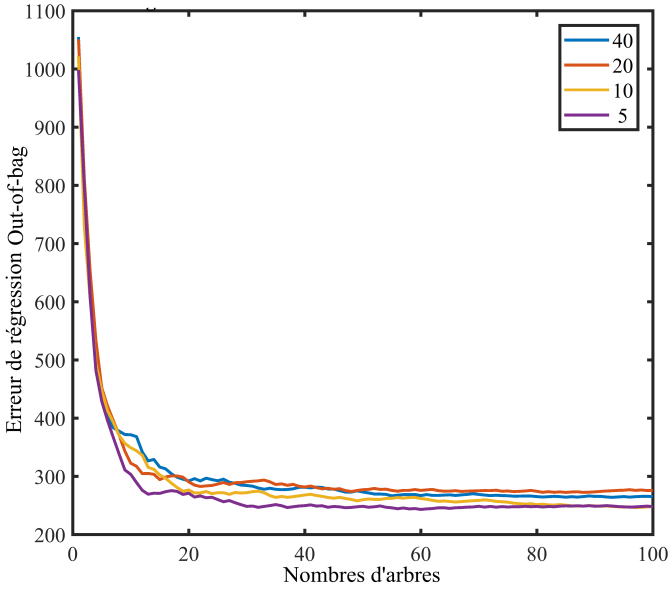

Fig. 3. Évolution de l'erreur de régression en fonction des nombres d'arbres et de feuilles pour des modèles horaires.

3. RÉSULTATS

3.1. Modèles mensuels

Dans le but d'apprécier les choix faits sur les paramètres, nous avons testé quatre configurations et retenu un modèle pour chacune d'elles. Nous injectons dans ces modèles les données des prédicteurs issues du jeu de test (année 2019) et obtenons des prédictions de consommation mensuelles. Ces prédictions sont comparées graphiquement sur la fenêtre du haut de la Fig. 4 aux données réelles de consommation (en bleu), et les résidus des modèles sont tracés sur la fenêtre du bas. Les résidus représentent l'écart entre la valeur réelle est la valeur prédite.
Les quatre modèles sont :

— 1 : une forêt aléatoire de 20 arbres 10 feuilles, en rouge ;
— 2 : une forêt aléatoire de 20 arbres 05 feuilles, en orange ;
— 3 : une forêt aléatoire de 40 arbres 05 feuilles, en violet ;
— 4 : une forêt aléatoire de 60 arbres 05 feuilles, en vert.

Sur le graphique du haut, on voit que la courbe qui est globalement la plus éloignée des données réelles (en bleu) est celle du modèle 1, il s'agit de la forêt de 20 arbres de 10 feuilles comme on pouvait s'y attendre. On voit aussi que les trois autres modèles sont proches et qu'il est difficile de juger quel est le meilleur. Si on regarde le graphique des résidus, en bas de la Figure 4, on voit que les écarts entre les valeurs des modèles et les valeurs réelles sont importants puisqu'ils se situent majoritairement entre -10 et 10 MWh par mois. Notons que le mois de janvier est celui qui présente le plus grand écart, et ce, pour tous les modèles. Cela s'explique du fait que le mois de départ est construit simplement à partir des données historique et n'a pas

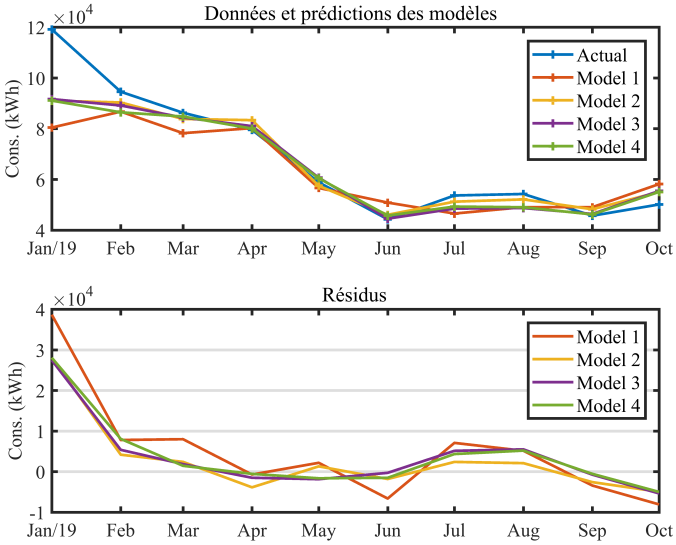

Fig. 4. Données réelles et prédites, plus résidus : différence entre données réelles et prédites. Modèles : 1. 20 arbres 10 feuilles, 2. 20 arbres 5 feuilles, 3. 40 arbres 5 feuilles et 4. 60 arbres 5 feuilles.

de prédicteurs de consommation passé (cf. Fig 1).

3.1.1. Validation des modèles

Puisqu'il est difficile de classer les modèles à cinq feuilles, voyons quelle est leur performance. La performance des modèles est estimée à l'aide des indices Mean Bias Error (MBE), Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), et est présentée dans le Tableau 1.
Voici la définition respective des indices :

1. Mean Bias Error : moyenne du biais du modèle, soit la moyenne des différences entre la valeur vraie et la valeur prédite.
   $MBE = mean(Yreel - Ypredit)$
2. Mean Absolute Error : moyenne des valeurs absolues des différences entre la valeur vraie et la valeur prédite.
   $MAE = mean(abs(Yreel - Ypredit))$
3. Mean Absolute Percent Error : moyenne des pourcentages en valeur absolue des différences entre la valeur vraie et la valeur prédite.
   $MAPE = mean((abs(Yreel - Ypredit)/Yreel) * 100)$

Alors que l'indice MBE indique la tendance (surestimation, sous-estimation ou absence de biais), les indices MAE et MAPE informe sur la précision du modèle.

Le Tableau 1 présente les valeurs de ces indices en fonction des modèles proposés.

Tableau 1. Indices de performance des modèles

Modèles	**MBE (kWh)**	**MAE (kWh)**	**MAPE (%)**
20 arbres 10 feuilles	5042	8774	11,57
20 arbres 05 feuilles	2761	5373	6,58
40 arbres 05 feuilles	3593	5505	6,84
60 arbres 05 feuilles	3824	5630	6,92

Étudions dans un premier temps les conséquences de la réduction du nombre de feuilles de 10 à 5, soient les deux premières lignes du tableau 1 : nous voyons que réduire le nombre de feuilles réduit aussi les erreurs, ce qui confirme les résultats de l'étude de l'erreur de régression vue sur la Figure 2.

Ensuite nous étudions les lignes 2 à 4. Du premier indice, MBE, nous déduisons que les modèles ont tendance à sous-estimer la valeur de consommation puisque l'indice est toujours positif. Ceci est correct parce que les résidus positifs ont l'air plus grand et majoritaires (cf. Figure 4). Cependant la surestimation des modèles au mois de janvier tire fortement cette moyenne vers le haut. Nous remarquons que l'indice MBE augmente lorsque le nombre d'arbres augmente, ce qui signifie que pour les cas présentés, plus il y a d'arbres plus la valeur prédite s'éloigne de la valeur réelle. Il pourrait s'agir d'un surentraînement.

Du deuxième indice, MAE, nous remarquons qu'il évolue aussi comme le nombre d'arbres. Ainsi si le nombre d'arbres augmente, l'écart en valeur absolue augmente. Ce qui éloigne la valeur prédite de la valeur réelle. Nous faisons la même remarque pour le troisième indice, MAPE. En effet, nous pouvons voir que le pourcentage augmente lorsque le nombre d'arbres augmente.

Nous déduisons des indices qu'augmenter le nombre d'arbres augmente le biais et les écarts en valeur absolue et en pourcentage. Aussi, les valeurs d'écart (MBE et MAE) sont élevées en kWh mais les valeurs qu'elles représentent en MAPE sont acceptables, de l'ordre de 7%.

Regardons maintenant les mêmes indices mais calculés sans le mois de janvier (Tableau 2) :

Tableau 2. Indices de performance des modèles sans le mois de janvier.

Modèles	**MBE (kWh)**	**MAE (kWh)**	**MAPE (%)**
20 arbres 10 feuilles	1304	5450	9,25
20 arbres 05 feuilles	-55	2848	4,69
40 arbres 05 feuilles	937	3063	5,03
60 arbres 05 feuilles	1127	3134	5,07

Nous y voyons exactement les mêmes variations que précédemment mais les valeurs sont plus basses. Avec des MAPE de l'ordre de 5%, nous sommes plutôt satisfaits des prédictions faites. Notons que sans le mois de janvier, le biais du modèle 2 est négatif et très faible. Le modèle 2 est d'ailleurs le meilleur modèle parmi ces quatre.

De ces valeurs, nous retenons qu'elles sont très élevées (plusieurs MWh et de 6 à 11%), et qu'elles sont certainement dues à l'horizon de prédiction, considéré comme long terme puisqu'il est de l'ordre du mois. Les valeurs de ces erreurs indiquent que le modèle de 20 arbres de 5 feuilles est le meilleur ; plus la valeur de l'erreur est faible meilleur est le modèle.

3.1.2. Boîtes à moustaches

Une étude sur la répartition des erreurs au moyen de boîtes à moustaches permet de voir l'évolution de l'erreur de prédiction en fonction des mois de l'année alors que le tableau précédent n'affiche que des moyennes. Avec les « boîtes à moustaches », connues aussi sous les noms de graphique « zones et valeurs », nous allons observer statistiquement comment sont distribuées les prédictions [19][20].

Une « boîte à moustache », ou « *box plot* » en anglais, montre la répartition des données au sein de quartiles, en mettant en valeur la moyenne et les valeurs hors norme. Il décrit les informations suivantes : la plus petite observation (valeur minimum), le quartile 1 (25%), la médiane (50%), le quartile 3 (75%) et la plus grande observation (valeur maximum). Des lignes appelées moustaches peuvent s'étendre verticalement à partir des zones. Ces lignes indiquent la variabilité en dehors des quartiles inférieurs et supérieurs. Les points situés à l'extérieur de ces lignes ou moustaches sont considérés comme des valeurs hors norme.

Regardons d'abord sur la fenêtre du haut de la Figure 5 comment sont distribuées les erreurs de prédiction de 2019 par modèle ; puis passons à la distribution des erreurs en nous intéressant aux valeurs mensuelles visibles sur la deuxième fenêtre.

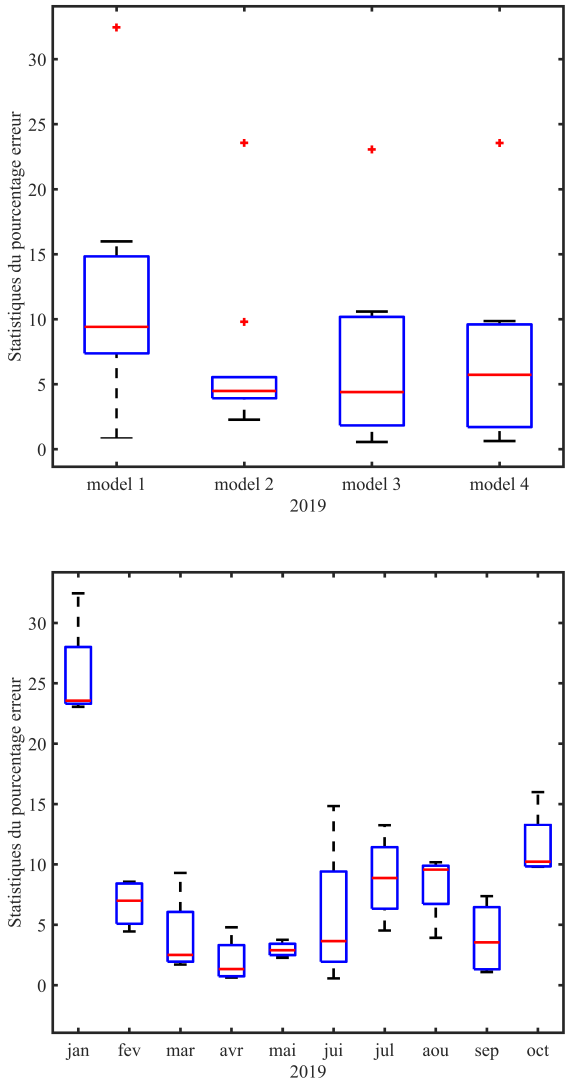

Fig. 5. Répartition statistique de l'erreur de prédiction par modèle (en haut) puis par mois (en bas) sur l'année 2019.

De cette figure, nous pouvons dire que le modèle 2 est celui dont la répartition est la moins étendue ; c'est-à-dire qu'il a tendance à mieux « cibler » la valeur réelle. Les trois autres modèles ont une répartition à peu près égale. Notons que chaque modèle a une erreur de prédiction jugée hors norme (points rouge en dehors du rectangle de répartition) et que le modèle 2 en a même une deuxième à 10%. Ce qui conforte notre déduction que le modèle 2 est le meilleur. Remarquons que la moyenne des erreurs des modèles (barre rouge à l'intérieur du rectangle de répartition) est sensiblement égale à 5% pour les modèles à cinq feuilles et proche de 10% pour le modèle à 10 feuilles. Nous retombons sur les valeurs MAPE calculées sans le mois de janvier, affichées dans le Tableau 2 puisque les valeurs hors norme ne sont pas comprises dans la répartition.

Les boîtes à moustaches de la fenêtre du bas nous permettent de savoir comment sont réparties les erreurs de prédiction par mois tous modèles confondus. Sans surprise d'après les gra-

phiques obtenus précédemment, le mois de janvier est celui que les modèles ont le plus de mal à prédire avec un taux d'erreur proche de 24%. Ensuite, les autres taux d'erreur varient entre 0 et 10%. On pourrait placer les mois dans deux classes de taux d'erreur : la moyenne haute et la moyenne basse.
Dans la moyenne haute, qui indique de moins bons résultats de prédiction, nous plaçons les mois de février, juillet, août et octobre parce qu'ils sont au-dessus des 5%. Notons qu'à l'intérieur de cette classe, le mois de février est bien meilleur que les autres puisque sa moyenne est d'environ 7% alors que les trois autres mois sont très proche des 10%.
Dans la moyenne basse, nous retrouvons les mois de mars, avril, mai, juin et septembre. Leur taux d'erreur moyen est inférieur à 5% et le plus bas est pour le mois d'avril avec 1,5%. Le mois de mai est le mieux ciblé par les modèles alors que le mois de juin est le plus étendu. Il suffit de regarder à nouveau le graphique des résidus de la Figure 4 pour voir que c'est le modèle 1 qui « vise » le plus mal et étend les quartiles, surtout aux mois de mars et juin.

3.1.3. Performance des prédicteurs

On mesure l'importance pour améliorer un modèle parce que certaines variables explicatives peuvent être non informatives et le dégrader. De plus, elle peut aussi permettre d'obtenir un modèle performant avec peu de variables explicatives mais grandement informative. L'importance est calculée par la permutation des observations des prédicteurs « out-of-bag (OOB) » [-]. On désigne par OOB_k l'échantillon Out Of Bag associé au kème arbre de la forêt. L'échantillon OOB d'un arbre est formé par les observations qui ne figurent pas dans le kème échantillon bootstrap. Il est important de comprendre que chaque modèle est unique puisque chaque forêt l'est tout autant. En conséquence, l'importance des prédicteurs est dépendante de la forêt ; mais en faisant cette étude pour plusieurs forêt nous parvenons à présenter une tendance d'importance des prédicteurs.

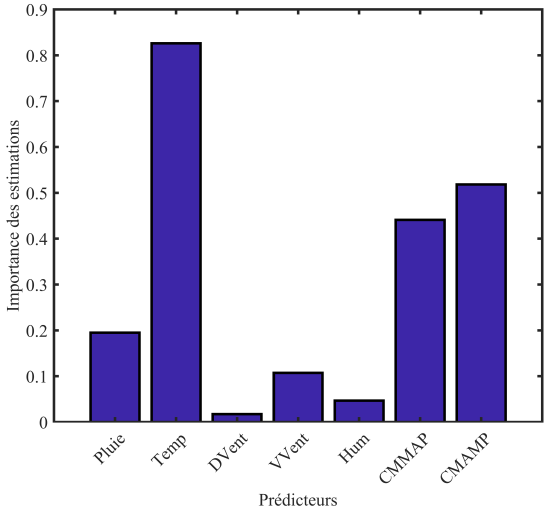

Fig. 6. Histogramme présentant l'importance des prédicteurs.

De la Figure 6, nous constatons en premier que la température est la variable explicative la plus importante pour chacun des modèles, suivie par les consommations passées : Consommation Même Mois Année Précédente (CMMAP) et Consommation Même Année Mois Précédent (CMAMP). Ensuite, les autres variables n'ont pas une importance suffisamment significative pour être analysées.

3.2. Modèles horaires

De la même manière que le modèle mensuel, nous avons testé quatre configurations et retenu un modèle pour chacune d'elles. Nous injectons dans ces modèles les données des prédicteurs issues du jeu de test (août - novembre 2019) et obtenons des prédictions de consommation horaire. Ces prédictions sont comparées graphiquement sur la fenêtre du haut de la Figure 7 aux données réelles de consommation (en bleu), et les résidus des modèles sont tracés sur la fenêtre du bas. Les quatre modèles sont :

— 1 : une forêt aléatoire de 20 arbres 10 feuilles, en rouge ;
— 2 : une forêt aléatoire de 20 arbres 05 feuilles, en orange ;
— 3 : une forêt aléatoire de 40 arbres 05 feuilles, en violet ;
— 4 : une forêt aléatoire de 60 arbres 05 feuilles, en vert.

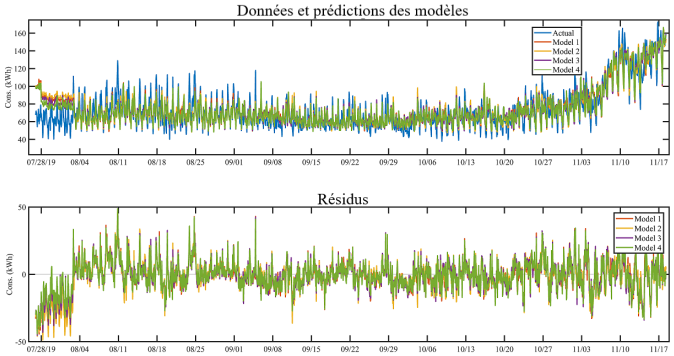

Fig. 7. Données réelles et prédites, plus résidus : différence entre données réelles et prédites. Modèles : 1. 20 arbres 10 feuilles, 2. 20 arbres 5 feuilles, 3. 40 arbres 5 feuilles et 4. 60 arbres 5 feuilles.

Sur le graphique du haut, on voit que les modèles sont proches et qu'il est difficile de juger quel est le meilleur. Si on regarde le graphique des résidus, en bas de la Figure 7, on voit que les écarts entre les valeurs des modèles et les valeurs réelles ne sont pas « trop » importants puisqu'ils se situent majoritairement entre -50 kWh et +50 kWh.
Deux grands changements de la part des modèles sont nettement observables dès le début de la phase de test, la Figure 8 est un zoom sur cette partie.

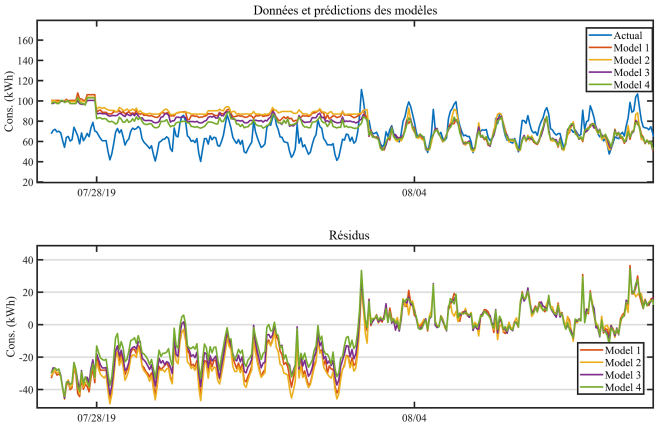

Fig. 8. Zoom sur le départ des données réelles et prédites, plus résidus.

Nous distinguons en premier une variation des modèles qui évolue autour de 100 kWh et qui est très éloignée des données réelles ; ensuite un changement améliore le comportement des modèles qui varient autour des 80 kWh ; un dernier changement recalibre les modèles pour que ceux-ci varient autour des bonnes valeurs. Le premier changement a lieu 24 h après le début de la

comparaison et le deuxième se produit au bout de sept jours. En relevant ces indices temporels, nous établissons le lien avec les prédicteurs que nous avons créés à partir des consommations passées : Consommations du Jour Précédent à la Même Heure (CJPMH) et Consommation Semaine Précédente Même Heure (CSPMH).
Ainsi, nous sommes certains que ces prédicteurs sont très importants puisqu'ils permettent de « corriger » significativement les modèles. De plus, cela nous confirme nos explications quant aux écarts des valeurs de prédiction du mois de janvier lors de la création des modèles mensuels.

3.2.1. Validation des modèles

La performance des modèles est encore estimée à l'aide des indices Mean Bias Error (MBE), Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), et est présentée dans le Tableau 3.

Tableau 3. Performance des modèles horaires.

Modèles	**MBE (kWh)**	**MAE (kWh)**	**MAPE (%)**
20 arbres 10 feuilles	-1,36	8,36	12
20 arbres 05 feuilles	-1,21	8,51	12,15
40 arbres 05 feuilles	-0,92	8,24	11,73
60 arbres 05 feuilles	-0,74	7,99	11,30

Étudions dans un premier temps les conséquences de la réduction du nombre de feuilles de 10 à 5, soient les deux premières lignes du tableau 3 : nous voyons que réduire le nombre de feuilles réduit le MBE mais augmente les erreurs.
Ensuite nous étudions les lignes 2 à 4. Du premier indice, MBE, nous déduisons que les modèles ont tendance à légèrement surestimer la valeur de consommation puisque l'indice est toujours négatif. Ceci est difficile à confirmer sur la Figure 7 parce que le grand nombre de points sur le graphique des résidus ne permet pas de trouver une tendance majoritaire à l'œil nu.
D'après nos essais affichés dans le Tableau 3, si nous généralisons, nous pouvons dire que les erreurs diminuent quand le nombre d'arbres augmente. C'est ce que nous pouvions voir sur la Figure 3. Ces baisses d'erreur ne sont toutefois pas très conséquentes pour 20 arbres de plus dans la forêt aléatoire.
Contrairement aux modèles mensuels, les valeurs d'écart (MBE et MAE) sont très faibles en kWh mais les valeurs qu'elles représentent en MAPE sont élevées, de l'ordre de 12%. Nous considérons ces modèles comme satisfaisants mais désirons réduire le taux d'erreur MAPE autour de 5%.

3.2.2. Boîtes à moustaches

Dans cette sous-section, nous regardons comment sont distribuées les erreurs de prédiction par jours de la semaine et par modèle ; puis par heures et par modèle. Et nous constatons immédiatement en les observant que toutes les répartitions sont identiques ; le comportement de chaque modèle est semblable aux autres. Ainsi, un seul graphique est présenté pour chaque étude.

Jours de la semaine Sur la Fig. 9 nous voyons graphiquement que la répartition des erreurs, taille des boîtes, est plutôt homogène entre les différents jours. En effet, les répartitions se situent toutes entre 3 et 17 %. Concernant les moyennes des erreurs, elles semblent visuellement très proches. À partir du Tableau 4 qui en dresse les nombres exacts, nous pouvons dire que le jour le moins bien prédit par les modèles est le lundi pour tous ; et que le jour le mieux prédit est le jeudi pour les modèles de 40 arbres et plus mais que globalement le samedi est le jour le moins erroné. La dernière colonne du tableau nous indique que le modèle de 40 arbres se classe bon dernier avec une moyenne

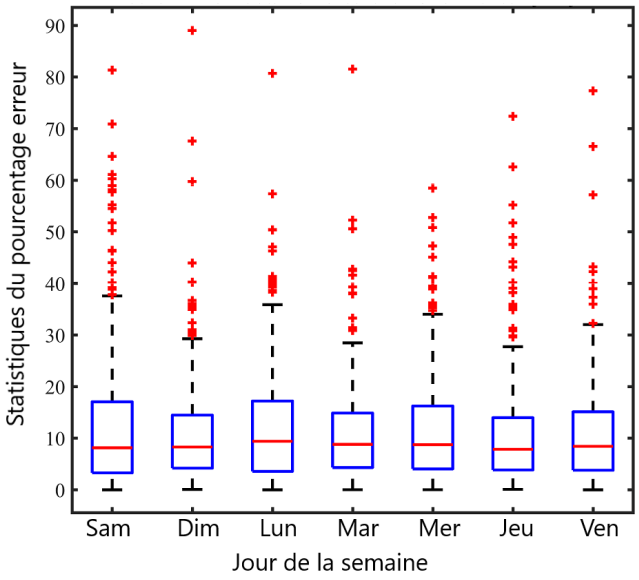

Fig. 9. Répartition statistique de l'erreur de prédiction par jours de la semaine pour le modèle 60 arbres 05 feuilles.

plus élevée que les autres. Les trois modèles restant se valent et sont difficiles à départager vu le faible écart entre leur moyenne. La Figure 10 affiche graphiquement les erreurs de prédiction que contient le tableau.

Tableau 4. Valeurs moyennes des erreurs de prédiction (en %) par jour et par modèle.

Modèles	**Sam.**	**Dim.**	**Lun.**	**Mar.**	**Mer.**	**Jeu.**	**Ven.**	**Moy.**
20a10f	8,1	8,4	9,2	8,3	8,9	8,6	8,3	8,54
20a05f	8,0	8,2	9,7	8,6	8,8	8,3	8,3	8,55
40a05f	8,0	8,5	9,6	9,4	9,4	7,7	9,0	8,80
60a05f	8,1	8,3	9,4	8,8	8,7	7,9	8,4	8,51
Moyenne	8,05	8,35	9,47	8,77	8,95	8,12	8,5	8,6

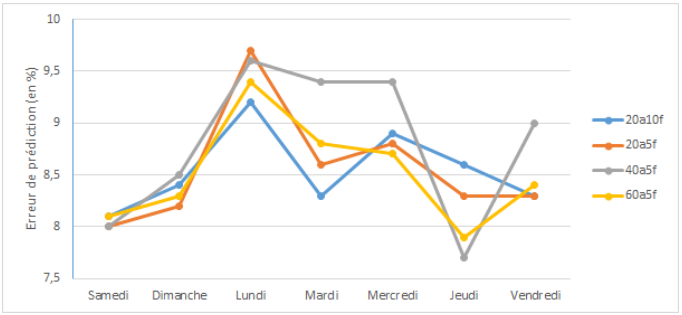

Fig. 10. Évolution des erreurs de prédiction par jour et par modèle.

Et de cette Figure, nous observons les remarques faites juste précédemment : les valeurs maximales et minimales des erreurs ont effectivement lieu les lundi et jeudi, alors que le taux d'erreur le plus bas pour tous les modèles est bien le samedi. Cette Figure nous permet également de distinguer des tendances dans l'évolution des erreurs des modèles. Ainsi, nous observons une décroissance du lundi au samedi et une détérioration des prédictions du dimanche au lundi. Nous confirmons que le modèle 40 arbres est le moins bon. Nous pensons que le comportement de

la charge les deux jours de week-end est bien différent du comportement des cinq jours de la semaine et que pour cette raison la journée du lundi a du mal à être prédite.

Heures du jour Tout comme les boîtes à moustaches des journées, celles des heures sont similaires d'un modèle à l'autre, alors un seul graphique est présenté, voir la Figure 11.

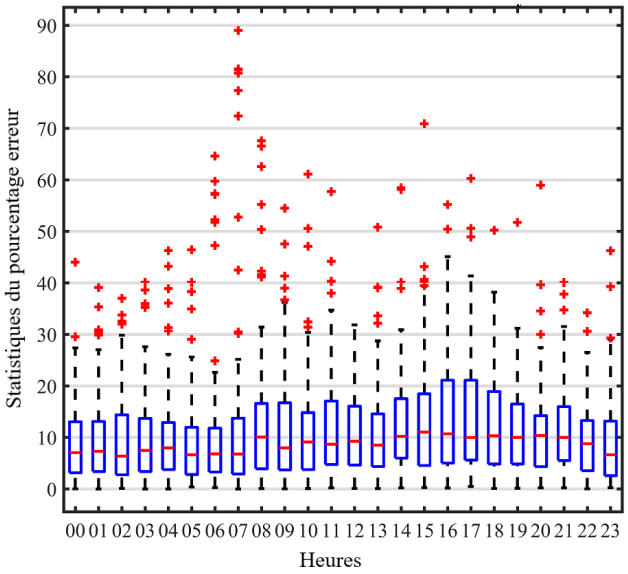

Fig. 11. Répartition statistique de l'erreur de prédiction par heures de la journée pour le modèle 60 arbres 05 feuilles.

Les points aberrants sont éparpillés de la même manière aux mêmes heures pour tous les modèles (6-8h et 14-17h) et la taille des boîtes varie de manière commune, les boîtes sont plus étendues à partir de 8 h et le sont encore plus entre 15h et 19h.
Si nous regardons uniquement les valeurs moyennes des erreurs de prédiction, barre rouge à l'intérieur des boîtes, nous nous apercevons que les modèles ont là aussi le même comportement global. La moyenne des erreurs est proche de 10 % entre 8 et 21h voire 22h ; sinon elle avoisine les 6,5 %. Ce qui signifie que les modèles commettent moins d'erreur de prédiction durant les heures de nuit que de jour. Ce qui semble logique puisque la consommation électrique dépend de la présence des bateaux à quai ; et que cette présence fluctue en fonction des entrées sorties des plaisanciers, lesquels naviguent évidemment moins la nuit que le jour.

3.2.3. Performance des prédicteurs

Comme précédemment, les histogrammes d'importance des prédicteurs pour nos quatre modèles sont semblables et un seul est présenté via la Fig. 12 :
Nous constatons encore une fois que la température et les consommations antérieures sont de forte influence sur la qualité des prédictions. Quelques fois la température est même dépassée par les consommations : Consommation Jour Précédent Même Heure (CJPMH) et Consommation Semaine Précédente Même Heure (CSPMH). La variable d'humidité étant absente des données fournies, ce sont les variables relatives au vent qui se classent juste après.

4. CONCLUSIONS ET PERSPECTIVES

Nous sommes satisfaits des performances de nos modèles long et court terme. Les modèles de prédiction court terme, de l'ordre de l'heure sont utilisés pour l'aide à la stabilité du réseau électrique de la marina. Ils permettent de mieux répartir

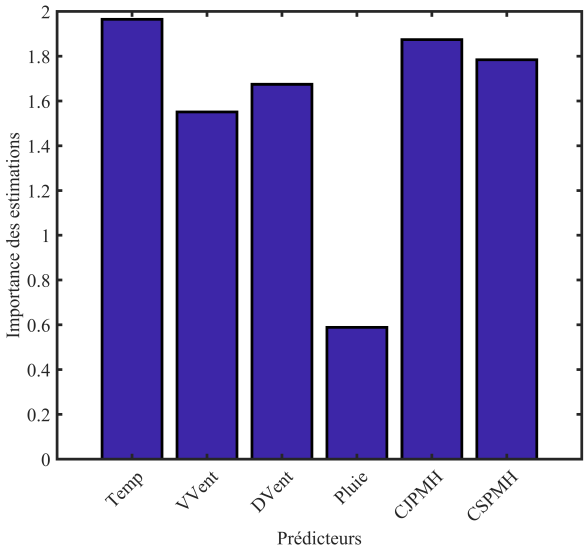

Fig. 12. Histogramme présentant l'importance des prédicteurs pour les modèles horaires.

la charge électrique globale sur les trois transformateurs basse tension en changeant le nombre de bateaux à quai de la zone affectée au transformateur. Les modèles mensuels permettent en premier lieu d'anticiper la facturation du fournisseur d'énergie électrique. Mais ils font surtout parti d'une perspective : ils sont réalisés pour être utilisés avec des modèles de prédiction de production d'électricité à partir de sources d'énergie renouvelable. Prédire la consommation mensuelle permet de mieux connaître ses besoins et dimensionner un champs de panneaux photovoltaïques qui devrait bientôt être installé ; cela permet d'établir la cohérence consommation/production et d'étudier la possibilité du mode d'autoconsomamtion pour la marina.

D'après notre travail, les meilleurs modèles de prédiction pour la marina sont ceux de 20 arbres 5 feuilles et 60 arbres 5 feuilles. Nous pensons que pour les départager il faut regarder du côté de la complexité et de la vitesse de calcul des valeurs estimées.

Nous validons l'utilisation des méthodes CART et forêts aléatoires pour prédire la consommation d'électricité de la marina. Ces modèles statistiques paramétriques sont plus simples à mettre en place que des modèles mathématiques qui prennent en compte des caractéristiques physiques du réseau électrique.

5. REMERCIEMENTS

Nous remercions l'Union Européenne et spécialement le Fond de Développement Régional et Européen qui finance en partie ce projet avec la région Occitanie ainsi que nos partenaires sur le projet SGMC : le Groupe HBF, la société Wattlet, et la régie de Port Camargue. Nous remercions également MétéoFrance pour les données météorologiques qui nous ont été fournies.

6. RÉFÉRENCES

[1] Climate Change Act 2008, Chapter 27.

[2] Clean Energy For All Europeans, European Commission, Brussels, 2016.

[3] The EU climate and energy package. Brussels, 2010.

[4] Capros, Pantelis, et al. Analysis of the Eu Policy Package On Climate Change and Renewables. Energy policy, v. 39,.3 pp. 1476-1485.

[5] Jean-Claude Sabonnadière, Gestion de l'énergie et efficacité énergétique, Éd. Lavoisier, 2007. Séries Nouvelles technologies de l'énergie.

[6] Anderson, Roger & Ghafurian, Reza & Gharavi, Hamid. Smart Grid The Future of the Electric Energy System, 2018.

[7] Shabanzadeh, Morteza & Moghaddam, Mohsen. What is the Smart Grid ? Definitions, Perspectives, and Ultimate Goals. Power System Conference, 2013.

[8] Hatziargyriou, Nikos & Asano, Hiroshi & Iravani, Reza & Marnay, Chris. (2007). Microgrids. Power and Energy Magazine, IEEE. 5. 78 - 94.

[9] Guerassimoff, G., Microgrids : pourquoi, pour qui ?, 2017.

[10] Lopes, João Abel Peças, Nikos D. Hatziargyriou, Joseph Mutale, Predrag Djapic and Nicholas Jenkins. Integrating distributed generation into electric power systems : A review of drivers, challenges and opportunities, 2007.

[11] Mashhour, Elaheh & Moghaddas-Tafreshi, S.M. A review on operation of micro grids and Virtual Power Plants in the power markets. ICAST 2009 - 2nd International Conference on Adaptive Science and Technology. 273 - 277.

[12] Mehrizi-Sani, Ali, Amir H. Etemadi, Claudio A. Cañizares, Reza Iravani, Mehrdad Kazerani, Amir H. Hajimiragha, Oriol Gomis-Bellmunt, Maryam Saeedifard, Rodrigo Palma-Behnke, Guillermo Jimenez-Estevez and Nikos D. Hatziargyriou. Trends in Microgrid Control IEEE-PES Task Force on Microgrid Control, 2014.

[13] Pierre-André Cornillon, Nicolas Hengartner, Vincent Lefieux, Eric Matzner-Lober. Prévision de la consommation d'électricité par correction itérative du biais. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009.

[14] Benjamin Goehry. Prévision multi-échelle par agrégation de forêts aléatoires. Application à la consommation électrique. Méthodologie. Université Paris-Saclay, 2019.

[15] L. Breiman,« Classification and regression trees », 2017.

[16] L. Breiman,« Random forests », Machine Learning,vol. 45, n1, p.5-32, 2001.

[17] Liaw, Andy & Wiener, Matthew. (2001). Classification and Regression by RandomForest.

[18] Robin Genuer, Jean-Michel Poggi. Arbres CART et Forêts aléatoires, Importance et sélection de variables, 2017.

[19] Williamson DF, Parker RA, Kendrick JS. The Box Plot : A Simple Visual Method to Interpret Data. Ann Intern Med. 1989 ;110 :916–921.

[20] Michael Frigge, David C. Hoaglin & Boris Iglewicz (1989) Some Implementations of the Boxplot, The American Statistician, 43 :1, 50-54.

[21] Vincent Lefieux. Modèles semi-paramétriques appliqués à la prévision des séries temporelles. Cas de la consommation d'électricité.. Mathématiques [math]. Université Rennes 2, 2007.

[22] Anestis Antoniadis, Xavier Brosat, Jairo Cugliari, Jean-Michel Poggi. Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. Journal de la Société Française de Statistique, Société Française de Statistique et Société Mathématique de France, 2014, 155 (2), pp.202- 219.