Analyse des courbes de charge d'électricité et prédiction à court terme dans les secteurs résidentiel et tertiaire

Fatima Fahs

Thèse

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE

UMR 7501

Strasbourg

présentée pour obtenir le grade de docteur de l'Université de Strasbourg
Spécialité MATHÉMATIQUES APPLIQUÉES

Fatima Fahs

Analyse des courbes de charge d'électricité et prédiction à court terme dans les secteurs résidentiel et tertiaire

Soutenue le 19 juin 2023
devant la commission d'examen

M. Etienne Birmele, président
M. Jairo Cugliari, rapporteur
Mme Ndeye Niang, rapporteuse
Mme Mitra Fouladirad, examinatrice
Mme Sophie Dabo, examinatrice
M. Frédéric Bertrand, directeur de thèse
Mme Myriam Maumy, directrice de thèse
M. Nicolas Lachiche, directeur de thèse

irma.math.unistra.fr

IRMA
Institut de Recherche Mathématique Avancée

cnrs

Université de Strasbourg

Remerciements

Je tiens tout d'abord à remercier et à exprimer ma gratitude à mes deux directeurs de thèse M. Frédéric Bertrand et Mme Myriam Maumy-Bertrand qui m'ont guidée tout au long de ce projet de recherche. Merci pour cette belle opportunité que vous m'avez offerte et qui m'a permis de développer mes compétences en tant que chercheuse tout en me confrontant aux problématiques du monde industrielle. Ce travail n'aurait pu être mené à bien sans votre aide pertinente et votre soutien.

Un grand merci également à M. Nicolas Lachiche, mon codirecteur de thèse qui a accepté de co-diriger mon travail du côté de l'école doctorale MSII de Strasbourg. Je souhaite remercier également M. Philippe Helluy le directeur de l'IRMA.

Je souhaiterais exprimer ma profonde gratitude envers M. Christophe Prud'homme, mon professeur de master, qui a été l'un des principaux moteurs de mon parcours académique. Grâce à lui j'ai eu la chance de bénéficier d'une bourse d'études pour mon master et d'accéder à la possibilité de poursuivre en thèse. Je suis infiniment reconnaissante pour les opportunités qu'il m'a offertes et pour l'impact positif qu'il a eu sur ma vie professionnelle.

Sincères remerciements à l'ensemble des membres du jury. Vous me faites un immense honneur d'accepter d'être de le jury de la soutenance de cette thèse. Je suis extrêmement reconnaissante envers Mme Ndeye Niang et M. Jairo Cugliari pour avoir accepté d'être les rapporteurs de ce manuscrit. Je remercie également Mme Mitra Fouladirad et Mme Sophie Dabo-Niang les examinatrices de ce travail. Je voudrais également exprimer mes sincères remerciements envers M. Étienne Birmelé, qui a accepté de présider le jury de ma soutenance de thèse.

Merci, à Mme Nathalie Kostmann la secrétaire de l'école doctorale MSII, pour toutes les facilitations administratives. Je remercie également Mme Céline Caldini-Queiros qui m'a aidé à monter le dossier CIFRE de la thèse.

Cette thèse n'aurait pas été possible sans le financement de l'ANRT (Association nationale de la recherche et de la technologie) et le fournisseur d'électricité.

Je remercie chaleureusement mes parents et mes frères qui m'ont toujours soutenue

dans les moments de découragement et m'ont accompagné tout au long de mes études supérieures.

Je remercie mon cher époux pour son soutien quotidien et sa croyance en ma réussite. Je lui exprime toute ma gratitude et mon amour. À ma fille qui m'a donné de l'énergie pour avancer, et qui a supporté mes absences durant ces dernières années.

Finalement, un grand merci à toutes les personnes qui ont de près ou de loin contribué à cette thèse. Je leur suis infiniment reconnaissante.

Résumé

Le déploiement massif des compteurs intelligents dans le secteur résidentiel et tertiaire a permis de récolter des données de consommation électrique de haute fréquence à l'échelle des consommateurs (particuliers, professionnels, ...). Ces données constituent une matière première pour les recherches sur la prévision de la consommation de l'électricité à cette échelle. La majorité de ces recherches visent largement à répondre aux besoins du milieu industriel comme les applications dans le contexte des maisons intelligentes ainsi que les programmes de la gestion et de la réduction de la consommation. L'objectif de ce travail est de déployer ou mettre en œuvre des modèles de prévision de la charge électrique à court terme $(J + 1)$ à l'échelle des consommateurs qui seront intégrés dans des applications à des fins industrielles. La complexité du sujet réside dans le fait que les données de consommation à cette échelle sont très volatiles. En effet, elles comprennent une grande quantité de bruit et dépendent du mode de vie du consommateur et de ses habitudes de consommation. De plus, le déploiement de ces modèles dans des applications industrielles ajoute une certaine complexité supplémentaire au sujet. Durant cette thèse, plusieurs modèles ont été déployés pour la prévision de charge des ménages en tenant compte de différentes contraintes industrielles. Les modèles ont été testés et évalués sur un grand échantillon de courbes de charges disparates du secteur résidentiel et tertiaire. Une approche a été proposée également pour la prévision des courbes de charge les plus volatiles. Ensuite, nous avons étudié l'impact de l'utilisation des prévisions à l'échelle des ménages sur la prévision de la charge agrégée. Les modèles testés ont été également adaptés pour répondre aux besoins en prévision de l'entreprise pour d'autres portefeuilles.

Mots clés : Prévision de la charge électrique, court terme, secteur résidentiel, secteur tertiaire, approche industrielle, échelle du consommateur, courbes de charge volatiles, prévision probabiliste, charge agrégée.

Abstract

The widespread deployment of smart meters in the residential and tertiary sectors has enabled the collection of high-frequency electricity consumption data at the consumer level (individuals, businesses, ...). This data constitutes a raw material for research on predicting electricity consumption at this scale. The majority of this research is largely focused on meeting the needs of the industrial sector, such as applications in the context of smart homes and programs for managing and reducing consumption. The goal of this work is to deploy or implement short-term (D + 1) electricity load forecasting models at the consumer level that will be integrated into industrial applications. The complexity of the subject lies in the fact that consumption data at this scale is very volatile, including a high level of noise and depending on the consumer's lifestyle and consumption habits. Additionally, the deployment of these models in industrial applications adds further complexity to the subject. In this thesis, several models were deployed for forecasting household load, taking into account various industrial constraints. The models were tested and evaluated on a large sample of diverse residential and tertiary sector load curves. An approach was also proposed for forecasting the most volatile load curves. Then, we studied the impact of using household-level predictions on the prediction of aggregated load. The tested models were also adapted to meet the needs of the company's forecasting for other portfolios.

Keywords : Electric load forecasting, short term, residential sector, tertiary sector, industrial approach, consumer scale, volatile load curves, probabilistic forecast, aggregate load.

Table des matières

Liste des figures

Liste des tableaux

Listes des acronymes et anglicismes

ADEME	Agence de l'Environnement et de la Maîtrise de l'Énergie
ANN	*Artificiel Neural Network*
ARIMA	*AutoRegressive Integrated Moving Average*
CART	*Classification And Regression Trees*
CDD	*Cooling Degree Days*
CNN	*Convolutional Neural Network*
CRE	Commission de Régulation de l'Énergie
DNN	*Deep Neural Network*
DRNN	*Deep recurrent neural network*
DSHW	*Double seasonal Holt-Winters*
DSR	*Demand Side Response*
ELF	*Electric Load Forecasting*
GAM	*Generalized Additif Model*
HC/HP	Tarif Heures Creuses/Heures Pleines
HDD	*Heating Degree Days*
KWF	*Functional Wavelet Kernel*
kWh	kilowatt-heure
LS-SVM	*Least-squares support vector machine*
LSTM	*Long Short Term Memory*
LTLF	*Long Term Load Forecasting*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MASE	*Mean Absolute Scaled Error*
MTLF	*Medium Term Load Forecasting*
MW	Mégawatt
NMAE	*Normalized Mean Absolute Error*
NRMSE	*Normalized Root Mean Square Error*
PDL	Point De Livraison
RF	*Random Forest*
RMSE	*Root Mean Square Error*
RNN-LSTM	Réseau de neurones récurrent à mémoire court et long terme
RTE	Réseau de transport d'électricité français
sMAPE	*Symmetric Mean Absolute Percentage Error*
SSM	*State Space Model*

STLF	*Short Term Load Forecasting*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
SWT	*Stationary Wavelet Transform*
W	Watt, une unité de puissance

Chapitre 1

Introduction générale

Le travail présenté dans ce manuscrit a été réalisé dans le cadre d'une thèse CIFRE chez un fournisseur d'énergie et à l'Institut de Recherche Mathématique Avancée (IRMA) sous la direction de M. Frédéric Bertrand, Mme Myriam Maumy-Bertrand et M. Nicolas Lachiche. Il a été financé par l'Association Nationale de la Recherche et de la Technologie (ANRT) et le fournisseur d'énergie.

L'objectif principal de la thèse est d'adapter et de mettre en œuvre des modèles statistiques et/ou des méthodes d'apprentissage automatique qui permettraient de prédire client par client la consommation électrique du lendemain (à court terme $(J+1)$[1]) à partir des historiques de consommation de chaque client, des données météorologiques et d'autres variables affectant la consommation électrique dans les secteurs résidentiel et tertiaire (comme les jours de la semaine, les heures de la journée, les jours fériés, ...). Ces prévisions permettent de procurer à chaque client une prévision à court terme $(J+1)$ de sa propre consommation, de détecter et de lui communiquer d'éventuelles anomalies tels que des écarts anormaux de la consommation. Cette approche prédictive doit prendre en considération les différents défis et contraintes du monde industriel. Ensuite, l'objectif a évolué pour répondre aux besoins de l'entreprise en terme de prévision de la charge électrique[2] pour d'autres portefeuilles et pour d'autres types d'applications (comme la prévision pour la gestion de la maintenance, la compensation des pertes, ...).

1. La prévision à $J+1$ désigne la prévision de la veille pour le lendemain.
2. Dans ce manuscrit la charge électrique est équivalente à la consommation et à la demande d'électricité.

1.1 Contexte général : une nouvelle ère pour les clients d'électricité

La libéralisation du marché de l'électricité en France ces vingt dernières années a transformé le paysage traditionnel du secteur d'électricité en mettant fin à la situation du monopole public. Concrètement, cela signifie que le marché d'électricité français s'est ouvert à la concurrence.

Le fournisseur d'électricité, EDF, qui détenait le monopole du marché d'électricité de 1946 à 2004 depuis la production jusqu'à la fourniture, partage désormais le marché avec d'autres fournisseurs, appelés « fournisseurs d'énergie alternatifs ». Depuis 2017, les clients particuliers comme les entreprises bénéficient du choix de leur fournisseur et par suite l'offre et la tarification qui conviennent le mieux à leurs besoins (MaPetiteEnergie, 2021). Dans ce contexte de concurrence libre et ouverte, les entreprises d'électricité misent de plus en plus sur la satisfaction du client pour se différencier. Certes, le prix est une source de satisfaction particulièrement intense mais loin d'être unique.

En effet, le déploiement des nouvelles technologies notamment le numérique et la transformation digitale ont complètement bouleversé le marché d'électricité et continuent de participer à sa transformation (voir la figure 1.1). La révolution digitale a également modifié les attentes des clients ainsi que leurs habitudes de consommation. Les clients d'aujourd'hui s'attendent de la part de leur fournisseur à ce qu'il leur offre des services sur mesure entièrement adaptés à leurs besoins suivant la tendance dans d'autres marchés de services (services bancaires, e-commerce, assurances, ...), quitte à payer un peu plus cher. Ils sont mieux renseignés et attendent des offres ultra-personnalisées et facilement accessibles. Ces clients passent de consommateurs passifs à consommateurs actifs et souhaitent d'être impliqués dans la gestion de leur consommation. Ils demandent que la communication avec leur fournisseur devient bilatérale avec une plus grande capacité d'action.

La personnalisation du service client devient donc un nouveau champ de bataille concurrentiel et un pilier de la satisfaction client. Les entreprises qui investissent dans la personnalisation de leurs services consolideront leur position, fidéliseront mieux leurs clients et amélioreront l'acquisition client, tandis que celles qui ne le font pas seront plus vulnérables aux perturbations du marché. Il est donc temps de commencer à mettre en œuvre des techniques de personnalisation intelligentes.

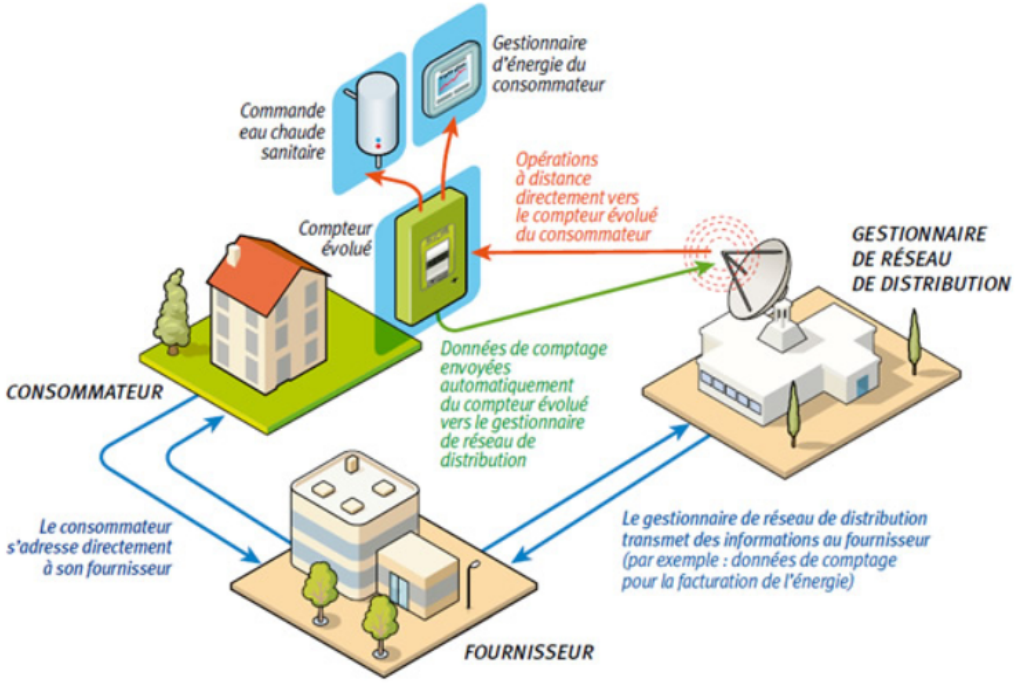


FIGURE 1.1 – La relation entre le consommateur, le fournisseur et le gestionnaire de réseau par l'entremise des compteurs intelligents [Source CRE].

Les exigences des clients se sont renforcées et ont considérablement augmenté par le déploiement massif des compteurs intelligents. En mars 2018, neuf millions de compteurs Linky [3] étaient déjà installés en France (SELECTRA, 2021) par Enedis [4]. Quotidiennement, 30 mille compteurs ont été installés dans l'objectif d'atteindre 35 millions compteurs intelligents résidentiels déployés à travers la France en 2020 (SELECTRA, 2021). En parallèle, l'Union européenne avait prévu qu'en 2020 près de 72% des consommateurs européens seront équipés de compteurs intelligents (SELECTRA, 2021). Il est important de noter qu'aucune information n'est disponible sur l'achèvement de ces objectifs.

Le **compteur intelligent** ou *smart meter* est un compteur numérique qui mesure, stocke et transfert de manière précise et en temps réel les données de consommation d'électricité à haute fréquence. Le compteur transmet directement, les diverses informations vers un système de sauvegarde permettant d'analyser les données recueillies. Il peut être installé sur des circuits électriques appartenant à des ménages ou à des bâtiments, et sont également utilisés pour mesurer la consommation d'énergie au niveau des appareils électriques. D'ailleurs, le compteur intelligent est doté d'une capacité de communication bidirectionnelle (transmission et réception des informations) entre le compteur et le fournisseur. Donc contrairement aux compteurs analogiques, il est capable d'agir et de réagir,

3. Linky est le nom du compteur électrique communicant développé par Enedis https://fr.wikipedia.org/wiki/Linky

4. https://fr.wikipedia.org/wiki/Enedis

sans intervention humaine directe [5] (voir la figure 1.2).

FIGURE 1.2 – Le compteur intelligent LINKY [Source Enedis].

Le déploiement massif de ces compteurs communicants [6] a constitué le premier pas vers l'intégration des fonctionnalités et des services plus évolués. Dans un premier temps, la collecte de grandes quantités de données de consommation à haute fréquence et à l'échelle d'un ménage a permis de fournir le retour d'information sur la consommation d'électricité aux clients en temps quasi réel au lieu d'une consommation cumulative tous les deux mois. Ces informations permettront au client de mieux gérer sa consommation, d'économiser de l'énergie et potentiellement d'agir sur sa facture (éco-gestes au quotidien, remplacement des équipements énergivores, l'isolation du logement, ...). Différentes études et projets suivis par l'ADEME [7] ont montré que ces économies peuvent aller jusqu'à 10% à 15% pour les clients particuliers. Grâce à ces télé-relèves, [8] les clients sont facturés sur leur consommation réelle et non pas sur une estimation de cette dernière.
Généralement, ces services sont accessibles via un espace client en ligne mais aujourd'hui certains fournisseurs d'électricité proposent à leur clientèle une application, accessible sur différents supports, permettant de suivre leur consommation et leur dépense énergétique en temps réel. Par exemple, Enedis et EDF disposent notamment de leurs propres applications mobiles (voir la figure 1.3). Les fonctionnalités les plus courantes proposées dans ces applications sont celles qui permettent d'identifier les pics de consommation d'électricité, de visualiser en euros et en kWh les consommations avec la possibilité de les comparer par périodes et de fixer des alertes sur un seuil de consommation défini par le consommateur

5. Les compteurs intelligents ne nécessitent pas l'intervention d'un technicien sur place puisque les réglages se font généralement à distance et le relevé du compteur électrique s'effectue automatiquement et non pas manuellement comme dans le cas des compteurs analogiques.
6. Les compteurs communicants sont également les compteurs intelligents
7. https://www.ademe.fr/lademe/presentation-lademe
8. Le télé-relève est un accès à distance aux informations collectées par un compteur communicant https://pro.engie.fr/faq/compteur-releve/releve/tele-releve

et des objectifs à atteindre pour pouvoir effectuer des économies sur sa facture d'électricité (MaPetiteEnergie, 2021).

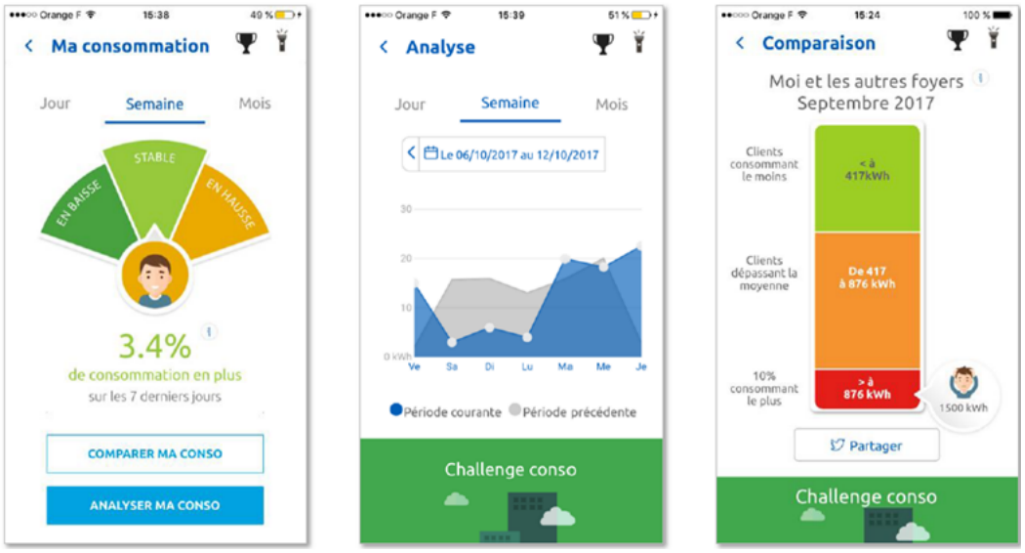


FIGURE 1.3 – Les fonctionnalités de l'application mobile de suivi de la consommation « Enedis à mes côtés » [Source Enedis] [9].

1.2 Problématique industrielle et motivations

Bien que l'arrivée des compteurs intelligents ait produit un afflux de données facilement accessibles par le client, ces données demeurent une condition nécessaire à l'émergence de services innovants mais pas suffisante. La seule communication de la consommation en temps réel n'est pas suffisante pour aider le client à comprendre et agir sur sa consommation. L'information reste assez brute et extrêmement bruitée et l'interprétation par le client n'est pas toujours aisée. Par conséquence, les bénéfices des services clients proposés sur la base des données brutes restent encore théoriques pour le consommateur en termes de maîtrise de la consommation. En plus, le client a tendance à remettre en cause les services de pilotage de la consommation d'électricité à la mode en ce moment qui se basent sur la comparaison de sa propre consommation avec des ménages équivalents pour juger son niveau de consommation.

Le défi principal pour les fournisseurs d'énergie actuellement consiste alors à transformer ces données sur les consommations d'énergie en informations plus claires et plus exploitables par l'entremise des outils d'apprentissage automatique et d'intelligence artificielle. La piste la plus prometteuse est de prédire la consommation de chaque client à partir de ses propres données de consommation et évaluer son niveau de consommation par auto-comparaison c'est à dire la comparaison de ses consommations réelles avec les

9. https://selectra.info/energie/guides/conso/

valeurs prédites par son propre modèle de prévision. Pour cela, il est nécessaire d'avoir une analyse prévisionnelle fiable et interprétable de la consommation d'électricité pour mener à bien ces objectifs et par suite de promouvoir les services conçus dans l'optique de maîtrise de la demande d'électricité.

Dans le secteur résidentiel, fournir au client une analyse prévisionnelle détaillée de sa consommation d'électricité à partir de ses propres données permet de l'aider à optimiser sa consommation et à modifier son comportement énergétique si nécessaire (pratiquer des éco-gestes, isoler sa maison, remplacer ses appareils électriques énergivores, ...) et par conséquent, de faire des économies sur ses factures. Ces prévisions sont également intéressantes pour les clients qui produisent leur électricité (à partir des panneaux solaires). Une telle information permet aux clients de gérer leur production et la vente du surplus d'électricité à leurs voisins et/ou à leurs fournisseurs. Les prévisions de charge peuvent également être utilisées pour détecter des anomalies en comparant la consommation réelle avec la charge prédite basée sur l'historique de la consommation du client. Les anomalies se définissent comme des cas particuliers, dans lesquels le consommateur a volontairement modifié ses habitudes de consommation, ou bien involontairement comme par exemple des problèmes de fuite électrique. Ces prévisions peuvent aussi aider à la prise en charge des personnes âgées et/ou malades dans leurs maisons. Par exemple, si la consommation d'activité surveillée d'une personne âgée ou d'un malade est inférieure aux prévisions, alors cette situation peut alerter les services d'aide sociale ou les équipes médicales d'une situation inhabituelle. L'envoi de tels alertes dans ces situations peut prévenir des circonstances dangereuses.

Dans le secteur tertiaire (boulangeries, hôtels, agences, ...), le client est potentiellement motivé par la détection des anomalies en cas d'absence et la surveillance en cas de dysfonctionnements des équipements professionnels. Donc une simple comparaison entre les valeurs prédites et la consommation réelle lui permettra d'évaluer son niveau de consommation (surconsommation, chute anormale des puissances,...) et par suite agir pour éviter une surconsommation coûteuse en optimisant le remplacement des machines défaillantes ou intervenir rapidement dans le cas d'une chute anormale de la puissance électrique.

Cette analyse prédictive à l'échelle du client peut être plus avantageuse pour les clients que pour les fournisseurs. Cependant, les fournisseurs cherchent par la proposition de ces services à se différencier de leurs concurrents, fidéliser leur clientèle et conquérir des nouveaux clients qui s'intéressent à ce type de services. Les fournisseurs peuvent également profiter de ces prévisions individuelles pour améliorer la prévision à des niveaux agrégés (bâtiments, quartiers, zones géographiques, ...). et par suite améliorer la gestion de leurs moyens de production (notamment pendant les périodes de pointe).

1.3 Objectif de la thèse

Cette thèse émane d'un besoin réel de la part du fournisseur d'énergie d'une analyse prévisionnelle fiable et interprétable de la charge électrique à l'échelle d'un ménage. Par conséquent, l'objectif de ce travail est de prédire la consommation d'électricité d'un ménage un jour (24 heures) à l'avance (c'est-à-dire prédire la veille pour le lendemain) à partir de son propre historique de consommation et des données météorologiques en tenant compte des besoins et des contraintes industrielles que nous détaillerons dans la section suivante.

Le modèle chargé d'effectuer cette tâche doit être capable de modéliser et prédire automatiquement des milliers de courbes de charge disparates sans intervention humaine. Les données que nous disposons pour cette étude sont des données privées d'un fournisseur d'électricité local en France.

L'objectif de la thèse a ensuite évolué avec le temps pour inclure la prévision de la charge pour d'autres portefeuilles de l'entreprise et pour des niveaux agrégés de la demande des ménages. Durant ces trois années de thèse nous avons tenté de trouver des réponses à ces questions :

1. Comment profiter du succès des modèles de prévision de la charge nationale pour mener à bien les objectifs ?
2. Dans quelle mesure les facteurs d'influence sur la charge nationale affectent-ils la charge à l'échelle des ménages ?
3. Comment introduire l'effet de la température extérieure dans les modèles de prévision de la charge des ménages ?
4. Comment évaluer les performances de ces modèles de prévision ? Que signifie un modèle performant dans un contexte de prévision à l'échelle des ménages ?
5. Comment réduire l'irrégularité dans les courbes de charge des ménages pour améliorer la performance des modèles de prévision à cette échelle ?
6. Comment quantifier les incertitudes des prévisions et les intégrer dans les prévisions ponctuelles de la charge ?
7. Comment profiter des modèles de prévision à l'échelle des ménages pour améliorer les prévisions à des niveaux agrégés ou pour d'autres portefeuilles ?

1.4 Enjeux entre les besoins et les contraintes

La charge d'électricité [10] est souvent représentée par une courbe de charge dite encore une courbe de puissance exprimée généralement en watts (W) (voir la figure 1.4). Cette

10. Dans ce manuscrit les trois expressions suivantes sont équivalentes : la charge d'électricité, la demande d'électricité et la consommation d'électricité.

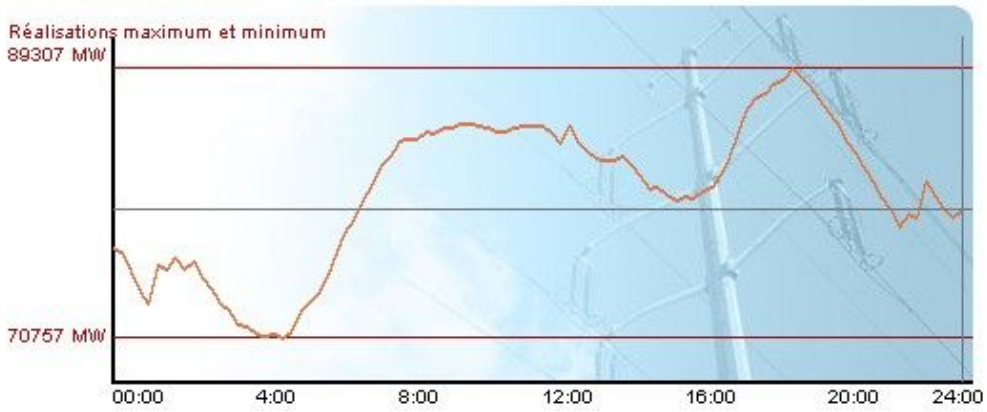


FIGURE 1.4 – Exemple d'une courbe de charge de la journée de 16/12/2009 de la demande d'électricité en France » [Source RTE].

courbe représente des relevés de la consommation d'électricité sur une période précise à des intervalles de temps réguliers (souvent toutes les 10 minutes ou 30 minutes) à différents niveaux (d'un appareil électrique, d'un ménage, d'un quartier) jusqu'à la courbe de charge nationale.

Contrairement à la charge nationale qui est caractérisée par sa régularité, la charge à l'échelle des ménages est très volatile et dynamique. La forme des courbes quotidiennes est parfois complètement différente d'un jour à un autre. Cela est principalement dû à l'irrégularité du comportement des occupants notamment le caractère aléatoire de l'usage de certains appareils tels que le chauffage ou certains appareils électroménagers. De même l'impact de la température extérieure sur la charge des ménages n'est pas claire. Par conséquent, il est plus difficile de prévoir la charge d'électricité à l'échelle d'un ménage qu'à l'échelle nationale.

À cette difficulté liée à la nature des données s'ajoute les contraintes industrielles (voir la section 3.3). En effet, la majorité des modèles de prévision de la charge des ménages dans la littérature sont conçus dans le contexte des applications industrielles et donc la solution envisagée doit également être compatible avec ces contraintes industrielles ce qui n'était pas malheureusement toujours le cas.

Le modèle de prévision de la charge à l'échelle des ménages doit être facilement adaptable pour de nombreuses typologies de ménages et pour différents profils de courbe de charge. Le modèle doit fonctionner avec précision et fiabilité à des niveaux aussi critiques comme les périodes de pointe, les jours fériés ou pendant les vacances. Le modèle doit être rapide pour effectuer des prévisions quotidiennes dans un délai raisonnable pour un nombre élevé de ménages, en se limitant à l'infrastructure de calcul existante et les ressources en mémoire. En plus, les modèles très gourmands en données ont des impacts négatifs sur leur facilité d'utilisation et leur facilité de mise à jour. En général, les modèles nécessitant moins de données sont préférables.

En plus et malgré l'impact positif de l'intégration des informations sur le mode de vie du client (nombre d'occupants dans le logement, les périodes de vacances, les appareils électriques, le rythme de vie des occupants, leurs âges, ...) dans les modèles de prévision de sa charge électrique le fournisseur d'électricité pour lequel cette étude est menée ne préfère pas inclure des données à caractère personnel sur ses clients par respect de leur vie privée, en plus le client est souvent réticent à donner ce type d'informations qui de toute façon ne reste pas pertinent dans le temps.

1.5 Plan de la thèse

Ce manuscrit est divisé en cinq chapitres :

Le premier chapitre représente une introduction générale sur le sujet. Dans ce chapitre nous abordons le contexte de la thèse, la problématique ainsi que nos objectifs.

Le deuxième chapitre est dédié à l'état de l'art. Dans ce chapitre, nous représentons les différents travaux de recherche qui ont porté sur la prévision de la charge d'électricité à court terme dans la littérature à l'échelle nationale et locale. Nous y trouvons d'abord une courte introduction sur le sujet avec une section qui décrit brièvement les trois catégories de l'horizon de prévision. Ensuite, deux sections sont dédiées à l'état de l'art sur les modèles de prévision à l'échelle nationale et locale respectivement. Enfin, dans la dernière section, un bilan des modèles de prévision de la charge électrique est établi afin de positionner les travaux de la thèse par rapport à la littérature.

Le troisième chapitre présente les fondements théoriques des différents modèles de prévision que nous avons déployés pour la prévision de la charge électrique à l'échelle des ménages à court terme. En plus, nous présentons les différentes métriques utilisées généralement dans la littérature pour mesurer la précision des modèles de prévision et celles utilisées en particulier pour la prévision de la charge à l'échelle des ménages. Finalement, nous avons présenté les critères de sélection des modèles de prévision adoptés dans notre approche.

Le quatrième chapitre présente l'ensemble des modèles que nous avons adaptés et testés pour la prévision de la charge électrique à l'échelle des ménages. La première partie est dédiée à la description et l'exploration du jeu de données qui a été mis à notre disposition par le fournisseur d'énergie. Ensuite, une étude comparative a été présentée dans la deuxième partie entre un modèle de prévision appelé *KWF* et deux modèles de référence. Le modèle *KWF* s'est révélé plus performant que les deux autres modèles pour la prévision de la charge électrique thermosensible et non-thermosensible [11]. Ensuite, un

11. Par thermosensibilité, nous désignons la sensibilité de la charge électrique à la variation des températures extérieures.

modèle qui combine le modèle *KWF* avec une méthode de *clustering* des courbes de charge journalières a été testé dans l'objectif d'améliorer la performance du modèle *KWF* notamment pour la prévision de la charge thermosensible. Les résultats obtenus montrent une amélioration globale dans la précision de la prévision de cette dernière. Dans la troisième partie du chapitre, et dans l'objectif d'améliorer la qualité de la prévision, nous avons testé trois autres modèles tels que le modèle additif généralisé (*GAM*), le modèle de régression multivariée par spline adaptative (*MARS*) et le modèle de réseau de neurones récurrents (*RNN-LSTM*) (voir le chapitre 3). La performance des modèles testés a été évaluée selon plusieurs critères notamment la thermosensibilité et la volatilité. Les résultats obtenus ont montré une grande hétérogénéité dans la prévision de la charge des ménages. Cette hétérogénéité s'explique par la différence des modes de vie des consommateurs. En effet, plus les données de consommation sont régulières plus la qualité de la prévision est satisfaisante. Dans la dernière partie de ce chapitre, nous présentons une approche de prévision des courbes de charge volatiles qui consiste à prédire l'énergie électrique consommée pendant la journée au lieu de prédire la courbe de charge en puissance. Cette approche permet de réduire l'irrégularité que nous observons dans les courbes de charge liée résultant du décalage des pointes (ou pics) de consommation dans la journée. De plus, elle permet de gagner en explicabilité par rapport à l'approche de prévision de la courbe de charge des puissances. Cette explicabilité est jugée intéressante pour aider le client à comprendre l'utilité des prévisions qui lui sont fournies et par conséquent, décider et agir en situations d'exception comme une surconsommation ou sous-consommation électrique inhabituelle.

Le cinquième chapitre est dédié aux applications autour de la prévision de la charge des ménages. Dans la première partie, nous avons étudié l'effet du *clustering* des données de consommation des ménages sur l'amélioration de la prévision de la charge agrégée de ces derniers. Ensuite, les modèles de prévision de la charge à l'échelle des ménages ont été adapté pour la prévision de la charge électrique des clients du secteur tertiaire. En raison de la régularité de cette dernière, les résultats obtenus ont été plus satisfaisants que ceux obtenus pour la prévision de la charge à l'échelle d'un ménage. Plusieurs méthodes de calcul des intervalles de prévision ont également été testées afin de quantifier les incertitudes des prévisions ponctuelles des données de consommation électrique dans le secteur résidentiel et tertiaire. Finalement, les modèles de prévision ont été adaptés et testés pour la prévision des données des pertes du réseau électrique et de la charge électrique du réseau de distribution du fournisseur d'énergie.

La dernière partie de ce manuscrit est dédiée à la conclusion et les perspectives.

Chapitre 2

État de l'art des modèles de prévision de la consommation électrique à court terme

Objectifs

Ce chapitre est une présentation générale de la littérature sur les modèles de prévision de la charge électrique à l'échelle nationale et locale. Un tableau résumant plusieurs études sur la prévision de la consommation à l'échelle des ménages est présenté également avec un bilan de la littérature servant d'une ligne directrice de nos travaux.

Sommaire

2.1 Introduction

La prévision de la charge électrique (ELF) est un sujet de recherche qui a attiré l'attention des chercheurs depuis des décennies (Jahan et al., 2020). Étant donné que l'électricité ne peut pas être stockée [1], l'équilibre production-consommation doit toujours être assuré par les gestionnaires de réseau pour éviter les pannes, la surcharge du réseau électrique ou les coupures de courant électrique. La prévision de la demande d'électricité joue un rôle crucial à cet égard. Elle est indispensable également pour d'autres besoins comme la gestion des moyens de production, la planification, la tarification, la maintenance du réseau de transport d'électricité, la réduction des pics et l'évaluation du marché, ... (Jahan et al., 2020).

Malgré l'existence d'une abondante littérature sur la prévision de la charge électrique, cette littérature ne permet pas une classification ni une comparaison facile de tous les modèles existants. Par conséquent, il existe différentes possibilités pour classer les différentes techniques. En effet, la différence dans les objectifs et les besoins de la prévision a conduit à une différence dans la manière dont les chercheurs traitent ce sujet que ce soit au niveau de la nature des données utilisées (privées ou publiques), ou à leurs échelles (nationale, différents niveaux d'agrégation spatiale et temporelle, bâtiments, foyers, ...) ainsi qu'à leurs granularités (des données demi-horaires, horaires, quotidiennes, mensuelles, ...), ou également au niveau des méthodes et techniques utilisées (modèles de prévision non intrusives, l'approche de prévision *top-down* et *bottom-up* (voir annexe du chapitre 6.2) (Capasso et al., 1994 ; Chin, 2016), ...) mais aussi au niveau de l'horizon temporel de la prévision (court, moyen et long terme qui seront définis dans la sous-section 2.2).

Dans ce chapitre nous présentons les différentes méthodes de prévision qui ont été utilisées dans la littérature pour la prévision de la charge électrique à l'échelle nationale ou régionale ainsi qu'à l'échelle locale (région, quartier, une immeuble ou un ménage) à court terme. Le lecteur souhaitant avoir plus d'informations sur les méthodes de prévision de la demande électrique à moyen et long terme peut se référer à Esteves et al. (2015) et à Khuntia et al. (2016). Il est important de noter que nous ne prétendons pas présenter une liste exhaustive de toutes les études bibliographiques publiées sur ce sujet mais une liste suffisante qui nous a permis de positionner nos travaux de thèse.

2.2 Horizon temporel de la prévision

Généralement, la prévision de la charge électrique peut être divisée en trois grandes catégories principalement déterminées par l'horizon temporel (Hong et Fan, 2016). Chaque

1. L'électricité est une énergie très volatile, elle ne peut pas être stockée que pour des quantités limitées et des usages très spécifiques. https://alumni.ifp-school.com/fr/revue/article/la-problematique-du-stockage-de-l-electricite/11.

catégorie répond à des besoins précis :

1. La prévision à court terme, *Short Term Load Forecasting* (STLF) correspond généralement à un horizon de prévision allant de quelques heures à un jour à l'avance. Ces prévisions permettent aux fournisseurs d'électricité de gérer la demande journalière ainsi que les arrêts de production, et l'optimisation du débit du réseau. Les modèles de prévision à court terme sont généralement basés sur des données à haute fréquence, mesurées toutes les heures ou toutes les demi-heures (Hong et Fan, 2016).
2. La prévision à moyen terme, *Medium Term Load Forecasting* (MTLF) consiste généralement à prédire quelques semaines jusqu'à un an à l'avance. Ces prévisions sont utilisées pour établir des composantes économiques telles que les tarifs, les plans tarifaires et les réglementations auxquels s'ajoutent la planification des maintenances et la commercialisation de l'énergie (Hong et Fan, 2016).
3. La prévision à long terme, *Long Term Load Forecasting* (LTLF) correspond généralement à un horizon de prévision supérieur à un an et peut aller jusqu'à 20 ans à l'avance. Ces prévisions jouent un rôle fondamental dans la planification du système énergétique (les choix d'investissements, la préparation de l'infrastructure à l'évolution de la demande, ...) (Hong et Fan, 2016).

2.3 Prévision à court terme à l'échelle nationale

La prévision de la charge électrique au niveau national à court terme (STLF) a fait l'objet de plusieurs recherches depuis des décennies. La première publication sur ce sujet remonte au 1918 (Huck et al., 1980). Au cours de ces années, une vaste littérature a été mise en œuvre décrivant et testant des méthodes et techniques de modélisation et de prévision de la demande d'électricité à court terme. La prévision de la charge à ce niveau-là est considérée comme relativement facile en raison de la régularité et la périodicité des données contrairement à la prévision à l'échelle locale. Ces techniques peuvent être classées en deux grandes classes bien que la frontière entre les deux devient de plus en plus ambiguë :

1. Les approches statistiques : nous pouvons citer les modèles de séries temporelles ((S)ARIMA et (S)ARIMAX) (Mohamed et al., 2010 ; Nowicka-Zagrajek et al., 2002) puis les modèles de lissage exponentiel (Abd Jalil et al., 2013 ; J. W. Taylor, 2003), les modèles de régression paramétrique (Bianco et al., 2009), les modèles de régression semi-paramétrique (*Generalized Additif Model* (GAM) (Pierrot et al., 2011 ; Erişen et al., 2017)) ainsi que les modèles dynamiques (*State Space Model* (SSM) (Dordonnat et al., 2008)).
2. Les approches basées sur l'intelligence artificielle et en particulier celles issues de l'apprentissage automatique ainsi que les réseaux de neurones (Metaxiotis et al.,

2003 ; Kaytez et al., 2015), les machines à vecteurs de support (*Support Vector Machine*) (Kaytez et al., 2015) et les algorithmes évolutionnaires [2] (Hinojosa et al., 2010).

Les méthodes statistiques fournissent des équations mathématiques explicites où la charge est représentée en fonction de plusieurs variables en entrée comme les variables calendaires (jour de la semaine, heures de la journée, jours fériés, ...) et les variables météorologiques (température, humidité, vent, ensoleillement, ...). Ce sont les premières méthodes qui ont été testées dans la littérature pour la prévision de la charge électrique. Ainsi, pendant des années ces méthodes ont été des modèles de référence. Les modèles de moyenne mobile autorégressive (ARIMA) ainsi que leurs versions saisonnières nommées ((S)ARIMA) et ((S)ARIMAX)(X) ont été aussi utilisés pour le même objectif (Nowicka-Zagrajek et al., 2002 ; Mohamed et al., 2010). Les modèles SARIMA ont été déployés pour modéliser la composante saisonnière de la demande. Ces modèles ont montré de bonnes performances en dehors des périodes de haute volatilité comme les jours fériés et les périodes avec fortes amplitudes thermiques [3] (Nowicka-Zagrajek et al., 2002). D'autres exemples de méthodes statistiques paramétriques utilisées dans ce domaine sont les modèles de lissage exponentiel (Abd Jalil et al., 2013 ; J. W. Taylor, 2003) et les modèles de régression multiple (Bianco et al., 2009). Ces modèles méritent l'attention parce qu'ils sont faciles à mettre en œuvre, nécessitent moins de puissance de calcul que d'autres approches (algorithmes génétiques, réseaux de neurones, machine à vecteurs de support) et donnent des bonnes performances avec une erreur de prévision (MAPE [4]) comprise entre 1% et 3% (J. W. Taylor, De Menezes et al., 2006). Par contre, ils sont très peu évolutifs cela tient au fait qu'ils sont des modèles paramétriques. Il est intéressant de noter également que leur principale faiblesse sont les hypothèses de linéarité et de normalité sous-jacentes qui en réduisent la portée. En effet, la relation entre la demande d'électricité et les variables d'influence n'est pas toujours linéaire (voir la partie 4.2.2.2) et ces méthodes ne sont pas conçues pour modéliser des relations non linéaires.

J. W. Taylor (2003) a introduit la méthode Holt-Winters double saisonnière (DSHW), dont l'avantage important est qu'elle ne nécessite que la durée des deux cycles saisonniers à spécifier. Dans sa thèse, Dordonnat et al. (2008) a proposé un modèle de régression dynamique périodique linéaire et non linéaire (*State Space Model*) pour la prévision de la demande en France à l'échelle nationale et obtient une erreur (MAPE) autour de 1,5%, le terme dynamique du modèle signifie que le modèle permet aux paramètres de varier dans le temps. Certains chercheurs ont travaillé sur des approches semi et non-paramétriques (voir le chapitre 3) comme Fan et al. (2011) qui ont proposé un modèle additif généralisé (voir la partie 3.2.1.2) pour prévoir la demande d'électricité au pas demi-horaire jusqu'à

2. Ce sont des algorithmes stochastiques qui se basent sur l'idée de faire évoluer un ensemble de solutions dans l'objectif de trouver les meilleures solutions.

3. https://fr.wikipedia.org/wiki/Amplitude_thermique

4. L'erreur MAPE est l'erreur absolue moyenne en pourcentage sa formule mathématique est donnée par l'équation 3.34.

sept jours à l'avance pour les systèmes électriques du marché national australien de l'électricité et obtiennent un résultat satisfaisant par rapport aux techniques de prévision de la charge proposées dans la littérature avec une erreur de prévision (MAPE) de 1,88%. Pierrot et al. (2011) ont appliqué les modèles additifs généralisés (GAM) sur la demande d'électricité en France pour la première fois. Le point fort des modèles additifs généralisés est qu'ils sont capables d'intégrer des relations non linéaires et ils sont en même temps facilement interprétables. Antoniadis, Brossat et al. (2014) ont mis en œuvre un modèle de prévision (*KWF*) pour les données chronologiques fonctionnelles en présence de non stationnarité (voir la sous-section 3.2.2), le modèle a été testé sur les données de la demande d'électricité en France. Le problème de non stationnarité peut être remédié par la détection et l'intégration des classes de stationnarité (voir la sous-section 3.2.2). Cugliari (2011) a proposé dans sa thèse des algorithmes de classification non supervisée pour la détection de ces classes.

Les méthodes d'apprentissage automatique (*machine learning*) ont été largement utilisées pour la prévision de la demande d'électricité à court terme et ont montré de bonnes performances. Les trois modèles de *machine learning* les plus utilisés pour la prévision de la charge électrique sont : le réseau de neurones artificiels (ANN), la machine à vecteurs de support (SVM) et les forêts aléatoires (RF). Ces modèles ont été testés dans le principal but de trouver des alternatives aux modèles statistiques pour modéliser les relations non-linéaires entre les facteurs d'influence et la charge électrique. H. Hippert et al. (2005) dans leur article évaluent et comparent la performance de prévision de la charge électrique par des réseaux de neurones (ANN) et des méthodes traditionnelles basées sur le lissage et la régression. Ils concluent que les réseaux de neurones sont plus performants pour leur jeu de données (une erreur MAPE autour de 2,5%). Dans Guo et al. (2006), un modèle de machine à vecteurs de support est proposé pour prévoir la charge électrique la charge électrique de la province du Hebei en Chine. Kaytez et al. (2015) ont testé plusieurs techniques telles que les machines à vecteurs de support (SVM), les machines à vecteurs de support des moindres carrés (LS-SVM) et les réseaux de neurones artificiels (ANN) pour la prévision de la demande d'électricité de la Turquie. Les résultats indiquent que le modèle (LS-SVM) proposé est le plus performant et le plus rapide parmi les modèles testés.

Diverses structures [5] de réseaux de neurones artificiels (ANN) ont trouvé leur place dans la prévision de la charge électrique à court terme afin d'améliorer la précision des modèles historiques, comme les réseaux de neurones flous (Papadakis et al., 1998), les réseaux de neurones à ondelettes (Bashir et al., 2009), les réseaux de neurones à ondelettes floues (Kodogiannis et al., 2013). Quelques modèles des réseaux de neurones profonds (DNN) sont également proposés dans la littérature : ce sont des réseaux (ANN) ayant plus d'une couche cachée. Cette structure à plusieurs couches de calcul augmente la capacité

5. La structure d'un réseau de neurones désigne son architecture qui peut être définie par le nombre de couches dans le réseau, le nombre de neurones par couche ainsi que les relations définies entre ces couches.

du réseau à modéliser des relations non linéaires complexes (Bengio et al., 2013). Bedi et al. (2019) ont proposé un cadre analysant les dépendances à long terme dans les données historiques et les modèles à court terme dans les données segmentées. (LSTM) a ensuite été appliqué en incluant une fenêtre mobile utilisant les données de demande d'électricité de l'Inde. Le modèle développé avait une meilleure performance par rapport au modèle (DRNN), les réseaux de neurones artificiels (ANN) et la régression des vecteurs de support (SVR).

L'intégration des **variables météorologiques** (comme la température, la nébulosité, la vitesse du vent, l'humidité ...) dans les modèles de prévision de la charge électrique est également un critère intéressant pour la classification des études sur ce sujet. En effet, les conditions météorologiques sont responsables d'importantes variations dans la charge électrique. La température est sans doute la variable le plus important parmi les variables météorologiques et donc celle la plus utilisée dans la littérature. Plusieurs études ont montré une relation entre la charge électrique et la température (Rajbhandari et al., 2021 ; Chapagain et al., 2020 ; Fung et al., 2006 ; Dryar, 1944). En effet, Dryar (1944) a été le premier a formulé mathématiquement la relation entre la charge électrique et les variables météorologiques. Selon (Dryar, 1944) la charge est séparée en deux composants principaux :

— une valeur fixe qui correspond à la charge de base.

— une quantité variable qui reflète l'effet de la météo.

Une étude réalisée en France (Bouktif et al., 2019) a montré que la température a une corrélation négative de 0,94 avec la charge, alors qu'une étude réalisée dans la métropole de Bangkok a constaté que la demande d'électricité pour la charge résidentielle avait tendance à augmenter de 6,79% pour une augmentation de 1 ℃ de la température (Wangpattarapong et al., 2008). Cette relation est, en quelque sorte, spécifique à chaque pays et dépend fortement de la zone climatique du pays surtout qu'une grande partie de la consommation finale de l'électricité est liée aux besoins de chauffage et de climatisation. Gupta et al. (1972) ont proposé une méthode de prévision adaptative qui se base à la fois sur les données historiques de la charge électrique et sur les prévisions météorologiques. Cette méthode combine des modèles stochastiques de charge et des modèles météorologiques de charge adaptatifs. Al-Zayer et al. (1996) ont intégré la température sous forme de degrés-jours [6] dans un modèle de régression par morceaux pour modéliser l'effet de la température. Certains auteurs comme Peirson et al. (1994) ont montré également que la relation entre la la charge électrique et température peut être différente en fonction de l'heure de la journée et du jour type de jour. Par conséquent, des modèles ont été conçus pour modéliser cette relation pour différentes heures de la journée et types de jour indépendamment (Ramanathan et al., 2001). Dordonnat et al. (2008) ont proposé un modèle d'espace d'état guassien linéaire qui intègre la température sous la forme de

6. https://energieplus-lesite.be/theories/climat8/degres-jours-d2/

degrés jours de chauffage (HDD) [7] et de degrés jours de climatisation (CDD) [8] pour prédire la charge électrique française. Pierrot et al. (2011) ont également proposé des modèles additifs généralisés pour la prévision de la charge française qui intègrent des variables météorologiques notamment les températures extérieures, les températures retardées [9] ainsi que le maximum et le minimum de ces dernières.

En raison du coût ou de l'indisponibilité des prévisions météorologiques, voire même parfois pour éviter les incertitudes liées à ces dernières (J. W. Taylor et Buizza, 2003), des modèles n'utilisant que des données historiques sans entrées météorologiques ont été développés et sont toujours en cours d'utilisation. J. W. Taylor (2003) a proposé un double lissage exponentiel saisonnier pour prédire la demi-heure suivante jusqu'au lendemain (48 prévisions en avance) des données demi-horaires britanniques avec une erreur de prévision (MAPE) comprise entre 1,25 et 2%. J. W. Taylor et McSharry (2007) ont testé plusieurs méthodes de prévision appliquées sur dix séries chronologiques de la demande d'électricité intra-journalière de dix pays européens sans intégration des données météorologiques. Les auteurs soulignent que l'intégration des données météorologiques peut améliorer la précision des modèles de prévision de la charge au-delà de quatre à six heures d'avance. Pour des délais plus courts, leur méthode proposée donne des résultats compétitifs aux résultats obtenus par des modèles intégrant des données météorologiques. Cugliari (2011) ont proposé également une méthode de prévision appelée *KWF* pour des données de séries chronologiques fonctionnelles et qui s'est avérée concurrente aux autres méthodes de prévision de la charge française sans intégration de la température.

2.4 Prévision à court terme à l'échelle locale et individuelle

La prévision de la *charge électrique locale* peut aller de la prévision de la charge d'un ménage en passant par la prévision de la consommation agrégée d'un groupe de ménages (bâtiment, quartier, ...) jusqu'à la prévision de la charge d'une petite région géographique (village, ville, ...), dont la charge horaire moyenne varie de quelque kW à plusieurs MW. Dans ce qui suit, nous nous concentrons principalement sur la prévision au niveau d'un ménage.

Malgré la maturité des recherches sur la prévision de la charge électrique à l'échelle nationale ou à des niveaux d'agrégation élevés (des grandes villes, un département ou une région géographique, ...) et leur succès global, les recherches sur la prévision de la charge à

7. La variable HDD est définie comme HDD $= max(T^* - T_t, 0)$ où T^* est la température de seuil de chauffage.

8. La variable CDD est définie comme CDD $= max(T_t - T^*, 0)$ où T^* est la température de seuil de climatisation.

9. Soit T_t la température à l'instant t alors la température retardée ou décalée d'une période de 24h est alors T_{t-24}. Ces températures sont souvent utilisées dans la prévision de la charge électrique pour tenir compte de l'inertie thermique des bâtiments.

l'échelle d'un ménage ne sont pas encore au rendez-vous. En effet, la demande d'électricité au niveau national est considérée comme un processus stochastique avec un certain degré de régularité du et maintenu par des facteurs socio-économiques comme la périodicité du cycle économique qui se traduit par une différence de niveau de consommation entre les jours ouvrables et les week-ends, les habitudes de consommation et les cycles du jour et de la nuit. Cette régularité est à la base de la bonne performance des méthodes de prévision à cette échelle et donc la prévision de la charge à ce niveau est considérée relativement évidente et facile. Bien que cette régularité caractérise aussi la charge électrique à des niveaux d'agrégation élevés, les courbes de charge à l'échelle d'un ménage montrent une grande variabilité car elles dépendent de plusieurs variables telles que le nombre d'occupants, leur mode de vie, les équipements électriques, la taille du logement ainsi que les caractéristiques thermiques du bâtiment et le climat de la région. De plus, la consommation d'électricité au niveau du ménage peut varier considérablement d'un jour à l'autre et donc contrairement à la charge globale, la charge du ménage est caractérisée par son irrégularité. Ainsi, la prévision de la demande électrique à l'échelle d'un ménage est plus difficile qu'à d'autres échelles comme les bâtiments, les quartiers ou les charges résidentielles régionales.

La collecte des données de haute fréquence (intra-journalières) de la consommation d'électricité à l'échelle des ménages ces dernières années grâce au déploiement massif des compteurs intelligents dans le secteur résidentiel et l'émergence des applications dans le contexte des maisons intelligentes, des programmes de la gestion et de la réduction de la consommation ont relancé les recherches sur la prévision des courbes de charge à cette échelle aussi bien dans le domaine académique qu'industriel.

Certains auteurs ont testé et évalué les techniques classiques de la prévision de la charge nationale directement sur les données brutes de consommation d'électricité des ménages. Veit et al. (2014) ont appliqué un certain nombre de méthodes de prévision, notamment des modèles ARIMA, des réseaux de neurones et des modèles de lissage exponentiel en utilisant plusieurs stratégies pour la sélection des données d'entraînement. L'évaluation a été effectuée sur deux ensembles de données, le premier concernait un seul ménage en Allemagne et le second, six ménages aux États-Unis. Les résultats indiquent que la précision des prévisions varie considérablement en fonction du choix de la méthode et de la stratégie de prévision ainsi que la configuration des paramètres. L'erreur de prévision (MAPE) variait entre 5% et plus de 100% pour l'ensemble du jeu de données. Wijaya, SFRJ Humeau et al. (2014) ont également proposé une évaluation quantitative de différentes méthodes d'apprentissage automatique (comme la régression linéaire, la régression basée sur (SVM) et perceptron multicouche [10]) pour la prévision de la charge électrique à court terme au niveau individuel et agrégé. Ghofrani et al. (2011) ont étudié le modèle de filtre de Kalman dans le contexte de la prévision de la consommation à l'échelle des ménages. La méthode proposée est évaluée pour différents horizons de prévision et

10. https://fr.wikipedia.org/wiki/Perceptron_multicouche

différentes granularités [11] de données pour un seul ménage. Les résultats montrent que la disponibilité des données en temps réel améliore la précision de la prévision de charge significativement, par contre, cette amélioration de la précision vient à l'encontre du coût de calcul. Les auteurs ont conclu qu'un choix judicieux du taux d'entraînement des données permet d'éviter une charge de calcul élevée tout en obtenant des résultats satisfaisants.

D'autres auteurs considèrent que les méthodes classiques de prévision de la charge ne conviennent pas tout à fait à la prévision à l'échelle des ménages. Dang-Ha et al. (2017) expliquent que plusieurs approches existantes ne sont pas applicables pour la prévision de la consommation à l'échelle locale soit en raison d'un long temps d'entraînement, soit d'un processus d'optimisation instable ou d'une sensibilité aux hyperparamètres [12]. Ensuite, ils proposent cinq modèles automatiques adaptés à la prévision de la charge locale qui nécessitent une intervention humaine très limitée. Les résultats montrent que la version modifiée du modèle additif de Holt-Winters [13] saisonnier proposée dans cet article rivalise avec des méthodes plus complexes avec seulement trois mois de données d'entraînement.

De nouvelles approches de prévision qui vont être présentées ci-dessous ont été conçues spécifiquement pour la charge des ménages. Des chercheurs estiment nécessaire de précéder la prévision par une étape de pré-traitement de données comme l'analyse spectrale (analyse par ondelettes, la transformation de Fourier, ...). Une architecture de réseau de neurones profond qui combine des caractéristiques de transformée en ondelettes stationnaires et des réseaux de neurones convolutifs est présentée dans Eneyew et al. (2020). Cette approche utilise des caractéristiques automatiquement extraites du jeu de données de consommation des ménages en appliquant des opérations de décomposition en ondelettes, de convolution et de *clustering*. Les résultats de cette étude ont montré l'avantage d'intégrer des caractéristiques d'ondelettes avec des réseaux de neurones convolutifs (voir annexe 6.2) pour améliorer la précision des prévisions. Un autre modèle dans Yan et al. (2019) est proposé combinant un réseau de neurones récurrents à mémoire court-terme et long terme (LSTM) avec la technique de transformée en ondelettes stationnaire (SWT) (voir annexe du chapitre 6.2). L'utilisation du (SWT) a comme objectif de réduire la volatilité et d'augmenter les dimensions des données, ce qui peut potentiellement selon les auteurs aider à améliorer la précision des prévisions par le réseau (LSTM). Les résultats de l'application de ce modèle sur les données de cinq maisons à Londres, au Royaume-Uni montrent que le modèle proposé a une performance supérieure à la méthode persistante, la régression vectorielle de support (SVR), le réseau de neurones récurrents à mémoire court-terme et long terme (LSTM) et le réseau de neurones convolutifs combinant mémoire à long court terme (CNN-LSTM), pour différentes granularités de données.

11. La granularité des données dans notre cas signifie la fréquence des données (demi-horaires, horaires, journalières, mensuelles, ...).

12. Un hyperparamètre d'un modèle est fixé manuellement et contrairement aux paramètres ne peut pas être estimé à partir des données mais contribue à l'estimation de ces derniers.

13. https://docs.oracle.com/cloud/help/fr/pbcs_common/CSPPU/holt-winters_additive.htm

Des auteurs ont eu recours à une autre approche, le *clustering* (voir annexe 6.2) afin d'améliorer la performance des modèles classiques de prévision. Certains d'entre eux ont choisi le *clustering* des courbes de charge journalières d'un même ménage. D'après Yildiz et al. (2018) l'utilisation d'un seul modèle de prévision pour toute la période peut entraîner des performances du modèle limitées, en raison de la grande variabilité des profils journaliers des courbes de charge d'un jour à un autre. Par conséquent, les auteurs ont proposé une étape de pré-*clustering* (voir annexe 6.2) des jours similaires en fonction de leur écart-type. Par suite, au lieu d'appliquer un modèle de prévision à l'ensemble de jeu de données d'un ménage particulier, un modèle de prévision est alors appliqué à chacun des *clusters*. Ensuite, des modèles d'apprentissage automatique bien connus tels que les réseaux de neurones artificiels et les machines à vecteurs de support sont testés et évalués sur des données de 14 ménages de la Nouvelle-Galles du Sud, en Australie. Les résultats ont montré que l'étape préliminaire de *clustering* a augmenté la complexité de l'analyse, mais elle a entraîné des améliorations significatives des performances de prévisions. De même, Abreu et al. (2012) ont travaillé sur l'identification des caractéristiques du profil quotidien d'un seul ménage par l'analyse en composantes principales ensuite le *clusternig* des profils journaliers similaires. D'autres chercheurs ont choisi le *clustering* des données de consommation des ménages afin de définir des groupes de ménages ayant des profils de consommation d'électricité similaires dont le but est souvent d'améliorer la prévision de la charge agrégée. Wijaya, SFRJ Humeau et al. (2014) ont présenté différentes techniques de *clustering* qui peuvent être utilisées pour regrouper les données de consommation d'électricité à l'échelle des ménages dans l'objectif d'améliorer la précision de la prévision de la charge électrique agrégée. Les résultats obtenus montrent une amélioration de la précision de la prévision de la charge électrique agrégée de tous les ménages dans le jeu de données et que cette amélioration devient plus importante avec l'augmentation de la taille ce dernier. De même, Franklin L Quilumba et al. (2014) ont étudié également l'application du *clustering* sur les données de consommation des ménages dont le but d'améliorer la performance de la prévision de la charge globale. La méthode proposée a été appliquée à deux ensembles de jeu de données de deux entreprises d'électricité différentes, la première aux États-Unis et la deuxième en Irlande. Cerquitelli et al. (2018) adopte aussi une approche de *clustering* de profils des courbes de charge domestiques qui a été utilisée avec succès au Portugal et l'applique aux données du Royaume-Uni. Dans Laurinec et al. (2017), une nouvelle méthode de prévision de la charge des ménages est utilisée. Les données de tous les ménages sont utilisées pour le *clustering*. Après celui-ci, les courbes de charge des différents ménages sont prétraitées par la normalisation, ensuite chaque courbe de charge normalisée est représentée à l'aide des coefficients d'un modèle de régression linéaire multiple. Les prévisions finales basées sur le centroïde de chaque *cluster* sont mises à l'échelle par des paramètres de normalisation stockés pour générer des prévisions pour chaque ménage. Cette méthode est comparée à l'approche de la prévision pour chaque ménage séparément. L'évaluation de cette méthode a été menée sur deux ensembles de données de compteurs intelligents provenant d'une part de résidences en Irlande et d'autres part

d'usines en Slovaquie. Les résultats obtenus ont prouvé que cette méthode basée sur le *clustering* améliore la précision des prévisions. Elle est également plus évolutive puisque le modèle n'est pas conçu pour chaque ménage.

D'autres auteurs ont proposé des techniques comme la modélisation de l'activité des ménages (la durée et les horaires de fonctionnement des équipements électriques) associée à des modèles de prévision classiques. Les impacts des activités quotidiennes des résidents et de l'utilisation des appareils sur la consommation électrique de l'ensemble du ménage sont intégrés pour améliorer la précision du modèle de prévision dans Gajowniczek et al. (2017). Javed et al. (2012) recommandent l'utilisation de données anthropologiques et structurelles (la surface du foyer, le nombre des occupants, le nombre des enfants, les caractéristiques du foyer,...) pour la prévision de la charge à l'échelle des ménages dans le contexte des applications de réseau intelligent comme la réponse à la demande (DSR). Dans Ding et al. (2015), les auteurs ont intégré une variable qui modélise la séquence d'activités du ménage dans un modèle de régression vectorielle de support (SVR). L'étude a révélée que l'intégration de ce type d'information pourrait améliorer la précision de la prévision de la charge 15 minutes à l'avance pour les ménages individuels, par contre, l'accessibilité à cette information reste difficile et donc la méthode est classée comme peu pratique.

Contrairement à l'échelle nationale, la sensibilité de la consommation électrique à la température extérieure, « la thermosensibilité », à l'échelle locale est plus particulièrement à l'échelle des ménages est encore un point de désaccord entre les experts. Certes, la consommation d'électricité d'un ménage augmente avec l'augmentation ou la diminution de la température extérieure en raison du déclenchement des appareils énergivores (comme le chauffage ou la climatisation). D'ailleurs certaines études sur ce sujet ont montré que l'intégration des température extérieure dans les modèles de prévision de la charge à cette échelle n'avait aucune valeur ajoutée sur la performance de ces derniers. Par conséquent, il n'est pas évident que les variables météorologiques jouent un rôle important dans les prévisions de charge à l'échelle des ménages. Haben, Giasemidis et al. (2019) expliquent dans leur article que la température extérieure et la demande sont fortement liées à la saisonnalité annuelle. Ainsi, l'effet de la température extérieure sur la charge peut être intégré dans les modèles de prévision implicitement à travers la variable modélisant la saisonnalité annuelle, ce qui peut justifier l'invariabilité du résultat de la prévision par rapport à l'intégration ou pas de la température extérieure. D'après J. W. Taylor et Buizza (2003), l'intégration de la charge retardée comme variable d'entrée dans les modèles de prévision peut remplacer l'effet de la température, si cette dernière change d'une manière lisse. Rodrigues et al. (2014) ont proposé un modèle de réseaux de neurones artificiels de la consommation d'électricité quotidienne et horaire à l'échelle des ménages, en utilisant uniquement les données physiques et démographiques des ménages en combinaison avec des effets de calendrier annuel dans le but de montrer qu'il est possible d'obtenir de bons résultats même sans tenir compte des données météorologiques. Wijaya, SFRJ Humeau

et al. (2014) ont conclu dans leur étude qu'au niveau des clients individuels, l'intégration de la température n'améliore pas significativement la précision des prévisions.

Par contre, d'autres auteurs estiment nécessaire l'intégration des données météorologiques dans les modèles de prévisions de la charge des ménages. Par exemple, les auteurs dans Gerossier et al. (2017) ont proposé un modèle de prévision probabiliste de la charge journalière des ménages individuels. Ce modèle utilise à la fois les données de consommation individuelle des ménages ainsi que les prévisions de la température extérieure. Beccali et al. (2008) ont introduit « l'indice humidex » qui prend compte de la demande de chauffage et de refroidissement due à l'inconfort thermique ressenti par les résidents des ménages. Taieb et al. (2021) ont abordé l'incertitude liée à la météo et à la demande d'électricité en proposant des prévisions probabilistes basées sur la régression quantile.

L'agrégation est une autre approche populaire dans le domaine de la prévision de la charge qui vise à éviter les irrégularités dans les courbes de charge des ménages. Elle se concentre sur la prévision de la consommation de groupes de ménages, généralement regroupés par localisation géographique (bâtiments, quartiers, ...). L'agrégation réduit généralement la variabilité de la charge, ce qui se traduit par des formes de charge plus lisses et plus prévisibles. Sevlian et al. (2014) ont montré que pour différentes méthodes de prévision et horizons, l'agrégation de plus de clients améliore les performances de prévision relatives jusqu'à un jusqu'à un certain nombre de ménages. Au-delà de ce nombre, plus aucune amélioration des performances relatives ne peut être obtenue. Dans Samuel Humeau et al. (2013), les auteurs considèrent une variété de méthodes et montrent également une forte relation entre les erreurs de la prévision (MAPE) et la taille d'agrégation. Idem dans Hayes et al. (2015), les auteurs appliquent les modèles (ANN) et (ARIMA) à 4 niveaux différents d'agrégations des données de compteurs intelligents irlandais et danois et ils sont arrivés aux mêmes conclusions.

2.5 Bilan de la littérature et positionnement des travaux de la thèse

Au début de ce chapitre, nous avons cité les différentes méthodes et techniques issues de la statistique ou de l'intelligence artificielle qui ont été élaborées pour la prévision de la charge électrique aux niveaux national ou local. La littérature sur les modèles de prévision de la charge électrique nationale est bien mature comme nous avons pu le constater et les travaux développés dans ce cadre sont assez satisfaisants. Cependant, il existe encore de nombreux obstacles à l'accélération des travaux de recherche concernant la prévision à l'échelle des ménages.

Plusieurs articles ont été publiés sur ce sujet afin de trouver « la meilleure technique »

de prévision à cette échelle. Certains chercheurs ont appliqué les modèles classiques de la prévision de la charge nationale à l'échelle locale et montrent qu'avec quelques adaptations, les modèles donnent des résultats satisfaisants. D'autres pensent que les modèles classiques ne sont pas adaptés à cette tâche en raison de leur complexité et optent pour des modèles hybrides [14]. Cependant, rien ne confirme la contribution de ces modèles à la littérature ni celle des modèles classiques. L'auteur dans Hong et Fan (2016) parle du mythe de la meilleure technique et explique que pratiquement tous les articles montrent la supériorité de diverses techniques sur des jeux de données particuliers. Cela rend alors les conclusions difficiles à généraliser sur d'autres jeux de données. Il ajoute qu'il est important de comprendre qu'il n'y a pas une technique de prévision idéale qui fonctionne pour tous types de données et que ce sont les données et les juridictions qui déterminent la meilleure technique plutôt que l'inverse. L'auteur estime important que les chercheurs commencent par comprendre les besoins de l'entreprise puis analyser les données et ensuite passer au processus d'essais pour déterminer la meilleure technique qui répond à leurs propres besoins. L'auteur cite également les différents points faibles et forts des différentes classes de techniques de prévision.

Un article recensant les méthodes les plus récentes d'utilisation des données de compteurs intelligents telles que la prévision, la classification et l'optimisation a été publiée par Yildiz et al. (2017). Dans cet article, les auteurs expliquent qu'il existe des différences remarquables entre les études publiées à ce sujet non seulement en raison de la différence des applications envisagées mais aussi en raison de la diversité des données utilisées, ce qui rend difficile la comparaison de ces études entre elles. En effet, la différence des applications envisagées par les prévisions entraîne une différence dans le choix de l'horizon de la prévision, la granularité et l'échelle des données, le choix des facteurs d'influence, la durée de l'évaluation (période test) ainsi que les méthodes d'évaluation de la performance des prévisions, ce qui justifie la difficulté de la comparaison de ces études.

En plus, dans la majorité des articles la qualité des prévisions est évaluée sur des données des ménages pré-sélectionnées selon des critères souvent non détaillés et dans des périodes stables (exclusion des jours fériés) et moins sensibles à la variation de la température. Certes ce type de sélection est nécessaire pour mettre en évidence l'intérêt des modèles de prévision mais ne permet pas de tester leurs robustesses aux différents profils de clients et dans les périodes les moins stables. Par suite, les approches mises en œuvre pour la prévision de la charge à cette échelle ne sont pas toutes industriellement viables en raison de leurs complexités (difficile à adapter sur de nouvelles données, non prise en compte des contraintes de temps de calcul et des ressources en mémoire).

Compte tenu du manque dans la littérature d'études comparatives des performances de différentes méthodes de prévision appliquées à un ensemble standard de données de ré-

14. Un modèle hybride de prévision est un modèle qui combine plusieurs modèles ou méthodes de prévision

férence selon des critères comparables (la granularité des données, l'horizon de la prévision, mesure d'erreur...), le positionnement de nos travaux par rapport aux études publiées à ce sujet était compliqué. Le tableau 2.1 présente un résumé des différentes études publiées sur la prévision à l'échelle d'un ménage en termes de nombre des ménages, granularité des données, d'horizon de prévision, la durée de l'évaluation des modèles, l'intégration de la température, le type des données ainsi que le résultat de l'étude en termes d'erreurs (NRMSE, RMSE, MAPE). Le nombre de ménages utilisés dans ces études est un facteur très intéressant qui permet de juger la performance des modèles proposés ainsi que leur robustesse [15] et leur adaptabilité aux différents profils de courbes de charge (les personnes âgées, les personnes en télétravail, les personnes ayant une activité professionnelle régulière, les étudiants qui ont une vie dynamique, ...). En plus, les études qui cherchent à prédire la consommation avec un horizon intra-journalier (moins de 24h) n'ont pas le même intérêt économique que les prévisions d'un jour à l'avance. En général, dans le contexte des applications industrielles où le modèle doit effectuer des prévisions pour des milliers de clients les prévisions intra-journalières sont assez coûteuses et la prévision pour le lendemain $(J+1)$ est plus appropriée [16]. Un autre facteur aussi intéressant à prendre en compte dans cette comparaison est la durée d'évaluation. Les études utilisant des périodes plus courtes pour leur évaluation pourraient présenter des résultats de précision peu fiables.

Dans notre travail, nous avons tenu compte des défis mentionnés précédemment pour mettre en œuvre une étude comparative de plusieurs modèles de prévision de la charge des ménages à court terme (le lendemain). L'étude est réalisée sur des données réelles de consommation de 1000 clients de profils disparates d'un fournisseur d'électricité locale en France. Les modèles choisis pour cette étude sont des modèles sélectionnés parmi une grande variété de modèles de la littérature selon des critères que nous détaillerons plus tard (voir la section 3.3).

15. Le terme « robustesse » employé dans notre contexte désigne un critère qualitatif et non quantitatif du modèle de prévision.

16. La prévision à $(J+1)$ est très importante car elle permet aux fournisseurs d'énergie d'ajuster leurs achats/ventes pour le lendemain et de réduire au maximum les écarts offre/demande et les coûts associés.

Article	Nombre de ménages	Granularité des données	Horizon de prévision	Durée d'évaluation	Température	Les données	NRMSE	RMSE
Samuel Humeau et al. (2013)	782	1h	24 h	6 mois	oui	Irish Dataset [a]	0,80	-
Wijaya, SFRJ Humeau et al. (2014)	782	1h	24h	6 mois	oui	Irish Dataset	0,61	-
Gerossier et al. (2017)	226	1h	24 h	3 mois	oui	Portuguese Dataset	0,43	-
Lusis et al. (2017)	27	30 min	30 min	28 jours	oui	Austrelian Dataset	-	0,52
Shi et al. (2017)	920	30 min	1 h	1 mois	Non	Irish Dataset	-	0,45
Yildiz et al. (2018)	14	5,15,30 et 60 min	24 h	-	oui	Austrelian Dataset	-	0,80
Voss et al. (2018)	200	-	24 h	6 mois	oui	Pecan Street Dataset	0,53	-
Yan et al. (2019)	5	5 min	30 min	3 mois	Non	UK-DALE Dataset [b]	-	0,05

Table 2.1 – Résumé des différentes études publiées sur la prévision de la charge à l'échelle des ménages.

a. (ISSDA, 2020)
b. (Kelly et al., 2015)

Chapitre 3

Le cadre théorique de la prévision

Objectifs

Dans ce chapitre, nous présenterons plusieurs modèles les plus couramment utilisés dans le domaine de la prévision de la charge électrique. Nous décrirons brièvement le cadre théorique de chaque modèle, ainsi que ses avantages et limites. Nous exposerons ensuite les critères de sélection que nous avons adoptés pour choisir les modèles les plus adaptés à notre objectif de prévision de la charge électrique à l'échelle des ménages. Enfin, nous présenterons les mesures d'erreur les plus couramment utilisées pour évaluer et comparer la précision des modèles de prévision en général, ainsi que celles spécifiquement développées pour évaluer la prévision à l'échelle des ménages.

Sommaire

3.1 Introduction

La prévision est un élément crucial pour les entreprises qui cherchent à améliorer leur efficacité opérationnelle et leur compétitivité sur le marché. En effet, l'utilisation des prévisions permet à ces entreprises d'anticiper la demande future et de gérer leurs moyens de production d'une manière plus efficace (Makridakis, Wheelwright et al., 1998). De nombreuses études ont été menées sur le cadre théorique de la prévision au fil des années. Ces études ont abordé différents sujets tels que les modèles de prévision, les critères de sélection de ces modèles, l'évaluation de la précision des prévisions, ainsi que les diverses applications possibles de la prévision dans différents domaines (Hyndman et Athanasopoulos, 2018).

La mise en place de la prévision dans l'industrie consiste à adapter et tester plusieurs modèles de prévision et à sélectionner le modèle de prévision adapté à chaque type de données. Cette sélection se fait en fonction de certains critères, tels que la précision, l'interprétabilité, la complexité, la stabilité, la vitesse de calcul, ... (Hyndman et Athanasopoulos, 2018). Ces critères de sélection dépendent des besoins spécifiques de l'entreprise et de l'application envisagée. L'évaluation de la précision des modèles de prévision est également un aspect important de la mise en place de la prévision dans l'industrie. Elle permet de vérifier la performance des modèles de prévision, d'identifier les éventuels problèmes et d'améliorer les modèles afin de réduire les coûts associés à une mauvaise prévision (Hyndman et Koehler, 2006).

Le sujet de notre étude, la prévision de la charge électrique est un exemple concret et important de l'application de la prévision dans un contexte industriel. Les modèles de prévision usuels ont été adaptés à la prévision de la charge électrique et utilisés pour aider les entreprises d'électricité à planifier leur production et leur distribution d'électricité afin de répondre à la demande de leurs clients. Dans ce chapitre, notre objectif est de présenter plusieurs modèles de prévision largement utilisés dans le domaine de l'électricité, en exposant brièvement leur fondement théorique ainsi que les avantages et inconvénients respectifs de chaque modèle. Nous allons également présenter les différents critères utilisés pour la sélection et la mise en œuvre de ces modèles de prévision dans le contexte spécifique de la prévision de la charge électrique à l'échelle des ménages, ainsi que les métriques d'évaluation utilisées pour mesurer leur performance.

3.2 Modèles de prévision

Pendant de nombreuses années, la précision de la prévision a été privilégiée par rapport à l'interprétabilité. Aujourd'hui, les besoins ont évolué aussi bien du côté utilisateur que du côté consommateur et l'interprétabilité est devenue dans certains contextes, comme

les secteurs de la santé, de la banque ou de l'assurance, un objectif éthique et juridique (Lipton, 2016 ; Molnar, 2021). L'interprétabilité est définie dans l'article Molnar et al. (2018) comme étant le degré auquel un humain peut comprendre la cause d'une décision. Selon ce critère, les modèles de prévision peuvent être classées en deux catégories principales en fonction de leur niveau de transparence et d'intérprétabilité : les modèles de type boîte noire (appelé plus communément *black-box*) ou le modèle de type boîte blanche (appelé plus communément *white-box*). Les modèles de boîte noire sont des modèles qui utilisent des techniques complexes pour capturer les relations entre les variables d'entrée et de sortie. Cependant, ils sont souvent difficiles à interpréter puisque nous ne pouvons pas comprendre comment les variables sont liées les unes aux autres pour parvenir à une prévision finale. De plus, ils ne permettent pas à l'utilisateur de disposer au final d'une expression analytique du modèle identifié. Par contre, les modèles de type boîte blanche sont des modèles facilement interprétables et les relations entre les variables sont souvent explicites, les raisons pour lesquelles ces modèles sont largement utilisés surtout dans des contextes industriels. Il est important de noter également que la distinction entre les deux types de modèles n'est pas toujours facile et qu'il existe des modèles qui se situent entre les deux, en termes de complexité et d'interprétabilité (Molnar, 2021).

L'ensemble des modèles de prévision peut être divisé également en deux catégories : les modèles issus de la statistique mathématique et les modèles issus de l'apprentissage automatique (appelé plus communément *machine learning*). Les modèles issus de la statistique mathématique s'appuient généralement sur les deux hypothèses suivantes : les modèles sont paramétriques et de type additif. L'hypothèse paramétrique permet de réaliser des tests d'hypothèses et d'incertitude qui sont nécessaires pour la validation du modèle. Ces postulats restreignent considérablement l'étendue des applications de ces modèles. Cependant, l'amélioration constante de la puissance de calculs des ordinateurs a permis d'obtenir des modèles plus flexibles et moins contraignants. Les modèles issus du *machine learning* sont un bon exemple de la contribution des outils informatiques dans la progression des méthodes statistiques (James et al., 2013). En effet, ces derniers sont basés sur des approches non paramétriques où la structure du modèle n'est pas spécifiée voire inconnue. Aucune forme spécifique n'est imposée pour l'estimateur et l'additivité n'est pas attendue. Ainsi les hypothèses sur la distribution normale de l'erreur, la linéarité, par exemple, ne sont pas alors utilisées pour la modélisation.

Comme abordé précédemment dans le chapitre 2, la littérature sur la prévision de la charge électrique regorge de modèles de prévision appartenant à diverses catégories, depuis les modèles linéaires simples jusqu'aux modèles de réseaux de neurones les plus complexes. Dans les paragraphes suivantes, nous présenterons les fondements théoriques d'une sélection de modèles de prévision utilisés dans la littérature et considérés comme pertinents pour la prévision de la charge électrique. Nous espérons ainsi fournir une vue d'ensemble des approches théoriques les plus courantes dans le domaine de la prévision de la charge électrique, ainsi que des avantages et inconvénients associés à chaque modèle.

Cette présentation nous permettra de justifier le choix des modèles que nous avons utilisés dans le contexte spécifique de la prévision de la charge à l'échelle des ménages. Il est important de souligner que la liste de modèles présentée n'est pas exhaustive et qu'il existe de nombreux autres modèles de prévision sur ce sujet.

Dans l'annexe 6.2, vous trouverez une liste de méthodes de régression univariée qui complète le chapitre. Ces méthodes sont des techniques fondamentales en analyse de données, largement utilisées dans différents domaines pour modéliser des relations non linéaires entre les variables. En outre, ces méthodes constituent la base des théories fondamentales pour des modèles plus complexes qui ont été utilisés dans notre étude. La régression à noyau est la base du modèle *KWF* que nous présentons dans ce chapitre, ainsi que les méthodes des *splines* de lissage et de régression qui sont à la base des deux modèles *GAM* et *MARS*.

3.2.1 Modèles de régression

L'analyse de régression est une technique largement utilisée pour la prévision de charge électrique (Papalexopoulos et al., 1990 ; Ramanathan et al., 1997 ; Hong, 2010 ; Hong, Pu Wang et al., 2015). Ces modèles utilisent des données historiques de consommation d'électricité pour prédire la charge future. Ils utilisent généralement des variables explicatives, telles que la température, la saisonnalité, les jours fériés, ... pour expliquer les variations de la consommation d'électricité. Les modèles de régression les plus couramment utilisés pour la prévision de la charge électrique sont les modèles de régression linéaire multiple (Atan et al., 2014), les modèles de régression non paramétrique tels que la régression à noyau (Shao et al., 2019) et les forêts aléatoires (J. Zhang et al., 2017).

La régression est une méthode de modélisation statistique permettant d'établir une relation mathématique entre une variable appelée **variable expliquée** ou **variable à expliquer** ou encore **réponse**, notée généralement Y, et un ensemble de p variables appelées **variables explicatives**, notées $X_1, X_2, \ldots, X_p$ (Fox, 2015). La variable réponse Y est alors décomposée en deux parties : une fonction notée m qui dépend de l'ensemble des p variables explicatives et une autre partie absorbée par l'erreur, notée ε qui est inobservable. Cela se traduit mathématiquement par l'équation suivante :

$$Y = m(X_1, \ldots, X_p) + \varepsilon. \tag{3.1}$$

Le modèle de régression cherche à déterminer la variation de l'espérance conditionnelle de la variable aléatoire Y sachant $\mathbf{X} = (X_1, \ldots, X_p)$ en fonction d'un ensemble de p variables explicatives. En d'autres termes, le modèle étudie comment la variable aléatoire Y évolue « en moyenne » en fonction des p variables explicatives. Ceci signifie donc que la fonction m peut être s'exprimer ainsi : $m(X) = \mathbb{E}(Y|X)$. Dans la théorie des modèles

de régression, l'hypothèse suivante est très souvent émise : l'erreur ε est un processus aléatoire indépendant des p variables explicatives. Le modèle le plus utilisé de régression est le modèle linéaire. Ce dernier, lorsqu'il est approprié au contexte, est très pratique puisqu'il propose un modèle facile à interpréter puisqu'il s'écrit alors sous une combinaison linéaire des p variables explicatives :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{3.2}$$

3.2.1.1 Modèle de régression linéaire

Le modèle de régression linéaire simple (respectivement multiple) est le modèle le plus connu parmi les modèles de régression paramétrique. Il a pour objectif d'expliquer la variation d'une variable aléatoire notée Y, à valeurs réelles, à partir d'une variable notée X de $\mathbb{R}^p$, dont chacune des p coordonnées peut être observée ou fixée par l'expérimentateur et qui sont notées X_1, ..., X_p. Il est souvent utilisé dans sa forme multiple et s'écrit alors de la façon suivante :

$$\mathbb{E}(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{3.3}$$

Nous pouvons réécrire ce modèle linéaire pour un échantillon aléatoire de taille n de la façon suivante :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.4}$$

où y_i est la valeur de l'i-ème observation du vecteur $\mathbf{Y}$, x_{ij} l'i-ème observation de la j-ème variable explicative X_j. Les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ sont des paramètres réels inconnus appelés paramètres de régression ou coefficients de régression. Les ϵ_i sont les n variables aléatoires d'erreur, non observées, indépendantes et identiquement distribuées qui vérifient les conditions suivantes :

1. $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ et $\mathbb{V}ar(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 I_n$.
2. Le vecteur aléatoire $\boldsymbol{\epsilon}$ est indépendant de la distribution conjointe de $X_1, \ldots, X_p$. Par conséquent, nous avons :

$$\mathbb{V}ar(Y|X_1, \ldots, X_p) = \sigma_\epsilon^2.$$

L'équation (3.4) peut s'écrire matriciellement de la manière suivante :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.5}$$

où

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Les paramètres inconnus du modèle sont représentés par le vecteur $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ et la variance σ_ϵ^2. Ils sont souvent estimés par la méthode des moindres carrés ordinaires puisqu'elle ne nécessite pas d'hypothèse supplémentaire sur la distribution de l'erreur $\boldsymbol{\epsilon}$, contrairement à la méthode du maximum de vraisemblance qui se base sur l'hypothèse de normalité de cette dernière. L'estimation par la méthode des moindres carrés ordinaires consiste à trouver $\hat{\beta} \in \mathbb{R}^{p+1}$ qui minimise la somme des erreurs quadratiques :

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

Le problème d'optimisation s'écrit alors de la façon suivante :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right]^2.$$

Ce modèle présente certains avantages tels que sa facilité d'utilisation et son interprétabilité. Cependant, il est limité dans sa capacité à modéliser des relations non linéaires. L'hypothèse de linéarité du modèle est très restrictive. En pratique, la relation entre la variable Y et les variables explicatives $X_1, X_2, X_3, \ldots, X_p$ n'est pas souvent linéaire. Dans ce cas, les chercheurs et les utilisateurs ont recours habituellement à des modèles plus complexes comme la régression polynomiale, qui reflètent mieux la relation entre Y et les variables explicatives $X_1, X_2, X_3, \ldots, X_p$ ou bien d'effectuer des transformations de données qui permet de rétablir des relations linéaires entre les variables (Trevor Hastie et al., 2009).

Le modèle ***Hong's Vanilla Benchmark*** (Hong, Peng Wang et al., 2011) est un exemple d'un modèle de régression linéaire multiple connu comme modèle de référence dans le domaine de prévision de la charge électrique qui exprime la charge électrique à un instant donné (Y_t) en fonction de la température extérieure (T_t), la variable de temps (H_t) et les variables calendaires telles que le jour (J_t) et le mois (M_t). La relation entre la température et la charge électrique est modélisée par un polynôme du troisième ordre. L'équation du modèle s'écrit de la façon suivante :

$$\begin{aligned} Y_t = {} & \beta_0 + \beta_1 t + \beta_2 J_t \times H_t + \beta_3 M_t + \beta_4 M_t \times T_t \\ & + \beta_5 M_t \times T_t^2 + \beta_6 M_t \times T_t^3 + \beta_7 H_t \times T_t \\ & + \beta_8 H_t \times T_t^2 + \beta_9 H_t \times T_t^3 + \epsilon_t. \end{aligned} \tag{3.6}$$

Ce modèle a été utilisé dans la compétition *Global Energy Forecasting Competition* en 2012. Cette compétition a pour objectif d'évaluer les techniques de prévision de la consommation d'énergie et de la production d'énergie renouvelable. Les résultats ont montré que l'utilisation de modèles plus flexibles que le modèle *Vanilla Benchmark* a permis d'obtenir jusqu'à 40% de d'amélioration en termes de précision de la prévision (Hong, Pinson et al., 2014).

L'avancement de la technologie a permis la conception de nouvelles méthodes de régression plus flexibles, qui n'imposent pas une forme prédéterminée de la fonction m, mais qui permettent de la construire selon les informations provenant des données. Ces méthodes sont connues sous le nom de **régression non paramétrique**. Les méthodes non paramétriques de régression les plus connues sont la régression loess (régression par polynômes locaux et polynômes locaux pondérés) (voir l'annexe 6.2), l'estimation par noyau de la fonction de régression (voir l'annexe 6.2), et la régression spline (voir l'annexe 6.2). Ces méthodes permettent de contrôler la flexibilité de l'estimateur de la fonction de régression m. Une version multivariée de la régression non paramétrique existe également. Elle permet à la fois d'obtenir un estimateur flexible et capable de modéliser la relation entre Y et plusieurs variables explicatives $X_1, \ldots, X_p$.

Les modèles de régression linéaire peuvent être étendus aux modèles autorégressifs (Box et al., 1970), où la variable à prédire est une fonction linéaire de ses propres valeurs passées. Les courbes de charge sont souvent considérées comme des séries temporelles autocorrélées car les valeurs actuelles de la charge sont généralement corrélées avec les valeurs passées. Cela signifie que la structure temporelle des données est importante pour la prévision de la charge électrique et que les modèles doivent prendre en compte ces corrélations temporelles. Les modèles autorégressifs sont l'une des méthodes couramment utilisées pour modéliser cette structure temporelle et prévoir les valeurs futures de la charge électrique. La forme générale d'un modèle autorégressif d'ordre p (noté $AR(p)$) est la suivante :

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t \tag{3.7}$$

où y_t est la variable à prédire à l'instant t, ϵ_t est un terme d'erreur aléatoire à l'instant t, p est l'ordre du modèle autorégressif et ϕ_0 une constante. Les coefficients $\phi_1, \ldots, \phi_p$ représentent les pondérations de la variable à prédire à des instants passés. Cependant, ce modèle ne prend pas en compte les erreurs passées dans la modélisation, ce qui peut conduire à des prévisions imprécises. C'est là qu'entre en jeu le modèle *ARMA* (Box et al., 1970). En plus des valeurs passées de la série, le modèle *ARMA* utilise également les erreurs passées pour prédire les valeurs futures (Box et al., 1970). Cela permet de mieux prendre en compte les variations imprévisibles dans la série et d'obtenir des prévisions

plus précises. Le modèle $ARMA$(p,q) s'écrit de la façon suivante :

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Les paramètres p et q du modèle $ARMA(p, q)$ peuvent être estimés à l'aide de la méthode de l'autocorrélation partielle (PACF) et de l'autocorrélation simple (ACF) (Aragon, 2011) ou la méthode de *Box-Jenkins* (Box et al., 1970) alors que l'estimation des coefficients du modèle peut être effectuée à l'aide de la méthode du maximum de vraisemblance. Contrairement au modèle $ARMA$ qui ne considère pas de saisonnalité dans les données, le modèle $SARMA$ prend en compte cette saisonnalité. De plus, le modèle $SARMAX$ permet d'intégrer des covariables ou variables exogènes pour tenir compte d'autres facteurs qui peuvent influencer la série chronologique étudiée. L'équation d'un modèle $SARMAX(p, q, P, Q)_s$ avec K covariables est :

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{K} \beta_k x_{t,k} + \sum_{i=1}^{P} \phi_{is} y_{t-is} + \sum_{j=1}^{Q} \theta_{js} \epsilon_{t-js} + \epsilon_t$$

où p est le nombre de termes autorégressifs ; q est le nombre de termes de moyenne mobile ; P est le nombre de termes autorégressifs saisonniers ; Q est le nombre de termes de moyenne mobile saisonniers ; s est la période saisonnière. Bien que les modèles de la famille $ARMA$ soient largement utilisés pour modéliser et prévoir les séries temporelles, ils présentent également certains inconvénients. L'un des principaux inconvénients de ces modèles est leur capacité limitée à modéliser des données complexes, en particulier celles qui présentent des tendances non linéaires ou des relations non stationnaires. En effet, ces modèles reposent sur des hypothèses strictes telles que l'hypothèse de stationnarité de la série temporelle, l'indépendance des erreurs et la normalité de la distribution de ces dernières. Si ces hypothèses ne sont pas satisfaites, ce qui est souvent le cas pour les données économiques, météorologiques et de l'énergie, les prévisions du modèle peuvent être biaisées ou inexactes. Un autre inconvénient à ces modèles est leur complexité de calcul, qui peut être un défi dans certaines applications. En effet, les modèles de famille $ARMA$ peuvent être relativement complexes et nécessitent des calculs intensifs pour estimer les paramètres et générer des prévisions. Cela peut entraîner des temps de calcul élevés, en particulier pour les séries temporelles de long historique ou pour les modèles de haut ordre.

Les modèles de la famille $ARMA$ notamment les modèles $SARIMA$ et $SARIMAX$ (Box et al., 1970) ont été largement utilisés dans la littérature pour modéliser et prévoir la charge électrique (Fathi, 2019 ; Leite Coelho da Silva et al., 2022). Ces modèles sont souvent utilisés comme des modèles de référence dans la prévision de la charge électrique, car ils sont simples à interpréter. De plus, étant donné qu'ils sont largement utilisés, ils permettent de comparer facilement la performance d'autres modèles plus sophistiqués. En utilisant les modèles de famille $ARMA$ comme point de référence, les chercheurs évaluent si les autres modèles de prévision ont une performance supérieure ou non. Cela permet

également de déterminer si la complexité supplémentaire des autres modèles est justifiée par une amélioration significative de la performance de la prévision.

3.2.1.2 Modèle additif généralisé

Le modèle additif généralisé, abrégé en *GAM* pour *General Additive Model*, est une extension du modèle linéaire généralisé, abrégé en *GLM* pour *Generalized Linear Model* (Nelder et al., 1972), permettant de traiter des relations non linéaires. Ces modèles, introduits par Trevor J Hastie et Tibshirani (1990), se caractérisent à la fois par leur flexibilité et leur simplicité. Dans cette partie, nous supposons que nous avons p variables explicatives notées $X_1, \ldots, X_p$. Le modèle additif généralisé le plus simple prend la forme suivante :

$$Y = \beta_0 + \sum_{j=1}^{p} m_j(X_j) + \epsilon, \tag{3.8}$$

ou encore

$$\forall \quad i = 1, \ldots, n, \quad \mathbf{y}_i = \beta_0 + \sum_{j=1}^{p} m_j(\mathbf{x}_{ij}) + \epsilon_i, \tag{3.9}$$

où $\mathbf{y}_i$ est la valeur de la $i^{\text{ème}}$ observation de Y, $\mathbf{x}_{ij}$ la $i^{\text{ème}}$ observation de la $j^{\text{ème}}$ variable explicative X_j, m_j sont les fonctions lisses des variables X_j. Dans ce modèle la relation entre Y et chaque variable explicative X_j est additive (chaque variable X_j est modélisée séparément par la fonction m_j). Cette propriété d'additivité constitue un point fort de ce modèle du point de vue de l'interprétation puisqu'elle permet d'isoler l'effet de chaque variable explicative et de l'analyser indépendamment des autres variables. Le modèle *GAM* permet également d'intégrer des interactions entre les variables explicatives [1]. Un autre point fort de ce modèle par rapport aux autres modèles est qu'il donne la possibilité d'ajuster un modèle spécifique paramétrique ou non paramétrique à chaque variable explicative. La partie paramétrique du modèle peut prendre n'importe quelle forme de modèle paramétrique standard. Quant à la partie non paramétrique chaque fonction m_j peut être estimée par l'une des méthodes d'estimation présentées dans l'annexe 6.2. Par contre, il est préférable d'utiliser la même méthode d'estimation pour toutes les parties non paramétriques du modèle. Les fonctions m_j peuvent ainsi prendre plusieurs formes différentes comme les fonctions *splines*, les fonctions polynomiales ou *loess* (voir l'annexe 6.2) ainsi que les fonctions linéaires. Dans le cas où certaines fonctions m_j prennent une forme linéaire le modèle est dit semi-paramétrique. Un exemple d'une forme générale du modèle *GAM* est donné par l'équation suivante :

1. Il est possible d'intégrer des interactions entre deux variables explicatives sous la forme $m_k(X_j, X_{j'})$ dans le modèle *GAM* par l'intermédiaire des fonctions de lissage bivariées. Ils existent plusieurs fonctions de lissage bivariées d'efficacité variable dans la littérature capables d'estimer cette relation. Le lecteur intéressé par plus d'informations sur les fonctions de lissages bivariées peut se référer à la page 264 dans Trevor J Hastie (2017).

$$\mathbf{y}_i = \overbrace{\underbrace{\beta_0 + \beta_1 \mathbf{x}_{i1}}_{\substack{m_1(\mathbf{x}_{i1})\text{linéaire} \\ \text{(paramétrique)}}} + \underbrace{m_2(\mathbf{x}_{i2}) + \underbrace{m_3(\mathbf{x}_{i3}, \mathbf{x}_{i4})}_{\substack{\text{Interaction} \\ \text{entre } \mathbf{x}_{i3} \text{ et } \mathbf{x}_{i4}}} + \cdots + m_p(\mathbf{x}_{ip})}^{\text{(non paramétrique)}} + \epsilon_i}^{\text{modèle semi-paramétrique}}, \quad i = 1, \ldots, n. \tag{3.10}$$

L'estimation des paramètres du modèle *GAM* implique la recherche simultanée des fonctions m_j dans l'équation (3.9). Plusieurs méthodes sont disponibles pour estimer ces fonctions. L'une des plus connues est l'algorithme de *backfitting* proposé par Hastie et Tibshirani (Trevor J Hastie et Tibshirani, 1990), qui est implémenté dans le `package gam` du logiciel R. Les étapes de cet algorithme ont été résumées par Hastie et Tibshirani (RJ, 1990) de la manière suivante :

1. **La première étape** est l'étape de l'initialisation qui consiste à estimer la constante β_0 par $\hat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$ et les fonctions m_j par $\hat{m}_j^{(0)} = 0$.
2. **La deuxième étape** est l'étape itérative pendant laquelle à l'itération i la fonction $\hat{m}_j^{(i+1)}$ est estimée par une fonction lisse appliquée aux résidus de la façon suivante :

$$\hat{m}_j^{(i+1)} = Sm\left(Y - \hat{\beta}_0 - \sum_{k \neq j} \hat{m}_k^{(i)}(X_k)\right), \quad \text{où } j = 1, \ldots, p, \tag{3.11}$$

   où Sm est une fonction de lissage.
3. **La troisième étape** consiste à mettre à jour les fonctions $\hat{m}_j$ jusqu'à ce que $\hat{m}_j^{(i+1)} = \hat{m}_j^{(i)}$ pour tout j variant de $1, \ldots, p$. Sinon, retourner à la deuxième étape jusqu'à ce que cette condition se vérifie.

Cependant, le *backfitting* n'est pas la seule méthode utilisée pour estimer les paramètres du modèle *GAM*. Une autre technique couramment employée est la méthode des moindres carrés pondérés itérativement pénalisés abrégée en *P-IRLS* notamment parce qu'elle inclut l'estimation du degré de lissage pour les fonctions m_j contrairement à la méthode précédente. Dans cette méthode les fonctions m_j sont représentées à l'aide de *splines* de régression (voir l'annexe 6.2) et les paramètres du modèle *GAM* peuvent être ajustés en utilisant une approche similaire à celle des modèles linéaires généralisés (*GLM*). Cette méthode est implémentée dans le `package mgcv` du logiciel R. Elle permet de contrôler la régularité pour chaque terme grâce à un ensemble de pénalités appliquées à la vraisemblance du modèle *GLM*. Les paramètres du modèle *GAM* sont alors estimés par minimisation itérative du problème suivant :

$$-\ell(\Theta) + \sum_j \lambda_j \Theta^T S_j \Theta, \tag{3.12}$$

où Θ est le vecteur de paramètres à estimer qui contient β_0 (voir l'équation (3.8)) et les coordonnées β_j de m_j dans sa base de *splines*, S_j sont des matrices qui dépendent uniquement des fonctions de base, ℓ est la log-vraisemblance pour Θ. Les paramètres λ_i sont des paramètres de lissage qui contrôlent le compromis entre l'ajustement et la régularité, et ils peuvent être sélectionnés par minimisation du critère de validation croisée généralisée *GCV* (voir Wood (2001)) ou par l'approche de modèle mixte via le maximum de vraisemblance restreinte (*REML*) (voir Wood (2006)).

Malgré la flexibilité et les points forts de ce modèle, il reste très dépendant de l'avis de l'expert qui le bâtit. En plus, il demande une bonne connaissance sur les relations entre les variables explicatives puisque un ajustement inapproprié des interactions entre ces dernières peut donner des résultats non satisfaisants.

Les modèles *GAM* ont été utilisés pour prédire la charge électrique à différents niveaux d'agrégation, de la charge nationale (Pierrot et al., 2011) à la charge régionale et locale (Gerossier et al., 2017). Les performances du modèle *GAM* ont été évaluées sur des données de différents marchés de l'électricité, tels que le marché de l'électricité en France (Pierrot et al., 2011), le marché de l'électricité en Australie (Fan et al., 2012).

3.2.1.3 Régression multivariée par *spline* adaptative

La méthode de régression multivariée par *spline* adaptative, abrégée en *MARS* pour *Multivariate Adaptive Regression spline* a été développée par Friedman (1991) utilisant des *splines* de régression multivariées. Le terme « adaptative » signifie que le modèle est capable de s'adapter automatiquement aux données en déterminant le nombre de fonctions de base à utiliser, leur degré, ainsi que le nombre et l'emplacement des nœuds. Cela permet de construire un modèle non paramétrique flexible qui peut capturer des relations complexes entre les variables explicatives et la variable à prédire sans avoir à spécifier à l'avance un modèle fonctionnel ou une forme de distribution pour les données. En d'autres termes, la méthode est capable de s'adapter automatiquement à la complexité des données, ce qui la rend très utile dans les cas où la relation entre les variables est inconnue ou difficile à modéliser avec un modèle paramétrique traditionnel. Ce modèle permet également d'introduire des termes d'interaction entre les variables.

Cette méthode généralise la régression par partitionnement récursif et la modélisation additive (Friedman, 1991). La régression par partitionnement récursif consiste à diviser l'espace des variables explicatives en régions pour construire un modèle local approximatif pour chaque région à l'aide des fonctions polynomiales. Le modèle final est alors la somme de tous les modèles locaux construits dans chaque région. En utilisant cette méthode itérative, l'algorithme trouve le meilleur endroit pour couper le domaine et construire le modèle local qui décrit au mieux la relation entre la variable à expliquer et les variables

explicatives.

Supposons que $\mathbf{X} = (X_1, \ldots, X_p)$ représente les p variables explicatives de régression sur un certain domaine $D \subset \mathbb{R}^p$. La régression par partitionnement consiste alors à décomposer le domaine D des variables explicatives en r sous-domaines, puis à estimer une fonction de base sur chacun de ces sous-domaines. L'estimateur final de la fonction de régression m est alors formé par l'estimation des différentes fonctions de base sur tous les sous-domaines de D. L'estimateur de la fonction de régression m au point $\mathbf{x}$ prend alors la forme :

$$\hat{m}(\mathbf{x}) = \mathbf{P}_m(\mathbf{x}|(\alpha_i)_{i\in\{1,\ldots,p\}}) \quad \text{si} \quad \mathbf{x} \in D_m, \tag{3.13}$$

où les D_m sont les sous-domaines disjoints de D, tels que $D = \bigcup_{m=1}^{r} D_m$ et les $\alpha_i, \forall i \in \{1, \ldots, p\}$ sont les p coefficients de la fonction polynomiale $\mathbf{P}_m$ à estimer sur chaque sous-domaine D_m. De manière plus générale, les fonctions $\mathbf{P}_m$ peuvent adopter n'importe quelle forme paramétrique simple y compris des fonctions en escalier, des fonctions linéaires ou des constantes comme dans le cas des arbres de régression.

En imposant une contrainte qui assure la continuité de l'estimateur et de ses dérivées, il est possible d'utiliser cette technique pour obtenir un estimateur de type *spline* de régression (voir l'annexe 6.2). Les informations présentes dans la suite sont tirées de la référence Friedman et Roosen (1995). Pour développer la méthode *MARS*, Friedman et Roosen (1995) ont appliqué ce principe en utilisant des bases de fonctions linéaires tronquées. Cette méthode comporte deux étapes.

La première est une étape itérative dite **étape d'addition** dans laquelle le modèle introduit pas à pas un ensemble de fonctions de base. Ces fonctions prennent la forme des fonctions linéaires tronquées et de leurs produits tensoriels (voir l'équation (3.14)). Ces produits tensoriels permettent d'intégrer les interactions entre les variables explicatives. Ces bases peuvent être présentées par :

$$B_k(\mathbf{x}) = \prod_{j=1}^{J_k} [\pm(x_{v(j,k)} - \xi_{jk})]_+ \quad \text{avec} \tag{3.14}$$

$$(x - \xi)_+ = \begin{cases} x - \xi & \text{si } x \geq \xi \\ 0 & \text{si } x < \xi, \end{cases}$$

où B_k est la $k^{\text{ème}}$ fonction de base, J_k est le nombre de fonctions linéaires tronquées multipliées, ou autrement dit, le degré d'interaction entre les variables explicatives. $x_{v(j,k)}$ représentent les variables explicatives avec $1 \leqslant v(j,k) \leqslant n$ et ξ_{jk} sont les nœuds de coupure sélectionnés pour chaque variable explicative $x_{v(j,k)}$. À chaque itération le modèle ajoute

deux nouvelles fonctions de base. Ces deux fonctions de base sont choisies de façon à améliorer l'ajustement du modèle aux données. Cette procédure itérative se poursuit jusqu'à atteindre un nombre limite de bases. Dans la deuxième étape dite **étape d'élimination**, un ensemble final de fonctions de base est sélectionné parmi toutes les bases calculées dans la première étape. L'étape d'addition commence par l'ajustement du modèle par une seule base $B_0(\mathbf{x}) = 1$ pour chaque variable explicative $X_j, j \in \{1, \ldots, p\}$. Ensuite, le modèle sélectionne le meilleur nœud ou point de coupure, parmi tous les nœuds possibles d'une façon itérative ainsi que les deux nouvelles bases à intégrer dans le but d'améliorer son ajustement aux données. Un exemple du passage de la $M^{\text{ème}}$ itération à la $M+1^{\text{ème}}$ itération permet de clarifier cette procédure.
Après la $M^{\text{ème}}$ itération, le modèle est formé de $2M+1$ fonctions de base pour chaque variable explicative chacune ayant la forme de (3.14).

$$(B_k(\mathbf{x}))_0^{2M} = (B_k(\mathbf{x}))_{0 \leqslant k \leqslant 2M}. \tag{3.15}$$

L'objectif dans la $M+1^{\text{ème}}$ itération est de sélectionner les deux bases à intégrer dans le modèle parmi les bases candidates suivantes :

$$\begin{aligned} B_{2M+1}(\mathbf{x}) &= B_{l(M+1)}(\mathbf{x})[+(x_{v(M+1)} - \xi_{M+1})]_+, \\ B_{2M+2}(\mathbf{x}) &= B_{l(M+1)}(\mathbf{x})[-(x_{v(M+1)} - \xi_{M+1})]_+, \end{aligned} \tag{3.16}$$

où $B_{l(M+1)}(\mathbf{x})$ est l'une des $(2M+1)$ fonctions de base déjà choisies dans les M itérations précédentes alors $0 \leqslant l(M+1) \leqslant 2M$, $v(M+1)$ est une des variables explicatives qui ne figure pas dans les bases $B_{l(M+1)}(\mathbf{x})$ et ξ_{M+1} est le nœud accordé à cette variable. Les trois paramètres $v(M+1)$, le nœud ξ_{M+1} et la base $B_{l(M+1)}(\mathbf{x})$ qui définissent les deux nouvelles bases sont déterminés en minimisant la somme des résidus au carré du modèle suivant :

$$\begin{aligned} (l(M+1), v(M+1), \xi_{M+1}) = \arg\min_{l,v,t} \sum_{i=1}^{N} \Big[\mathbf{y}_i - \sum_{k=0}^{2M} a_k B_k(\mathbf{x}) \\ - a_{2M+1} B_l(\mathbf{x})[+(x_v - \xi)]_+ \\ - a_{2M+2} B_l(\mathbf{x})[-(x_v - \xi)]_+ \Big]^2, \end{aligned} \tag{3.17}$$

où l varie de 0 à M, v varie de 1 à p (le nombre de variables explicatives) et ξ appartient au vecteur des nœuds candidats de la variable $\mathbf{x}_v$. Supposons que l^*, v^* et ξ^* les valeurs estimées des trois paramètres l, v et ξ dans (3.17) alors le modèle ajoute les deux bases suivantes dans cette itération aux bases déjà existantes :

$$\begin{aligned} B_{2M+1}(\mathbf{x}) &= B_{l^*}(\mathbf{x})[+(x_{v^*} - \xi^*)]_+, \\ B_{2M+2}(\mathbf{x}) &= B_{l^*}(\mathbf{x})[-(x_{v^*} - \xi^*)]_+, \end{aligned} \tag{3.18}$$

Cette procédure est répétée jusqu'à atteindre la taille maximale du modèle définie par défaut par $\max(21, 2p+1)$.

Ensuite, la deuxième étape de l'élimination commence une fois que le modèle est obtenu avec une dimension maximale. Durant cette étape, toutes les bases sont candidates pour l'élimination d'une façon itérative à l'exception de la base $B_1(\mathbf{x})$. La base retirée à chaque fois est celle qui son élimination du modèle minimise le critère de validation croisée généralisée (GCV) défini de la façon suivante :

$$GCV(\hat{m}_K(\mathbf{x}), \lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y_i} - \hat{m}_K(\mathbf{x}))^2}{\left[1 - \frac{K+\lambda(K-1)}{n}\right]^2}, \tag{3.19}$$

où λ est le paramètre de lissage qui pénalise la complexité du modèle, K est la dimension du modèle (nombre de bases). Le modèle final est alors le modèle ayant la plus petite valeur de GCV parmi toutes les possibilités. Le modèle *MARS* peut alors être représenté par la forme suivante dite décomposition ANOVA :

$$\hat{m}(\mathbf{x}) = \hat{m}(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \hat{m}_0 + \sum_{i=1}^{p}\hat{m}_i(\mathbf{x}_i) + \sum_{i<j}\hat{m}_{ij}(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i<j<k}\hat{m}_{ijk}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \cdots \tag{3.20}$$

En plus de sa capacité à détecter les relations non linéaires et de prendre en compte les interactions entre les variables explicatives, le modèle *MARS* est relativement rapide à entraîner et à utiliser pour la prévision, ce qui le rend attractif pour les applications en temps réel. Contrairement à certains modèles de prévision plus complexes, le modèle *MARS* est relativement facile à interpréter car il produit des équations de régression simples et explicites.

Des études ont montré que le modèle *MARS* peut fournir des prévisions précises de la charge électrique à court terme et qu'il peut être plus efficace que d'autres modèles de prévision tels que la régression par vecteurs de support (SVR) ou les modèles autorégressifs (ARIMA) (Al-Musaylh et al., 2018).

Les informations présentées dans cette partie ont été extraites de Friedman et Roosen (1995). Pour en savoir plus sur le modèle *MARS*, le lecteur intéressé peut consulter cette référence.

3.2.1.4 Forêts aléatoires

Les forêts aléatoires sont une méthode de statistique non-paramétrique très performante introduite par Breiman (2001) qui permet de résoudre des problèmes de régression

ou des problèmes de classification. Nous présentons, dans ce chapitre, cette méthode uniquement dans le contexte de la régression. Les informations représentées dans cette partie sont tirées de Genuer et al. (2016). Cette méthode fait partie de la famille des méthodes d'ensemble (Dietterich, 2000), tout comme le *Bagging* (Breiman, 1996), le *Boosting* (Freund et al., 1996), le *Randomizing Outputs* (Breiman, 2000) ou encore le *Random Subspace* (Ho, 1998). Le principe général de ces méthodes consiste à construire une collection de prédicteurs, puis à agréger l'ensemble de leurs prévisions.

Considérons une collection de prédicteurs par arbre, représentée par $\left(\hat{h}(.,\Theta_1),\ldots,\hat{h}(.,\Theta_q)\right)$. Dans cette collection, les variables aléatoires $\Theta_1,\ldots,\Theta_q$ sont indépendantes et identiquement distribuées (i.i.d) à partir de l'échantillon de données $\mathcal{D}_n = (X_1,Y_1),\ldots,(X_n,Y_n)$. Chaque paire (X_i,Y_i) est considérée comme une réalisation de n variables aléatoires i.i.d., toutes ayant la même distribution que la paire (X,Y). Ici, X représente les variables explicatives, qui sont des éléments de l'espace vectoriel $\mathbb{R}^p$, tandis que Y est la variable cible à prédire. Le prédicteur des forêts aléatoires noté $\hat{h}_{RF}$ est obtenu en agrégeant cette collection d'arbres aléatoires de la façon suivante :

$$\hat{h}_{RF}(\mathbf{x}) = \frac{1}{q}\sum_{l=1}^{q}\hat{h}(\mathbf{x},\Theta_l). \tag{3.21}$$

Une catégorie particulière de forêts aléatoires se démarque des autres en raison de sa remarquable performance sur de nombreux jeux de données. Cette catégorie est connue sous le nom de forêts aléatoires à variables d'entrée aléatoires, abrégé en *Random Forests-RI* pour *Random Forests with Random Inputs* (Breiman et Cutler, 2004). D'ailleurs, le terme « forêts aléatoires » est souvent utilisé spécifiquement pour faire référence à cette famille dans de nombreuses études.

Le principe de construction d'une forêt *Random Forests-RI* est de générer plusieurs échantillons *bootstrap* $\mathcal{D}_n^{\Theta_1},\ldots,\mathcal{D}_n^{\Theta_q}$ (comme dans le *Bagging*). $\mathcal{D}_n^{\Theta_l}$ est obtenu en tirant n observations avec remise dans l'échantillon $\mathcal{D}_n$. La variable aléatoire Θ_l représente alors le $l^{\text{ème}}$ tirage de l'échantillon *bootstrap*. Ensuite, sur chaque échantillon $\mathcal{D}_n^{\Theta_l}$, une variante des arbres de décision CART est appliquée (voir Breiman, Friedman et al. (2017)). Dans ce contexte, chaque arbre est construit en sélectionnant au hasard un ensemble de m variables et en définissant la meilleure séparation parmi ces m variables. Finalement, la collection d'arbres obtenus est agrégée par moyenne pour donner le prédicteur *Random Forests-RI* (voir l'équation (3.21)).

En outre de leur capacité à modéliser les interactions non linéaires, les forêts aléatoires ont plusieurs avantages. Tout d'abord, elles peuvent gérer efficacement les problèmes de données manquantes ou de bruit dans les données. De plus, les forêts aléatoires sont robustes face aux valeurs aberrantes, ce qui en fait un choix approprié pour les données de charge électrique qui peuvent souvent contenir des anomalies. Elles peuvent également

fournir une estimation de l'importance des variables, ce qui peut aider à comprendre les facteurs clés qui influencent la prévision. Cependant, les forêts aléatoires peuvent être considérées comme une « boîte noire » puisqu'elles ne fournissent pas une relation explicite entre les variables. Cela peut rendre les forêts aléatoires moins utiles dans les contextes où l'interprétabilité est essentielle, ou lorsqu'une compréhension rigoureuse de la relation entre les variables est nécessaire.

Les forêts aléatoires sont largement utilisées pour la prévision de la charge électrique (Lahouar et al., 2015 ; Dudek, 2015 ; N. Huang et al., 2016 ; Q. Huang et al., 2017). Les références citées présentent différentes applications des modèles de forêts aléatoires pour la prévision de la charge électrique dans lesquelles les auteurs ont utilisé des approches innovantes telles que la sélection des caractéristiques d'entrée pertinentes à l'aide d'un ensemble de règles expertes, la décomposition en ondelettes et l'utilisation de la mesure de l'importance de permutation pour améliorer la précision de la prévision. Les résultats obtenus ont montré que les modèles de forêts aléatoires peuvent fournir des résultats précis pour la prévision à court terme de la charge électrique.

Pour plus d'informations sur les forêts aléatoires, le lecteur intéressé peut se référer à Trevor Hastie et al. (2009).

3.2.2 Modèle fonctionnel *KWF*

Le modèle *KWF* pour *Kernel Wavelet Functional* a été développé pour la prévision des séries chronologiques fonctionnelles en présence de non-stationnarités (Antoniadis, Brossat et al., 2014). Sa singularité réside dans son utilisation exclusive des informations des courbes de charge pour générer les prévisions. Il identifie les contextes historiques similaires à la situation actuelle, et calcule des prévisions en moyennant et pondérant les trajectoires futures de ces contextes par des poids de similarité. Cette notion de similarité s'appuie sur l'analyse multirésolution par ondelettes (voir annexe 6.2) (Strang et al., 1996 ; Stephane Georges Mallat, 1988). Ce modèle s'inscrit dans la continuité des travaux antérieurs. En particulier, Poggi (1994) avait proposé pour un problème similaire un prédicteur de nature analogue, mais destiné à un processus multivarié. Par la suite, Antoniadis, Paparoditis et al. (2006) ont établi un cadre pour les processus fonctionnels stationnaires, s'appuyant sur la transformée en ondelettes. Ce cadre a été ensuite adapté aux processus non-stationnaires par Cugliari (2011). Ainsi, le modèle actuel aborde spécifiquement les caractéristiques non-stationnaires des séries temporelles, comme les variations du niveau moyen et l'existence de groupes représentant différentes classes de stationnarité. Il intègre à la fois l'analyse multirésolution (voir annexe 6.2) (Strang et al., 1996 ; Stephane Georges Mallat, 1988) et la régression à noyau fonctionnel (voir annexe 6.2) pour anticiper le futur en tant que combinaison linéaire des données passées. Son approche se base sur des segments de données discrètes, considérés comme des

courbes dans l'espace fonctionnel, avec une décomposition en ondelettes pour générer les prévisions. Le modèle *KWF* offre plusieurs avantages par rapport aux autres méthodes de prévision, notamment en termes de flexibilité, d'adaptabilité.

Afin de mieux comprendre les bases théoriques du modèle *KWF*, nous allons présenter les composantes principales de ce modèle décrites plus en détails dans la publication de Antoniadis, Brossat et al. (2014) et la thèse de Cugliari (2011). Ces deux sources ont également influencé la façon dont le modèle *KWF* est présenté et les notations utilisées.

Considérons un processus stochastique supposé en premier temps stationnaire $(Z_i)_{i\in\mathbb{Z}}$ à valeurs dans un espace fonctionnel H (par exemple $H = L_2([0,1])$) et que nous disposons d'un échantillon de n courbes $Z_1, \ldots, Z_n$. L'objectif du modèle *KWF* est alors de prédire Z_{n+1}.

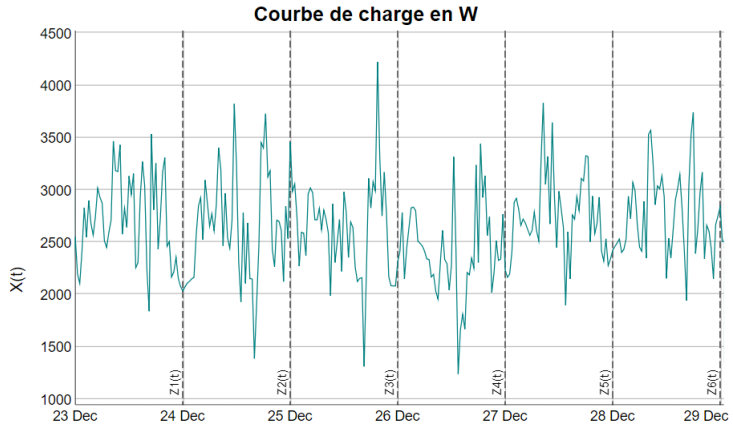


FIGURE 3.1 – Décomposition de la courbe de charge en courbes de charge journalières.

Le modèle comprend deux étapes distinctes. La première étape consiste à identifier parmi les segments du passé $Z_1, \ldots, Z_{n-1}$ ceux qui ressemblent le plus au dernier segment observé Z_n. Cette étape est suivie du calcul des poids de similitude entre ces segments passés et le segment Z_n. La prévision du segment Z_{n+1} est obtenue en effectuant une somme pondérée des segments de $Z_1, \ldots, Z_n$ en utilisant les poids de similitude calculés.

1. **Première étape**. Le modèle *KWF* représente chaque segment $Z_i, i = 1, \ldots, n$, par son développement sur une base d'ondelettes tronqué à une échelle $J > j_0$ afin de prendre en compte dans le calcul de dissimilarité la dimension infinie des segments à comparer. Pour une échelle J supérieure à un certain rang j_0, chaque segment Z_i est représentée par une version tronquée de son développement par la transformée en ondelettes discrète de la façon suivante :

$$Z_{i,J}(t) = \sum_{k=0}^{2^{j_0}-1} a_{j_0,k}^{(i)} \phi_{j_0,k}(t) + \sum_{j=j_0+1}^{J} \sum_{k=0}^{2^j-1} d_{j,k}^{(i)} \psi_{j,k}(t) \quad \text{avec} \quad t \in [0,1] \tag{3.22}$$

avec $a_{j,k} =< g, \phi_{j,k} >$, $d_{j,k} =< g, \psi_{j,k} >$ où $\phi_{j,k}$ est une famille de fonctions d'échelles et $\psi_{j,k}$ une famille d'ondelettes. La première somme dans l'équation de décomposition en ondelettes représente l'approximation à la résolution j_0 de la trajectoire Z_i. Cette approximation est obtenue en utilisant la projection orthogonale de Z_i sur l'espace de fonction généré par les fonctions d'ondelette à la résolution j_0 (voir annexe 6.2). Cette approximation est associée aux basses fréquences de la courbe et donnent une représentation lisse de la courbe Z_i à une résolution donnée.

La deuxième somme dans l'équation représente la somme des coefficients de détails à chaque niveau de résolution allant de $j_0 + 1$ à J. Ces coefficients sont obtenus en utilisant la projection orthogonale de Z_i sur l'espace de fonction généré par les fonctions d'ondelette à chaque résolution allant de $j_0 + 1$ à J (voir annexe 6.2). Ces coefficients donnent des informations sur les changements localisés en temps dans la courbe Z_i à chaque résolution, et sont associés aux hautes fréquences. En d'autres termes, les coefficients d'ondelettes décrivent les variations brusques et locales de la courbe Z_i à chaque résolution.

Les coefficients obtenus de cette décomposition seront utilisés pour construire la prévision de la courbe Z_{n+1} à partir des courbes $Z_1, \ldots, Z_n$ en utilisant la méthode de régression à noyau . En effet, la décomposition en ondelettes permet de représenter les courbes de manière plus compacte en utilisant un nombre réduit de coefficients de base, tout en préservant les informations importantes telles que les changements locaux et globaux de la série temporelle.

2. **Deuxième étape**. Le modèle utilise la méthode de régression à noyau pour modéliser la relations entre les valeurs fonctionnelles passées $Z_1, \ldots, Z_n$ et la future valeur Z_{n+1}. Pour cela, une fonction noyau est utilisée pour mesurer la similarité entre les valeurs fonctionnelles ou segments. Ce noyau est appliqué à chaque paire de valeurs fonctionnelles passées et le dernier segment observé, permettant de calculer un poids qui sera utilisé pour estimer la fonction de régression. La prévision des futurs segments est obtenue en calculant une moyenne pondérée des segments passés, utilisant ces poids. Pour ce faire, une distance de dissimilarité, notée D, est définie afin de mesurer la distance entre les deux segments à Z_l et Z_m à partir des coefficients de détails $d_{j,k}$ de manière unique de la façon suivante :

$$D(Z_l, Z_m) = \sum_{j=j_0+1}^{J} 2^{-j/2} \left(\sum_{k=0}^{2^j-1} \left(d_{j,k}^{(l)} - d_{j,k}^{(m)} \right)^2 \right)^{1/2}. \tag{3.23}$$

Comme le processus Z est supposé ici stationnaire, les coefficients d'approximation ne contiennent pas d'information utile pour la prévision (ils fournissent des moyennes locales) et sont donc considérés comme inutiles dans le calcul de la dissimilarité entre deux segments Z_l et Z_m. Cette mesure de dissimilarité D permet alors d'identifier les segments ayant des motifs similaires même en ayant des approximations différentes (Antoniadis, Paparoditis et al., 2006).

Notons $\Xi_i = \left(a_{J,k}^{(i)}\right)_{0\leqslant k\leqslant 2^J-1}$ le vecteur des coefficients d'échelle du i-ème segment Z_i à la résolution J, la plus fine. La prévision des coefficients d'échelle (à l'échelle J), notée $\widehat{\Xi}_{n+1}$, de Z_{n+1} est calculée par :

$$\widehat{\Xi}_{n+1} = \frac{\sum_{m=1}^{n-1} K_{h_n}(D(Z_{n,J}, Z_{m,J}))\Xi_{m+1}}{\frac{1}{n} + \sum_{m=1}^{n-1} K_{h_n}(D(Z_{n,J}, Z_{m,J}))}. \tag{3.24}$$

où $K_{h_n}(.) = h_n^{-1}K(./h_n)$, K est le noyau de probabilité et h_n la largeur de la fenêtre associée à la régression à noyau. L'application de la transformée inverse de la transformée d'ondelettes sur $\widehat{\Xi}_{n+1}$ permet de prédire le segment Z_{n+1} de la façon suivante par :

$$\widehat{Z}_{n+1}(t) = \sum_{k=0}^{2^J-1} \widehat{a_{J,k}^{(n+1)}}\phi_{J,k}(t). \tag{3.25}$$

Il est également possible de réécrire le prédicteur comme un barycentre des futurs des segments du passé grâce aux poids $w_{n,m}$:

$$\widehat{Z}_{n+1}(t) = \sum_{m=1}^{n-1} w_{n,m}Z_{m+1}(t) \quad \text{où} \quad w_{n,m} = \frac{K_{h_n}(D(Z_{n,J}, Z_{m,J})}{\sum_{m=1}^{n-1} K_{h_n}(D(Z_{n,J}, Z_{m,J}))}. \tag{3.26}$$

Le choix de la distance de dissimilarité décrite ci-dessus ainsi que la méthode de calcul des prévisions reposent sur l'hypothèse que le processus Z est stationnaire. Cependant, il est souvent difficile de vérifier cette hypothèse sur des données réelles, la raison pour laquelle Cugliari (2011) a proposé dans sa thèse des correctifs afin de permettre au modèle *KWF* de traiter des données non-stationnaires. L'efficacité des correctifs proposés a été testée pour deux types de non-stationnarités : l'évolution du niveau moyen des approximations et l'existence des classes similaires des segments Z_{m+1} dans les données de consommation d'électricité en France.

En effet, dans le cas où les segments Z_{m+1} présentent des niveaux moyens très différents, cela peut affecter la qualité de la prévision. Pour remédier à cela, l'auteur dans la thèse Cugliari (2011) propose de centrer les segments avant de procéder au calcul de la prévision. L'auteur explique que le centrage est présent d'une façon implicite dans la partie du modèle qui consiste à chercher les segments similaires du passé puisque les coefficients d'approximation ne sont pas pris en compte dans le calcul de la distance de dissimilarité alors que le centrage doit être effectué d'une manière explicite dans le calcul de la prévision. Autrement dit, si nous reprenons la transformation d'ondelettes discrète du segment Z_i :

$$Z_i(t) = \sum_{k\in\mathbb{Z}} a_{j_0,k}^{(i)}\phi_{j_0,k} + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} d_{j,k}^{(i)}\psi_{j,k} = S_i(t) + D_i(t), \tag{3.27}$$

où $S_i(t)$ la partie qui représente l'approximation et $D_i(t)$ celle des détails.

Si Z est un processus centré alors $S_i(t) = 0$ et $Z_i(t) = D_i(t)$ et par suite le modèle *KWF* de base est applicable. Sinon, la prévision du segment Z_{n+1} s'écrit alors comme

la somme de la prévision de S_{n+1} et de D_{n+1} de la façon suivante :

$$\widehat{Z}_{n+1}(t) = \widehat{S}_{n+1}(t) + \widehat{D}_{n+1}(t). \tag{3.28}$$

$\widehat{D}_{n+1}$ est calculé dans le modèle KWF de base comme étant :

$$\widehat{D}_{n+1}(t) = \sum_{m=1}^{n-1} w_{n,m} D_{m+1}(t) \tag{3.29}$$

Pour le calcul $\widehat{S}_{n+1}$, Poggi (1994) propose une méthode qui consiste à ajuster la transition entre les niveaux moyens d'un jour et de son lendemain par :

$$\widehat{S}_{n+1}(t) = S_n(t) + \sum_{m=1}^{n-1} w_{n,m} \triangle (S_{m+1})(t), \tag{3.30}$$

où $\triangle(S_{m+1})(t) = S_{m+1}(t) - S_m(t)$.

Cette méthode de correction s'est avérée la plus performante parmi plusieurs méthodes testées sur les données de consommation d'électricité en France dans Cugliari (2011).

Pour résoudre le problème de l'existence des groupes de similitude dans les segments Z_m, l'auteur de l'étude Cugliari (2011) a proposé une redéfinition du poids $w_{n,m}$ calculé entre les deux segments Z_n et Z_m. Cette redéfinition permet de restreindre le calcul de la similitude aux segments appartenant au même groupe. Le poids $w_{n,m}$ est alors redéfini par $\tilde{w}_{n,m}$ de la façon suivante :

$$\tilde{w}_{n,m} = \frac{w_{n,m} \mathbb{1}_{\{gr(m)=gr(n)\}}}{\sum_{m=1}^{n} w_{n,m} \mathbb{1}_{\{gr(m)=gr(n)\}}}, \tag{3.31}$$

où $gr(n)$ représente le groupe du segment Z_n. Par suite, le poids de similitude est calculé uniquement pour les deux segments Z_n et Z_m si elles appartiennent au même groupe. Ces groupes peuvent être des groupes déterministes connus ou des groupes à déterminer à partir des méthodes de *clustering*.

L'application du modèle KWF à la prévision de la charge électrique a donné des résultats prometteurs Cugliari (2011). Selon l'auteur, le modèle KWF s'est montré compétitif par rapport aux autres modèles de prévision utilisés pour la prévision de la charge électrique à l'entreprise EDF, en termes de précision et d'efficacité. En effet, l'utilisation de la technique de régression à noyau fonctionnel combinée à la décomposition en ondelettes dans le modèle présente plusieurs avantages sur la prévision de la charge électrique. Tout d'abord, la prévision des courbes de charge sous forme des données fonctionnelles permet de prendre en compte leur structure intrinsèque plutôt que de les traiter comme des observations indépendantes. De plus, la décomposition en ondelettes permet d'identifier les changements locaux dans la structure de ces données fonctionnelles, ce qui peut aider à modéliser et prévoir les variations locales avec plus de précision. Cela est particulièrement utile pour

la prévision de la charge électrique, où les variations locales peuvent être importantes et doivent être prises en compte pour obtenir une bonne précision de la prévision.

Ce modèle de prévision peut être utilisé également comme un outil important pour l'imputation des valeurs manquantes dans les données de courbe de charge. En effet, les méthodes classiques d'imputation comme la moyenne ou la régression peuvent conduire à des biais importants et à une perte de précision. En revanche, les modèles de prévision des processus fonctionnels peuvent prendre en compte la structure temporelle des données et exploitent les informations des segments passés pour estimer les valeurs manquantes de manière plus précise. Cependant, il est important de noter que le modèle *KWF* a également des limites. Par exemple, il peut être plus complexe que certains d'autres modèles de prévision, puisqu'il nécessite une compréhension approfondie de la théorie des ondelettes et de la théorie de la régression à noyau fonctionnel, ce qui peut rendre son utilisation plus difficile. En outre, le modèle est limité dans sa capacité à intégrer des données de variables exogènes telles que la température.

Pour plus d'informations sur le modèle *KWF* le lecteur peut se référer à Cugliari (2011) et à Antoniadis, Brossat et al. (2014).

3.2.3 Réseaux de neurones

Les réseaux de neurones, également appelés réseaux de neurones artificiels (ANN), sont un type particulier d'algorithmes d'apprentissage automatique dont l'architecture et le fonctionnement sont inspirés du mode de fonctionnement du cerveau humain (Goodfellow et al., 2016).

Ces algorithmes sont devenus très populaires ces dernières années, en raison de leur capacité à résoudre des problèmes complexes de manière très efficace. En effet, les réseaux de neurones ont montré des performances élevées dans de nombreuses applications d'apprentissage automatique comme le traitement du son et de l'image, notamment la reconnaissance faciale et la classification de texte. Contrairement à de nombreuses approches statistiques, ils n'exigent aucune connaissance préalable sur la nature des relations entre les variables. Ils sont également capables de détecter automatiquement des relations non linéaires et complexes entre les variables d'entrée et de sortie. Il existe plusieurs types d'architectures des réseaux de neurones :

1. **les perceptrons multicouches (MLP)**, sont les premiers et les plus simples types de réseaux de neurones. Ce sont des réseaux à propagation directe (*feedforward*), ce qui signifie que le flux d'informations se déplace de la couche d'entrée vers la couche de sortie. Chaque couche dans ce réseau est entièrement connectée (*fully connected*), ce qui signifie que chaque neurone dans une couche est connecté à tous les neurones de la couche suivante. En revanche, il n'y a pas de connexion entre les neurones d'une

même couche (Nielsen, 2015).

2. **les réseaux de neurones convolutifs (CNN)**, ils sont conçus particulièrement pour le traitement d'images. Il s'agit également des réseaux à propagation directe (*feedforward*) qui consistent en un empilage multicouche de neurones. Le réseau CNN le plus simple est constitué de trois principaux types de couches : la couche convolutive, la couche de regroupement (*pooling*) et la couche entièrement connectée (*fully connected*). La couche convolutive permet d'extraire les caractéristiques ou *features* présentes dans les données. La couche de regroupement sert à réduire la taille des données (sous-échantillonnage) dans l'objectif de réduire le surapprentissage (Le Cun et al., 1989).
3. **les réseaux de neurones récurrents (RNN)**, ils conviennent en particulier au traitement des séries temporelles. Contrairement, aux réseaux *feedforward*, les réseaux RNN peuvent avoir des flux d'informations dans les deux sens en introduisant des boucles dans le réseau (Goodfellow et al., 2016).

Les réseaux de neurones sont constitués essentiellement des **neurones** appelées « neurones formels » qui jouent le rôle des unités de calcul. Les neurones sont empilés dans des couches. La structure générale d'un réseau de neurones est formée d'une couche d'entrée, une ou plusieurs couches cachées, ainsi qu'une couche de sortie (voir la figure 3.2).

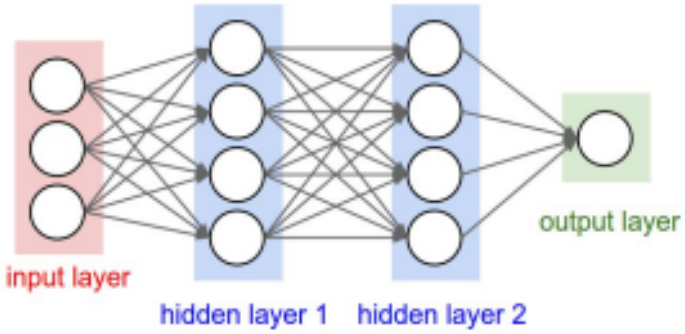


figure 3.2 – Architecture d'un réseau de neurones artificiel (MLP) [Source WikiStat].

La couche d'entrée apporte les données initiales dans le réseau et les envoie aux couches suivantes. Un poids est affecté à chaque variable d'entrée permettant ainsi de déterminer le degré d'importance de cette dernière. Les poids les plus importants contribuent d'une façon plus significative à la sortie par rapport aux autres entrées. Les entrées sont multipliées par leurs poids respectifs, puis additionnées. Ensuite, ces valeurs passent à travers une fonction dite **fonction d'activation**. Cette fonction sert à introduire une non linéarité dans les relations entre les variables. Si ces valeurs calculées par les neurones sont supérieures à la valeur seuil spécifiée, ces neurones sont activés et ces valeurs sont envoyées à la couche suivante du réseau. Sinon, le neurone ayant une valeur inférieure à la valeur seuil est désactivé et aucune donnée n'est transmise de sa part à la couche suivante. Les valeurs

transmises deviennent les données d'entrée de la nouvelle couche et ainsi de suite jusqu'à la couche finale. Le résultat final est obtenu à partir de la couche de sortie (voir la figure 3.3).

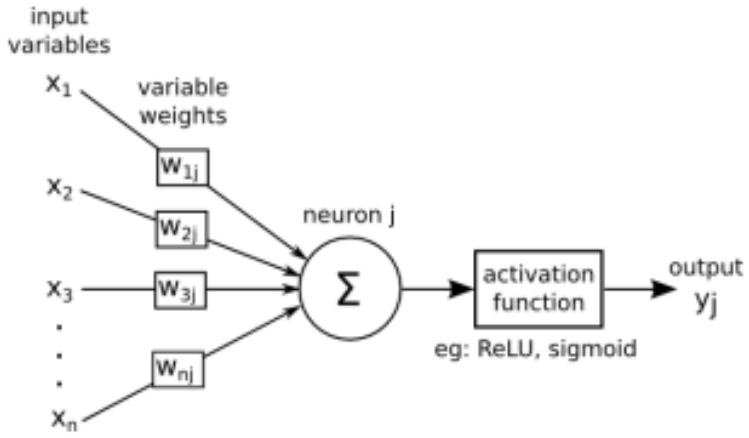


FIGURE 3.3 – Propagation de l'information dans un réseau de neurones classique[Source WikiStat].

Pour simplifier, prenons l'exemple d'un seul neurone artificiel j (voir la figure 3.3). Le neurone j prend comme entrée le vecteur $x = (x_1, \ldots, x_n)$, sa sortie y_j est définie en fonction d'un vecteur de poids $w_j = (w_{j,1}, \ldots, w_{j,n})$, d'une fonction d'activation σ ainsi que d'un biais noté b_j de la façon suivante :

$$y_j = \sigma(\sum_{i=1}^{n} w_{j,i}x_i + b_j).$$

La fonction σ peut prendre les formes suivantes :

1. la fonction identité : $\sigma(u) = u$.
2. la fonction sigmoïde : $\sigma(u) = \frac{1}{1+e^{-u}}$.
3. la fonction unité linéaire rectifiée (ou ReLU) : $\sigma(u) = max(0, u)$.
4. la fonction hyperbolique (tanh) : $\sigma(u) = \frac{e^{2u}-1}{e^{2u}+1}$.

L'étape d'entraînement du réseau, consiste à ajuster les poids des connexions en fonction des erreurs de prévision après le passage de chaque instance de données dans le réseau. Le réseau évalue donc ses performances après chaque prévision. Si la performance n'est pas supérieure à celle de l'étape précédente, les poids du réseau sont alors ajustés afin d'améliorer le résultat. L'algorithme utilisé pour l'ajustement des poids est l'algorithme de rétropropagation du gradient (*backpropagation*), basé sur la descente du gradient (RUMELHART et al., 1986).

Malgré leur succès, l'interprétabilité des réseaux de neurones reste un enjeu majeur dans le domaine de l'apprentissage automatique. En effet, les réseaux de neurones sont souvent considérés comme des modèles de « boîte noire » car il est difficile de comprendre

comment ils prennent des décisions à partir de leurs entrées. Cela peut constituer un obstacle à leur adoption dans certaines applications où la justification des résultats est essentielle vis à vis des utilisateurs telles que les services personnalisés dans le domaine de la santé, de l'assurance et de la sécurité.

La prévision de la charge électrique par les réseaux de neurones est un sujet de recherche actif dans la littérature. De nombreuses études ont été publiées sur ce sujet, proposant différentes approches pour améliorer la précision des prévisions. Par exemple, certaines études ont utilisé des architectures de réseau de neurones spécifiques, telles que les réseaux de neurones récurrents (RNN) (Biswas et al., 2016 ; Shi et al., 2017) ou les réseaux de neurones convolutionnels (CNN) (Voss et al., 2018) pour modéliser les données de la charge électrique. Certaines études ont également combiné les réseaux de neurones avec d'autres techniques de prévision telles que les séries chronologiques floues (Sadaei et al., 2019) et les machines à vecteurs de support (Niu et al., 2005). Dans l'ensemble, les résultats de ces études montrent que les réseaux de neurones peuvent fournir des prévisions précises de la charge électrique. Cependant, il est important de noter que les performances de la prévision dépendent de la quantité et de la qualité des données d'entrée ainsi que l'architecture du réseau de neurones utilisé.

Pour plus d'informations sur les réseaux de neurones, le lecteur peut se référer à Trevor Hastie et al. (2009) et à Goodfellow et al. (2016).

3.3 Critères de sélection des modèles de prévision

La procédure générale de modélisation et de prévision de la charge électrique est décrite par Gross et al. (1987). Tout d'abord, elle commence par la sélection des modèles de prévision appropriés selon des critères qui permettent de répondre essentiellement aux objectifs recherchés par la prévision. Ces objectifs peuvent être différents pour les mêmes données et à la même échelle en fonction des applications envisagées. C'est la raison pour laquelle des modèles proposés pour certaines applications ne sont pas appropriés pour d'autres applications. Selon Makridakis, Wheelwright et al. (1998), les critères de sélection des modèles de prévision dans un contexte industriel doivent être déterminés en fonction des besoins et des objectifs de l'entreprise ainsi que du contexte commercial spécifique dans lequel elle opère. Les auteurs mettent en avant l'importance de l'avis des experts métiers qui peuvent contribuer efficacement à l'identification des critères les plus pertinents. Dans notre cas, ces critères ont été identifiés par des experts métiers en réponse aux besoins spécifiques de l'entreprise. Ces critères se résument par :

1. **La précision** : ce critère est très intéressant puisqu'en premier lieu notre objectif est de donner une analyse prédictive fiable et interprétable pour chaque client à partir de ses propres données de consommation. Par contre, les modèles qui assurent une

bonne précision n'ont pas forcément les meilleures performances puisque souvent la notion de la performance est contrainte par d'autres critères.

2. **La rapidité** : le temps de calcul du modèle de prévision joue un rôle crucial dans notre cas puisque le modèle doit effectuer des prévisions quotidiennes dans un délai raisonnable pour un nombre élevé de ménages. Comme le temps de calcul dépend du nombre de ménages, de l'horizon de prévision ainsi que du modèle de prévision, un modèle ayant une bonne précision mais très lent ne peut pas être utilisé.
3. **La flexibilité** : ce modèle doit être capable de modéliser des relations non-linéaires notamment entre la charge électrique et la température. Une raison pour laquelle la littérature sur ce sujet est orientée vers les modèles de régression non-paramétriques et les modèles de l'apprentissage automatique comme les réseaux de neurones qui sont plus flexibles que les modèles paramétriques et moins contraignants.
4. **L'adaptabilité et la robustesse** : le modèle doit être facilement adaptable pour de nombreuses typologies de ménages et pour différents profils de courbes de charge sans aucune intervention humaine. Par exemple, le mode de vie dynamique des jeunes ne ressemble pas au mode de vie des personnes âgées retraitées ou celui d'une famille avec des enfants par rapport aux heures de présence au domicile et aux habitudes de consommation. En plus, le modèle doit être robuste face aux multiples sources de changements dans les données comme les périodes de vacances, l'augmentation du niveau de consommation due par exemple à l'augmentation du nombre des occupants ou au changement des équipements électriques ainsi que le changement des habitudes de consommation (faire du télétravail, passage à la retraite, ...).
5. **L'interprétabilité** : bien que la plupart des recherches académiques se concentrent sur la précision comme critère principal de sélection du modèle de prévision, dans le milieu industriel, l'interprétabilité et la facilité d'utilisation sont autant importantes que la précision puisqu'elles facilitent la prise de décision. La raison pour laquelle les modèles simples comme les modèles de persistance, les modèles de régression paramétrique et les arbres de décision sont largement utilisés dans l'industrie. Nous avons alors une préférence pour les modèles compréhensibles qui peuvent être facilement utilisés et maintenus par des non spécialistes au lieu d'un modèle « boite noire » est incontestable. Dans notre approche, nous avons testé également des modèles type « boîte noire » dans l'objectif de vérifier si une probable amélioration justifie leur utilisation au détriment de l'interprétabilité dans le contexte de la prévision de la charge électrique à l'échelle des ménages.
6. **L'autonomie** : le modèle doit être capable de fournir automatiquement des prévisions quotidiennes pour tous les ménages sans aucune intervention des experts.

La stratégie que nous avons adoptée est courante, elle consiste à appliquer plusieurs techniques de prévision de différents degrés de complexité et de les comparer à des modèles de référence simples de la littérature comme les méthodes de prévision naïve qui consistent à attribuer la valeur des données de la période précédente à la prévision de la période

suivante. L'objectif que nous visons par la diversité des techniques et des modèles proposés est de pouvoir mettre en œuvre une approche qui tient en considération le plus possible l'ensemble de ces critères. En plus, compte tenu de l'hétérogénéité des courbes de charge et de profils des consommateurs la diversité des approches proposées permet l'exploitation de la possibilité d'attribution d'un modèle par classe ou par profil de consommateurs puisque rien ne garantit l'existence d'un seul modèle capable de répondre à nos besoins en tenant compte de tous ces critères.

3.4 Évaluation des modèles de prévision

Une fois que les modèles de prévision sont mis ont place, il est nécessaire de définir des indicateurs de performance qui permettent de les évaluer. Bien que les critères comme la flexibilité, l'adaptabilité et la robustesse, l'interprétabilité et l'autonomie soient des critères qualitatifs non mesurables, la précision de la prévision peut être mesurée avec des outils statistiques. Elle permet d'évaluer la pertinence des prévisions générées d'une part, et de comparer la performance des différents modèles testés d'autre part. En général, les mesures d'erreur sont couramment utilisées pour estimer la précision des méthodes de prévision. Dans cette section, nous présentons en premier temps, un ensemble de mesures d'erreur « traditionnelles » commun à tous les domaines de prévision. Les formules de calcul, les avantages et les inconvénients de chaque mesure sont aussi présentés. Ensuite, nous présentons les mesures d'erreur qui sont utilisées dans le domaine de la prévision de la charge électrique à l'échelle des ménages.

3.4.1 Indicateurs de précision

Différentes métriques ont été utilisées pour mesurer la précision des modèles de prévision. Ces métriques mesurent l'erreur entre les valeurs prédites par le modèle de prévision et les observations réelles. Généralement, le modèle le plus performant par rapport au critère de précision est celui qui a la plus petite valeur d'erreur de prévision. Les métriques d'évaluation de la prévision les plus connues dans la littérature sont l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) et le pourcentage d'erreur absolu moyen (MAPE).

Soit $(\hat{y}_t)_{t=1,\dots,n}$ le vecteur des valeurs prédites de la série temporelle $(y_t)_{t=1,\dots,n}$ par un modèle de prévision, et $e_t = y_t - \hat{y}_t$ l'erreur de la prévision à chaque instant t. L'erreur quadratique moyenne (RMSE) est définie donc comme :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(e_t)^2}. \tag{3.32}$$

Étant donné que les erreurs sont mises au carré avant d'être moyennées, la métrique d'erreur (RMSE) donne un poids relativement élevé aux grands écarts d'erreurs. Cela signifie que la métrique d'erreur (RMSE) est plus utile lorsque de grands écarts sont particulièrement indésirables comme dans le cas de la prévision des pointes de consommation (Alduailij et al., 2021). Par contre, si cela n'est pas souhaitable, d'autres métriques seront plus adaptées pour donner une meilleure idée sur la performance globale du modèle de prévision sans influence supplémentaire des grands écarts exceptionnels.

L'erreur absolue moyenne (MAE) est définie par :

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|e_t|. \tag{3.33}$$

C'est une métrique linéaire ce qui signifie que toutes les différences individuelles sont pondérées de manière égale dans la moyenne, et qu'elle pénalise autant les grandes erreurs que les petites erreurs. Les deux métriques (MAE) et (RMSE) peuvent être utilisées ensemble pour analyser la performance et la qualité de la prévision. L'erreur (RMSE) sera toujours supérieure ou égale à l'erreur (MAE) et plus l'écart entre les deux est grand, plus la variance des erreurs individuelles dans l'échantillon est grande. Si l'erreur (RMSE) est égale à l'erreur (MAE), alors toutes les erreurs sont de même grandeur. Les deux métriques prennent des valeurs de 0 à $+\infty$. Ce sont des métriques orientées négativement c.à.d les valeurs proches de 0 sont meilleures.

Le principal inconvénient de ces deux métriques est la dépendance à l'échelle (Hyndman et Koehler, 2006). Par conséquent, si les données à prévoir ont des différentes échelles, ces deux métriques ne peuvent pas être utilisées. Pour éviter la dépendance à l'échelle, des versions normalisées de ces deux métriques ont été proposées dans la littérature. Les deux métriques (NRMSE) et (NMAE) sont calculées de la façon suivante :

$$\text{NRMSE} = \frac{\text{RMSE}}{\bar{y}} \quad \text{NMAE} = \frac{\text{MAE}}{\bar{y}}, \tag{3.34}$$

où $\bar{y}$ est le facteur de normalisation, qui est généralement égal soit à la moyenne de la série à prédire, soit à la valeur maximale mesurée sur l'horizon de prévision, soit à la différence entre les valeurs maximale et minimale. La normalisation par la moyenne de la série temporelle est celle la plus privilégiée pour l'évaluation de la prévision de la demande d'électricité à l'échelle des ménages.

L'erreur (MAPE) est calculée de la façon suivante :

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|. \tag{3.35}$$

Elle est l'une des mesures les plus largement utilisées dans le domaine de prévision, en raison de ses avantages d'indépendance d'échelle et d'interprétabilité (Bowerman et al., 2005 ; Hanke et al., 2005 ; Makridakis, Andersen et al., 1982). Cependant, l'erreur (MAPE) présente l'inconvénient majeur de produire des valeurs infinies ou indéfinies pour des valeurs réelles nulles ou proches de zéro (Kim et al., 2016). Plusieurs alternatives ont été proposées pour résoudre ce problème comme le pourcentage d'erreur absolu moyen symétrique (sMAPE) proposée par Makridakis (1993), qui est une (MAPE) modifiée dans laquelle le diviseur est la moitié de la somme des valeurs réelles et prévues.

L'erreur (sMAPE) est calculée de la façon suivante :

$$\text{sMAPE} = \frac{100}{n} \sum_{t=1}^{N} \frac{|e_t|}{(\frac{|y_t|+|\hat{y}_t|}{2})}. \tag{3.36}$$

Une autre importante métrique est l'erreur d'échelle absolue moyenne (MASE). Il s'agit de l'erreur absolue moyenne des valeurs de prévision, divisée par l'erreur absolue moyenne de la prévision naïve. Elle peut surmonter le problème de génération des valeurs infinies ou indéfinies de l'erreur (MAPE). La mise à l'échelle de cette erreur signifie que la qualité du modèle est calculée par rapport à une méthode naïve de prévision [2]. Si l'erreur (MASE) est supérieure ou égale à 1, le modèle n'est pas plus performant que la méthode naïve. Plus l'erreur (MASE) est faible, plus le modèle dépasse en performance la méthode naïve de prévision. Il existe deux versions de l'erreur (MASE) une pour les séries temporelles non saisonnières et une pour les séries temporelles saisonnières. La version saisonnière est calculée par la formule suivante :

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|e_t|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|y_t - y_{t-m}|} \quad \text{où } m \text{ est la période saisonnière.} \tag{3.37}$$

Il existe d'autres métriques d'évaluation qui peuvent être utilisées et qui n'ont pas été répertoriées ci-dessus. Le lecteur intéressé par plus d'informations sur les mesures d'erreur de prévision peut se référer à Shcherbakov et al. (2013). En raison des inconvénients que présentent chacune de ces mesures d'erreur, le choix d'une seule métrique risque à conduire à une évaluation inexacte des modèles de prévision. Pour pallier ces inconvénients, plusieurs mesures d'erreur sont souvent utilisées simultanément. Certes, cette stratégie permet de garantir une meilleure évaluation de la précision du modèle mais une certaine complexité peut être ajoutée à l'interprétabilité des résultats. En effet, un modèle peut se révéler précis pour une mesure d'erreur, et moins précis pour une autre. Donc, l'utilisation de

2. La méthode naïve attribue à la valeur à prédire l'observation précédente. Il existe également une version saisonnière de la méthode naïve qui attribue à la valeur à prédire l'observation de la saison précédente.

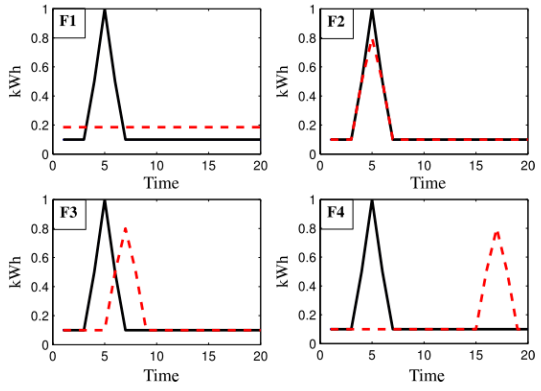
plusieurs métriques nécessitent une connaissance des limites de chacune afin de pouvoir tirer des conclusions pertinentes sur la précision des modèles.

3.4.2 Métriques de la prévision de la charge des ménages individuels

La métrique d'évaluation est un indicateur qui permet de quantifier la performance d'un modèle de prévision. Par conséquent, le choix d'une métrique adaptée au problème de prévision est aussi important que le choix du modèle puisque la qualité de ce dernier dépend directement de la métrique utilisée pour l'évaluer. Des différentes métriques ont été utilisées historiquement pour quantifier les erreurs de prévision et par suite mesurer la précision des modèles de prévision.

Cette diversité dans les métriques d'erreur est destinée à satisfaire la différence des besoins de prévision. En effet, l'évaluation de la qualité de la prévision doit dépendre du besoin et contribuer à la prise de décision. Dans le domaine d'électricité par exemple, la prévision de la demande d'électricité pour des fins de planification diffère de la prévision de la demande pour la gestion de la distribution. Dans le premier cas, les planificateurs s'intéressent à la prévision des pics de la demande tandis que les fournisseurs s'intéressent plutôt à la prévision de la demande journalière. Par conséquent, la métrique dont l'utilisation est jugée utile dans un cas n'est pas forcément utile dans l'autre cas.

Les métriques « traditionnelles » telles que l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (RMSE) sont souvent utilisées pour l'évaluation de la prévision de la demande d'électricité à l'échelle nationale. En revanche, l'utilisation de ces dernières pour la prévision de la charge des ménages présente certains inconvénients. En effet, la majorité des études réalisées sur ce sujet utilisent des jeux de données de différents ménages ayant des différents niveaux de consommation. Par conséquent, l'utilisation des métriques dépendantes de l'échelle telles que la métrique (RMSE) et (MAE) n'a pas d'utilité. En plus, ces métriques comme expliqué dans Hoffman et al. (1995) et Brooks et al. (1996) peuvent conduire à des conclusions erronées sur les performances réelles des modèles. En effet, le décalage de prévision d'un motif particulier notamment un pic de consommation dans la journée, même s'il est bien prédit par le modèle en termes d'amplitude, il subit ce qui s'appelle l'effet de double pénalité (Keil et al., 2009). Afin de faire face à ce problème, Haben, Ward et al. (2014) ont proposé une nouvelle métrique de mesure d'erreur qui convient à l'évaluation des prévisions ayant des motifs qui peuvent être déplacés dans le temps comme dans le cas des prévisions de la demande d'électricité à l'échelle des ménages. L'idée se base sur la recherche d'une permutation de la prévision dans l'espace spatio/temporel qui minimise l'erreur calculée selon une métrique « traditionnelle » donnée telle que l'erreur moyenne absolue (MAE) et l'erreur quadratique moyenne (RMSE). Cette mesure de l'erreur ajustée est alors calculée à partir des déplacements discontinus des prévisions dans l'espace de temps. L'amplitude des déplacements de prévisions est limité



Error	Forecast			
	F1	F2	F3	F4
Absolute error	0.82	0.20	0.99	1.00
Adjusted error ($w = 1$)	0.82	0.20	0.79	1.00
Adjusted error ($w = 2$)	0.82	0.20	0.48	1.00
Adjusted error ($w = 3$)	0.82	0.20	0.20	1.00

FIGURE 3.4 – Les quatre graphiques représentant les différentes prévisions F_1, F_2, F_3 et F_4 (ligne en pointillés rouge) d'une courbe de charge (ligne noire continue) ainsi que les erreurs absolues (w=0) et ajustées (w=1,2 et 3) de la p-norme [source (HABEN, WARD et al., 2014)].

par un paramètre w qui restreint le pas de déplacement en unité de temps à w maximum.

La figure 3.4 montre un exemple de comparaison entre quatre prévisions différentes F_1, F_2, F_3 et F_4 d'un profil simple de courbe de charge représentant un pic de consommation à l'instant $t = 5$ par rapport aux erreurs absolues et ajustées de la p-norme [3]. Pour plus d'information sur la nature des quatre prévisions le lecteur peut se référer à HABEN, WARD et al. (2014). L'effet de double pénalité est illustré par la médiocrité de la prévision F_4. La qualité des prévisions F_1, F_2 et F_4 n'a pas changé avec l'ajustement de l'erreur pour $w = 1, 2$ et 3. Ce résultat est dû au fait que la prévision F_1 est plate donc aucune permutation de cette dernière permet de minimiser l'erreur. Quant à la prévision F_2, l'ajustement de l'erreur n'a pas améliorée sa qualité puisque la meilleure permutation de la prévision qui permet de minimiser l'erreur pour les trois valeurs de $w = 1, 2$ et 3 est la prévision F_2 elle même. Étant donné que la distance entre le pic de la prévision F_4 et le pic de la consommation réelle dépasse la limite d'ajustement w, la qualité de cette dernière n'a pas pu être améliorée par ces ajustements, contrairement à la prévision F_3.

En résumé, cette métrique d'erreur ajustée permet de tolérer des valeurs prévues qui sont légèrement décalées dans le temps. Elle peut être considérée comme une généralisation de l'erreur p-norme (lorsque le paramètre de déplacement w, est égale à zéro, la métrique se réduit aux erreurs p-norme). Malgré ses avantages, cette nouvelle métrique proposée ne peut pas être utilisée dans toutes les applications parce qu'elle déforme la prévision de manière discontinue. De plus, elle n'est pas indépendante de l'échelle, ce qui

3. https://fr.wikipedia.org/wiki/Espace_de_suites_âĎŞp

la rend inadaptée pour comparer la précision des prévisions de la charge électrique de différents ménages. Finalement, le calcul de cette métrique est très chronophage par rapport aux autres métriques traditionnelles puisqu'il nécessite une permutation des prévisions Wijaya, Vasirani et al. (2015).

L'erreur absolue moyenne en pourcentage (MAPE) peut être considérée comme la plus largement déployée pour la prévision de la consommation à l'échelle des ménages (Rodrigues et al., 2014 ; X. Zhang et al., 2018 ; Eneyew et al., 2020) en raison de son indépendance de l'échelle. Par contre, l'utilisation de cette métrique a été critiquée parce qu'elle produit des valeurs infinies ou indéfinies lorsque les valeurs réelles sont nulles ou proches de zéro. Si les valeurs réelles sont très petites (généralement inférieures à un), cette erreur produit des erreurs en pourcentage extrêmement importantes (valeurs aberrantes), tandis que des valeurs réelles nulles entraînent des valeurs infinies ce qui cause un problème d'interprétabilité. Les travaux de Rossi et al. (2013) ont montré que l'erreur (MAPE) de leurs modèles de prévision a augmenté à 10% à 500% et plus en passant de la prévision à l'échelle nationale à la prévision à l'échelle des ménages individuels en raison d'une surestimation des petites charges de base.

Certaines études ont utilisé les métriques normalisées comme l'erreur quadratique moyenne normalisée (NRMSE) et l'erreur absolue moyenne normalisée (NMAE). En effet, la normalisation rend les résultats plus comparables entre les ménages ayant des différents niveaux de consommation (Singh et al. (2012), Wijaya, SFRJ Humeau et al. (2014) et Gerossier et al. (2017)). Plusieurs méthodes de normalisation à partir des données sont également proposées pour ces métriques (voir la sous-section 3.4.1). L'utilisation de plusieurs métriques d'évaluation est une pratique aussi courante dans les approches de prévision de la charge électrique à l'échelle des ménages (Yildiz et al., 2018 ; Gerossier et al., 2017 ; Wijaya, SFRJ Humeau et al., 2014).

3.5 Conclusion

Dans ce chapitre, nous avons présenté les fondements théoriques des modèles les plus fréquemment utilisés dans le domaine de la prévision de la charge électrique. Nous avons également présenté les avantages et les limites théoriques de chaque modèle ainsi que leur importance dans le contexte de la prévision de la charge électrique.

En effet, la prévision de la charge électrique à l'échelle des ménages présente des exigences particulières qui la différencient des autres niveaux de prévision de la charge électrique. En plus des critères généraux tels que la précision et la complexité de calcul, la prévision de la charge électrique à l'échelle des ménages nécessite également une grande interprétabilité pour que les utilisateurs puissent comprendre et interpréter les résultats de manière claire et intuitive. De plus, le temps de calcul doit être rapide pour permettre une

utilisation pratique dans un environnement opérationnel. Enfin, la flexibilité est également importante pour permettre une adaptation facile aux différents profils de de consommation électrique des ménages. Toutes ces exigences doivent être prises en compte lors de la sélection des modèles de prévision pour la charge électrique à l'échelle des ménages. Nous avons donc exposé les critères de sélection que nous avons utilisés pour choisir les modèles les mieux adaptés à nos besoins en termes de prévision de la charge électrique à l'échelle des ménages, en prenant en compte des aspects tels que la qualité de la prévision, l'interprétabilité et la rapidité d'exécution pour une utilisation industrielle.

Notre revue de la littérature et la présentation des différents modèles de prévision, de leurs avantages et de leurs limites nous ont conduits à conclure qu'il n'existe pas de modèle unique qui puisse être considéré comme le meilleur pour la prévision de la charge électrique à l'échelle des ménages. Nous considérons donc que déployer et comparer plusieurs modèles de prévision permet non seulement de choisir celui qui permet de prédire au mieux la charge électrique à l'échelle des ménages de notre jeu de données, mais également de fournir une approche plus globale sur la performance de différents modèles et leur contribution possible à nos besoins.

Par ailleurs, nous avons présenté les indicateurs qui permettent de mesurer la qualité des prévisions générées par les modèles, à la fois dans un cadre général et dans le cadre spécifique de la prévision de la charge à l'échelle des ménages. Cette analyse nous a permis de conclure qu'il n'existe pas d'indicateur unique capable d'évaluer exhaustivement la précision de la prévision, et que l'utilisation de plusieurs indicateurs combinés est la méthode la plus appropriée pour évaluer la qualité des prévisions.

Chapitre 4

Prévision de la consommation d'électricité des ménages

Objectifs

À la lueur des études de la littérature introduites dans le chapitre 2 ainsi que des modèles de prévision introduits dans le chapitre 3, nous présenterons les travaux que nous avons menés dans le contexte de la prévision de la charge électrique à l'échelle des ménages. Tout d'abord, nous commençons par présenter une description générale du jeu de données qui a été mis à notre disposition pour cette étude. Nous décrivons également les différentes caractéristiques des courbes de charge des ménages en les comparant à celles des courbes de charge à l'échelle nationale. Cette description permettra de comprendre les choix qui seront faits ultérieurement dans ce manuscrit pour les modèles de prévision ainsi que les variables d'entrée de ces derniers. Ensuite, nous présentons les différents modèles que nous avons mis en œuvre pour la prévision de la charge électrique à l'échelle des ménages. La performance de ces modèles appliqués au jeu de données est évaluée en fonction de la précision mesurée par plusieurs métriques, du degré d'interprétabilité ainsi que du temps de calcul. Les résultats obtenus sont analysés en fonction de plusieurs caractéristiques présentes dans les données comme la thermosensibilité et la volatilité. Enfin, nous proposons une nouvelle approche de prévision des courbes de charge qui consiste à prédire l'énergie électrique consommée pendant la journée au lieu des courbes de charge des puissances. Cette approche de prévision permet de gagner en précision et en interprétabilité.

Sommaire

4.1 Introduction

La disponibilité des données de consommation électrique à l'échelle des ménages grâce au déploiement des compteurs intelligents a motivé les recherches sur la prévision de la charge électrique à cette échelle dans les deux milieux académique et industriel. La particularité de ce sujet de prévision réside dans le type de données à traiter. En effet, les modèles de prévision conçus historiquement pour la prévision de la charge électrique se basent tous sur l'hypothèse de la régularité de cette dernière.

Par contre, les données de consommation à l'échelle des ménages sont caractérisées par leur irrégularité et leur volatilité, puisqu'elles dépendent du mode de vie des occupants qui n'est pas souvent régulier. Cette irrégularité est à l'origine de la difficulté de la prévision de charge électrique à cette échelle. Malgré cette irrégularité, les courbes de charge des ménages présentent des motifs récurrents qui reflètent des habitudes de consommation régulières comme l'utilisation de certains appareils électriques à des horaires précis de la journée. La qualité de la prévision de la charge à l'échelle des ménages dépend alors du degré de régularité de ces habitudes et de la capacité du modèle de prévision à les détecter. Par conséquent, une attention particulière doit être portée au choix des modèles de prévision de la charge à cette échelle surtout que ce choix doit également tenir compte du type de l'application envisagée par la prévision (voir la section 3.3).

Comme évoqué précédemment dans le chapitre 1, l'objectif principal de cette thèse consiste à mettre en œuvre des modèles de prévision de la consommation électrique à l'échelle des ménages. Ces prévisions seront utilisées pour fournir des services personnalisés en temps réel autour de la consommation électrique de ces ménages. L'étude bibliographique nous a permis de prendre connaissance des différents modèles qui ont été mis en œuvre pour la prévision de la charge électrique à l'échelle des ménages. Ces études malgré leur nombre limité montrent que c'est le type d'application envisagée par la prévision qui influence le plus le choix des modèles de prévision les plus appropriés. En effet, les modèles qui se révèlent performants pour la prévision des ménages dans un contexte ne sont forcément appropriés dans un autre. Pour cela, nous avons adopté une approche pour la problématique de la prévision de la charge électrique à l'échelle des ménages qui consiste à :

1. sélectionner des modèles de la littérature de prévision de la charge électrique que nous estimons capables de répondre à nos besoins en tenant compte des contraintes détaillées dans la section 3.3 ;
2. adapter ces modèles pour la prévision de la charge électrique des ménages ;
3. tester et évaluer les modèles sur une grande quantité de données de courbes de charge des ménages ;
4. comparer les résultats et les performances des modèles en fonction des caractéristiques présentes dans les données (thermosensibilité, périodicité, volatilité, ...) d'une

part et des caractéristiques des modèles (rapidité, complexité, interprétabilité, explicabilité, ...) d'autre part.

Dans ce qui suit, nous présenterons cette approche en détails. Nous introduirons également une approche pour la prévision de la charge électrique des ménages qui permet de réduire la volatilité et l'irrégularité dans les courbes de charge des ménages et par conséquent, améliorer la précision de la prévision à cette échelle.

4.2 Données de la consommation électrique dans le secteur résidentiel

4.2.1 Description des données

Les données utilisées dans cette étude sont des données privées anonymes de la consommation d'électricité d'un millier de clients particuliers en France. Ces clients avaient un compteur intelligent d'électricité installé dans leurs logements qui collectait des mesures de consommation d'électricité (W) toutes les demi-heures pendant la période allant de janvier 2017 jusqu'à janvier 2019.

L'ensemble de données contient différents types de logement tels que les maisons individuelles, les petits collectifs (moins de cinq appartements) et les grands collectifs (plus de cinq appartements). En plus, nous distinguons dans le jeu de données deux types de tarification. L'option tarifaire de base qui propose un prix du kWh unique tout au long de la journée et l'option heures pleines/heures creuses qui permet au client de bénéficier d'une tarification qui évolue en fonction de l'heure de la journée (LeLynx.fr, 2022). Nous détaillerons plus tard l'impact de la tarification sur le profil de la courbe de charge.

Variable	**Description**
Nombre des ménages	1000 avant pré-traitement
Période	de 01-01-2017 à 01-01-2019
Localisation	France
Resolution	30 minutes
Type de logement	Appartements/ Maisons
Tarif	Tarif de base et Tarif HC/HP [1]

Table 4.1 – Description de l'ensemble du jeu de données.

Aucune information n'est fournie sur les équipements des foyers ou le mode de vie de ses occupants ce qui restreint l'étude aux données brutes des courbes de charge. Nous disposons également pour cette étude des relevés de la température extérieure extrapolées au pas demi-horaire pour la même période d'une station météorologique locale en alsace. Les tableaux 4.1 et 4.2 montrent quelques caractéristiques des données utilisées dans notre

étude. En comparaison avec les logements grand collectif et petit collectif, les logements individuels ont la moyenne de consommation électrique quotidienne la plus élevée. Cette différence peut s'expliquer par le fait que les logements individuels ont généralement une surface habitable plus grande, plus d'appareils électriques, de systèmes de chauffage et de climatisation, ainsi que d'autres équipements nécessitant de l'électricité. En revanche, les logements grand collectif et petit collectif peuvent avoir une consommation électrique quotidienne inférieure grâce à la mise en commun de certaines ressources telles que l'éclairage et le chauffage, ainsi qu'à l'utilisation de systèmes de chauffage centralisé.

Type logement	**Puissance Moyenne**	**Pourcentage**	**Tarif HC/HP** [2]	**Tarif base**
Individuel	877 W	56%	47,2%	52,8%
Grand Collectif	640 W	24%	61,2%	38,8%
Petit Collectif	660 W	20%	25,7%	74,3%

TABLE 4.2 – Statistiques sur les données utilisées dans notre étude.

La majorité des logements en grand collectif ont souscrit à un contrat avec option tarifaire HC/HP contrairement aux logements individuels et petit collectif (voir tableau 4.2). Le choix d'un contrat avec option tarifaire HC/HP dépend de plusieurs facteurs tels que la taille du logement, le nombre d'occupants, le mode de vie et les habitudes de consommation d'électricité. Dans le cas des logements en grand collectif, il est possible que le gestionnaire de l'immeuble ou le syndicat des copropriétaires ait négocié un contrat collectif pour tous les logements, offrant ainsi une tarification avantageuse pour l'ensemble de l'immeuble. Pour les logements individuels et petit collectif, les occupants peuvent avoir des habitudes de consommation plus variées, ce qui peut rendre difficile la planification de la consommation pendant les heures creuses. De plus, il est possible que les occupants des logements individuels préfèrent avoir une plus grande flexibilité dans l'utilisation de leur électricité et optent pour des tarifs plus flexibles et moins contraignants.

D'après la figure 4.1, la consommation électrique moyenne quotidienne de la plupart des ménages dans l'ensemble de données se situe entre 500 W et 1500 W. Au-delà de cette limite, les ménages sont souvent équipés de chauffages électriques ou correspondent à des logements de grande surface. Par ailleurs, les ménages avec une consommation électrique moyenne inférieure à cette plage peuvent être équipés de systèmes de chauffage au gaz ou au fioul, ou correspondent à des logements de petite surface tels que des studios étudiants ou des appartements de type F1 ce qui réduit leur consommation électrique quotidienne.

2. heure creuse (HC)/heure pleine (HP) : une option tarifaire qui permet de bénéficier d'un prix réduit du kWh pendant les heures où la demande en électricité est la plus faible (généralement la nuit d'où l'appellation parfois par tarif jour/nuit)(PINON, 2022)

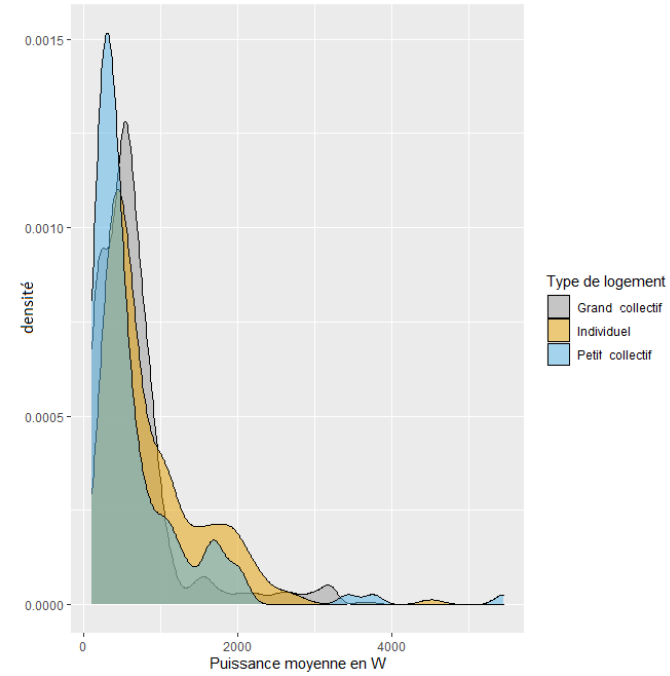


FIGURE 4.1 – La distribution de la puissance moyenne en W des ménages appartenant à chaque type de logement.

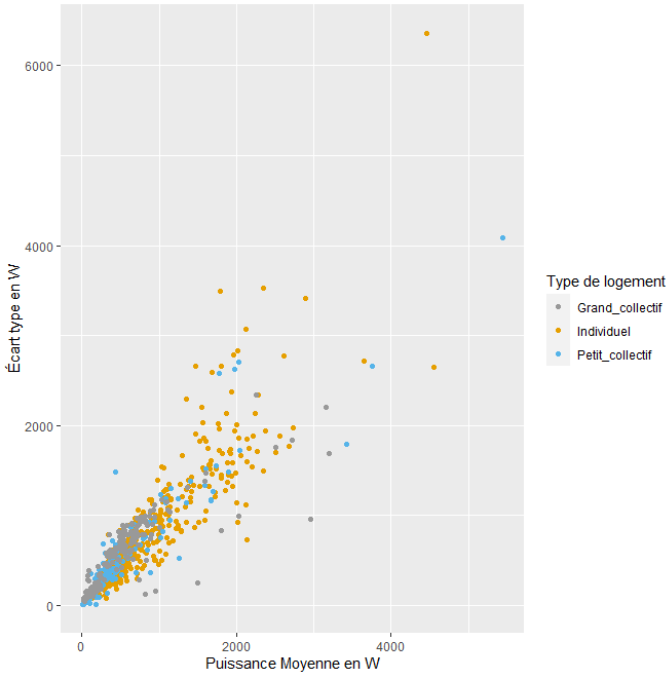


FIGURE 4.2 – Nuage de points de l'écart-type en W (axe y) par rapport à la puissance moyenne en W de la demande d'électricité des ménages dans les trois types de logement (axe x).

La figure 4.2 montre que l'écart type augmente avec l'augmentation de puissance

moyenne journalière des ménages. Cela suggère que les ménages ayant une consommation électrique plus élevée ont une plus grande variabilité dans leur consommation d'électricité. De plus, cette figure montre une très forte disparité dans les moyennes de consommation électrique des ménages dans le jeu de données. Cette grande variation souligne l'importance de comprendre les facteurs qui influencent la consommation électrique dans les ménages afin de pouvoir mettre en place des stratégies efficaces pour la prévision de consommation électrique à l'échelle des ménages.

4.2.2 Caractéristiques de la consommation électrique des ménages

La consommation électrique dans le secteur résidentiel à l'échelle d'un ménage peut varier considérablement en fonction de nombreux facteurs. La courbe de charge d'un ménage représente les relevés de puissance (W) mesurés par un compteur intelligent toutes les 10, 15 ou 30 minutes. Grâce à l'émergence des compteurs intelligents dans le secteur résidentiel, ces données sont devenues accessibles à la fois par les fournisseurs d'électricité et le consommateur. Ces courbes de charge constituent un outil précieux qui permet de comprendre le comportement de consommation électrique propre à chaque ménage et par conséquent, développer des stratégies de gestion personnalisées et efficaces de la consommation électrique.

La courbe de charge au niveau des ménages (voir figure 4.3) contrairement à la courbe de charge nationale (voir figure 4.4) ou agrégée (voir figure 4.5) montre une haute volatilité, car elle dépend de plusieurs facteurs propre à chaque ménage tels que le comportement des occupants et leur mode de vie, leurs équipements électriques ainsi que les caractéristiques du bâtiment et le climat local (Tascikaraoglu et al., 2014). Dans la plupart des cas les ménages affichent de très petites charges de base entre 0 et 500 W avec une succession des pics[3] et des creux qui varie selon la puissance électrique des appareils relativement énergivores, tels que les chauffes eaux électriques, les chauffages électriques, les sèche-linges, les pompes de piscine,... et d'une manière erratique de sorte que les motifs sont difficilement visibles lors d'une inspection rapide (Yildiz et al., 2017) (voir figure 4.3). Au contraire, la forme de courbe de charge nationale est plus lisse avec des motifs répétés régulièrement (voir figure 4.4) marqués par le cycle économique avec des différences de niveau de consommation entre les jours ouvrés et les week-ends et le jour et la nuit. Dans cet exemple, la figure 4.5 présente l'extrait d'une courbe de charge d'un ménage qui montre un pic de consommation à 01h30 qui correspond aux cycles programmés des électroménagers énergivores avec une vallée l'après-midi, et une consommation plus élevée le soir par rapport à la matinée. D'autres petits pics sont présents également, certains apparaissent entre 12h00 et 14h00 décrivant une activité particulière dans cette période

3. Les pics (ou pointes) de consommation d'électricité sont les moments où la consommation d'électricité atteint des sommets (maximums) pendant une période bien définie. https://www.energies.leclerc/actualite-energie/pic-consommation-electrique.

de la journée très probablement liée à l'utilisation des appareils de cuisson.

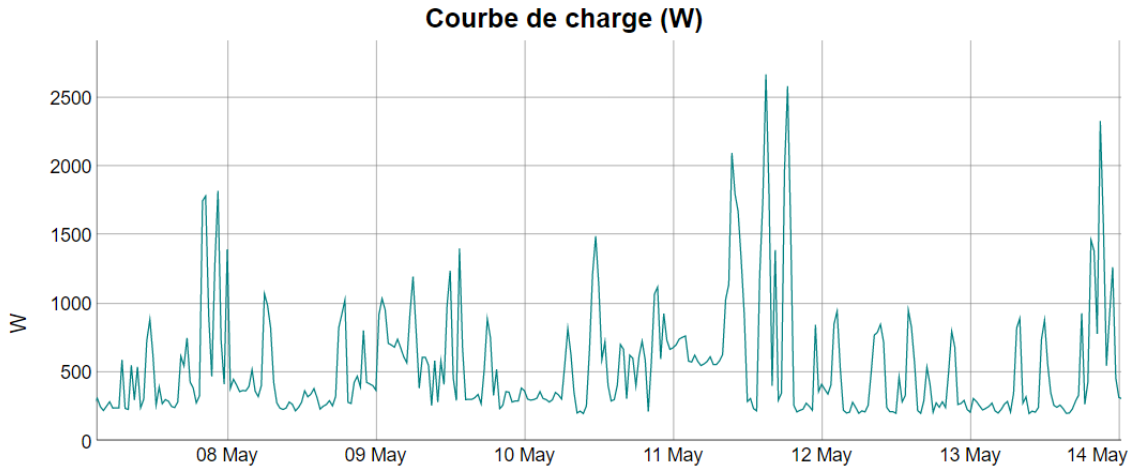


FIGURE 4.3 – Exemple d'une courbe de charge d'un ménage pendant une semaine du 7 mai au 14 mai 2018 (données privées du fournisseur).

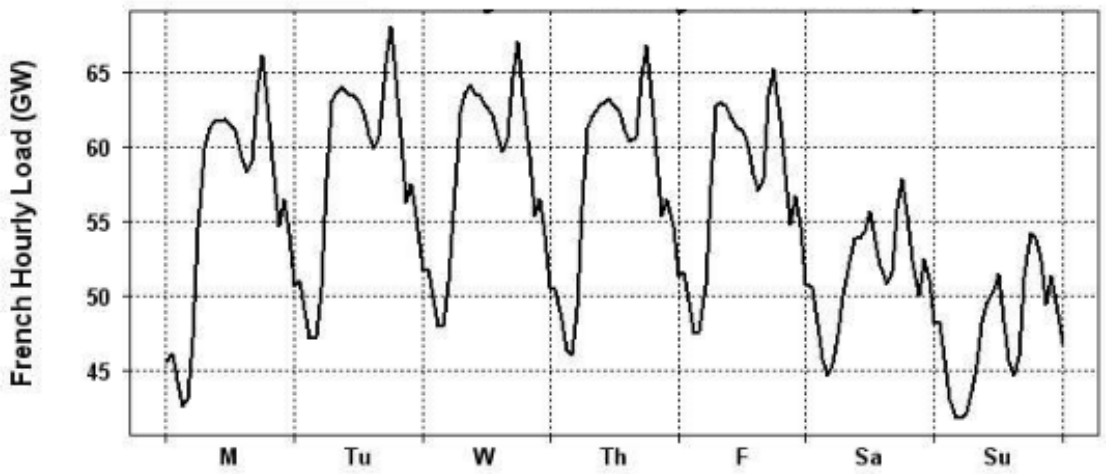


FIGURE 4.4 – Extrait de la courbe de charge nationale française pendant une semaine [Source (PIERROT et al., 2011)].

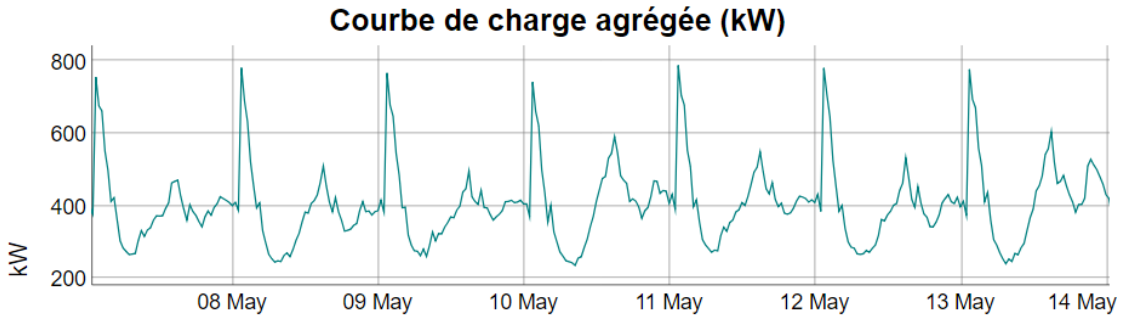


FIGURE 4.5 – Extrait de la courbe de charge agrégée de 900 clients résidentiels pendant une semaine du 7 mai au 14 mai 2018 (données privées de l'entreprise).

Les conditions météorologiques ont également un impact significatif sur les charges résidentielles, en particulier pour les besoins de chauffage et de climatisation. L'hiver, lorsque la température extérieure est basse, la demande de chauffage augmente, ce qui se traduit par une augmentation de la consommation d'énergie électrique. En été, lorsque la température est élevée, la demande de climatisation augmente, ce qui a également un impact significatif sur la courbe de charge résidentielle. Nous parlons alors dans les deux cas de la thermosensibilité des courbes de charge. Les courbes de charge non-thermosensibles correspondent aux courbes de charge des ménages qui ne dépendent pas du chauffage électrique. En effet, les variations de la consommation électrique de ces ménages sont principalement liées à l'utilisation des appareils électriques, électroménagers et de l'éclairage.

En France, la climatisation est généralement perçue comme un équipement de luxe et n'est pas encore très courante dans les foyers, selon une étude récente de l'Ademe (Agence de la transition écologique), environ 25% des ménages français étaient équipés d'un climatiseur en 2020 (GOUTY, 2022). C'est la raison pour laquelle la plupart des ménages sensibles à la température dans l'ensemble de données ont une sensibilité liée à l'utilisation du chauffage. Cependant, la demande de climatisation est en augmentation constante en raison du changement climatique et des vagues de chaleur estivales de plus en plus fréquentes.

Les charges résidentielles sont aussi affectées par le calendrier. Les effets de calendrier comprennent tous les changements de consommation de charge liés aux périodes calendaires tels que les jours de la semaine, les vacances, les jours fériés et les événements spéciaux sportifs ou culturels, ... (LUSIS et al., 2017). Les habitudes de consommation varient considérablement pendant ces périodes, ce qui a un impact sur les courbes de charge électrique. Par exemple, les jours fériés peuvent entraîner une baisse de la consommation pendant les heures de travail, tandis que les soirées de grands événements sportifs peuvent entraîner une augmentation de la consommation d'énergie électrique. Les vacances peuvent avoir un impact significatif sur la consommation électrique de deux manières différentes. D'une part, pour les foyers qui partent en vacances, nous pouvons observer une baisse

importante de la consommation électrique pendant la période d'absence, sans forcément s'annuler en raison des équipements pilotables à distance comme le système de vidéo-surveillance, les équipements connectés, les volets roulants, les serrures connectées ou les systèmes d'éclairage automatique, ... (voir figure 4.6). D'autre part, pour les foyers qui restent à la maison pendant les vacances, les habitudes de consommation peuvent changer, notamment avec des heures de réveil et de coucher différentes, une utilisation accrue de certains appareils électriques et une réduction de l'utilisation de certains autres.

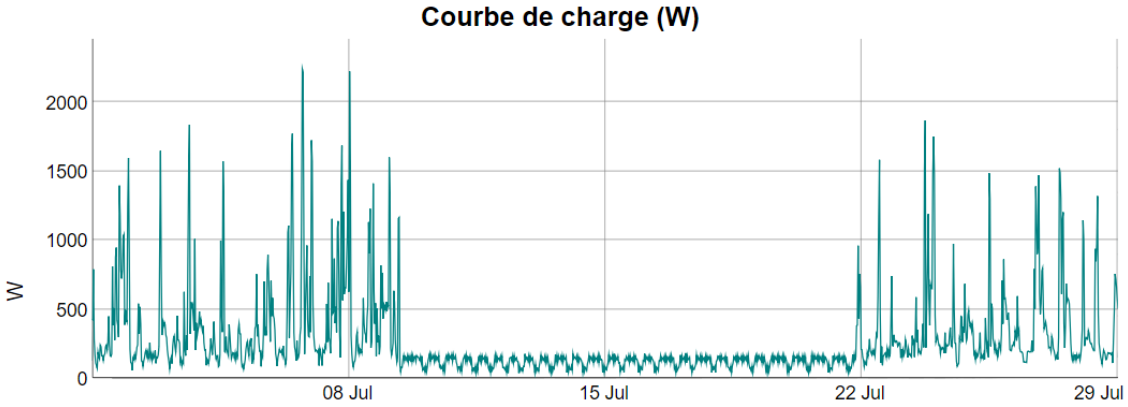


FIGURE 4.6 – Extrait d'une courbe de charge d'un ménage montrant une période de vacances entre le 09 juillet et le 21 juillet 2018.

La détermination de ces périodes de vacances à partir des courbes de charge est très difficile sans informations supplémentaires de la part du consommateur. Les données de consommation électrique brutes ne permettent pas toujours de déterminer de manière fiable les habitudes de consommation pendant les vacances, car ces périodes peuvent varier en fonction de nombreux facteurs, tels que les types de travail, les préférences personnelles et les événements familiaux. Il est donc important d'obtenir des informations sur les périodes de vacances directement auprès des consommateurs, afin de pouvoir intégrer ces informations dans les modèles de prévision de la charge électrique. Cependant les consommateurs peuvent être réticents à partager des informations sur leurs habitudes de consommation pendant les périodes de vacances, en raison de préoccupations liées à la vie privée ou à la sécurité des données.

Les chercheurs ont entrepris des efforts pour identifier les périodes de vacances à partir des données de consommation d'électricité, sans nécessiter des informations détaillées de la part des consommateurs. Par exemple, GEROSSIER (2019) a travaillé sur l'extraction des périodes de vacances des courbes de charge des ménages et il a conclu que les périodes de vacances prolongées sont considérées comme des événements rares : environ une semaine par an pour 15% des ménages. En plus, la courbe journalière de l'électricité durant ces jours ne se distingue pas facilement de toute autre journée pendant laquelle le foyer a moins consommé de l'électricité pour une raison ou pour une autre.

Dans les parties suivantes, nous allons examiner certaines des caractéristiques observées dans les courbes de charge électrique des ménages dans le jeu de données, en nous concentrant sur trois aspects : la volatilité, la thermosensibilité et la périodicité.

4.2.2.1 La volatilité

La prévision de la charge électrique à l'échelle des ménages est plus complexe que celle de la charge électrique agrégée ou nationale puisqu'elle est plus volatile, plus bruitée et mois régulière. Le terme volatilité est parfois utilisé pour désigner ces trois notions d'une façon inappropriée. En effet, en dehors du secteur de la finance, la volatilité décrit une tendance à des changements ou des fluctuations rapides et imprévisibles (Zareipour et al., 2007).

Dans le contexte de l'énergie, la volatilité des courbes de charge électrique fait référence à la fréquence et à l'amplitude des fluctuations de la consommation électrique d'un ménage sur une période donnée. Les fluctuations peuvent être causées par plusieurs facteurs, tels que les comportements de consommation, les équipements électriques, les conditions météorologiques. Les courbes de charge des ménages peuvent varier considérablement d'un jour à l'autre, voire d'une heure à l'autre, en fonction de ces facteurs. Le degré de cette variabilité, qui peut être mesuré en utilisant des caractéristiques statistiques telles que l'écart-type ou le coefficient de variation, indique le niveau de volatilité dans les données de consommation d'électricité. Dans Hou et al. (2021), les auteurs utilisent le coefficient de dispersion pour analyser la volatilité de la charge électrique quotidienne. Ils définissent alors la volatilité comme le rapport entre l'écart-type et la moyenne de la consommation électrique. Cet indice de volatilité quotidien que nous notons V_L (L pour *Load*) est alors calculé par la formule suivante :

$$V_L = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(P_i - P_{\text{moy}})^2}}{P_{\text{moy}}} \tag{4.1}$$

où P_{moy} désigne la puissance journalière moyenne en W, P_i la puissance au temps i de la journée, et n est le nombre d'observations par jour. Cet indice est indépendant de l'unité de mesure et par conséquent, il permet de comparer la dispersion autour de la moyenne des courbes de charge journalières des différents ménages. Plus la valeur de l'indice de volatilité est élevée, plus la charge journalière est volatile.

Nous avons calculé l'indice de volatilité de toutes les courbes de charge présentes dans le jeu de données en utilisant l'indice quotidien de volatilité défini dans l'équation (4.1) sur la période d'un an de l'historique de données. Pour quantifier la volatilité de chaque courbe de charge, nous avons calculé la moyenne des indices de volatilité quotidiens. Les résultats sont présentés sous la forme d'un histogramme sur la figure 4.7, qui montre

la distribution de l'indice de volatilité pour les différents ménages du jeu de données. Les indices de volatilité varient de 0 à 4, reflétant la disparité de la volatilité entre les ménages. Cette observation est cohérente avec la littérature, qui montre que les indices de volatilité des ménages peuvent varier considérablement, même s'ils ont des caractéristiques et des appareils électriques similaires. En effet, la manière dont les appareils électriques sont utilisés peut avoir un impact sensible sur la volatilité des courbes de charge (Hou et al., 2021 ; Yildiz et al., 2017).

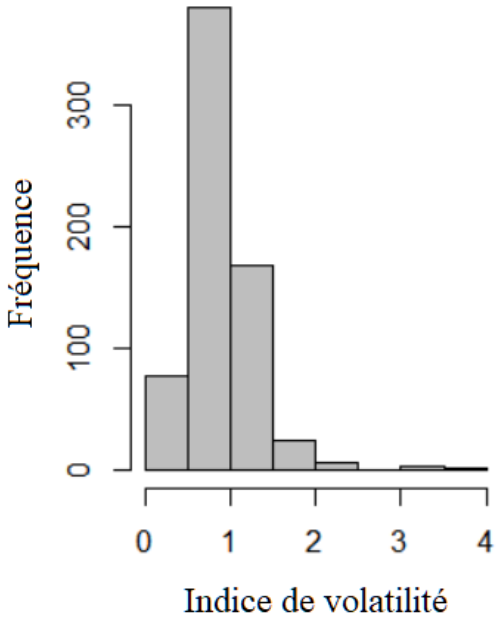


FIGURE 4.7 – Histogramme montrant la distribution de l'indice de volatilité des ménages dans le jeu de données.

4.2.2.2 La thermosensibilité

La relation entre la température extérieure et la demande de l'électricité au niveau national a fait l'objet de nombreuses recherches pendant plusieurs années (Dordonnat et al., 2008 ; Özkizilkaya, 2014). En effet, la thermosensibilté de la demande d'électricité est en grande partie le résultat de l'utilisation des appareils de chauffage électriques et de climatisation dans les périodes hivernales et estivales. Par conséquent, la forme de cette relation dépend fortement des caractéristiques climatiques du pays qui déterminent le type de l'équipement électrique à utiliser. Par exemple, dans certains pays chauds de l'Amérique, la demande d'électricité est beaucoup plus importante dans la période estivale en raison de l'utilisation massive de climatiseurs tandis qu'en Europe par exemple la demande est plus sensible aux températures froides qu'aux températures chaudes puisque l'utilisation de la climatisation est encore limitée.

Les travaux de Bessec et al. (2008) ont mis en évidence la relation non-linéaire entre la demande d'électricité et la température dans plusieurs pays de l'Europe. Les auteurs ont souligné que l'effet de la température est plus important dans la période hivernale que dans la période estivale contrairement aux études publiées dans des régions aux États-Unis (Amato et al., 2005) et l'Arabie Saoudite (Al-Zayer et al., 1996). D'autres études comme celles de Grafe et al. (2007) et d'Engle et al. (1986) décrivent une relation en forme de U entre la consommation d'électricité et la température extérieure. Un tel résultat est obtenu dans les pays du monde où à la fois le chauffage et la climatisation sont utilisés (voir figure 4.8).

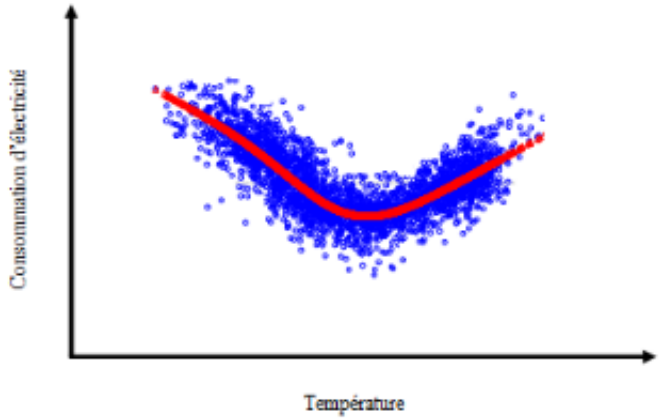


figure 4.8 – Graphique montrant la relation en forme de U entre la consommation d'électricité et la température extérieure [Source (Moral-Carcedo et al., 2005)].

En France, la demande dans le secteur résidentiel peut être divisée en deux parties : une partie due à la thermosensibilité liée principalement à l'utilisation des chauffages dans la période hivernale et des systèmes de ventilation ou de climatisation dans la période estivale et une autre partie indépendante de la température (voir figure 4.9).

La relation entre la demande et la température extérieure est alors non-linéaire. Dans certaines études comme celle de Dordonnat et al. (2008), elle est modélisée par trois fonctions linéaires par morceaux (voir figure 4.10). La première ligne ayant une pente négative représentée en rouge sur la figure 4.10 modélise l'effet du chauffage sur la demande en dessous d'une température seuil fixée au niveau national à $15°C$. La ligne ayant une pente positive représentée en vert modélise l'effet de refroidissement au dessus d'une deuxième température seuil fixée à $18°C$. La ligne bleue entre $15°C$ et $18°C$ représente l'absence de l'impact de la température sur la demande.

Cette relation entre la température extérieure et la demande d'électricité qui paraît simple, en réalité est bien plus compliquée. En effet, la demande de l'électricité au temps t ne dépend pas uniquement de la température au temps t mais également des températures des jours passés en raison de l'inertie thermique des bâtiments (Le Comte et al.,

1981 ; Özkizilkaya, 2014 ; Dordonnat et al., 2008). Certaines études intègrent ainsi les températures retardées dans la modélisation de la relation entre la demande d'électricité et la température (Goude et al., 2013 ; Papalexopoulos et al., 1990).

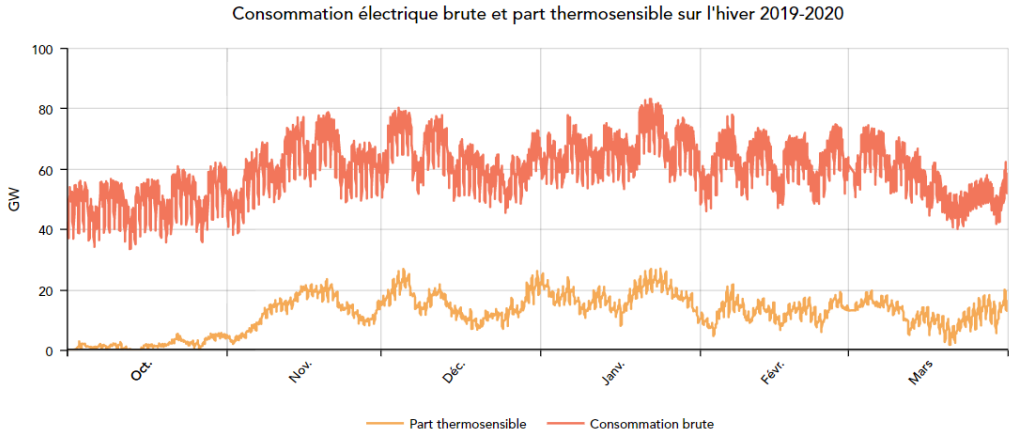


FIGURE 4.9 – Graphique montrant la consommation électrique brute et la part thermosensible de la demande (hiver 2019-2020). La courbe en rouge représente la consommation brute et la courbe en orange représente la partie thermosensible [Source Enedis].

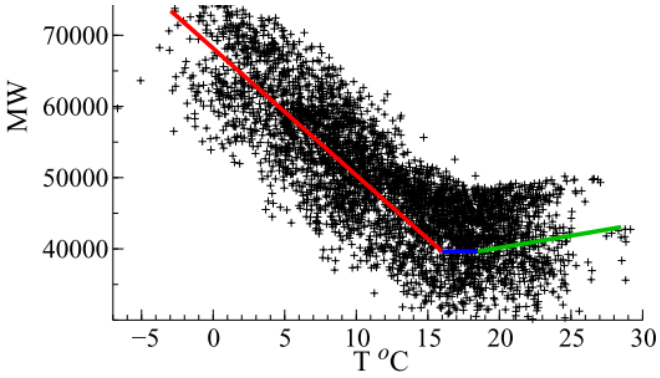


FIGURE 4.10 – Graphique montrant la relation entre la température moyenne journalière en $C°$ et la demande moyenne journalière de la France en MW. [Source (Dordonnat et al., 2008)].

Certains chercheurs expliquent également que cette relation est affectée par le cycle des jours ouvrés et des jours non ouvrés (Henley et al., 1997). En effet, la demande d'électricité liée à la thermosensibilité est différente entre le secteur tertiaire et le secteur résidentiel selon le type du jour (Özkizilkaya, 2014).

Après le déploiement massif des compteurs intelligents et la disponibilité des données de consommation à l'échelle des ménages la relation entre la température extérieure et la charge des ménages est devenue un centre d'intérêt des chercheurs. Contrairement à l'échelle nationale, l'impact de la température extérieure sur la prévision à l'échelle des

ménages est encore un point de désaccord entre les chercheurs. Tandis que certains chercheurs considèrent que l'intégration de la température dans les modèles de prévision de la charge des ménages n'a pas un impact significatif sur la précision de la prévision, d'autres signalent une amélioration. Yizhen Wang et al. (2021) ont montré dans leur article qui étudie l'influence des caractéristiques météorologiques sur la prévision de la charge électrique des ménages que l'intégration des variables météorologiques avait un effet positif sur l'amélioration de la précision des prévisions pour la plupart des ménages, mais il pourrait avoir un effet négatif sur les ménages moins sensibles aux changements météorologiques.

La thermosensibilité d'un ménage se réfère alors à la réactivité de la consommation d'électricité aux fluctuations de température. En d'autres termes, elle indique dans quelle mesure la consommation d'électricité d'un ménage varie en fonction des changements de température. Dans le jeu de données, nous avons constaté que tous les ménages ne réagissent pas de la même manière aux variations de température. Certains ménages ont une consommation d'électricité qui fluctue considérablement en fonction de la température (ménages thermosensibles), tandis que d'autres ont une consommation d'électricité relativement stable, quel que soit le niveau de température (ménages non-thermosensibles). En outre, le degré de thermosensibilité peut également varier au sein des ménages thermosensibles. Cette différence de sensibilité peut être due à plusieurs facteurs tels que l'isolation de la maison, le type de chauffage utilisé, les préférences de température des occupants, la surface du logement, ...

En effet, le degré d'isolation thermique d'une maison peut influencer la quantité d'énergie électrique nécessaire pour maintenir une température confortable. Si la maison est bien isolée, elle conservera mieux la chaleur ce qui réduira la demande d'énergie pour le chauffage. Le type de système de chauffage utilisé peut également jouer un rôle important. Par exemple, un logement équipé d'un système de chauffage centralisé peut avoir une courbe de charge plus sensible aux variations de température, car l'ensemble des pièces du logement est chauffé en même temps. En revanche, un logement équipé d'un système de chauffage électrique autonome, tel qu'un radiateur électrique, peut avoir une courbe de charge moins sensible, car chaque pièce peut être chauffée indépendamment. Les préférences de température des occupants de la maison peuvent également avoir un impact sur la sensibilité de la courbe de charge au facteur thermique. Si les occupants préfèrent maintenir une température plus élevée ou plus basse que la moyenne, cela peut se traduire par des pics de demande d'énergie pour le chauffage à des moments spécifiques de la journée. Enfin, la surface de logement à chauffer est également un facteur important. Un logement de plus grande surface nécessitera généralement plus de chauffage pour atteindre une température confortable, ce qui peut augmenter la thermosensibilité de la courbe de charge électrique.

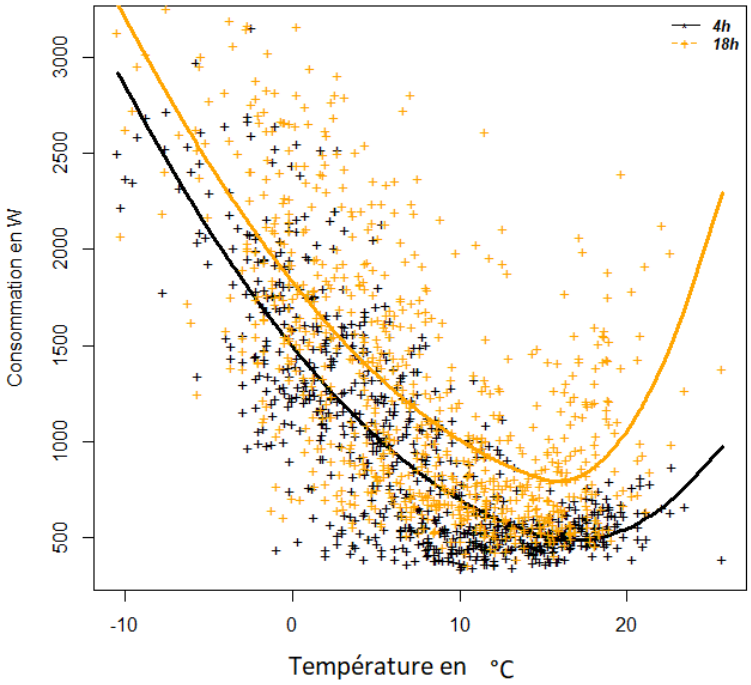
La thermosensibilité de la consommation électrique peut varier considérablement en fonction du type de jour de la semaine et des heures de la journée. En général, la ther-

mosensibilité de la consommation électrique est plus élevée pendant les heures où les occupants sont présents à domicile, comme le soir et le week-end. La thermosensibilité de la consommation électrique peut être influencée aussi par les options tarifaires proposées aux consommateurs. Par exemple, les consommateurs ayant des contrats d'électricité avec une option tarifaire « Heures Creuses/Heures Pleines » (HC/HP) ont tendance à consommer davantage d'électricité pendant les heures creuses où le tarif est moins élevé. Ces consommateurs peuvent donc adapter leur consommation d'électricité en reportant certaines tâches énergivores à ces heures creuses pour économiser de l'argent. Ils peuvent également utiliser des équipements électriques programmables y compris les chauffages pour ajuster automatiquement la consommation d'électricité de leur maison en fonction des plages horaires pendant lesquelles l'énergie électrique est moins chère. La figure 4.11 représente un nuage de points de la consommation d'électricité en (W) à 4h et 18h de deux ménages thermosensibles en fonction de la température extérieure. Nous remarquons que dans le cas du premier ménage ayant un contrat d'électricité avec l'option tarifaire de base (voir figure 4.11a) la variation de la température a un impact plus fort à 18h qu'à 4h du matin, tandis que dans le cas du deuxième ménage ayant un contrat d'électricité avec l'option tarifaire HC/HP (voir figure 4.11b) l'impact de la variation de la température est plus fort à 4h du matin qu'à 18h.

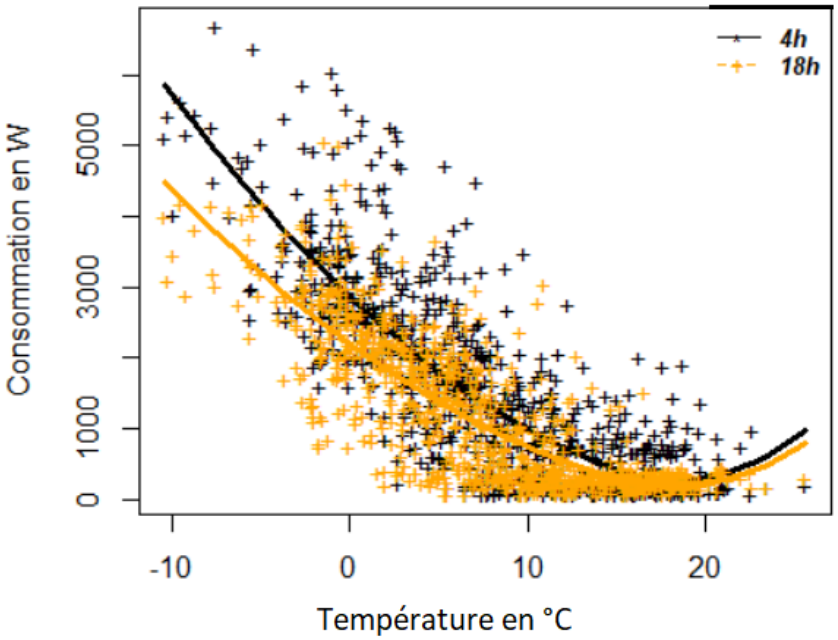
La figure 4.12 montre la relation de la consommation d'électricité au pas demi-horaire en (W) en fonction de la température extérieure en °C à différentes heures de la journée du ménage thermosensible 2 de la figure 4.11b. Les graphiques montrent clairement que la partie thermosensible de la consommation est plus sensible à la variation de la température extérieure de 00h00 à 08h00 du matin, les plages horaires qui correspondent aux heures creuses du ménage.

En plus, il existe une relation entre la moyenne de la consommation d'un ménage et sa thermosensibilité. La figure 4.13 montre la relation entre la moyenne de la consommation journalière et la thermosensibilité des ménages dans le jeu de données. Cette thermosensibilité est définie comme étant la corrélation linéaire au carré entre la demande de l'électricité d'un ménage au pas demi-horaire et la température extérieure (voir (Gerossier, 2019)). Nous remarquons que plus la thermosensibilité du ménage augmente plus la moyenne de sa consommation journalière augmente. Ce résultat attendu est justifié par le fait que le chauffage est l'appareil électrique le plus énergivore [4], par conséquent, les ménages équipés par des chauffages électriques consomment plus de l'énergie électrique par rapport à leurs homologues qui ne possèdent pas de chauffages électriques. Nous remarquons également que des ménages ayant la même thermosensibilté peuvent avoir des moyennes journalières de consommation très disparates en raison de la différence des modes de consommation des occupants, leurs nombres, leurs équipements électriques, ainsi que la surface de ces foyers (voir la figure 4.13).

4. https://ekwateur.fr/2019/04/05/appareils-electriques-consommation-electricite/



(a) Ménage thermosensible 1 (contrat tarif de base).



(b) Ménage thermosensible 2 (contrat tarif HC/HP).

FIGURE 4.11 – Nuage de points montrant la consommation d'électricité à 4h (points noires) et à 18h (points oranges) en fonction de la température extérieure pour deux ménages thermosensibles. Le premier ménage (a) ayant un contrat d'électricité d'option tarif de base et le deuxième ménage (b) ayant un contrat d'électricité d'option HC/HP. Les lignes oranges et noires sont estimées par des *splines* de régression.

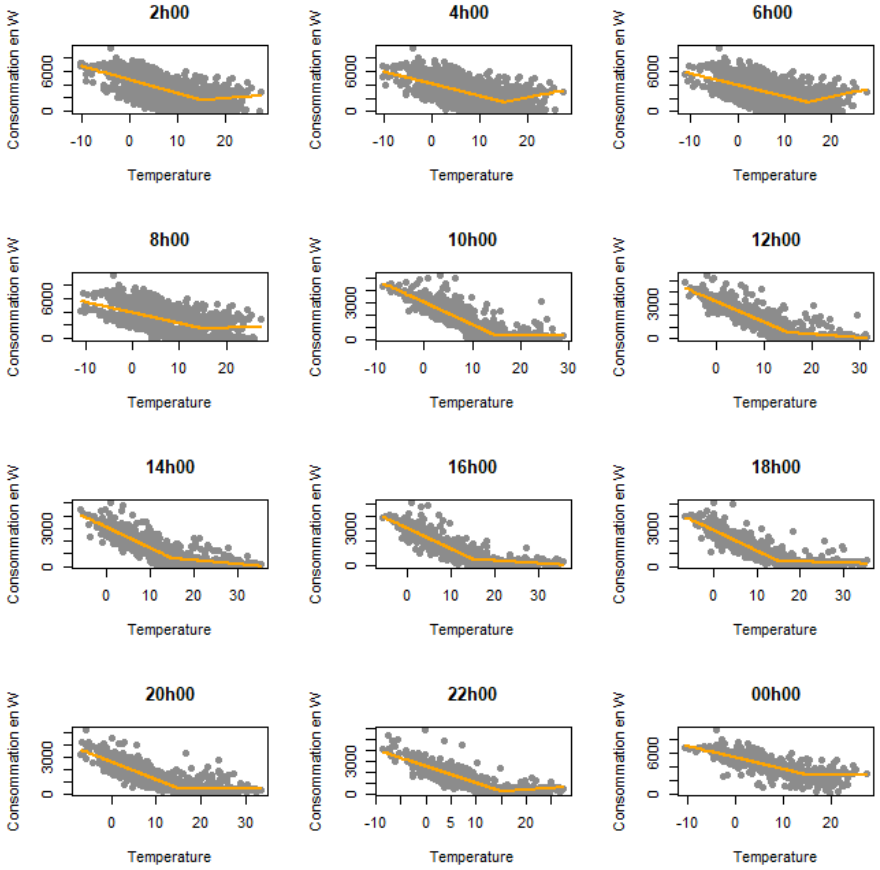


FIGURE 4.12 – Des nuages de points montrant la consommation d'électricité en W au pas demi-horaire (axe y) en fonction de la température extérieure en °C (axe x) à différentes heures de la journée du ménage thermosensible 2 de la figure 4.11b.

La figure 4.14 représente la relation entre la volatilité et la thermosensibilité des ménages. Nous remarquons que la figure 4.14 ne montre pas de corrélation linéaire entre la volatilité et la thermosensibilité des ménages. En effet, deux ménages ayant des degrés de thermosensibilité similaires peuvent avoir des indices de volatilité différents. Cette observation s'explique par le fait que la volatilité dépend de la manière dont les équipements électriques sont utilisés tout au long de la journée, que le ménage soit thermosensible ou non et qu'il dispose d'un chauffage électrique ou non.

Dans notre étude, nous avons distingué entre les courbes de charge thermosensibles (voir figure 4.15a) pour les clients possédant des chauffages électriques et les courbes de charge moins thermosensibles ou non-thermosensibles pour les clients qui ne possèdent pas de chauffages électriques (voir figure 4.15b) afin de pouvoir juger l'utilité de l'intégration de la température extérieure dans les modèles de prévision de la charge des ménages (voir la sous-section 4.2.4).

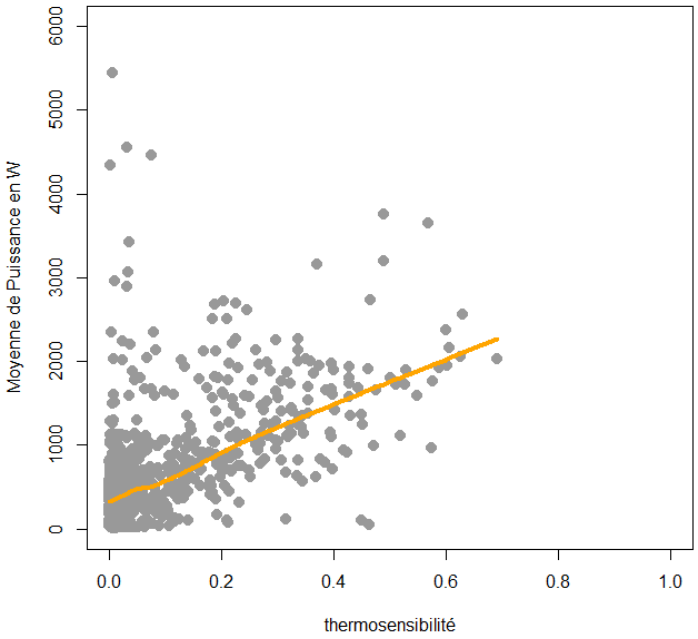


FIGURE 4.13 – Nuage de points montrant la moyenne de la consommation journalière des ménages dans le jeu de données en fonction de la thermosensibilité de ces derniers.

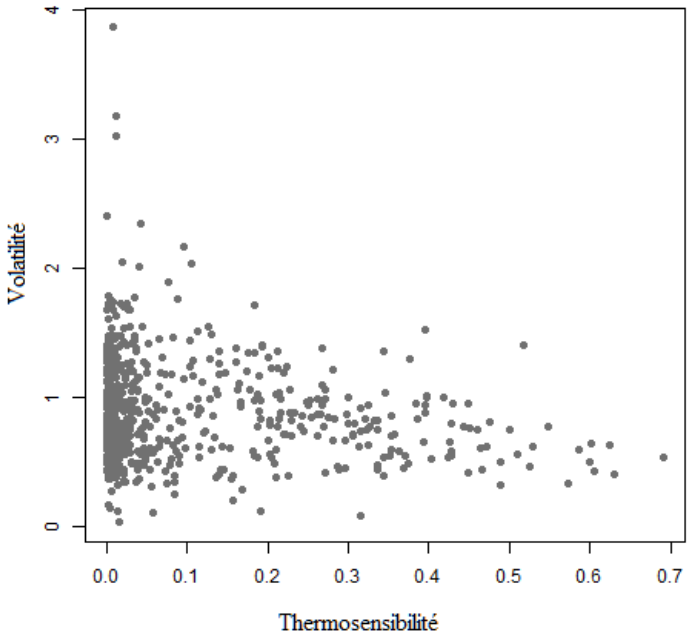
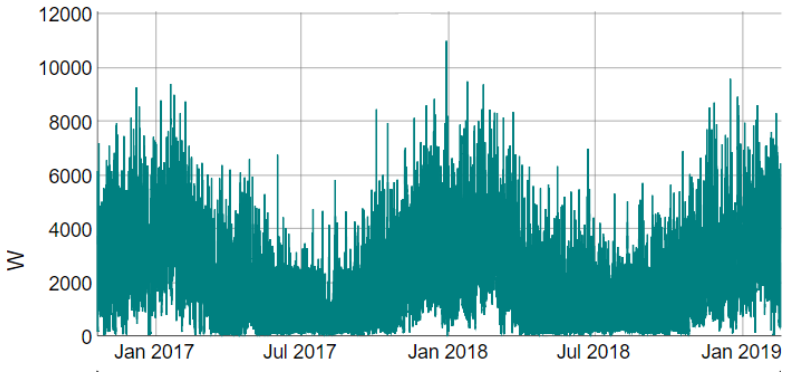
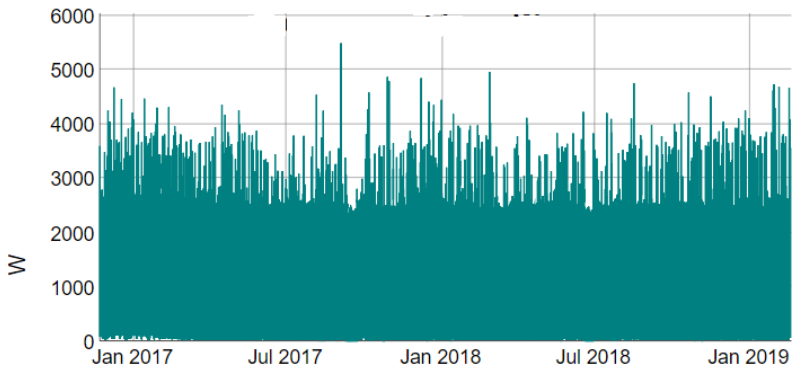


FIGURE 4.14 – Nuage de points montrant les indices de volatilité des ménages dans le jeu de données en fonction de la thermosensibilité de ces derniers.



(a) Ménage résidentiel thermosensible.



(b) Ménage résidentiel non thermosensible.

FIGURE 4.15 – Exemples de courbes de charge de 2017 à 2019.

Il est intéressant de souligner que d'autres facteurs météorologiques comme le rayonnement solaire, l'humidité, la vitesse du vent, la nébulosité et la précipitation ont également un impact sur les courbes de charge d'électricité. Par contre, la littérature montre que la température extérieure est la principale variable météorologique qui affecte la demande d'électricité (HABEN, GIASEMIDIS et al., 2019 ; GROSS et al., 1987 ; ENGLE et al., 1986 ; Henrique Steinherz HIPPERT et al., 2001). Par conséquent, et compte tenu de la complexité de la prévision de la charge à l'échelle des ménages, nous avons décidé d'étudier uniquement l'impact de la température extérieure comme variable météorologique sur la charge électrique des ménages.

4.2.2.3 La périodicité

Bien que la charge d'électricité à l'échelle des ménages est connue par son irrégularité, pour plusieurs raisons citées dans cette thèse dont la principale est la nature irrégulière du mode de vie des occupants, nous trouvons heureusement toujours des schémas ou des motifs répétés dans les courbes de charge des ménages sur lesquels se basent les prévisions. En effet, ces courbes de charge sont caractérisées par une périodicité journalière, hebdomadaire

et annuelle. La périodicité journalière est le résultat de la fidélité du consommateur à ses habitudes de consommation journalière comme ses heures de réveil, les heures d'utilisation de certains appareils électriques (machine à café, bouilloire, grille-pain, plaques de cuisson, lave-vaisselle, lave-linge, ...) ainsi que ses horaires de présence à domicile (voir figure 4.16).

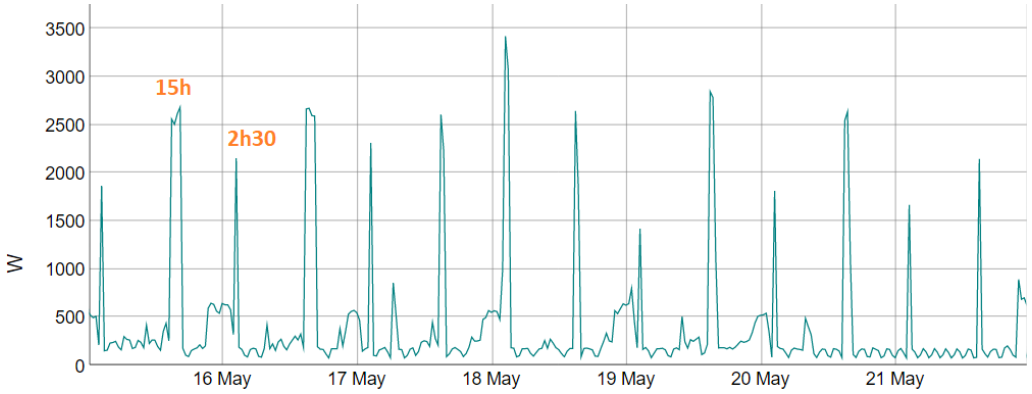


FIGURE 4.16 – Courbe de charge d'un ménage ayant des pics de consommation périodiques d'amplitudes différentes à 2h30 et 15h tous les jours pendant la semaine allant du 15 mai au 21 mai 2018. Ces pics sont liés au déclenchement du tarif « heure creuses » et la mise en route automatique du chauffe-eau.

Ensuite, la périodicité hebdomadaire est représentée par la différence des habitudes de consommation entre les jours de la semaine et ceux du week-end. En général, la consommation d'électricité augmente les jours de week-end puisque les occupants passent généralement plus d'heures à la maison pendant la plage horaire de 10h00 à 18h00 alors qu'ils passent ce temps au travail les autres jours de la semaine. En plus, nous remarquons un décalage dans les habitudes de consommation les jours de week-end puisque les individus du ménage se réveillent généralement plus tard par rapport aux autres jours de la semaine. La figure 4.17 montre les boîtes à moustaches de la consommation d'électricité au pas demi-horaire en fonction des jours de la semaine de deux ménages différents (M_1 et M_2). Cette figure montre la différence de la distribution de la consommation au pas demi-horaire entre les jours de la semaine et les jours de week-end. Pour chaque jour de la semaine, la ligne horizontale noire représente la médiane de la demande, la longueur des boîtes à moustaches correspond à l'intervalle interquartile et les points à l'extérieure des moustaches représentent en principe les valeurs extrêmes. D'après ce graphique, il est possible de remarquer que la majorité des distributions de consommation ont une asymétrie positive. Cette asymétrie s'explique par le fait que la consommation électrique des ménages est principalement composée de charges de base faibles (entre 0 et 500 W) la plupart du temps. Toutefois, l'utilisation occasionnelle d'appareils énergivores tels que les séchoirs, les sèche-linge et les pompes de piscine entraîne des pics de consommation d'électricité, affectant ainsi la symétrie de la distribution (YILDIZ et al., 2017). Nous remarquons également que pour les deux ménages M_1 et M_2 la médiane de consommation de deux jours de week-end (samedi et dimanche) est la plus élevée par rapport aux autres jours de

la semaine et que l'écart interquartile est plus étalé pour les jours de week-end que pour le reste des jours de la semaine.

La sensibilité de la consommation d'électricité aux conditions météorologiques et plus particulièrement à la température extérieure représente en grande partie la périodicité annuelle des courbes de charge (voir figure 4.15a). Cette périodicité dépend également des périodes de vacances et de la variation de la durée des jours (*daylight duration*).

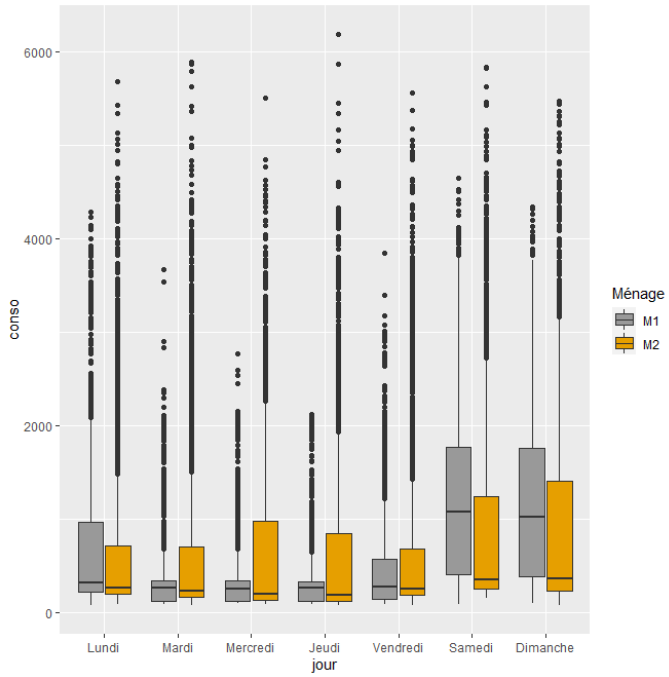


FIGURE 4.17 – Les boîtes à moustaches de la consommation d'électricité au pas demi-horaire de deux ménages M_1 et M_2 en fonction des jours de la semaine.

En conclusion, l'analyse des courbes de charge électrique des ménages dans le jeu de données révèle une grande variabilité dans la volatilité, la thermosensibilité et la périodicité de ces courbes. Cette variabilité est due aux différences dans le mode de vie des occupants, à l'isolation thermique du logement, au type de chauffage utilisé, aux préférences de température, aux habitudes quotidiennes et saisonnières des occupants. Toutefois, cette variabilité rend la prévision de la charge électrique à l'échelle des ménages difficile, car elle dépend de nombreux facteurs interdépendants et variables. Les modèles de prévision doivent donc prendre en compte ces facteurs pour être efficaces.

4.2.3 Prétraitement des données

L'une des étapes les plus importantes dans la modélisation et le développement des modèles de prévision est le pré-traitement des données. Ce processus prépare les données pour l'analyse en traitant ou en supprimant les données incorrectes ou dupliquées.

Dans notre étude les données ont été traitées pour remplacer les données manquantes lorsque moins de 10% de données manquaient. Les courbes de charge avec plus de 10% de données manquantes sont exclus de l'étude. Dans le cas des données manquantes intrajournalières, les données sont remplacées par une simple interpolation. Dans le cas où plus de heures manquaient, les valeurs manquantes sont remplacées par les données de la semaine précédente (même jour, même heure).

Les données dont nous disposons sont des relèves des compteurs par PDL [5] pour la période allant de janvier 2017 jusqu'à janvier 2019 et non pas par client ou par contrat. Cela signifie qu'il peut exister dans les données une courbe de charge de plusieurs clients successifs d'un même logement. Ces courbes de charge sont exclus également de l'étude afin de garantir que l'entraînement et le test du modèle de prévision s'effectuent sur les données d'un même client (voir figure 4.18). L'étape de prétraitement a réduit la taille de notre échantillon de 1000 à 720 courbes de charge.

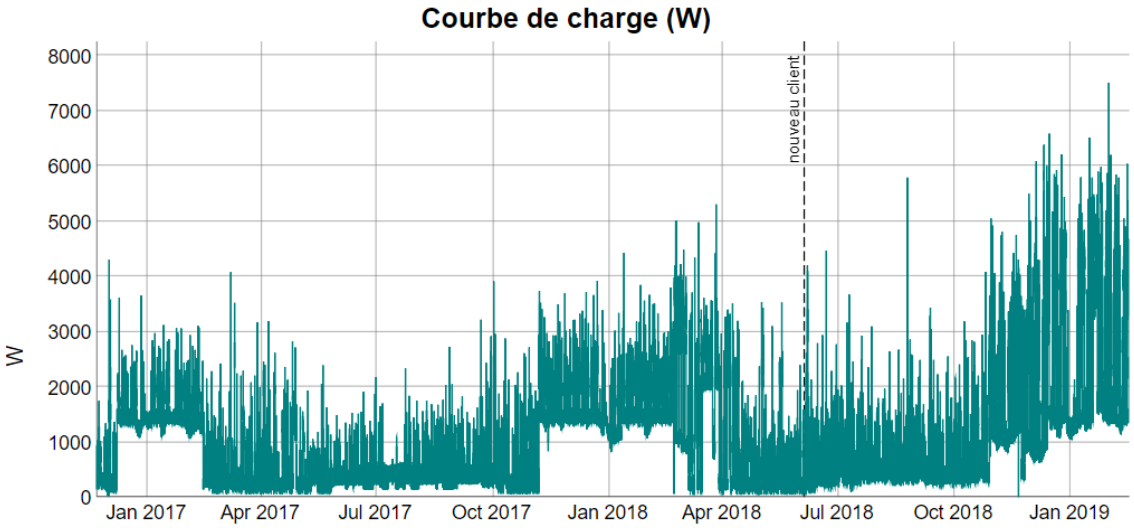


FIGURE 4.18 – Courbe de charge de deux clients successifs d'un même logement. La courbe de charge montre un changement dans le mode et le niveau de consommation à partir de 05-06-2018.

4.2.4 Segmentation des courbes de charge suivant le critère de thermosensibilité

Nous distinguons dans notre étude les courbes de charge thermosensibles des courbes de charge non-thermosensibles par le rapport que nous notons P_{relative} de la moyenne de consommation en hiver à la moyenne de consommation en été :

5. Le PDL, Point de Livraison, c'est un numéro sur le compteur électrique qui permet d'identifier le logement. Il ne faut pas le confondre avec le numéro du compteur qui peut changer lorsque le compteur est remplacé, le PDL, lui, reste toujours le même (NOÉMIE, 2018).

$$P_{\text{relative}} = \frac{P_{\text{moy}}^{\text{hiver}}}{P_{\text{moy}}^{\text{été}}}$$

où $P_{\text{moy}}^{\text{hiver}}$ est la moyenne des puissances en hiver (décembre, janvier et février) et $P_{\text{moy}}^{\text{été}}$ est la moyenne des puissances en été (juin, juillet et août). Si le rapport P_{relative} est supérieur à 1,5 la courbe de charge est considérée comme thermosensible. Suite à cette segmentation les données sont divisées en 312 courbes de charge thermosensibles et 408 courbes de charge non-thermosensibles. Cette segmentation est intéressante afin de pouvoir étudier l'intérêt de l'intégration de la température extérieure dans les modèles de prévision de la charge des ménages.

4.3 Prévision par le modèle *KWF*

Le modèle *KWF* (défini dans la sous-section 3.2.2) est un modèle de prévision des séries chronologiques fonctionnelles en présence de non stationnarités. Ce modèle a été appliqué avec succès à la prévision de la demande d'électricité en France, où il a montré des performances similaires à celles d'autres modèles de prévision conçus pour le même objectif (Cugliari, 2011). En revanche, il est capable de traiter les données d'une façon plus simple et parcimonieuse (Antoniadis, Brossat et al., 2014). De plus, ce modèle est capable de fournir des prévisions à plusieurs horizons d'une manière simultanée en se basant sur toutes les trajectoires similaires passées et non pas sur quelques points dans le passé comme c'est le cas de plusieurs modèles de prévision (Antoniadis, Brossat et al., 2014).

En prenant en compte les différents avantages théoriques du modèle *KWF* (voir la sous-section 3.2.2), il est possible de supposer que ce modèle peut apporter des résultats probants à la prévision de la charge électrique à l'échelle des ménages pour plusieurs raisons. Tout d'abord, l'approche fonctionnelle utilisée par le modèle *KWF* permet de modéliser la relation entre la charge électrique et le temps d'une manière plus précise que les méthodes de modélisation traditionnelles. En effet, la représentation des données de charge électrique sous forme de fonctions permet de prendre en compte les variations temporelles complexes et non-linéaires dans l'historique de données. Par conséquent, cela peut conduire à des prévisions plus précises de la charge électrique. Ensuite, la décomposition en ondelettes peut permettre de détecter des motifs de consommation récurrents à différentes échelles de temps ce qui peut améliorer la précision des prévisions de la charge électrique à l'échelle des ménages. Enfin, le modèle *KWF* peut capturer et traiter les changements dans les habitudes de consommation en ayant la capacité de traiter les non-stationnarités dans les données.

En outre, le modèle *KWF* présente un avantage supplémentaire dans l'industrie, car

il peut être utilisé pour l'imputation des données manquantes. En effet, lorsqu'il y a des valeurs manquantes dans les données de charge électrique à l'échelle des ménages, le modèle *KWF* est capable de prédire ces valeurs manquantes en se basant sur les trajectoires similaires passées. Cette capacité est particulièrement importante dans l'industrie de l'électricité.

En résumé, le modèle *KWF* peut être utilisé pour la prévision de la charge électrique à l'échelle des ménages ainsi que pour l'imputation des données manquantes. Pour plus d'informations sur les fondements théoriques du modèle ainsi que ses avantages et ses limites, l'auteur peut se référer à la sous-section 3.2.2.

4.3.1 Approche de prévision

Notre objectif est de réaliser des prévisions de consommation d'électricité pour chaque ménage du jeu de données pour une journée à l'avance en utilisant le modèle *KWF*. Ces prévisions sont fournies pour chaque demi-heure sur une période de 24 heures. Notre approche suppose que les données réelles de consommation d'électricité sont disponibles la veille de la journée de prévision et avant la période de prévision. Dans les paragraphes suivantes, nous décrirons en détail la mise en œuvre de notre approche de prévision par le modèle *KWF* ainsi que la méthode utilisée pour évaluer les résultats obtenus.

4.3.1.1 Préparation des données

Les étapes de préparation des données que nous avons suivies sont les suivantes :

1. **Décomposition en courbes de charge journalières.** Elle consiste à diviser la courbe de charge électrique de chaque ménage en courbes de charge journalières (segments). Chaque courbe de charge journalière est formée de 48 observations, correspondant aux mesures demi-horaires de la consommation électrique du ménage pour chaque jour. Cette étape de préparation des données permet de modéliser la consommation électrique à l'échelle journalière.
2. **Interpolation des courbes de charge journalières.** Cette étape s'applique à toutes les courbes de charge journalières dans les données de consommation de chaque ménage. Elle a pour but de transformer toutes les données de consommation de chaque ménage en échelle dyadique, c'est-à-dire en puissance de 2. Cette transformation permet l'application directe de l'algorithme pyramidal de Stephane Georges Mallat (1988) sur les données et leur décomposition en ondelettes (voir annexe 6.2). Cette étape permet alors d'obtenir des données interpolées à 2^J points par segment pour $J = 6$, au lieu des 48 points d'origine.
3. **Division des données en jeu d'entraînement et jeu de test.** Les relevés de

chaque ménage ont été divisés en deux ensembles : le jeu d'entraînement et le jeu de test. L'ensemble d'entraînement contient 70% des données, couvrant la période entre le 01-01-2017 et le 30-06-2018, tandis que l'ensemble de test contient les 30% restants, couvrant la période entre le 01-07-2018 et le 01-01-2019. Cette division permet d'évaluer la performance du modèle sur une période de temps distincte de celle sur laquelle il a été entraîné, et donc de mesurer sa capacité à généraliser à de nouvelles données.

4. **Détermination des groupes de calendrier.** Le modèle *KWF* dans Antoniadis, Paparoditis et al. (2006) est utilisé pour prédire la charge électrique à partir de l'historique de consommation d'électricité, sans tenir compte de l'effet de calendrier. Cependant, l'analyse des courbes de charge électrique a mis en évidence une variation sensible de la consommation en fonction des jours de la semaine et des jours fériés, tels que souligné par Lusis et al. (2017). Cette variation se traduit par des différences dans la forme des courbes de charge journalières et remet en question l'hypothèse de stationnarité des données de consommation électrique nécessaire pour l'application du modèle *KWF*. En général, les week-ends et les jours fériés sont souvent associés à une consommation d'électricité plus élevée en raison de la présence à domicile plus que les autres jours de la semaine. De plus, en France, le mercredi est souvent considéré comme un jour où la consommation d'électricité est plus élevée que les autres jours de la semaine pour les ménages ayant des enfants puisque c'est un jour de repos. Afin de prendre en compte l'effet de calendrier dans le modèle *KWF*, Cugliari (2011) a proposé d'intégrer des groupes de jours dans le modèle *KWF*. Cette intégration se fait lors de l'étape de prévision en utilisant un traitement spécifique pour le calcul de la ressemblance. Un vecteur de poids $\tilde{w}_{n,m}$ est défini entre les jours m et n, pour tout m variant de 1 à $n-1$ par

$$\tilde{w}_{n,m} = \begin{cases} w_{n,m} & \text{si } gr(m) = gr(n) \\ 0 & \text{sinon} \end{cases}$$

où $gr(n)$ indique le groupe du n-ème jour et $w_{n,m}$ a été défini à l'équation (3.26). Donc les valeurs de l'indice pour tous les jours m qui n'appartiennent pas au même groupe que le jour n sont mises à zéro. Dans notre approche, nous avons inclus neuf groupes déterministes, à savoir les sept jours de la semaine ainsi que les jours qui précèdent les jours fériés et les jours fériés eux-mêmes. L'intégration du groupe des jours qui précèdent les jours fériés sert à informer le modèle que le jour à prévoir est un jour particulier.

5. **Choix des paramètres du modèle.** Bien que l'optimisation des paramètres d'un modèle de prévision puisse contribuer significativement à améliorer sa performance, elle peut se révéler difficilement réalisable en raison de la taille de l'échantillon de données et de la complexité du modèle. Ainsi, pour économiser des ressources et du temps tout en obtenant des résultats satisfaisants, l'utilisation de paramètres par défaut ou de connaissances spécialisées peut être une alternative efficace. Ce qui

est le cas pour la prévision de la charge électrique à l'échelle des ménages, nous avons utilisé les recommandations expérimentales présentées dans Cugliari (2011) pour sélectionner les paramètres appropriés du modèle *KWF*, plutôt que d'optimiser chaque paramètre pour l'ensemble des données, ce qui aurait nécessité des ressources importantes en termes de temps et de puissance de calcul. Nous avons sélectionné la famille d'ondelettes *Daubechies least asymmetric wavelets* (DaubLeAsymm) avec un filtre de taille six pour notre modèle. Selon les recommandations de Cugliari (2011), cette famille est mieux adaptée pour capturer les variations saisonnières et prédire les jours fériés. Pour la sélection du noyau de régression, nous avons choisi le noyau gaussien.

6. **Application du modèle *KWF* aux données.** Pour effectuer les prévisions à l'aide du modèle *KWF*, nous avons utilisé le `package enercast` du logiciel R [6]. Les prévisions ont été réalisées pour chaque jour de l'ensemble de test de chaque ménage en ajoutant progressivement les nouvelles données collectées la veille du jour à prévoir à l'historique, de manière à éviter toute lacune dans les données. Ce processus permet de reproduire le protocole opérationnel de prévision qui sera déployé en industrie.

4.3.2 Évaluation et *Benchmarking*

4.3.2.1 Métriques d'évaluation

Afin de mieux comprendre et analyser les résultats, plusieurs métriques d'évaluation des erreurs ont été utilisées notamment la métrique (sMAPE), la métrique (NMAE), et la métrique (NRMSE) ainsi que la métrique (MASE). Une description de ces quatre métriques est fournie dans la sous-section 3.4.1. La métrique MASE que nous avons utilisée compare l'erreur MAE, également définie à la sous-section 3.4.1, du modèle de prévision à l'erreur MAE du modèle de prévision saisonnier qui définit la prévision du jour $(J+1)$ comme étant la consommation réelle du jour précédent (J).

4.3.2.2 Les modèles de référence

Une méthode courante pour évaluer les performances des modèles de prévision est de les comparer à des modèles de référence ou à des modèles présentés dans la littérature, appelée *Benchmarking*. Cette approche permet de démontrer l'efficacité des modèles proposés par rapport à des méthodes naïves souvent utilisées dans l'industrie ou à des modèles concurrents de la littérature. Dans notre étude, nous avons utilisé deux modèles de référence pour la comparaison.

1. **Un modèle de forêt aléatoire.** Nous avons traité chaque courbe de charge comme

6. https ://github.com/cugliari/enercast

48 séries chronologiques quotidiennes distinctes, puis ajusté un modèle pour chaque demi-heure de la journée pour capturer la saisonnalité quotidienne de la demande des ménages. En conséquence, nous avons ajusté 48 modèles de forêts aléatoires pour chaque ménage, pour prévoir la demande d'électricité à un jour d'avance. Pour modéliser la saisonnalité hebdomadaire, nous avons intégré une variable catégorielle qui prend en compte le type de journée. Les entrées du modèle sont résumées dans la table 4.3.

Caractéristique des données	**Entrée du modèle**
Saisonnalité hebdomadaire	les jours de la semaine/ week-end
Saisonnalité journalière	décalage d'un jour de la courbe de charge
La thermodépendance	décalage d'un jour de la température extérieure
Effet calendrier	jour férié

TABLE 4.3 – Tableau résumant les différentes entrées du modèle de forêt aléatoire proposé pour la prévision de la consommation des ménages.

Le modèle de forêt aléatoire est implémenté dans le `package randomForest` du logiciel R. Nous avons également paralléliser l'étape de l'entraînement du modèle sur différents cœurs de calcul (parallélisation sur des processeurs multicœurs) afin de réduire le temps de calcul majoritairement lié à l'ajustement de 48 modèles pour la prévision de chaque demi-heure.

2. **Un modèle climatologique.** Ce modèle a été introduit pour prédire les données météorologiques en 1977 par MURPHY (1977). Ensuite, il a été utilisé comme modèle de référence pour la prévision de l'énergie solaire photovoltaïque (KREUWEL et al., 2020) et la prévision de la demande d'électricité des ménages (GEROSSIER, 2019). Le modèle calcule des prévisions quantiles basées sur l'historique des données sans condition, ce qui signifie que le modèle calcule un profil de consommation pour un jour donné indépendamment de la variation de la température ou des effets du calendrier (jours fériés ou non). Dans notre approche, nous avons choisi de calculer les prévisions en utilisant les moyennes. Pour chaque demi-heure d'un jour donné de la semaine, la moyenne de la consommation est calculée sur la période d'entraînement, de sorte qu'un lundi à 17h a la même prévision toute l'année calculée comme étant la moyenne de toutes les observations des lundis à 17h dans la période d'entraînement.

4.3.3 Résultats

Le tableau 4.4 résume les performances moyennes des trois modèles de prévision à savoir *KWF*, le modèle de forêt aléatoire et climatologique à $(J+1)$ appliqués au jeu de données à la fois pour les courbes de charge thermosensibles et non-thermosensibles. Dans le cadre de ce manuscrit, nous utilisons le terme « performance moyenne » pour

désigner les moyennes des performances de plusieurs modèles évalués sur l'ensemble des courbes de charge dans le jeu de données. Pour obtenir ces performances, chaque modèle a été entraîné sur l'ensemble de données d'entraînement de chaque courbe de charge, puis évalué sur un ensemble de données de test distinct. Les performances de chaque modèle ont ensuite été calculées sur l'ensemble de données de test de chaque courbe de charge, et les valeurs présentées dans la table 4.3 sont les moyennes de ces performances pour toutes les courbes de charge dans le jeu de données.

Modèle	**Thermosensibilité**	NMAE	NRMSE	MASE	sMAPE
KWF	Non-thermosensible	0,47	0,72	0,81	44,57
	Thermosensible	0,44	0,76	0,79	54,46
Forêt aléatoire	Non-thermosensible	0,51	0,86	0,94	46,16
	Thermosensible	0,51	0,87	0,87	55,20
Climatologique	Non-thermosensible	0,50	0,85	0,94	47,15
	Thermosensible	0,79	1,14	1,44	79,00

TABLE 4.4 – La performance moyenne des modèles de prévision *KWF*, forêt aléatoire et climatologique à $(J+1)$ selon les quatre erreurs sélectionnées. Les meilleurs résultats sont affichés en bleu pour les courbes de charge non-thermosensibles, tandis que les meilleurs résultats pour les courbes de charge thermosensibles sont affichés en orange.

D'après les résultats présentés dans ce tableau, nous pouvons conclure que la performance moyenne du modèle *KWF* dépasse la performance moyenne des deux autres modèles de référence (forêt aléatoire et climatologique) pour notre échantillon test en termes des quatre métriques NMAE, NRMSE, MASE et sMAPE. D'après les valeurs obtenues de l'erreur MASE qui sont inférieures à 1 dans le cas du modèle *KWF*, nous pouvons conclure que la performance moyenne de ce modèle dépasse également la performance de la méthode de prévision naïve utilisée dans le calcul de la métrique MASE qui définit la prévision du jour $(J+1)$ comme étant la consommation réelle du jour précèdent (J) de l'ordre de 20%. Nous concluons également que la performance des modèles se dégrade d'une façon remarquable lorsqu'ils sont testés sur les données des courbes de charge thermosensibles surtout dans le cas du modèle climatologique, à l'exception, du modèle *KWF* qui prédit mieux en moyenne la charge thermosensible (erreur NMAE en moyenne égale à 0,44 pour la charge thermosensible contre 0,47 en moyenne pour la charge non-thermosensible) et le modèle de forêt aléatoire qui garde la même performance de prévision (l'erreur NMAE en moyenne est égale à 0,51 pour les courbes de charge thermosensibles et non-thermosensibles).

Le modèle *KWF* diminue l'erreur NMAE en moyenne de **6,3%** (**5,78%** pour l'erreur sMAPE, **18%** pour l'erreur NRMSE et **16%** pour l'erreur MASE) par rapport au modèle climatologique et de **8,5%** (**3,26%** pour l'erreur sMAPE, **19,4%** pour l'erreur NRMSE et **16%** pour l'erreur MASE) par rapport au modèle forêt aléatoire pour les courbes de

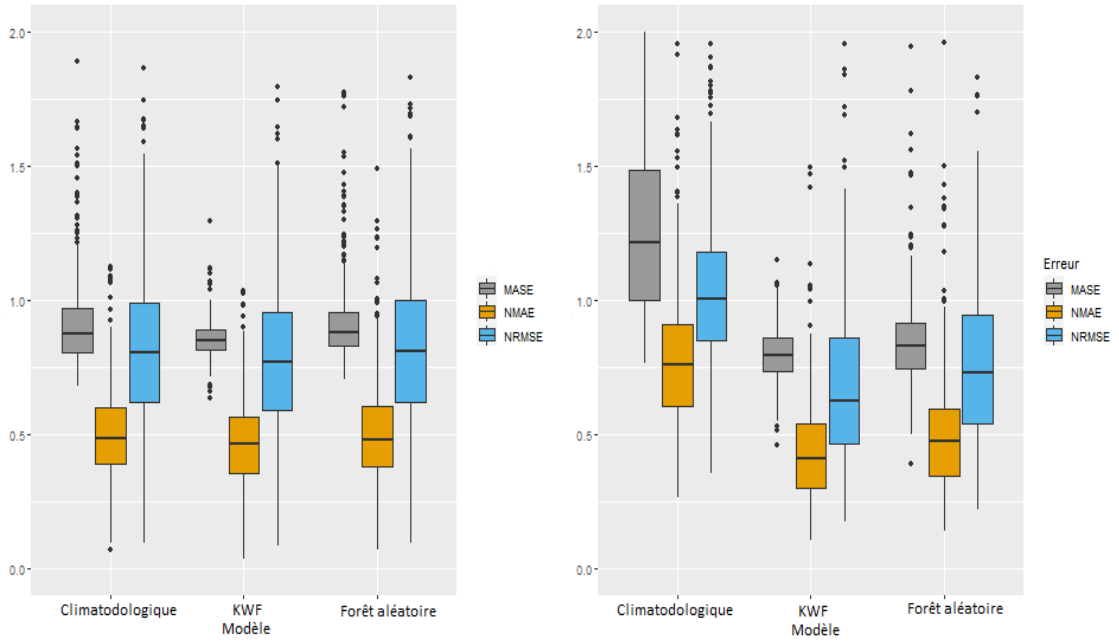
charge non-thermosensibles.

Cette diminution **en moyenne** est beaucoup plus remarquable au niveau des courbes de charge thermosensibles. En effet, elle est de l'ordre de **79,54%** en moyenne pour l'erreur NMAE par rapport au modèle climatologique (**50%** pour l'erreur NRMSE, **82,2%** pour l'erreur MASE et **45,06%** pour l'erreur sMAPE) et elle est de l'ordre de **15,9%** en moyenne pour l'erreur NMAE par rapport au modèle de forêt aléatoire (**14,4%** pour l'erreur NRMSE, **10,12%** pour l'erreur MASE et **1,35%** pour l'erreur sMAPE).

Étant donné que la moyenne des erreurs est très sensible aux valeurs extrêmes, nous avons fait le choix d'examiner également les distributions des erreurs de trois modèles à l'horizon $(J+1)$. La figure 4.19 montre les distributions des trois erreurs NMAE, NRMSE et MASE par modèle de prévision *KWF*, climatologique et forêt aléatoire. Les résultats obtenus renforcent les conclusions précédentes et prouvent que le modèle *KWF* est de meilleure performance prédictive à $(J+1)$ pour notre échantillon test à la fois en moyenne et en distribution des erreurs pour les quatre métriques testées. Les médianes des erreurs de prévision NMAE, NRMSE et MASE par le modèle *KWF* sont inférieures respectivement aux médianes des erreurs de prévision par le modèle climatologique et le modèle de forêt aléatoire à la fois pour les courbes de charge thermosensibles et non-thermosensibles. Pour les courbes de charge non-thermosensibles, les boîtes à moustaches des distributions de l'erreur MASE sont toutes en dessous de 1 indiquant ainsi que les trois modèles sont globalement plus performants que la méthode de prévision naïve. Par contre, la présence des valeurs aberrantes montre que dans certains cas la méthode de prévision naïve est plus performante que les trois modèles testés. En effet, certains ménages ont une consommation plus régulière que les autres et présentent une saisonnalité quotidienne très forte dans leurs courbes de charge qui fait qu'une méthode de prévision naïve est suffisante pour prédire leurs consommations. Nous remarquons également que le modèle *KWF* a moins de valeurs aberrantes que les deux autres modèles. Pour les courbes de charge thermosensibles, les mêmes conclusions peuvent être tirées des distributions de l'erreur MASE à l'exception du modèle climatologique qui montre une dégradation globale de la performance par rapport à la méthode de prévision naïve contrairement aux deux autres modèles. Un tel résultat a été attendu puisque le modèle climatologique calcule les prévisions en moyennant la consommation passée en fonction des jours de la semaine et des heures de la journée sans tenir compte des données météorologiques alors que les températures extérieures sont intégrées dans le modèle de forêt aléatoire. Par contre, le modèle *KWF* n'intègre aucune information sur la température extérieure mais il s'est révélé capable de s'adapter à la prévision de la charge thermosensible.

Les boîtes à moustaches des erreurs NRMSE de trois modèles dans le cas des courbes de charge thermosensibles et non-thermosensibles sont plus étalées que les boîtes à moustaches des erreurs NMAE et leurs moustaches supérieures dépassent la valeur 1. Étant donné que l'erreur NRMSE pénalise plus fortement les grandes erreurs, alors une erreur

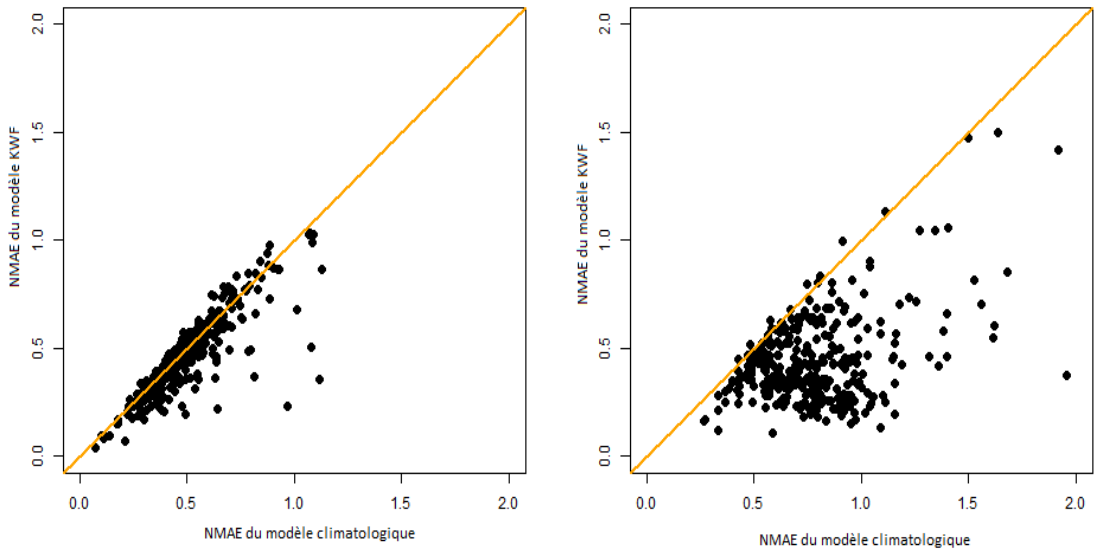
NRMSE élevée dans le contexte de la prévision de la charge électrique est le résultat d'une sous-estimation ou surestimation des pics de consommation ainsi qu'un déplacement de la prévision des pics de consommation dans la journée. Par conséquent, la disparité observée dans les erreurs NRMSE des trois modèles est due à la disparité de la régularité des pics des courbes de charge des ménages.



(a) Courbes de charge non-thermosensibles. (b) Courbes de charge thermosensibles.

FIGURE 4.19 – Les boîtes à moustaches des erreurs MASE, NMAE et NRMSE par modèle de prévision (climatologique, *KWF* et forêt aléatoire).

Pour avoir une idée plus précise de la performance des modèles à l'échelle des ménages individuellement, les erreurs de prévision NMAE du modèle *KWF* sont tracées en fonction des erreurs NMAE du modèle de forêt aléatoire et climatologique respectivement pour chaque ménage dans notre échantillon de données. Le résultat est présenté dans les figures 4.20 et 4.21. La figure 4.20a compare la performance du modèle *KWF* et celle du modèle climatologique par rapport à la métrique d'erreur NMAE pour les courbes de charge non-thermosensibles, tandis que la figure 4.20b compare la performance de ces deux modèles pour les courbes de charge thermosensibles. Nous remarquons que la plupart des points dans les deux cas sont en dessous de la première bissectrice, ce qui signifie que globalement le modèle *KWF* est plus performant que le modèle climatologique pour la prévision du jour $(J+1)$ de la charge électrique thermosensible et non-thermosensible des ménages. Cette différence dans la performance des deux modèles est plus visible dans le cas des courbes de charge thermosensibles.

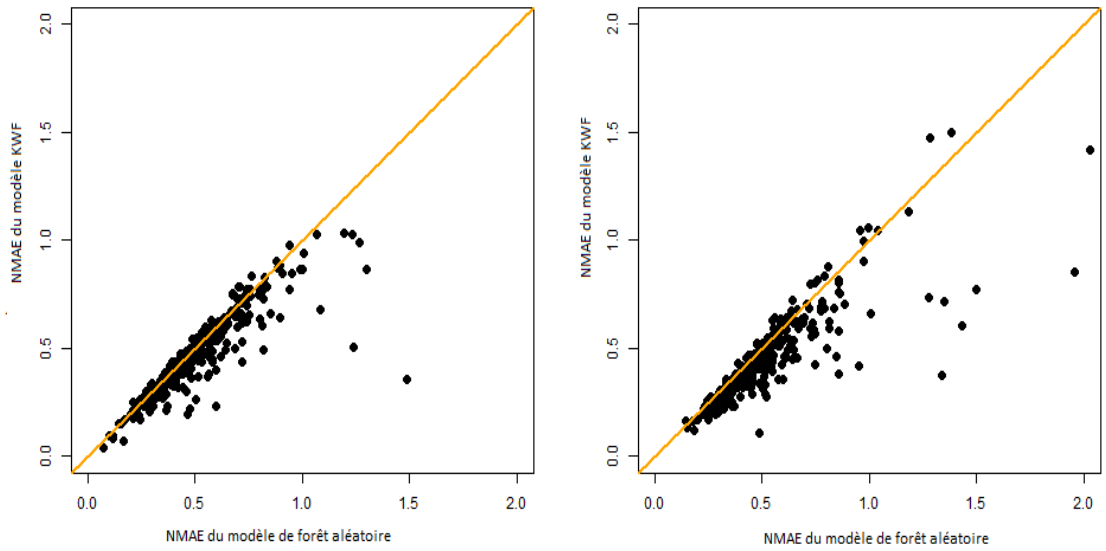


(a) Courbes de charge non-thermosensibles. (b) Courbes de charge thermosensibles.

FIGURE 4.20 – Nuage de points de l'erreur NMAE du modèle *KWF* en fonction de l'erreur NMAE du modèle climatologique pour les courbes de charge non-thermosensibles et thermosensibles.

La même conclusion peut être tirée de la figure 4.21 par rapport à la performance du modèle *KWF* vis-à-vis du modèle de forêt aléatoire. Il est important de souligner que le modèle *KWF* a une fréquence plus élevée de prévisions précises de la charge thermosensible que le modèle de forêt aléatoire malgré l'intégration de la température extérieure dans ce dernier.

Dans les deux figures 4.20 et 4.21 nous remarquons que certains points sont au-dessus de la première bissectrice. Cela indique que dans certains cas, les deux modèles climatologique et forêt aléatoire peuvent être légèrement plus efficaces que le modèle *KWF* pour la prévision de la charge de certains ménages. Ce résultat peut être expliqué par le fait que chez certains ménages la consommation d'électricité est très erratique ce qui rend la prévision par des modèles complexes comme le modèle *KWF* plus difficile et légèrement moins efficace que la prévision par le modèle de forêt aléatoire ou le modèle climatologique.



(a) Courbes de charge non-thermosensibles. (b) Courbes de charge thermosensibles.

FIGURE 4.21 – Nuage de points de l'erreur NMAE du modèle *KWF* en fonction de l'erreur NMAE du modèle de forêt aléatoire pour les courbes de charge non-thermosensibles et thermosensibles.

En conclusion, le résultat obtenu montre que la performance du modèle *KWF* dépasse celle des deux modèles climatologique et forêt aléatoire à la fois pour les charges thermosensibles et non-thermosensibles en moyenne et en distribution des erreurs. Les résultats obtenus montrent une hétérogénéité dans la prévision de la charge à l'échelle des ménages. Cette hétérogénéité est due à la différence du niveau de régularité dans les courbes de charge des ménages. La précision de la prévision des courbes thermosensibles par le modèle climatologique est médiocre. Par contre, le modèle de forêt aléatoire garde approximativement la même performance pour la prévision de la charge thermosensible et non-thermosensible à l'exception de l'erreur sMAPE. Le modèle *KWF* s'est révélé le plus performant et le mieux approprié parmi les modèles testés pour la prévision de la charge des ménages pour le jeu de données.

4.3.4 Pertinence du modèle *KWF*

Pour caractériser la pertinence individuelle du modèle *KWF*, nous avons examiné la relation entre les erreurs de prévision par le modèle *KWF* et quelques facteurs caractérisant la consommation d'électricité des ménages tels que la moyenne de consommation, la volatilité, ainsi que le type de tarification. Cette comparaison permettra de définir des critères d'application industrielle du modèle.

4.3.4.1 Relation entre l'erreur de la prévision et la moyenne de la consommation

Des études de la littérature soulignent l'existence d'une relation entre le niveau de l'agrégation de la charge électrique et l'erreur de la prévision de cette dernière Wijaya, SFRJ Humeau et al. (2014). Ces études ont montré que l'erreur de la prévision diminue avec l'augmentation du niveau d'agrégation de la charge électrique, ce qui justifie que la prévision de la charge électrique au niveau national ou agrégé est plus précise que la prévision de la charge des ménages. Par exemple Sevlian et al. (2014) ont introduit l'idée de l'effet d'agrégation sur la prévision de la charge électrique. Ils ont montré que la précision des prévisions, mesurée en erreur MAPE ou en NMAE s'améliore avec l'augmentation de la charge électrique moyenne jusqu'à un niveau critique au-delà duquel aucune amélioration ne peut être envisagée.

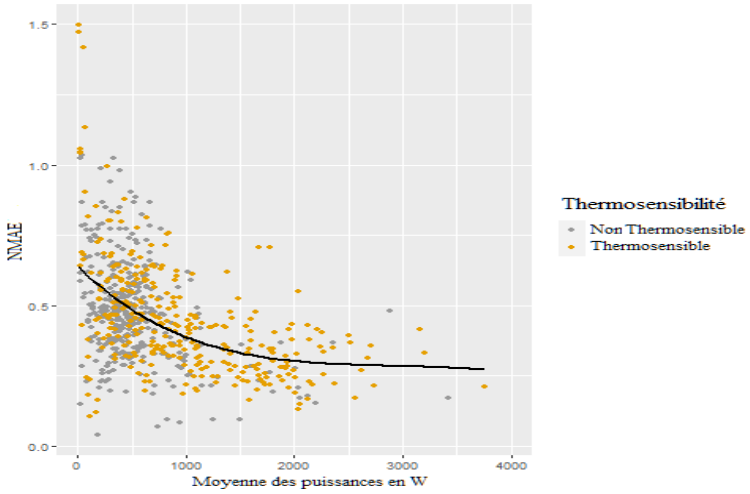


FIGURE 4.22 – Nuage de points de l'erreur NMAE du modèle *KWF* en fonction de la moyenne des puissances des ménages en W pour les courbes de charge non-thermosensibles et thermosensibles.

Dans notre cas, la performance du modèle *KWF* a été comparée à la puissance moyenne du ménage sur la figure 4.22. Cette dernière montre une diminution importante de l'erreur NMAE lorsque la puissance moyenne (en W) du ménage augmente à la fois pour les courbes thermosensibles et non-thermosensibles. Ce résultat rejoint ceux obtenus dans les deux articles cités précédemment. La figure 4.22 montre également que pour des puissances moyennes comparables ou similaires, la précision de la prévision par le modèle *KWF* varie significativement d'un ménage à un autre. Ce résultat est justifié par la différence de l'usage des équipements électriques et du mode de vie des occupants de ces ménages.

4.3.4.2 Relation entre l'erreur de la prévision et la volatilité

La figure 4.23 montre la relation entre l'erreur de la prévision sMAPE et la volatilité définie dans la sous-section 4.2.1 à la fois pour les courbes thermosensibles et les courbes de charge non-thermosensibles. Les résultats obtenus pour prévision de la charge électrique des ménages dans le jeu de données vérifient les hypothèses sur la relation entre la volatilité des courbes de charge à l'échelle des ménages et la précision de la prévision (Hou et al., 2021 ; Yildiz et al., 2017). Nous pouvons voir sur la figure 4.23 une corrélation linéaire positive entre l'erreur de la prévision sMAPE et l'indice de volatilité à la fois des courbes de charge thermosensibles et les courbes de charge non-thermosensibles.

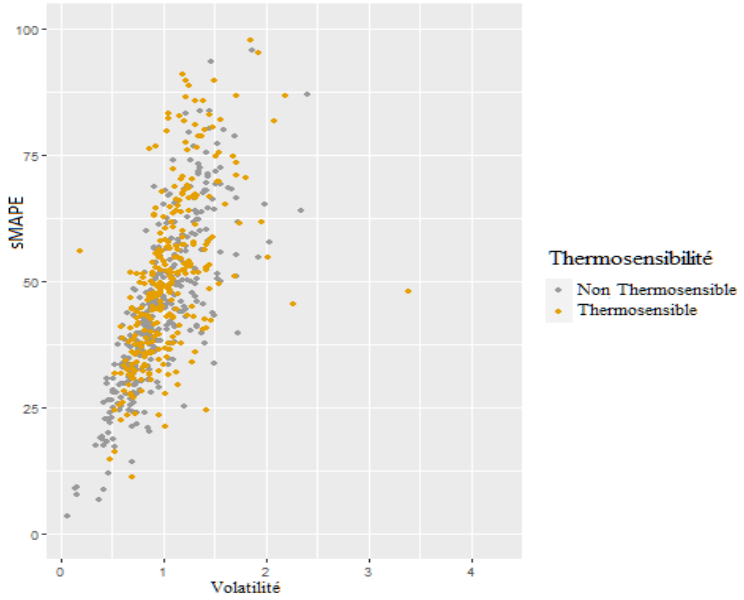
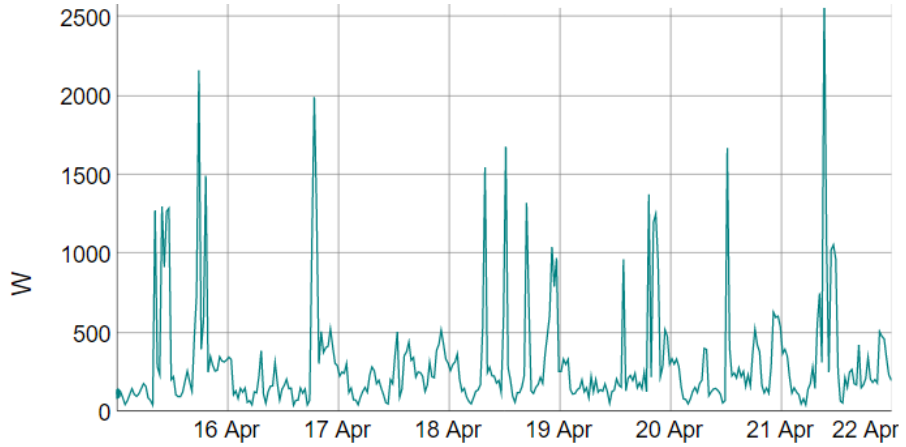


FIGURE 4.23 – Nuage de points de l'erreur sMAPE du modèle *KWF* en fonction de la volatilité des courbes de charge non-thermosensibles et thermosensibles.

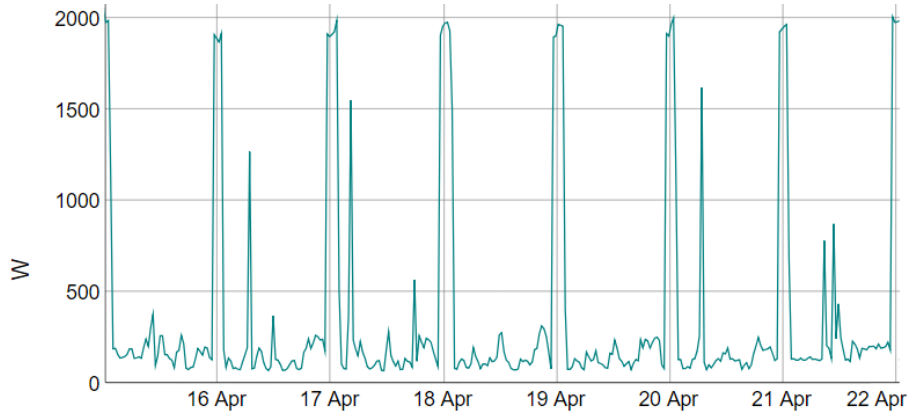
De plus, nous constatons que la précision de la prévision par le modèle *KWF* mesurée par l'erreur sMAPE varie considérablement entre les ménages ayant une volatilité similaire. Cela est dû premièrement au fait que la précision de la prévision de la charge d'un ménage ne dépend pas uniquement de sa volatilité. D'autres critères comme la régularité, le taux du bruit, la présence ou non des périodes de vacances pendant lesquelles le niveau de consommation est très bas et parfois quasiment nul ainsi que la thermosensibilité de la charge électrique ont également un impact sur la précision de la prévision de cette dernière.

4.3.4.3 Relation entre l'erreur de la prévision et l'option tarifaire

Les différentes options tarifaires de l'électricité ne donnent pas uniquement une information sur la manière dont l'électricité est facturée en fonction de l'heure de la journée mais



(a) Tarif de base.



(b) Tarif HC/HP.

6

FIGURE 4.24 – Extrait d'une semaine d'une courbe de charge de deux ménages M_3 et M_4.

également sur le mode de vie du consommateur. En effet, ces options ont été créées par les fournisseurs d'électricité pour encourager les clients particuliers à consommer pendant des plages horaires précises de la journée où la demande est généralement plus faible [7]. Par conséquent, les consommateurs qui optent pour l'option HC/HP ont tendance à consommer pendant les plages horaires des heures creuses (souvent huit heures par jour de minuit à 8h du matin). Ils programment leurs équipements électriques énérgivores dans des heures précises contrairement aux ménages qui optent pour l'option tarif de base qui ont tendance à étaler leur consommation sur la globalité de la journée (voir figure 4.24a). Par exemple, le démarrage du lave-linge à 20h plutôt qu'à 21h n'a aucun impact sur la facture d'électricité. Pour cette raison nous pouvons faire l'hypothèse que les ménages ayant un contrat avec option tarifaire HC/HP ont une consommation plus régulière (voir figure 4.24b) et par conséquent, moins difficile à prédire.

Le tableau 4.5 montre les moyennes des erreurs de prévision à l'horizon $(J+1)$ du modèle *KWF* appliqué à notre échantillon de données de 720 ménages. Les résultats obtenus valident l'hypothèse et montrent que la prévision par le modèle *KWF* des ménages

7. https://www.choisir.com/energie/articles/176212/bien-choisir-son-option-tarifaire

ayant un contrat avec option tarifaire HC/HP est plus précise et les résultats obtenus sont meilleurs en moyenne que ceux obtenus pour les ménages ayant un contrat avec option tarifaire de base.

Tarif	NMAE	NRMSE	MASE	sMAPE
Base	0,50	0,81	0,82	51,3
HC/HP	0,41	0,67	0,78	47,8

TABLE 4.5 – Les moyennes des erreurs du modèle de prévision KWF à $(J+1)$ par rapport aux options tarifaires (tarif de base et tarif HC/HP) des ménages.

La figure 4.25 montre les distributions des erreurs NMAE et NRMSE de prévision à $(J+1)$ du modèle KWF en fonction de l'option tarifaire. Nous pouvons constater que les médianes des erreurs NMAE et NRMSE de l'option de base sont supérieures à celles de l'option HC/HP. Nous pouvons également constater que les étendues pour l'option tarifaire de base sont plus larges que celles de l'option tarifaire HC/HP pour les deux métriques NMAE et NRMSE. De plus les distributions des erreurs pour les deux options tarifaires sont approximativement symétriques par contre pour l'option de base nous remarquons plus de valeurs aberrantes. Tous ces indices permettent de conclure que la qualité de la prévision des courbes de charge ayant l'option tarifaire HC/HP par le modèle KWF est meilleure que celle des courbes de charge ayant l'option tarifaire de base.

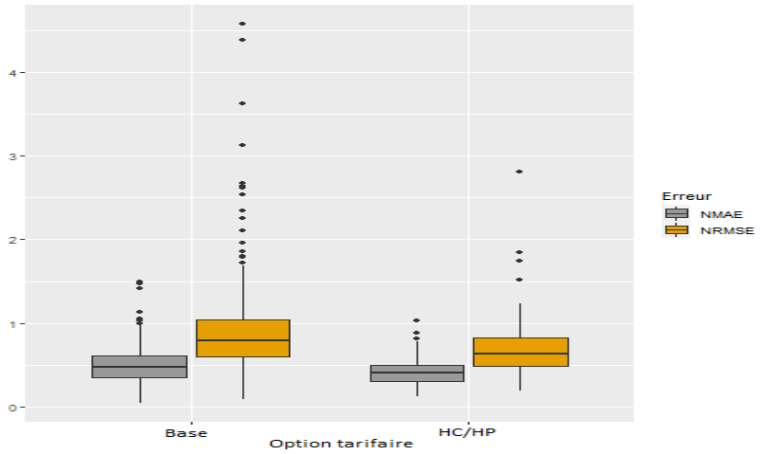


FIGURE 4.25 – Les boîtes à moustaches des erreurs NMAE et NRMSE du modèle de prévision KWF à $(J+1)$ par option tarifaire.

4.3.5 Intégration de l'impact de la température dans le modèle *KWF*

Jusqu'à présent, le modèle KWF testé sur les données ne prend pas en compte l'effet de la température extérieure. Il se base uniquement sur l'incorporation des groupes déterministes du calendrier. Dans sa thèse, (CUGLIARI, 2011) a présenté deux méthodes

de *clustering* pour les séries chronologiques fonctionnelles non stationnaires. Ces deux approches sont basées sur la transformée en ondelettes qui permet de décomposer les signaux en temps et en échelle, offrant ainsi la possibilité de regrouper les données fonctionnelles en groupes homogènes (voir l'annexe 6.2).

La première méthode proposée dans (Cugliari, 2011) consiste à regrouper les données fonctionnelles en utilisant des caractéristiques extraites de celles-ci, qui représentent la distribution de l'énergie à travers plusieurs échelles. Une étape de sélection de caractéristiques est appliquée avant le *clustering* visant à éliminer celles qui ne sont pas pertinentes et qui pourraient biaiser le résultat. La deuxième méthode utilise des mesures de similarité basées sur des outils de cohérence d'ondelettes.

Ces deux méthodes ont été appliquées aux courbes de charge journalières de la demande d'électricité en France, et les résultats obtenus ont été satisfaisants pour les deux méthodes. Les *clusters* obtenus par les deux méthodes ont été attribués avec succès à des jours spécifiques du calendrier qui présentent des caractéristiques précises en termes de température, tels que les jours froids ou chauds. La première méthode est généralement considérée comme à la fois facile à interpréter et rapide à calculer, ce qui peut être un avantage dans des situations où une analyse rapide et compréhensible est importante. En revanche, la deuxième méthode basée sur des mesures de similarité plus fines, telles que la cohérence d'ondelettes, peut donner des résultats plus raffinés.

Nous envisageons que ces méthodes de *clustering* soient utiles pour identifier des *clusters* regroupant des courbes de charge journalières homogènes en prenant en compte l'effet du calendrier et de la température à l'échelle des ménages. L'objectif est de vérifier si l'intégration de ces *clusters* dans le modèle de prévision *KWF* en remplacement les groupes déterministes du calendrier permet d'améliorer la précision de la prévision des courbes de charge thermosensibles en tenant compte de l'effet de la température.

Afin de vérifier si l'intégration des *clusters* identifiés à partir des courbes de charges homogènes selon la température peut améliorer la précision de la prévision des courbes de charge thermosensibles avec le modèle *KWF*, nous avons choisi d'implémenter la première méthode de *clustering*. Cette décision est motivée par des raisons d'interprétabilité ainsi que de rapidité, étant donné que la prévision par le modèle *KWF* doit être précédée par l'étape d'identification des *clusters* pour chaque ménage, ajoutant ainsi une charge de calcul supplémentaire. Il est donc préférable de privilégier la méthode la plus rapide et simple dans ce contexte.

Pour plus d'informations détaillées sur la méthode et les résultats, le lecteur intéressé peut se référer à la thèse de Cugliari (2011).

4.3.5.1 Description de la méthode de *Clustering*

Pour garantir l'efficacité de la prévision avec le modèle *KWF*, il est indispensable de vérifier l'hypothèse de stationnarité. Cependant, les données de consommation d'électricité à l'échelle nationale présentent deux sources de non-stationnarité en raison de l'effet saisonnier de la température et du calendrier. Pour répondre à ce défi, des méthodes de *clustering* des données fonctionnelles ont été développées dans l'objectif d'identifier des *clusters* qui peuvent satisfaire l'hypothèse de stationnarité. Ces méthodes de *clustering* cherchent à définir des *clusters* de courbes de charge journalières qui reflètent à la fois l'effet du calendrier et de la température sur la charge électrique. En d'autres termes, il s'agit de définir des *clusters* dont la stationnarité est vérifiée. Ces approches ont été décrites dans l'étude de Cugliari (2011).

Dans cette partie, nous présenterons la première méthode de *clustering* proposée dans Cugliari (2011), qui se base sur la décomposition en ondelettes discrètes des courbes de charge journalières. Cette méthode permet de calculer des caractéristiques reflétant la distribution de l'énergie à travers les échelles. Ces caractéristiques définiront à la suite les structures des *clusters*. Avant le *clustering*, une technique de sélection de caractéristiques est appliquée pour éliminer la redondance de l'information et améliorer les résultats et leur interprétation. Enfin, l'algorithme des *k-means* est utilisé pour déterminer les *clusters* à partir des caractéristiques calculées.

Pour décrire théoriquement la méthode, il convient de revenir à la décomposition en transformée d'ondelettes d'une fonction $z \in \mathcal{H}$, où $\mathcal{H}$ est un espace de Hilbert, qui s'écrit de la manière suivante pour tout $j_0 \geqslant 0$:

$$z(t) = \sum_{k=0}^{2^{j_0}-1} a_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t), \tag{4.2}$$

avec $a_{j,k} =< z, \phi_{j_0,k} >_{\mathcal{H}}, \quad d_{j,k} =< z, \psi_{j_0,k} >_{\mathcal{H}}$.

Grâce à la notion d'analyse multirésolution (AMR), le premier terme de l'équation (4.2) représente l'approximation de la fonction z à l'échelle j_0 et le second représente la somme des détails à toutes les échelles $j \geqslant j_0$. En pratique, chaque fonction z est généralement représentée sous la forme d'une fonction discrète définie sur une fine grille de taille N avec $N = 2^J$. Supposons alors que l'échantillon de dimension finie de la fonction z s'écrit $\mathbf{z} = (z(t_i) : i = 0, \ldots, N-1)$ nous pouvons écrire dans l'équation (4.2) la projection de l'approximation sur le sous-espace vectoriel d'approximation V_J (voir l'annexe 6.2) en utilisant la décomposition en $N = 2^J$ points et le niveau d'approximation le plus grossier $j_0 = 0$ de la manière suivante :

$$\tilde{z}_J(t) = a_0\phi_{0,0}(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t), \tag{4.3}$$

avec a_0 et $d_{j,k}$ les coefficients d'ondelettes calculés de la transformation discrète en ondelettes de l'échantillon de dimension finie de la fonction z. Par suite, en notant d_j l'ensemble de tous les coefficients détails de z à l'échelle j ($d_j = (d_{j,0}, \ldots, d_{j,2^j-1})$), une fonction peut être définie pour toute ondelette ψ qui associe à chaque $\mathbf{z}$ le vecteur $(d_0, \ldots, d_{J-1}, a_0)$ de la façon suivante :

$$\begin{aligned} \mathcal{W}_\psi : \mathbb{R}^N &\to \mathbb{R}^N \\ \mathbf{z} &\to (d_0, \ldots, d_{J-1}, a_0), \end{aligned}$$

Étant donné que la décomposition de $\mathbf{z}$ est faite dans une base orthonormée de l'espace d'Hilbert $\mathcal{H}$ (voir l'annexe 6.2), le théorème de Parseval permet d'obtenir la décomposition de $\mathbf{z}$ sous la forme :

$$\|\mathbf{z}\|_2^2 = a_0^2 + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}^2 = a_0^2 + \sum_{j=0}^{J-1}\|d_j\|_2^2, \tag{4.4}$$

Dans le contexte de la méthode de *clustering* proposée par Cugliari (2011), la représentation énergétique de la fonction z est utilisée pour extraire des caractéristiques importantes qui sont ensuite utilisées pour regrouper les courbes de charge journalières similaires en *clusters*. En effet, la manière dont l'énergie est distribuée entre les différentes échelles de variation dans la fonction z donne une indication sur la forme de la fonction, comme l'a souligné l'auteur. Par exemple, si une fonction a une forte concentration d'énergie aux grandes échelles, cela suggère une structure lisse et régulière, tandis qu'une forte concentration d'énergie aux petites échelles indique une structure plus complexe et oscillante.

Dans cette optique, Cugliari (2011) introduit la notion de contribution absolue (AC) et de contribution relative (RC) pour calculer des caractéristiques des courbes de charge journalières. Plus précisément, pour tout $j = 0, \ldots, J-1$, la contribution absolue de l'énergie à l'échelle j correspond à sa part dans l'énergie totale, tandis que la contribution relative est obtenue en divisant la contribution absolue par l'énergie totale. La contribution absolue (AC) et la contribution relative (RC) de l'échelle j sont définies ainsi :

$$\mathrm{cont}_j = \|d_j\|_2^2, \quad \text{et} \quad \mathrm{rel}_j = \frac{\mathrm{cont}_j}{\sum_{j=0}^{J-1}\mathrm{cont}_j}, \tag{4.5}$$

Il convient de noter que dans ces deux représentations, les différences de niveau moyen éventuelles des séries temporelles ne sont pas prises en compte, car le terme d'approximation a_0 n'est pas utilisé dans leur définition. L'auteur dans Cugliari (2011) explique que les courbes de charge journalières appartenant au même *cluster* sont considérées comme stationnaires et par conséquent, le coefficient d'échelle a_0 n'a pas de pouvoir discriminant. Ces représentations sont alors utilisées pour mesurer la différence de la répartition des énergies à travers les échelles.

Jusqu'à présent, chaque courbe de charge journalière ou fonction $\mathbf{z}$ est représentée par J caractéristiques, (cont_j ou rel_j $\forall j = 0, \ldots, J-1$) dépendant du nombre de points d'échantillonnage des données $J = log_2(N)$, où $\log_2$ désigne la fonction logarithme en base deux. Cependant, toutes ces J caractéristiques ne contiennent pas forcément des informations importantes pour le *clustering*. C'est pourquoi l'auteur utilise un algorithme de sélection de caractéristiques, qui permet de choisir les caractéristiques les plus informatives et d'éviter d'intégrer des informations redondantes qui pourraient biaiser les résultats.

L'algorithme de sélection employé est celui présenté par Steinley et al. (2008) qui se compose de deux étapes. La première étape consiste à normaliser les caractéristiques afin de les mettre sur une même échelle tout en préservant leur potentiel de contribution au *clustering*. Cela permet de calculer un indice de *clusterability* pour éliminer les caractéristiques qui n'ont que peu ou pas d'impact sur la structure des *clusters*. Bien que cet indice permette d'évaluer la contribution individuelle de chaque variable ou caractéristique, il ne permet pas de mesurer la contribution collective d'un groupe de variables à la définition des structures de *clusters*. Pour cette raison, la méthode de Steinley et al. (2008) propose une méthode de sélection d'un sous-ensemble de variables basée sur le pourcentage de variation totale expliquée par l'algorithme de *clustering*. Pour plus d'informations sur l'algorithme de sélection de caractéristiques, le lecteur peut se référer à l'article de Steinley et al. (2008).

Après avoir sélectionné les caractéristiques pertinentes, l'algorithme de *clustering k-means* est appliqué avec un nombre de *clusters* k déterminé à l'avance. Pour déterminer la valeur optimale de k, l'approche non-paramétrique de distorsion de saut proposée par Sugar et al. (2003) est utilisée. Cette méthode utilise la distorsion, qui mesure la distance moyenne par dimension entre chaque observation et son centre de *cluster* le plus proche. Elle consiste à tracer la courbe de la distorsion en fonction du nombre de *clusters*, et de détecter le saut le plus important sur cette courbe pour déterminer le nombre optimal k de *clusters*. Cependant, cette méthode ne se contente pas d'examiner la courbe brute de la distorsion. Au lieu de cela, elle utilise une version transformée de cette courbe, appelée courbe de distorsion transformée. Cette courbe est obtenue en élevant la distorsion à une certaine puissance et en la prenant comme une mesure de la distance entre les données et les centres de *clusters*. Cette transformation peut amplifier ces sauts et les rendre plus facilement détectables, ce qui permet de mieux identifier le nombre optimal de *clusters* dans

les données. L'idée est que les sauts dans la courbe de distorsion transformée correspondent à des valeurs de k qui séparent les *clusters* de manière significative, et donc ces valeurs de k peuvent être considérées comme des choix appropriés pour le nombre de *clusters*. En général, nous recherchons le plus grand saut dans la courbe de distorsion transformée, car cela correspond à une séparation optimale des *clusters*.

4.3.5.2 Implémentation

Le processus de prévision de la charge électrique à l'échelle de ménage que nous avons mis en œuvre en utilisant le modèle *KWF* combiné avec la méthode de *clustering* proposée par Cugliari (2011) peut être résumé par les étapes suivantes, chacune étant appliquée à chaque courbe de charge thermosensible dans le jeu de données :

1. **Décomposition en ondelettes** : nous avons décomposé les n segments de courbes de charge journalières $z_1(t), \ldots, z_n(t)$ par la série d'ondelettes tronquée à l'échelle J selon l'équation (4.3).
2. **Calcul des caractéristiques** : pour chaque courbe de charge journalière $z_i(t)$, nous avons calculé la contribution absolue (AC) plutôt que la contribution relative (RC), car des expériences menées dans l'étude de Cugliari (2011) ont montré que l'utilisation de la contribution absolue permettait de détecter plus de variabilité dans les *clusters*. L'auteur de cette étude explique également que la contribution absolue est invariante par décalage vertical et par échelle, tandis que la contribution relative n'est invariante que par décalage vertical.
3. **Sélection des caractéristiques** : nous avons implémenté en langage R et utilisé l'algorithme de Steinley-Brusco (Steinley et al., 2008) pour la sélection des caractéristiques.
4. ***Clustering* des données** : la méthode de Sugar et al. (2003) est utilisée pour déterminer le nombre optimal k de *clusters* parmi les valeurs allant de 1 à 10. Comme nous travaillons à l'échelle des ménages, la méthode de *clustering* doit être appliquée à chaque courbe de charge et le nombre optimal de *clusters* doit être déterminé automatiquement pour chaque ménage, sans intervention manuelle. Par conséquent, l'utilisation d'une méthode graphique pour sélectionner le nombre optimal de *clusters* à l'aide de la méthode de distorsion de saut peut ne pas être appropriée. Cependant, il est possible d'automatiser la méthode de saut de distorsion en utilisant la première différence de la transformée de distorsion. Cette première différence peut être calculée en soustrayant chaque valeur de distorsion transformée de la précédente. Pour calculer le saut à partir de la première différence de la courbe de distorsion transformée, il faut chercher la plus grande valeur positive de la première différence, qui correspond à la plus grande augmentation de la distorsion entre les valeurs de distorsion obtenues pour chaque nombre de *clusters* testé. Cela indique un change-

ment significatif dans la structure de regroupement et peut être interprété comme un saut dans la courbe de distorsion transformée. La valeur correspondant à cette plus grande première différence positive est le nombre optimal de *clusters* selon la méthode de saut. Après avoir déterminé le nombre k de *clusters*, l'algorithme de *clustering k-means* est initialisé de manière aléatoire 30 fois et le meilleur résultat est retenu.

5. **Prévision des courbes de charge journalières** : les *clusters* obtenus à l'issue sont introduits dans le modèle de prévision *KWF* par le vecteur de poids $\tilde{w}_{m,n}$ défini dans la sous-section 3.2.2.

Dans cette étude, seules les courbes de charge des ménages thermosensibles ont été utilisées pour déterminer si l'intégration des *clusters* obtenus par la méthode de *clustering* des courbes de charge journalières proposée par Cugliari (2011) dans le modèle de prévision *KWF* peut améliorer la précision des prévisions de courbes de charge pour les ménages thermosensibles.

4.3.5.3 Résultats

Le tableau 4.6 résume les performances moyennes des deux modèles de prévision *KWF* et *KWF* avec *clustering* à $(J+1)$ appliqués au jeu de données pour les courbes de charge thermosensibles.

Modèle	NMAE	NRMSE	MASE	sMAPE
KWF	0,44	0,76	0,79	54,46
KWF* avec *clustering	0,41	0,69	0,78	52,05
Pourcentage d'amélioration	6,82%	9,21%	1,27%	4,43%

Table 4.6 – La performance moyenne des modèles de prévision *KWF* et *KWF* avec *clustering* à $(J+1)$ selon les quatre erreurs sélectionnées pour les courbes de charge thermosensibles. Les meilleurs résultats sont affichés en orange.

D'après les résultats présentés dans le tableau 4.6, nous pouvons conclure que la combinaison de la méthode de *clustering* et le modèle de prévision *KWF* a permis d'améliorer la moyenne des quatre erreurs sélectionnées. En effet, la moyenne des erreurs NMAE a été améliorée de 6,82%, la moyenne de l'erreur NRMSE a été également améliorée de 9,21%, celle de l'erreur MASE de 1,27% alors que la moyenne de l'erreur sMAPE a été améliorée de 4,43% par rapport au modèle *KWF*.

Les boîtes à moustaches représentant les distributions des erreurs NMAE, NRMSE et MASE sur la figure 4.26 montrent que les médianes des trois erreurs de prévision par le

modèle *KWF* avec *clustering* NMAE, NRMSE et MASE sont inférieures respectivement aux médianes des erreurs de prévision par le modèle *KWF*. La boîte à moustaches des erreurs NRMSE du modèle *KWF* avec *clustering* est plus étalée vers le bas que celle du modèle *KWF* montrant ainsi une amélioration dans la précision de la prévision de certains ménages. Cependant, il est important de noter que cela ne garantit pas que tous les ménages bénéficieront d'une amélioration de la précision avec le *clustering*, car cela dépendra des caractéristiques des données de consommation de chaque ménage.

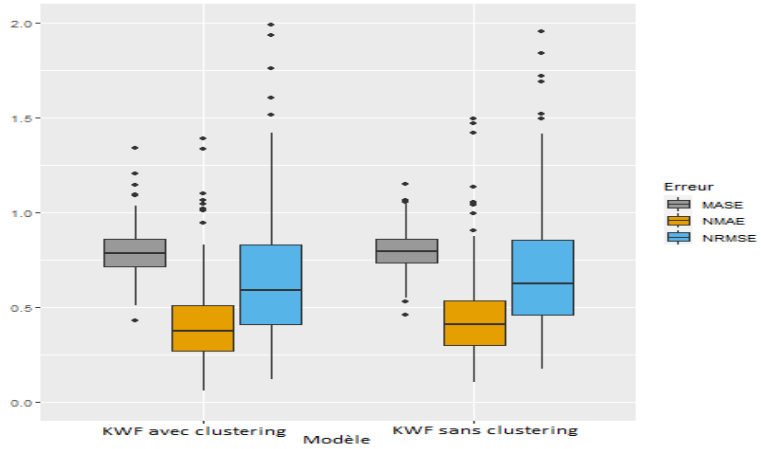
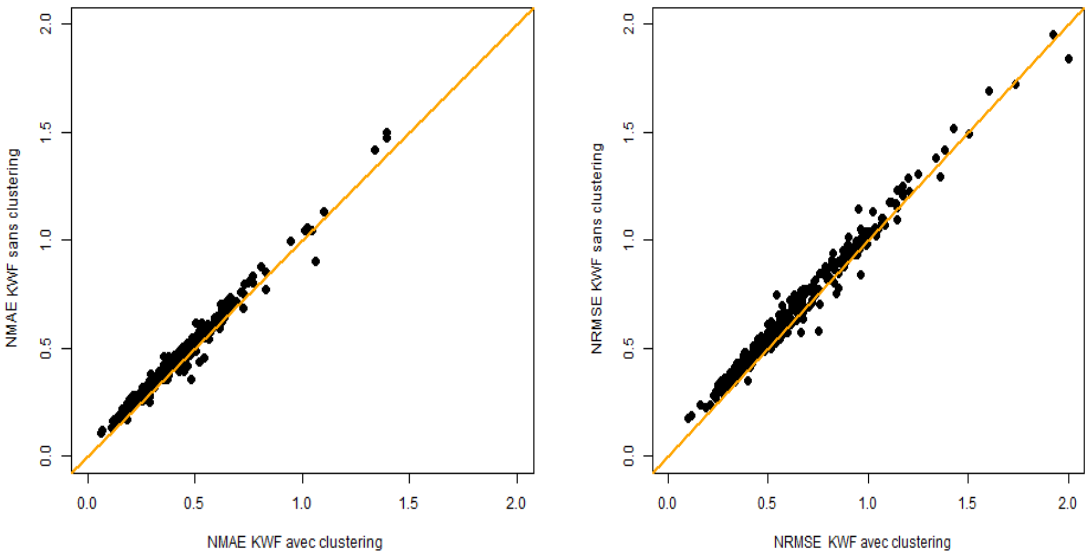


FIGURE 4.26 – Les boîtes à moustaches des erreurs de prévision par le modèle *KWF* et le modèle *KWF* avec *clustering* des ménages thermosensibles.

La comparaison de la performance des deux modèles à l'échelle des ménages individuellement, représentée sur la figure 4.27, montre les nuages de points de l'erreur NMAE et NRMSE du modèle *KWF* en fonction de l'erreur NMAE et NRMSE du modèle *KWF* avec *clustering*. Pour les deux métriques la majorité des points sont au dessus de la première bissectrice indiquant ainsi que le modèle *KWF* avec *clustering* est plus performant globalement pour la prévision de la charge thermosensible dans notre cas. Cela renforce les conclusions faites précédemment. Nous remarquons également que certains points sont en dessous de la première bissectrice. Cette dégradation de la performance du modèle *KWF* avec *clustering* en faveur du modèle *KWF* dans ces cas peut être due à deux raisons, le degré de la thermosensibilité du ménage et sa volatilité. Ce cas sera traité dans la partie 4.3.5.5.



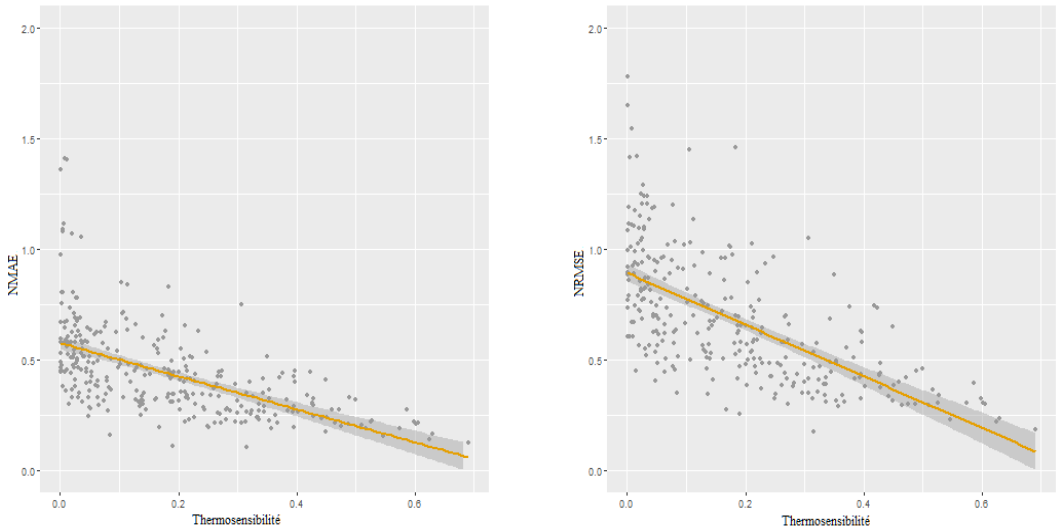
(a) L'erreur NMAE.

(b) L'erreur NRMSE.

FIGURE 4.27 – Nuage de points de l'erreur NMAE et l'erreur NRMSE du modèle *KWF* en fonction de l'erreur NMAE et NRMSE du modèle *KWF* avec *clustering*.

4.3.5.4 Relation entre l'erreur de la prévision et la thermosensibilité

La figure 4.28 montre clairement la relation entre les erreurs de prévision NMAE et NRMSE par le modèle *KWF* avec *clustering* et la thermosensibilité des ménages. Dans les deux cas nous pouvons conclure que les ménages thermosensibles sont plus faciles à prédire par le modèle *KWF* et que la précision de leur prévision est plus satisfaisante. De même deux ménages de même degré de thermosensibilité peuvent avoir des erreurs différentes en raison de la différence de leur mode de vie, la régularité de leur demande, ...

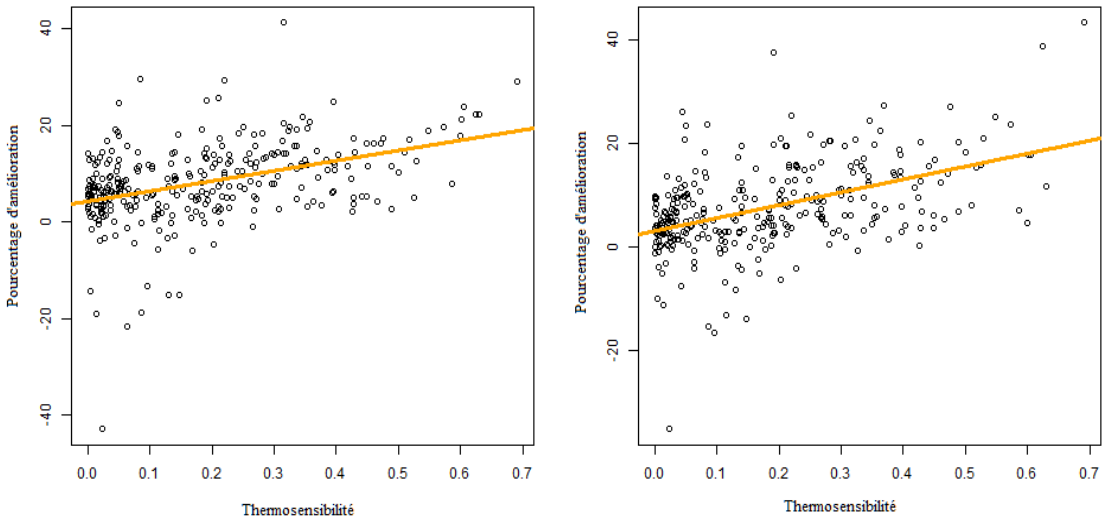


(a) L'erreur NMAE. (b) L'erreur NRMSE.

FIGURE 4.28 – Nuage de points de l'erreur NMAE et l'erreur NRMSE du modèle *KWF* avec *clustering* en fonction de la thermosensibilité du ménage.

4.3.5.5 Relation entre le pourcentage d'amélioration du modèle et la thermosensibilité

Dans cette section, nous allons étudier la relation entre le pourcentage d'amélioration de la précision de la prévision par le modèle *KWF* suite à l'intégration de l'étape de *clustering* et la thermosensibilité du ménage. Ceci permettra d'évaluer la contribution ou non de l'intégration des *clusters* dans la prévision de la part thermosensible de la charge électrique des ménages. La figure 4.29 montre la relation entre le pourcentage d'amélioration de la précision de la prévision mesurée par la métrique NMAE (voir la figure 4.29a) et l'erreur NRMSE (voir la figure 4.29b) de la prévision par le modèle *KWF* avec *clustering* par rapport au modèle *KWF* en fonction de la thermosensibilité des ménages. La figure permet de conclure que le pourcentage d'amélioration de la précision de la prévision mesurée par les deux métriques augmente avec l'augmentation de la thermosensiblité du ménage. Il varie entre 0% et 40%. Deux ménages ayant le même degré de thermosensibilité peuvent avoir des pourcentages d'amélioration différents. Cela revient comme nous avons expliqué précédemment à la différence de leur mode de vie et de la manière d'utilisation de leurs appareils électriques y compris le chauffage. Certains ménages montrent un pourcentage d'amélioration négatif indiquant alors une dégradation de la précision de la prévision par rapport au modèle *KWF*. Ces ménages dans la majorité des cas ont un petit degré de thermosensibilité situé entre $0,0$ et $0,2$.



(a) L'erreur NMAE. (b) L'erreur NRMSE.

FIGURE 4.29 – Nuage de points du pourcentage de l'amélioration de l'erreur NMAE et l'erreur NRMSE du modèle *KWF* avec *clustering* par rapport au modèle *KWF* en fonction de la thermosensibilité du ménage.

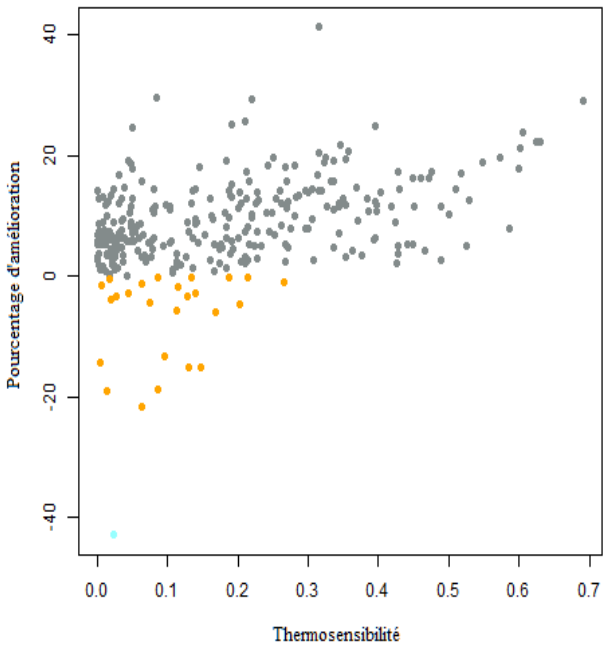


FIGURE 4.30 – Nuage de points du pourcentage de l'amélioration de l'erreur NMAE du modèle *KWF* avec *clustering* par rapport au modèle *KWF* en fonction de la thermosensibilité du ménage. Les points en orange représentent les ménages avec un pourcentage d'amélioration négative.

Prenons l'exemple du ménage M_5 représenté sur la figure 4.31 en bleu turquoise ayant un pourcentage de dégradation de la prévision par le modèle *KWF* avec *clustering* élevé par rapport à la prévision par le modèle *KWF* de base et un degré de thermosensibilité égal à 0,02. La méthode de *clustering* des courbes de charge journalières de ce ménage sera analysée plus en détail dans la suite de cette section dans l'objectif de donner une explication de cette dégradation de la performance dans ce cas. La figure 4.31 montre les *clusters* obtenus par la méthode de *clustering* et intégrés dans la prévision de la courbe de charge du ménage M_5. Étant donné que son degré de thermosensibilité est faible, les *clusters* de ses courbes de charge journalières ne reflètent pas des regroupements en fonction de la température. Nous avons alors essayé d'expliquer le résultat du *clustering* à travers les jours de la semaine et l'information du calendrier.

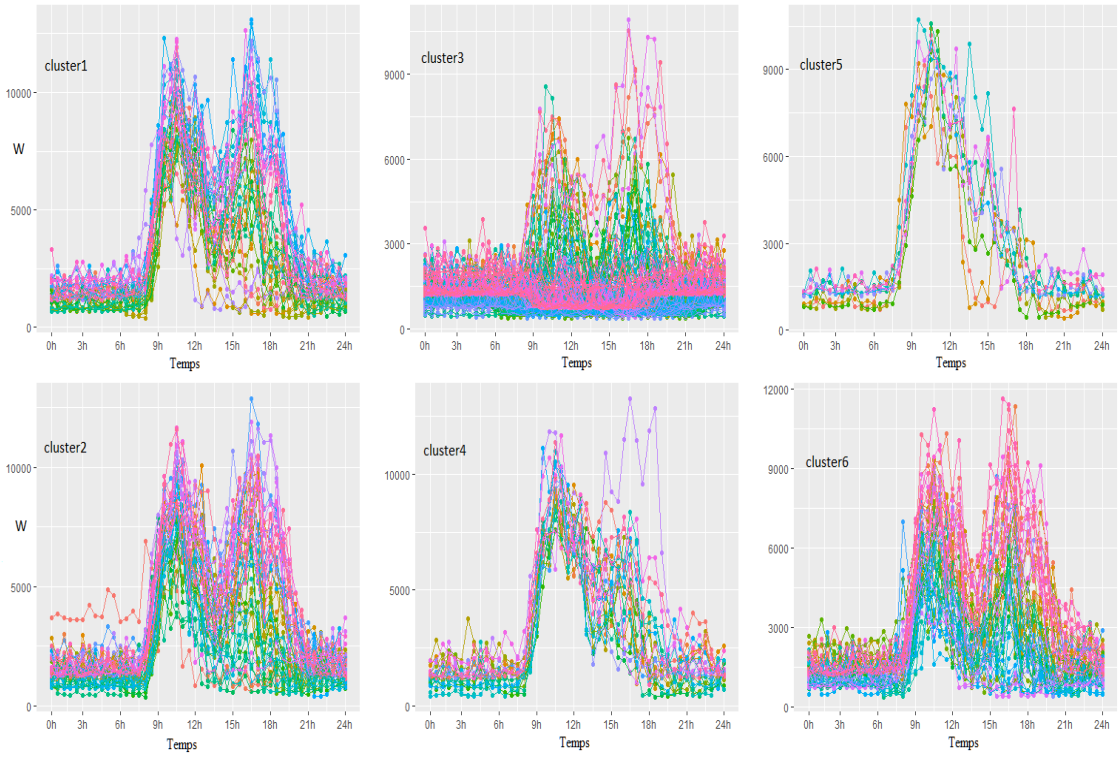


FIGURE 4.31 – Les *clusters* des courbes de charge journalières obtenus par la méthode de *clustering* et intégrés dans la méthode de prévision *KWF* du ménage M_5.

Le tableau 4.7 montre les différents *clusters* obtenus par le *clustering* des courbes de charge journalières du ménage M_5 ainsi que les jours du calendrier attribués à chacun. Les trois *clusters* 1, 2 et 6 regroupent les jours de la semaine lundi, mardi, jeudi et vendredi ayant des courbes de charge similaires. Ces courbes de charge se caractérisent par la présence de deux pointes de consommation : la première entre 9h et 12h et la seconde entre 15h et 18h. Ces dernières se distinguent par leurs formes et leurs amplitudes. Le *cluster* 3 contient globalement les courbes de charge des jours de week-end ainsi que les jours de vacances de la Toussaint et de Noël. Le *cluster* 4 regroupe des jours de mercredis et

de vendredis similaires. Le *cluster* 5 contient uniquement des jours de mercredis atypiques ayant une forme différente des mercredis présents dans le *cluster* 4.

cluster	**1**	**2**	**3**	**4**	**5**	**6**
Type du jour dans le *cluster*	lundi, mardi, jeudi et vendredi	lundi, mardi, jeudi et vendredi	samedi, dimanche, jours fériés et les deux semaines de vacances de la Toussaint et de Noël	mercredi et vendredi	mercredi atypique	lundi, mardi, jeudi et vendredi

TABLE 4.7 – L'interprétation des résultats de *clustering* des courbes de charge journalières du ménage M_5 à travers les informations sur le calendrier.

La figure 4.32 présente les groupes déterministes intégrés dans le modèle *KWF* de base. Une comparaison entre les figures 4.31 et 4.32 montre que les groupes déterministes des jours de week-ends et des jours fériés séparés sont plus homogènes que le *cluster* 3 proposé par la méthode de *clustering* et par conséquent, nous attendons à ce que la prévision de ces jours soit plus précise par le modèle *KWF* de base que par le modèle *KWF* avec *clustering*.

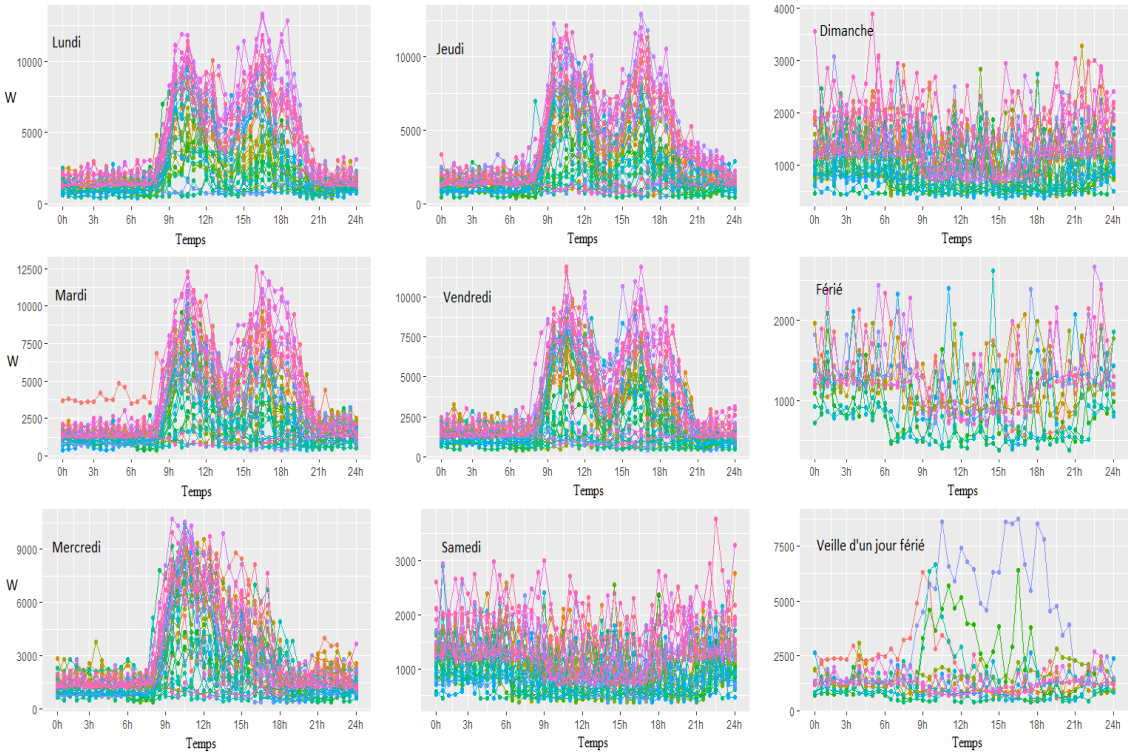


FIGURE 4.32 – Les groupes déterministes des courbes de charge journalières du ménage M_5 utilisés dans la méthode de prévision *KWF*.

Les résultats représentés sur la figure 4.33 montrent une dégradation globale de la

précision de la prévision du modèle *KWF* avec *clustering* mesurée par l'erreur NMAE et une dégradation plus significative conforme à nos attentes notamment pour les jours de week-ends et les jours fériés. Pour les lundis les boîtes à moustaches montrent que la distribution de l'erreur de prévision NMAE par le modèle *KWF* de base est plus dissymétrique que la distribution de l'erreur de prévision NMAE par le modèle *KWF* avec *clustering* et l'écart interquartile est plus étalé indiquant ainsi une distribution plus hétérogène.

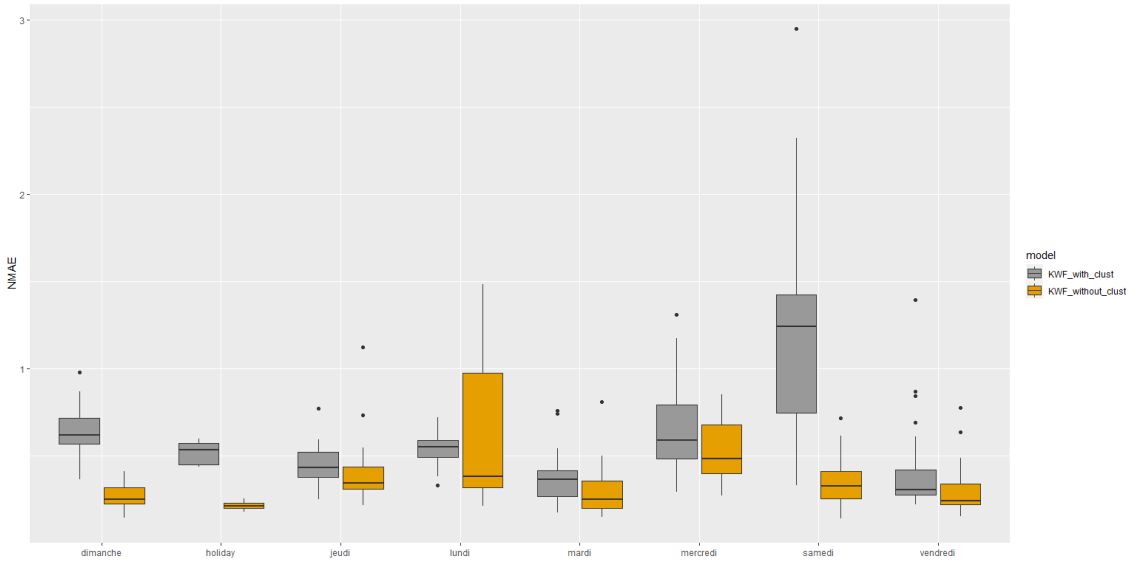


FIGURE 4.33 – Les boîtes à moustaches montrant la distribution de l'erreur NMAE de prévision du ménage M_5 par le modèle *KWF* sans *clustering* (avec des groupes déterministes) et le modèle *KWF* avec *clustering* en fonction des jours de la semaine et des jours fériés.

4.3.5.6 Exemples de prévision par le modèle *KWF* avec *clustering*

La figure 4.34 et la figure 4.35 montrent deux exemples de prévision des courbes de charge de deux ménages M_6 et M_7 par les deux modèles de prévision *KWF* et *KWF* avec *clustering* à $(J+1)$. La figure 4.34 montre que le modèle de prévision *KWF* avec *clustering* est plus apte à détecter les pics et les creux de consommation que le modèle *KWF*. La figure 4.35 montre que la performance du modèle *KWF* avec *clustering* dépasse celle du modèle *KWF* dans la détection des périodes de vacances. La courbe en vert représenteant la courbe prévisionnelle par le modèle *KWF* avec *clustering* s'adapte au bout d'un seul jour au niveau de la consommation réelle (représenté par la courbe en bleu canard) ce qui n'est pas le cas pour la courbe en violet qui représente la courbe prévisionnelle par le modèle *KWF*.

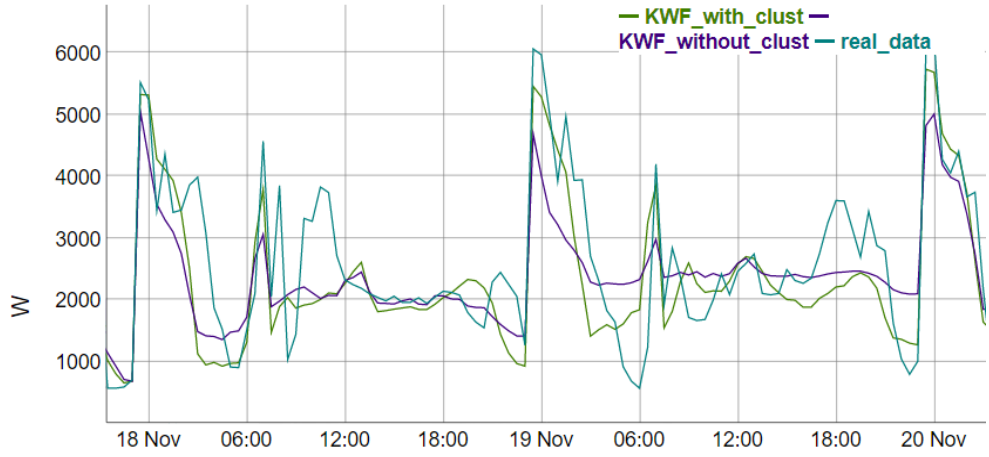


FIGURE 4.34 – Exemple de prévision à $(J+1)$ d'une courbe de charge d'un ménage M_6 par le modèle *KWF* (courbe en violet) et par le modèle *KWF* avec *clustering* (courbe en vert). La courbe en bleu canard représente la consommation réelle.

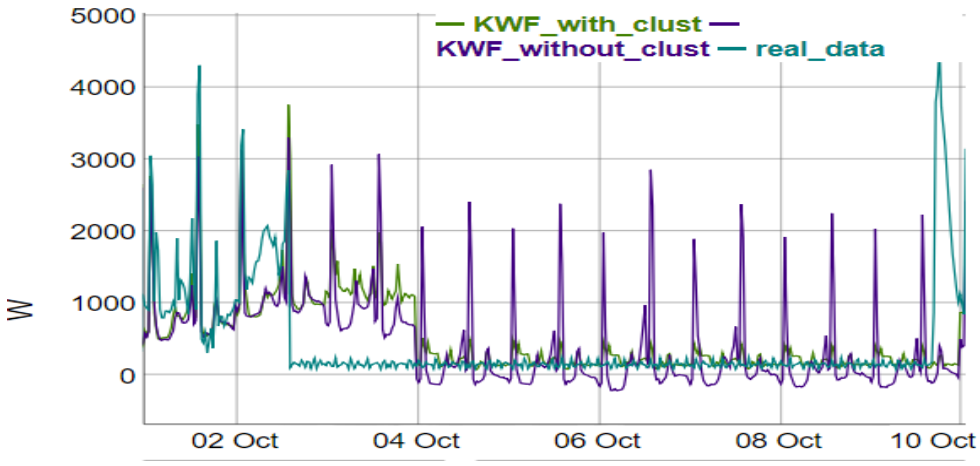


FIGURE 4.35 – Exemple de prévision à $(J+1)$ d'une période de vacances d'un ménage M_7 par le modèle *KWF* (courbe en violet) et par le modèle *KWF* avec *clustering* (courbe en vert). La courbe en bleu canard représente la consommation réelle.

4.3.6 Prévision de la charge électrique agrégée

Dans l'objectif de vérifier l'impact de l'intégration des *clusters* dans le modèle *KWF* sur la prévision de la charge électrique thermosensible, nous avons testé les deux modèles *KWF* et *KWF* avec *clustering* pour la prévision de la courbe de charge agrégée de tous les ménages thermosensibles dans le jeu de données (voir la figure 4.36). En effet, cette agrégation nous permet de juger la pertinence de cette approche dans la prévision de la charge électrique thermosensible à une échelle plus large que celle d'un seul ménage.

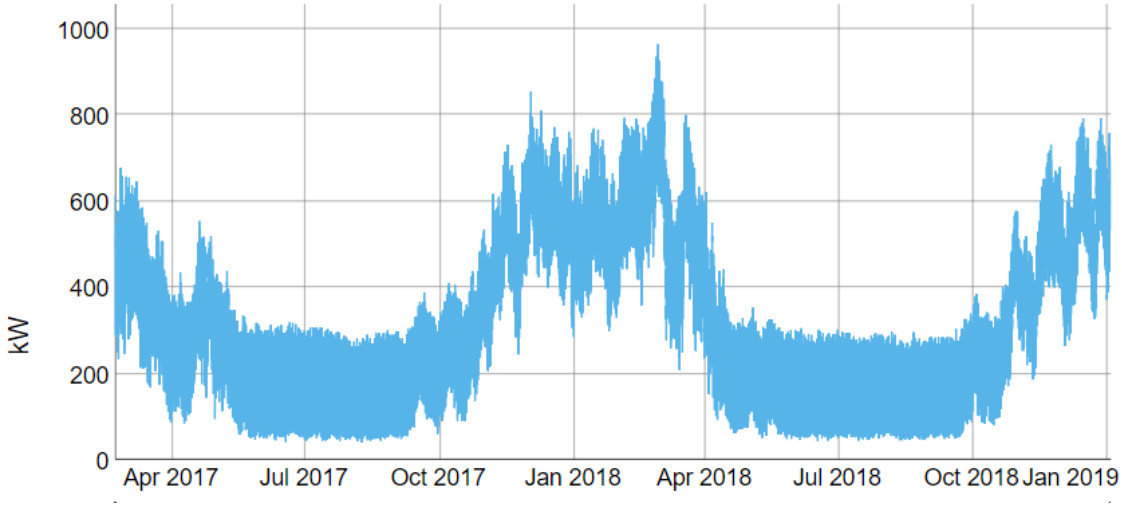


FIGURE 4.36 – La courbe de charge agrégée des courbes de charge thermosensibles.

4.3.6.1 Les résultats

Le tableau 4.8 montre les résultats obtenus par les deux modèles. Nous constatons d'après ces résultats que la prévision par le modèle *KWF* avec *clustering* est plus précise que la prévision par le modèle *KWF*. De plus, le pourcentage d'amélioration de deux erreurs NMAE et sMAPE de la prévision de la charge électrique agrégée (10% pour l'erreur NMAE et 22,69% pour l'erreur sMAPE) est supérieure à celui de la moyenne des erreurs de la prévision à l'échelle des ménages (6,82% pour l'erreur NMAE et 4,43% pour l'erreur sMAPE, voir la table 4.6). Cela indique clairement que l'intégration des *clusters* issus de la méthode de *clustering* proposée dans CUGLIARI (2011) dans le modèle *KWF* améliore la prévision de la charge électrique thermosensible à l'échelle des ménages ainsi qu'à l'échelle de l'agrégation de plusieurs ménages.

Modèle	NMAE	NRMSE	sMAPE
KWF	0,10	0,14	13,84
KWF* avec *clustering	0,09	0,13	10,70
Pourcentage d'amélioration	10%	7,14%	22,69%

TABLE 4.8 – La performance moyenne des modèles de prévision *KWF* et *KWF* avec *clustering* à $(J+1)$ pour la courbe de charge agrégée selon les trois métriques sélectionnées. Les meilleurs résultats sont affichés en orange.

La figure 4.8 montre les courbes de charge prévisionnelles par le modèle *KWF* et par le modèle *KWF* avec *clustering* de la charge électrique agrégée des ménages thermosensibles pour la période entre le 14 et le 16 décembre 2018.

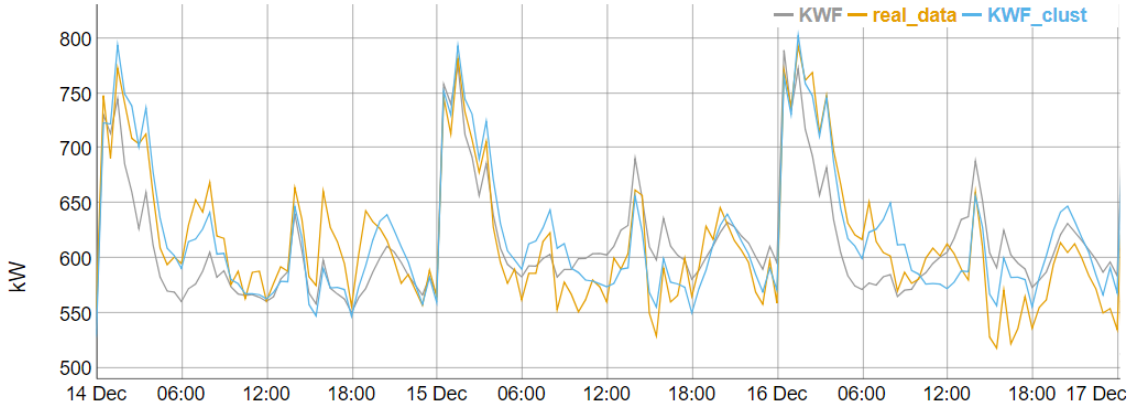


FIGURE 4.37 – Exemple de prévision à $(J+1)$ de la courbe de charge agrégée des ménages thermosensibles par le modèle *KWF* (courbe en gris) et par le modèle *KWF* avec *clustering* (courbe en bleu ciel). La courbe en orange représente la charge électrique agrégée réelle en kW.

4.3.7 Conclusion

D'après les résultats représentés dans la section précédente, nous pouvons conclure que l'intégration des *clusters* obtenus par la méthode de *clustering* proposée dans (CUGLIARI, 2011) dans le modèle *KWF* permet d'améliorer globalement la précision de la prévision des courbes de charge thermosensibles à l'échelle des ménages. Le pourcentage d'amélioration est très disparate, il dépend du degré de thermosensibilité de la charge électrique du ménage. En effet, l'intégration de ces *clusters* pour la prévision de deux courbes de charge de deux ménages ayant des degrés de thermosensibilité similaires peuvent donner des résultats très différents en raison de la différence du mode de vie et des habitudes de consommation électrique de chacun. Un effet négatif de l'intégration des *clusters* est observé chez certains ménages ayant un faible degré de thermosensibilité. L'impact de l'intégration des *clusters* pour la prévision de la charge agrégée des courbes de charge des ménages thermosensibles a été positif.

À ce stade, nous pouvons nous demander si l'intégration de la température comme données brutes dans un modèle de prévision de la charge électrique est plus efficace que l'utilisation des *clusters* qui intègrent l'effet de la variation de la température. Pour cette raison, et dans l'objectif de tester d'autres modèles de la littérature qui se sont révélés performants pour la prévision de la charge électrique à plusieurs échelles, nous présentons dans les sections suivantes une étude comparative entre le modèle *KWF* et trois autres modèles de prévision. Ces modèles ont des caractéristiques intéressantes qui les rendent adaptés à la prévision de la charge électrique à l'échelle des ménages.

4.4 Prévision par les modèles *GAM* et *MARS*

Le modèle *KWF* utilisé précédemment a permis d'effectuer des prévisions de la charge électrique à l'échelle des ménages en se basant sur l'historique de la consommation d'électricité et des groupes de calendrier déterministes ou des groupes obtenus par *clustering* qui intègrent l'impact de la variation de la température sur la charge électrique. Ce modèle se révèle plus performant en termes de précision qu'un modèle heuristique de forêt aléatoire, un modèle déterministe appelé climatologique et un modèle de persistance saisonnier. Toutefois, il serait intéressant de tester des modèles de régression non paramétrique multivariée comme le modèle *GAM* ou le modèle *MARS* qui permettent de modéliser des relations plus explicites entre la charge électrique, les variables de calendrier et les données de la température extérieure.

4.4.1 Prévision par le modèle *GAM*

Le modèle *GAM* a déjà été appliqué à la prévision de la charge électrique à court terme et à différents niveaux : national (Fan et al., 2011) et local (Gerossier et al., 2017). Il s'est alors révélé performant. En effet, le modèle *GAM* est une technique puissante et simple pour les trois raisons suivantes :

1. rapide et facile à interpréter puisqu'il s'agit d'un modèle additif, et par conséquent, l'interprétation de l'impact marginal d'une seule variable ne dépend pas des valeurs des autres variables du modèle. Par suite, une simple lecture de la sortie du modèle permet de comprendre la contribution de chaque variables à la prévision.
2. flexible et permet de découvrir des relations complexes dans les données puisque les relations entre les variables indépendantes et la variable dépendante ne sont pas supposées être linéaires.
3. permet d'avoir une équation explicite de la relation entre la variable dépendante et les variables indépendantes comme la température. Ce critère est assez important pour le fournisseur puisqu'il permet de gagner en termes d'interprétabilité.

En résumé, le modèle *GAM* établit un équilibre pertinent entre les modèles paramétriques qui sont faciles à interpréter et les algorithmes d'apprentissage extrêmement flexibles mais qui manquent d'interprétabilité comme les réseaux de neurones. Dans l'objectif de chercher à améliorer la précision de la prévision de la charge électrique des ménages dans l'échantillon test et d'étudier plus précisément l'effet de l'intégration de la température comme variable brute dans le modèle de prévision, nous avons proposé et testé un modèle *GAM* pour la prévision à court terme des données de courbes de charge des ménages. Il convient de souligner que de nombreux modèles *GAM* ont été proposés et évalués afin de déterminer les caractéristiques les plus appropriées pour un modèle destiné à la prévision de ce type

de données, ainsi que les variables indépendantes les plus pertinentes. Ces modèles sont inspirés des modèles proposés dans (Pierrot et al., 2011) pour la prévision de la charge totale française, c'est-à-dire au niveau national. Seul le modèle le plus performant est présenté dans ce manuscrit. Par rapport à l'intégration de la température dans le modèle de prévision de la charge thermosensible, nous avons testé trois modèles intégrant chacun la température sous une forme différente. Le modèle le plus performant est celui qui a été sélectionné pour la prévision de la charge thermosensible. Les fondements théoriques des modèles *GAM* sont présentés dans la partie 3.2.1.2.

4.4.1.1 Approche de prévision

Les prévisions sont fournies pour 24 heures sous forme de prévision au pas demi-horaire. Dans notre approche, nous supposons que les données réelles de consommation d'électricité sont disponibles la veille du jour à prédire et avant la prévision du jour $(J+1)$ (voir la figure 4.38). L'ensemble du jeu de données est divisé en une période d'entraînement (70% des données), qui correspond à la période entre le 01-01-2017 et le 30-06-2018, suivie d'une période de test (30% des données), qui correspond à la période entre le 01-07-2018 et le 01-01-2019 (voir la figure 4.39). Le modèle est alors entraîné sur les données d'entraînement. Puis il est stocké afin d'être utilisé pour générer les prévisions. La technique d'entraînement récursif, qui consiste à ré-entraîner le modèle d'une façon récursive avec des nouvelles données, n'est pas utilisée dans le cas présent. Cette technique est utile pour améliorer la qualité de la prévision en mettant à jour les paramètres du modèle d'une façon régulière. Cela est intéressant pour tenir compte des changements brusques qui peuvent avoir lieu suite à une modification du mode de vie des occupants du ménage (comme l'augmentation du nombre d'occupants, le changement du mode de travail, ...). En revanche, ce gain probable dans la précision va à l'encontre de la rapidité de l'approche, c'est la raison pour laquelle nous n'avons pas appliqué cette technique.

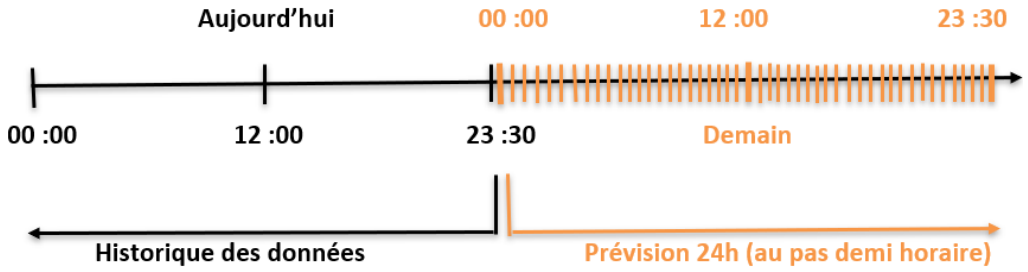


FIGURE 4.38 – Prévision la veille pour le lendemain.

FIGURE 4.39 – Division du jeu de données en ensemble d'entraînement et ensemble de test.

4.4.1.2 Variables d'entrée et ajustement du modèle

Dans la littérature, les variables explicatives généralement utilisées pour la prévision de la charge électrique des ménages sont classées en trois catégories : les variables de l'historique de la charge électrique, les variables de calendrier ainsi que les variables météorologiques (LUSIS et al., 2017). En effet, cette charge dépend aussi de plusieurs autres variables comme le nombre et le mode de vie des occupants, leurs appareils électriques, la classe énergétique du logement, ... (GAJOWNICZEK et al., 2017). Par contre, en raison de la difficulté à recueillir ce type de données, elles sont rarement utilisées dans les études sur la prévision de la charge électrique des ménages.

Les variables de l'historique de la charge électrique : en raison de sa périodicité, la courbe de charge est fortement corrélée à son historique, c'est-à-dire la consommation de l'électricité au temps t est fortement liée à la consommation au temps $(t-\gamma)$, avec $\gamma > 0$, donc passée. Par conséquent, l'utilisation des valeurs décalées d'une période γ dans le passé de la charge électrique comme variable d'entrée dans le modèle de prévision de cette dernière est une pratique connue dans la littérature (Franklin L QUILUMBA et al., 2014 ; LUSIS et al., 2017 ; GEROSSIER et al., 2017). La détermination de la période de décalage γ se fait généralement par le calcul de la fonction d'autocorrélation partielle de la courbe de charge d'un ménage représentée par la série temporelle $(y_t)_{t\geqslant 0}$ [8]. Cette fonction permet de mesurer l'autocorrélation d'une série temporelle pour un décalage donné indépendamment des autocorrélations pour les décalages inférieurs. En revanche, le calcul de l'autocorrélation pour chaque ménage séparément n'est pas vraiment une pratique recommandée dans notre cas puisqu'elle vient à l'encontre de l'adaptabilité du modèle à un grand nombre de ménages. Les variables représentant l'historique de la charge électrique que nous avons intégrées dans notre modèle sont alors :

1. la valeur de la charge électrique de la même demi-heure du jour précédent y_{t-48}, c'est-à-dire les valeurs de la courbe de charge décalées d'un jour. Cette variable est la plus importante parmi toutes les variables explicatives sélectionnées en raison de la forte saisonnalité journalière des données de consommation d'électricité (Franklin

8. Une courbe de charge peut être représentée par une série temporelle.

L Quilumba et al., 2014 ; Lusis et al., 2017 ; Gerossier et al., 2017). L'autocorrélation avec les valeurs décalées de la charge électrique d'une période $\gamma < 48$ peut être plus forte. Cependant, pour une prévision de la charge électrique journalière, ces données ne sont pas toutes disponibles au moment de la prévision.

2. la médiane de la même demi-heure observée les sept jours précédents que nous notons $\text{Me}_t = \text{médiane}(y_{t-48}, \ldots, y_{t-336})$. Cette variable permet d'ajuster la prévision dans le cas où le jour précédent représente une consommation irrégulière et atypique puisqu'elle intègre une information sur la consommation d'électricité lissée sur toute la semaine (voir (Gerossier, 2019)).

Les variables du calendrier : l'effet de calendrier est généralement intégré dans les modèles de prévision de la charge électrique pour permettre à ces derniers de prendre en compte les variations de la consommation d'électricité en fonction des jours de la semaine, des heures de la journée et des jours spéciaux comme les jours fériés. Les variables que nous avons utilisées pour intégrer l'effet de calendrier dans notre modèle sont alors :

1. le type du jour : les jours de la semaine {1 pour dimanche ,..., 7 pour samedi} et les jours fériés de la région hors les samedis et les dimanches {8},
2. l'heure de la journée au pas demi-horaire $\{1, \ldots, 48\}$.

Comme dans le cas du modèle *KWF*, deux modèles *GAM* ont été conçus l'un pour la prévision de la charge électrique non-thermosensible et l'autre pour la prévision de la charge thermosensible. Le modèle de prévision de la charge non-thermosensible prends comme variables d'entrée les quatre variables citées ci-dessus. La saisonnalité intrajournalière est modélisée en ajustant un modèle par demi heure. En effet, nous divisons la courbe de charge du ménage en 48 séries chronologiques pour chaque demi heure du jour. Notre modèle est alors formé de 48 modèles qui donnent chacun une prévision pour une demi heure précise du jour. La saisonnalité hebdomadaire et l'effet du calendrier sont modélisés à l'aide d'une variable catégorielle selon le type du jour. L'équation du modèle est alors donnée par :

$$\hat{y}_t^{dh} = m_1^{dh}(y_{t-48})\mathbb{1}_{\text{Type du jour}} + m_2^{dh}(\text{Me}_t) \quad \forall dh \in \{1, \ldots, 48\}. \tag{4.6}$$

Les deux fonctions de splines de lissage m_1^{dh} et m_2^{dh} sont ajustées indépendamment aux données d'entraînement de chaque ménage pour chaque demi heure dh. Nous avons fait le choix d'utiliser les bases de *splines* de régression à plaques minces (*thin plate regression spline bases*) puisqu'elles permettent la détermination automatique du nombre de nœuds et de leurs emplacements. Les dimensions des bases ont été fixées aux degrés de liberté maximaux autorisés pour chaque terme du modèle. Le nombre de degrés de liberté effectifs du modèle a été estimé à partir des données par la méthode de *GCV* (*Generalized Cross Validation*). L'estimation de chaque modèle propre à chaque demi heure de la journée a été faite séparément sur un cœur de calcul afin d'accélérer l'étape d'entraînement du modèle

(calcul parallèle multi-coeurs [9]). Ensuite, une fois que le modèle est entraîné, celui-ci est sauvegardé et utilisé pour générer les prévisions. Les `packages mgcv` [10] et `doParallel` [11] du logiciel R ont été utilisés pour mettre en œuvre ce modèle *GAM*.

Pour la prévision de la charge thermosensible, une variable liée à la température extérieure doit être ajoutée à l'ensemble des variables d'entrée du modèle de prévision de la charge non-thermosensible. En effet, **les variables météorologiques** utilisées dans la littérature pour la prévision de la charge électrique sont généralement la température extérieure, la vitesse du vent, l'humidité, la nébulosité. En revanche, la température extérieure est la variable qui a le plus d'influence sur la charge électrique et qui est la plus utilisée dans la littérature (Özkizilkaya, 2014). Elle est généralement intégrée sous plusieurs formes dans les modèles de prévision de la charge électrique (comme les températures décalées, les températures lissées, les degrés jours de chauffage, les degrés jours de climatisation, ...). Dans l'objectif d'étudier l'impact de l'intégration de la température comme variable d'entrée dans le modèle de prévision de la charge électrique thermosensible des ménages, nous avons testé les quatre modèles représentés ci-dessous sur le jeu de données des courbes de charge thermosensibles. Soit $(T_t)_t$ la série temporelle représentant les données de la température extérieure. Les quatre modèles testés sont alors :

1. le modèle *GAM* défini dans l'équation (4.6) qui n'intègre pas la température extérieure comme variable d'entrée.
2. le modèle *GAM* défini dans l'équation (4.6) avec la température décalée d'un jour [12]. L'équation du modèle est alors donnée par :

$$\hat{y}_t^{dh} = m_1^{dh}(y_{t-48})\mathbb{1}_{\text{Type du jour}} + m_2^{dh}(\text{Me}_t) + m_3^{dh}(T_{t-48}) \quad \forall dh \in \{1, \ldots, 48\}. \quad (4.7)$$

3. le modèle *GAM* défini dans l'équation 4.6 avec la température extérieure prédite au temps t ($\hat{T}_t$). En effet, puisque nous n'effectuons pas des prévisions en temps réel à ce stade, ces valeurs correspondent aux données de test des températures réelles. En revanche, nous supposons que les prévisions des températures du lendemain sont toujours disponibles au moment de la prévision de la charge électrique. Par conséquent, ces prévisions peuvent être intégrées sans problème pour la prévision de la charge du lendemain. L'équation du modèle est alors donnée par :

$$\hat{y}_t^{dh} = m_1^{dh}(y_{t-48})\mathbb{1}_{\text{Type du jour}} + m_2^{dh}(\text{Me}_t) + m_3^{dh}(\hat{T}_t) \quad \forall dh \in \{1, \ldots, 48\}. \quad (4.8)$$

4. le modèle *GAM* défini dans (4.6) avec la prévision de la température extérieure lissée

9. Il consiste à diviser la tâche de calcul en un ensemble de sous-tâches indépendantes à exécuter simultanément. Le calcul est ainsi réparti sur plusieurs cœurs de calcul.
10. https ://cran.r-project.org/web/packages/mgcv/index.html.
11. https ://cran.r-project.org/web/packages/doParallel/index.html.
12. C'est la valeur de la température extérieure de la même demi-heure du jour précédent T_{t-48}.

$(\hat{\mathbf{T}}_t)$ [13] donnée par la formule suivante :

$$\hat{\mathbf{T}}_t = \alpha \hat{T}_t + (1-\alpha)\hat{\mathbf{T}}_{t-1} \quad \alpha \in [0,1]. \tag{4.9}$$

Le paramètre de lissage α est estimé pour chaque ménage séparément de manière à maximiser la corrélation entre la série temporelle représentant la courbe de charge du ménage et la température $\hat{\mathbf{T}}_t$ (Gerossier, 2019). L'équation de ce modèle est alors donnée par :

$$\hat{y}_t^{dh} = m_1^{dh}(y_{t-48})\mathbb{1}_{\text{Type du jour}} + m_2^{dh}(\text{Me}_t) + m_3^{dh}(\hat{\mathbf{T}}_t) \quad \forall dh \in \{1,\ldots,48\}. \tag{4.10}$$

Ces quatre modèles ont été testé et évalué sur l'ensemble du jeu de données des ménages thermosensibles. Les résultats obtenus sont représentés par la figure 4.40. Ces résultats montrent les distributions des erreurs de prévision NMAE, NMRSE et MASE par les quatre modèles. Nous remarquons que le modèle *GAM* avec la température décalée d'un jour est globalement le plus performant par rapport aux distributions des trois erreurs et que l'intégration de la température sous n'importe quelle forme a amélioré globalement la précision du modèle de prévision par rapport au modèle qui n'intègre pas la température. Le modèle *GAM* retenu pour la prévision de la charge thermosensible est alors le modèle intégrant la température décalée d'un jour et ayant l'équation suivante :

$$\hat{y}_t^{dh} = m_1^{dh}(y_{t-48})\mathbb{1}_{\text{Type du jour}} + m_2^{dh}(\text{Me}_t) + m_3^{dh}(T_{t-48}) \quad \forall dh \in \{1,\ldots,48\}. \tag{4.11}$$

À ce stade, il convient de se demander si la division du modèle de prévision de la courbe de charge en 48 sous-modèles pour prendre en compte la saisonnalité journalière peut affecter la prise en compte des corrélations entre les différentes heures de la journée, ce qui pourrait potentiellement réduire la précision du modèle. Il est possible qu'un modèle conçu pour la prévision de la charge sur une journée entière soit plus approprié qu'un modèle pour chaque pas demi-horaire. Dans ce cas, comment peut-on détecter automatiquement ces corrélations et intégrer des interactions entre les variables d'entrée ? Dans l'optique d'explorer ces nouvelles pistes, nous avons mis en œuvre un modèle *MARS* et comparé ses performances avec celles des modèles *GAM* et *KWF*.

13. Les températures lissées ont été définies par Dordonnat et al. (2008) et intégrées dans un modèle de prévision de la charge électrique française afin de tenir en compte de l'effet de l'inertie thermique des bâtiments.

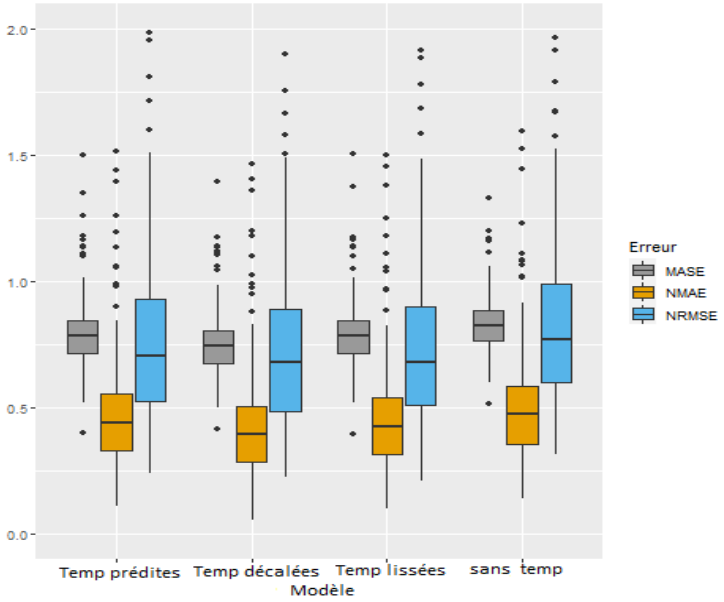


FIGURE 4.40 – Les boîtes à moustaches des erreurs de prévision à $(J+1)$ par les quatre modèles *GAM* : le modèle avec les prévisions de la température, le modèle avec la température décalée, le modèle avec les prévisions lissées de la température et le modèle sans température respectivement de gauche à droite.

4.4.2 Prévision par le modèle *MARS*

Le modèle *MARS* est un modèle de régression non paramétrique comme le modèle *GAM* qui permet de modéliser des relations linéaires et non linéaires entre les variables explicatives. En revanche, et contrairement au modèle *GAM* il introduit automatiquement des termes d'interaction entre ces dernières. Cette différence significative permet au modèle *MARS* de se démarquer du modèle *GAM*. Le modèle *MARS* permet également le recours à une procédure automatisée dans la sélection des variables explicatives les plus pertinentes. Des informations supplémentaires sur les bases théoriques de ce modèle sont présentées dans la partie 3.2.1.3.

Ce modèle est moins connu que le modèle *GAM* pour la prévision de la charge électrique (Al-Musaylh et al., 2018 ; Sigauke et al., 2010). Par contre, et compte tenu de ses avantages nous l'avons mis en œuvre pour la prévision de la charge des ménages.

L'approche de prévision adoptée pour ce modèle est la même que celle du modèle *GAM* (voir la partie 4.4.1.1). Par contre, notre modèle *MARS* prend en entrée les variables explicatives suivantes :

1. la valeur de la charge électrique de la même demi-heure du jour précédent y_{t-48} ;
2. la valeur de la charge électrique de la même demi-heure du même jour de la semaine précédente y_{t-336} ;

3. la médiane de la même demi-heure des sept jours précédents Me_t.
4. la valeur de la température extérieure de la même demi-heure du jour précédent T_{t-48} ;
5. le type du jour (jour de la semaine ou jour férié) ;
6. l'heure de la journée au pas demi-horaire, cette variable prend ses valeurs dans l'ensemble de $\{1, \ldots, 48\}$.

Contrairement au modèle *GAM*, un seul modèle *MARS* est conçu pour la prévision des courbes de charge non-thermosensibles et thermosensibles puisque le modèle est capable automatiquement de sélectionner les variables les plus pertinentes parmi les variables d'entrée en fonction de leur importance et leur contribution à l'amélioration du modèle de prévision. En plus, un seul modèle est conçu pour la prévision de la charge électrique à tout pas de temps de la journée contrairement, au modèle *GAM* que nous avons présenté précédemment qui propose un modèle par pas de temps. Le `package earth`[14] du logiciel R a été utilisé pour implémenter ce modèle *MARS*.

4.4.3 Résultats

Dans cette partie, nous comparons les résultats de la prévision de la charge électrique des ménages dans le jeu de données par les deux modèles *GAM* et *MARS* que nous avons décrit dans les deux parties précédentes.

Modèle	Thermosensibilité	NMAE	NRMSE	MASE	sMAPE
GAM	Non-thermosensible	0,46	0,82	0,84	45,26
	Thermosensible	0,44	0,80	0,82	52,51
MARS	Non-thermosensible	0,47	0,83	0,85	43,87
	Thermosensible	0,49	0,89	0,82	55,62

TABLE 4.9 – La performance moyenne des modèles de prévision *GAM* et *MARS* à $(J+1)$ selon les quatre erreurs (NMAE, NRMSE, MASE, sMAPE). Les meilleurs résultats sont affichés en bleu pour les courbes de charge non-thermosensibles, tandis que les meilleurs résultats pour les courbes de charge thermosensibles sont affichés en orange.

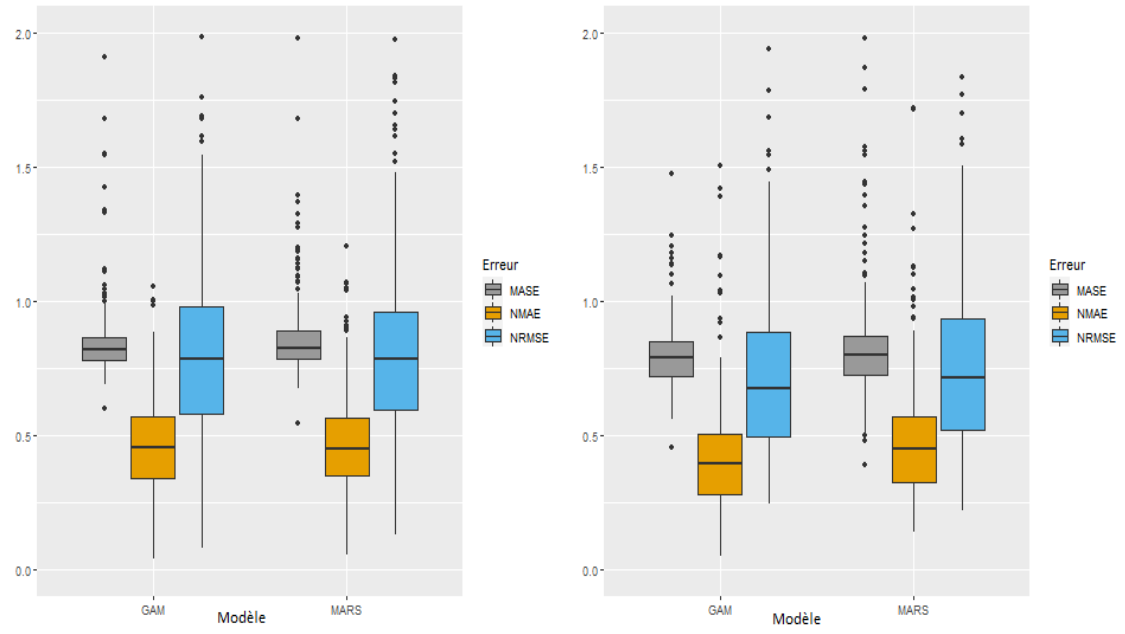
Le tableau 4.41 résume les performances moyennes des deux modèles de prévision *GAM* et *MARS* à $(J+1)$ appliqués au jeu de données à la fois pour les courbes de charge thermosensibles et non-thermosensibles. D'après ces résultats, nous pouvons conclure que la prévision par le modèle *GAM* est plus précise en moyenne selon les trois métriques NMAE, NRMSE, MASE et sMAPE pour les courbes de charge thermosensibles et non-thermosensibles que le modèle *MARS*, à l'exception de l'erreur sMAPE pour les courbes de

14. https ://cran.r-project.org/web/packages/earth/index.html.

charge non-thermosensibles. Étant donné que la moyenne des erreurs est très sensible aux valeurs aberrantes, nous avons examiné également la distribution des erreurs de prévision comme dans les étapes précédentes. Les résultats obtenus sont présentés sur la figure 4.41.

Les boîtes à moustaches sur la figure 4.41 montrent que pour les courbes de charge non-thermosensibles, les erreurs NMAE et NRMSE du modèle *MARS* ont une dispersion moindre par rapport à celles du modèle *GAM*. Cependant, la dispersion des erreurs MASE est plus grande pour le modèle *MARS* par rapport au modèle *GAM*. Cela suggère que le modèle *MARS* pourrait être meilleur pour prédire les valeurs absolues de la charge électrique, mais pourrait avoir une plus grande variabilité dans la prévision des changements relatifs de la charge par rapport au modèle *GAM*.

En revanche, les résultats de l'analyse des boîtes à moustaches pour les charges thermosensibles indiquent que le modèle *GAM* est plus performant que le modèle *MARS*, avec des erreurs de prévision moins étalées pour les trois métriques NMAE, NRMSE et sMAPE. De plus, les médianes des erreurs de prévision du modèle *GAM* sont plus petites que celles du modèle *MARS* pour chacune de ces métriques. En conséquence, il peut être conclu que le modèle *GAM* est plus fiable que le modèle *MARS* pour la prévision de charges électriques thermosensibles.



(a) Courbes de charge non-thermosensibles. (b) Courbes de charge thermosensibles.

FIGURE 4.41 – Les boîtes à moustaches des erreurs NMAE, NRMSE et MASE par le modèle *GAM* et le modèle *MARS*.

Bien qu'il soit difficile d'expliquer pourquoi les performances des modèles *GAM* peuvent être supérieures à celles des modèles *MARS* pour les charges thermosensibles et l'inverse pour les charges non-thermosensibles, il est généralement admis que les modèles *GAM* ont

une plus grande flexibilité que les modèles *MARS*, ce qui leur permet de mieux capturer les relations non linéaires dans les données. Ainsi, pour les charges thermosensibles, il est possible que les modèles *GAM* soient plus performants car ils sont capables de capturer les relations complexes entre la charge électrique et la température extérieure.

4.5 Prévision par le modèle *RNN-LSTM*

Les réseaux de neurones récurrents (*RNN*) sont largement déployés dans la littérature pour la prévision de la charge électrique à plusieurs échelles (Kong et al., 2017 ; Peñaloza et al., 2022 ; Hou et al., 2021). Ce vaste déploiement est due au fait que ces réseaux sont capables de traiter des données séquentielles. Contrairement, aux réseaux de neurones traditionnels, les unités de calcul, aussi appelées neurones, d'une couche donnée prennent en compte les informations calculées dans des étapes précédentes et les réinjectent dans le calcul des nouvelles sorties d'où le terme « réseau de neurones récurrents ». Cette particularité permet aux réseaux dans le contexte de la prévision de la charge électrique à l'échelle des ménages de détecter les corrélations présentes dans les données liées à la récurrence de certaines pratiques de consommation d'électricité (Kong et al., 2017).

4.5.1 Brève description du modèle

L'objectif principal de la mise en œuvre des réseaux de neurones récurrents est de permettre aux réseaux de traiter des données séquentielles et non pas uniquement des données isolées n'ayant aucun ordre chronologique. C'est grâce à ces neurones récurrents qui sont connectés à eux mêmes que l'information calculée au temps $(t-1)$ est réintégrée dans le calcul des sorties au temps (t) (voir la figure 4.42). Pour simplifier la description du fonctionnement du réseau récurrent nous prenons l'exemple d'un seul neurone récurrent j (voir la figure 3.3). Au temps t, le neurone récurrent j reçoit la donnée d'entrée $x^t = (x_1^t, \ldots, x_n^t)$ ainsi que sa propre sortie calculée à l'étape précédente y_j^{t-1}, il calcule une nouvelle sortie à partir de deux types de poids. Le vecteur de poids $W_j = (W_{1,j}, \ldots, W_{n,j})$ qui est le vecteur de poids classique accordé à la donnée d'entrée x^t lors de son passage dans neurone récurrent comme dans le cas d'un réseau de neurones classique (voir la section 3.2.3) tandis que le poids r_j est le poids accordé à la sortie précédente du neurone y_j^{t-1}. La sortie du neurone j calculée à l'étape t est alors donnée par :

$$y_j^t = \sigma\left(\sum_{i=1}^{n} W_{i,j} x_i^t + r_j y_j^{t-1} + b_j\right). \tag{4.12}$$

où σ est la fonction d'activation définie dans la sous-section 3.2.3, et b_j est le biais.

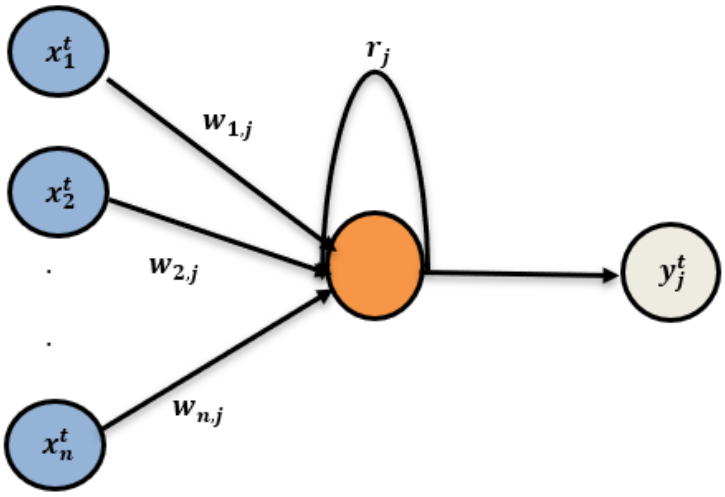


FIGURE 4.42 – Réseau de neurones avec un neurone récurrent. La flèche en arc représente la connexion ayant un décalage temporel (t-1).

Il est possible de construire un réseau de neurones récurrent avec plusieurs couches et de combiner des couches récurrentes avec des couches classiques, telles que des couches denses (A Multilayer Perceptron, *MLP*) ou des couches de convolution (*Convolutional Neural Networks*, *CNN*). Cette architecture du réseau permet de détecter des relations plus complexes dans les données. En revanche, cette pratique va à l'encontre de la rapidité de l'étape d'entraînement du réseau. Contrairement aux réseaux de neurones classiques, l'algorithme de rétropropagation du gradient qui sert à ajuster les poids dans l'étape d'entraînement n'est pas capable de prendre en compte la présence du cycle de récurrence dans ces réseaux.

La solution proposée à ce problème consiste à considérer une version dépliée dans le temps du réseau, qui élimine les cycles et par conséquent, rend l'utilisation de l'algorithme de rétropropagation standard possible. La figure 4.43 montre un exemple d'approximation du réseau récurrent par un réseau déplié deux fois dans le temps. L'algorithme de la rétropropagation du gradient appliqué au réseau déplié à travers le temps est appelé rétropropagation à travers le temps (*BackPropagation Through Time*) (WERBOS, 1990). Cet algorithme est le plus utilisé pour l'entraînement des réseaux de neurones récurrents.

En théorie, les réseaux de neurones récurrents sont censés détecter des dépendances à long terme dans les données. En revanche, l'approximation du réseau récurrent en réseau déplié K fois dans le temps ne permet de tenir compte que de K informations passées, ce qui limite sa capacité à modéliser des dépendances à long terme.

En plus, cette approximation rend le réseau récurrent plus profond (voir la figure 4.43) et par conséquent, il devient sujet à deux types de problèmes nommés respectivement dissi-

pation du gradient (*vanishing gradient*) [15] et exposition du gradient (*exploding gradient*) [16] (Hochreiter, Bengio et al., 2001). Ces deux problèmes sont connus dans le cas des réseaux de neurones profonds [17].

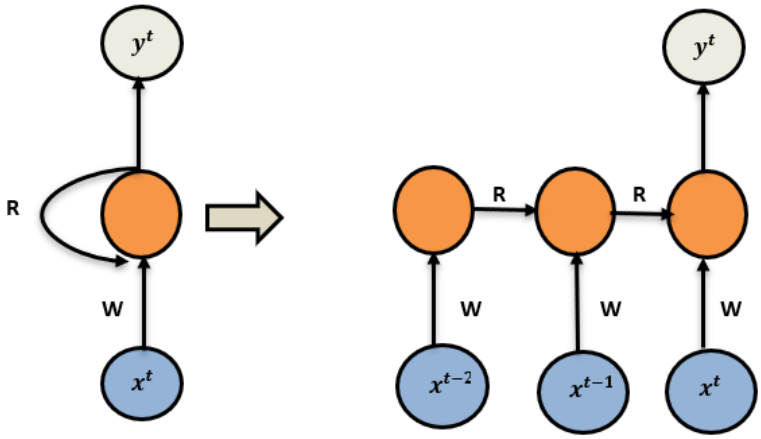


figure 4.43 – L'approximation d'un réseau de neurones récurrent non déplié par un réseau déplié deux fois dans le temps.

Pour pallier ces problèmes de calcul de gradients et permettre aux réseaux récurrents de modéliser des dépendances à long terme dans les données, une architecture des réseaux récurrents appelée ***Long-Short Term Memory* (*LSTM*)** ou réseaux de neurones récurrents à mémoire court-terme et long terme en français a été développée (Hochreiter et Schmidhuber, 1997). Les cellules *LSTM* à la base de ces réseaux possèdent une mémoire interne. Cette mémoire donne à la cellule la capacité de maintenir un état sur une longue période de temps. Une cellule *LSTM* des réseaux de neurones récurrents a une structure plus complexe que celle d'un neurone récurrent. Elle est formée d'un noyau central contenant l'état interne de la cellule, d'une porte d'oubli (*Forget Gate*), d'une porte d'entrée (*Input Gate*) et d'une porte de sortie (*Output Gate*) (voir la figure 4.44). Les portes contrôlent le flux d'informations. La porte d'entrée contrôle le flux d'informations entrant dans la cellule de mémoire. La porte de sortie contrôle le flux d'informations sortant de la cellule mémoire. La porte d'oubli filtre les informations contenues dans la cellule mémoire de l'étape précédente et décide quelle information doit être conservée et injectée dans le calcul de l'étape suivante et celle qui doit être jetée. L'architecture du réseau de neurones récurrents a évolué avec le temps et une architecture populaire a été introduite par Gers et al. (2002). Dans cette dernière des connexions dites « connexions *peephole* » sont ajoutées aux connexions du réseau récurrent classique. Ces connexions permettent à toutes

15. Le problème de **dissipation du gradient** est causé par la diminution rapide des valeurs des gradients calculés par la méthode de rétropropagation du gradient. Cette diminution peut provoquer l'arrêt de la phase d'entraînement. Ce problème s'aggrave avec l'augmentation du nombre des couches cachées dans le réseau.

16. Le problème **d'exposition du gradient** est causé par l'augmentation des gradients d'une façon exponentielle. Cette augmentation a pour effet de déstabiliser l'entraînement du réseau.

17. Le réseau de neurones profond est un réseau de neurones formé de plusieurs couches cachées.

les portes d'inspecter l'état actuel de la cellule même avant la fermeture ou l'ouverture de la porte de sortie. Une illustration de la variante *peephole* d'une cellule de *LSTM* est représentée sur la figure 4.44.

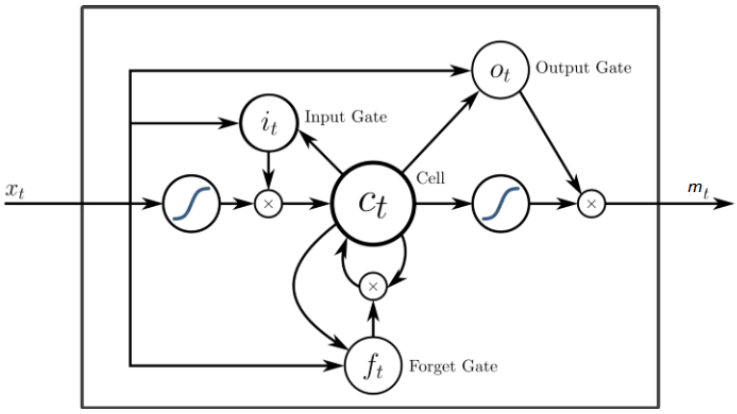


FIGURE 4.44 – Une représentation d'une cellule *LSTM* avec *peephole*, la porte d'entrée (i_t), la porte d'oubli (f_t) et la porte de sortie (o_t) [Source devoteam] [18].

Un réseau LSTM calcule les sorties $(y_1, \ldots, y_T)$ d'une séquence d'entrée $(x_1, \ldots, x_T)$ d'une manière itérative de $t = 1$ à T à l'aide des équations suivantes :

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i). \tag{4.13}$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f). \tag{4.14}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{cx}x_t + W_{cm}m_{t-1} + b_c). \tag{4.15}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o). \tag{4.16}$$

$$m_t = o_t \odot tanh(c_t). \tag{4.17}$$

$$y_t = \phi(W_{ym}m_t + b_y). \tag{4.18}$$

où W désignent des matrices de poids (W_{ix} est la matrice des poids de la porte d'entrée à l'entrée), W_{ic}, W_{fc}, W_{oc} sont des matrices de poids diagonales pour les connexions *peephole*. b_i, b_f, b_c, b_o et b_y désignent des vecteurs de biais, σ est la fonction sigmoïde logistique. i,f, o et c sont respectivement la porte d'entrée (*in*), la porte d'oubli (*forget*), la porte de sortie (*out*) et les vecteurs d'activation de la cellule (*cell*), qui ont tous la même taille que le vecteur d'activation m de la sortie de la cellule. $\odot$ est le produit élément par élément des vecteurs. σ est la fonction logistique sigmoïde. ϕ est la fonction d'activation de sortie du réseau.

18. https://tinyurl.com/3er9z9ur

Les matrices de poids sont ajustées comme dans le cas des réseaux récurrents classiques par l'algorithme de rétropropagation à travers le temps (*BackPropagation Through Time*).

4.5.2 Approche de prévision

La même approche proposée pour la prévision de la charge électrique des ménages par les modèles *GAM* et *MARS* décrite dans la partie 4.4.1.1 est adoptée pour la prévision par le modèle de réseau de neurones récurrent à mémoire court et long terme. Nous avons utilisé le modèle *RNN-LSTM* présenté dans (Kong et al., 2017) en l'adaptant à notre ensemble de données. Ce modèle s'est avéré le plus performant en moyenne en termes de précision pour la prévision de la charge électrique de 69 ménages de la Nouvelle Galles du Sud parmi plusieurs modèles de référence utilisés pour la prévision de la charge électrique.

Bien que le modèle de réseau de neurones récurrent à mémoire court et long terme soit largement utilisé dans la littérature pour la prévision de la charge électrique (Shi et al., 2017 ; Kong et al., 2017 ; Alonso et al., 2020), nous estimons que pour une utilisation industrielle, nous pouvons regretter leur manque d'interprétabilité. Cependant, nous avons fait le choix de développer un modèle *RNN-LSTM* dans l'objectif de vérifier si une probable amélioration dans la précision de la prévision justifie leur utilisation au détriment de l'interprétabilité dans le contexte de la prévision de la charge électrique à l'échelle des ménages. La première étape consiste à préparer l'ensemble de données pour la prévision par le modèle *RNN-LSTM*. Cela implique de définir le problème de prévision des données séquentielles comme étant un problème d'apprentissage supervisé. Notre approche consiste à prédire la séquence de la consommation d'électricité du jour $(J+1)$ compte tenu des données de consommation d'électricité du jour J, de la température extérieure (pour la prévision de la charge thermosensible) ainsi que du type de jour. Le modèle *RNN-LSTM* grâce à ces cellules *LSTM* permet ensuite de détecter les dépendances temporelles dans ces données. Par conséquent, la charge électrique au même pas de temps du jour précédent jusqu'à la dernière charge électrique mesurée avant la prévision peuvent toutes contribuer à prédire la charge électrique à un pas de temps donné du jour $(J+1)$.

Soit $(y_t)_t$ la série temporelle représentant la courbe de charge du ménage. Pour transformer le problème de prévision de la série temporelle $(y_t)_t$ en un problème d'apprentissage supervisé, nous supposons que $(y_{t-48}, \ldots, y_{t-1})$ le vecteur représentant la consommation d'électricité du jour J et que notre objectif est alors de prédire le vecteur de sortie $(y_t, \ldots, y_{t+47})$ du jour $(J+1)$ (voir la figure 4.45). La stratégie de prévision présentée dans figure 4.45 est dite la prévision glissante (*Rolling Forecast*) d'une fenêtre d'un jour. Dans la majorité des recherches sur la prévision de la charge d'électricité la taille de cette fenêtre est fixée à 1 (Kong et al., 2017). C'est à dire que la série temporelle est glissée d'un seul pas de temps. Cette stratégie permet d'intégrer les données les plus récentes dans le modèle et par conséquent, améliore significativement sa performance. Par contre, dans

notre cas puisqu'aucune information sur la consommation du jour $(J+1)$ n'est disponible au moment de prévision, nous intégrons uniquement les données de consommation du jour J pour prédire la consommation électrique du jour $(J+1)$. La taille de la fenêtre est alors fixée à 48. Le réseau *RNN-LSTM* construit est alors alimenté la veille par les nouvelles données recueillies pendant la journée pour la prévision du lendemain.

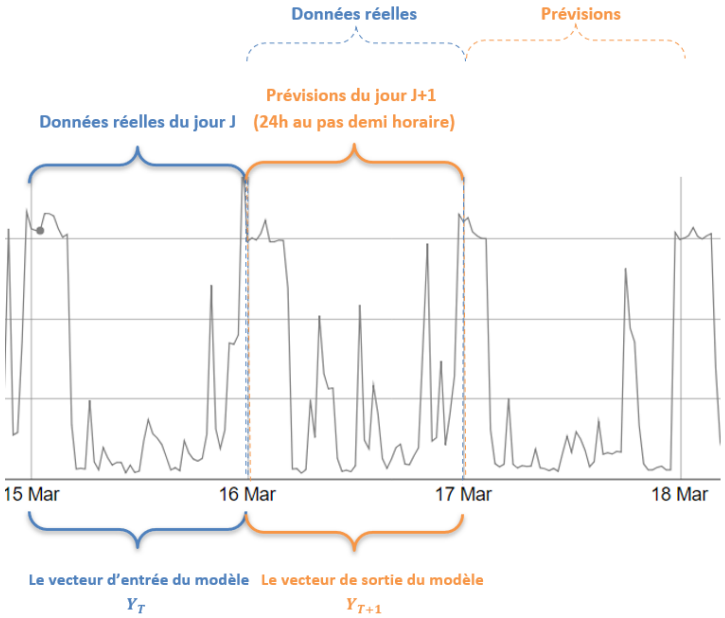


FIGURE 4.45 – Prévision par fenêtre glissante de la charge électrique d'un ménage.

4.5.2.1 Variables d'entrée et architecture du modèle

Avant de pouvoir adapter le modèle *RNN-LSTM* à l'ensemble des données, nous devons transformer les données en variables d'entrée du modèle. Les variables d'entrée que nous avons extraites des données et utilisées dans notre modèle de prévision sont les suivantes :

1. la séquence de la consommation d'électricité du ménage pour les 48 derniers pas de temps représentée par $\mathcal{Y} = (y_{t-48}, \ldots, y_{t-1})$.
2. la séquence de la température extérieure pour les 48 derniers pas de temps représentée par $\mathbf{T} = (T_{t-48}, \ldots, T_{t-1})$.
3. la variable correspondante au jour de la semaine dans $\mathcal{Y}$. Cette variable notée $\mathbf{D}$ est comprise entre 1 et 7 où 1 correspond à un dimanche et 7 à un samedi.
4. la variable binaire notée $\mathbf{H}$, indiquant si le jour dans $\mathcal{Y}$ est un jour férié ou non (1 pour férié et 0 sinon).

Après l'extraction des ces variables d'entrée, il est nécessaire de les mettre à l'échelle avant de les intégrer dans le modèle. Cette étape est primordiale pour la prévision par les techniques des réseaux de neurones en raison de leur sensibilité à l'échelle des données.

Elle consiste à transformer l'ensemble de variables d'entrée du modèle à une échelle similaire souvent entre 0 et 1 ou entre -1 et 1. Cette mise à l'échelle des variables d'entrée permet d'accélérer la convergence de l'algorithme de descente de gradient, et d'éviter que certaines variables d'entrée affectent d'une façon disproportionnée la phase d'entraînement du modèle en raison de la différence d'échelle.

Nous avons donc opté pour la normalisation $\min - \max$ pour les variables des courbes de charge ($\mathcal{Y}$) et de la température ($\mathbf{T}$). Cette méthode consiste à borner une variable z entre 0 et 1. Les nouvelles valeurs normalisées de z sont alors calculées par la formule suivante :

$$z_{norm} = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{4.19}$$

Les variables $\mathbf{D}$ et $\mathbf{H}$ sont mises à l'échelle par la méthode d'encodage dite *one-hot encoder* (Trevor Hastie et al., 2009). Cette méthode consiste à transformer chaque état de la variable en un vecteur de coordonnées (0 ou 1) ayant une dimension égale au nombre total des modalités de la variable. La seule coordonnée égale à 1 dans le vecteur est la coordonnée ayant le numéro de la modalité prise par la variable. Les variables d'entrée après la mise à l'échelle sont concaténées dans une matrice, notée $\mathbf{X}$, représentant la matrice d'entrée du modèle et $\mathbf{X}$ définie ainsi :

$$\mathbf{X} = [\widetilde{\mathcal{Y}}^T, \widetilde{\mathbf{T}}^T, \widetilde{\mathbf{D}}^T, \widetilde{\mathbf{H}}^T]. \tag{4.20}$$

où $\widetilde{\mathcal{Y}}, \widetilde{\mathbf{T}}, \widetilde{\mathbf{D}}$ et $\widetilde{\mathbf{H}}$ représentent les variables $\mathcal{Y}$, $\mathbf{T}, \mathbf{D}$ et $\mathbf{H}$ mises à l'échelle. La matrice $\mathbf{X}$ du modèle doit ensuite être redimensionnée en un tenseur à trois dimensions :

1. le nombre de valeurs de la première dimension correspond au nombre d'échantillons dans les données. Ici, cette dimension est égale au nombre du jour dans l'historique des données.
2. le nombre de valeurs de la deuxième dimension correspond au pas de temps ou *timestamp*. Cette dimension permet de définir la longueur de la série temporelle de l'historique de la charge électrique à intégrer comme données d'entrée à chaque pas de temps. Dans notre cas, cette dimension est fixée à 48 (48 valeurs pour une journée entière).
3. le nombre de valeurs de la troisième dimension correspond au nombre de caractéristiques ou *features* utilisées comme variables d'entrée du modèle. Dans notre cas, ce nombre est égal à 11 (une variable pour l'historique de la courbe de charge, une variable pour la température extérieure, sept variables pour les jours de la semaine et deux variables pour les jours fériés).

Comme dans le cas du modèle *GAM*, nous avons mis en œuvre deux modèles de réseaux récurrents pour la prévision de la charge électrique thermosensible et non-thermosensible qui se distinguent par l'intégration ou pas de la température dans la matrice des variables d'entrée $\mathbf{X}$. Il est important de souligner qu'il est également recommandé dans les modèles

de réseau de neurones récurrents de procéder à la transformation des séries temporelles afin de les rendre stationnaires. Cette pratique a pour impact d'améliorer la performance du modèle de prévision. Dans le jeu de données, les courbes de charge thermosensibles sont généralement celles qui présentent des problèmes de stationnarité en raison de la présence d'une forte saisonnalité annuelle. Cette saisonnalité reflète l'impact de la variation de la température sur la charge électrique. Tandis que, les problèmes de stationnarité liés à l'évolution de la tendance sont moins fréquents. Par conséquent, nous avons fait le choix de ne pas transformer les données pour supprimer la saisonnalité annuelle des séries temporelles puisque nous considérons que le modèle *RNN-LSTM* est capable de la capturer à travers la variation des données de la température extérieure.

Concernant l'architecture du modèle *RNN-LSTM* que nous avons mis en œuvre, nous avons fait le choix de développer un modèle qui se compose d'une couche d'entrée, de deux couches *LSTM* cachées et d'une couche dense de sortie qui fournira la prévision à 24h au pas demi-horaire. Cette architecture est utilisée dans plusieurs études sur la prévision de la charge électrique (Kong et al., 2017 ; Alonso et al., 2020 ; Hou et al., 2021). En effet, en raison de la nature séquentielle de la sortie des couches *LSTM*, un nombre arbitraire de couches *LSTM* peut être empilé pour former un réseau de neurones profond. Selon certaines études l'utilisation de plusieurs couches *LSTM* cachées a un impact positif sur l'amélioration de la performance des modèles de prévision (Kong et al., 2017).

L'optimisation des hyperparamètres est une étape très intéressante qui permet de trouver la configuration du modèle qui produit les meilleures performances. Cette étape consiste généralement à entraîner et à évaluer plusieurs modèles pour diverses combinaisons possibles afin de sélectionner les valeurs les plus optimales. En revanche, puisque nous traitons le sujet de prévision de la charge électrique à l'échelle des ménages individuellement, le réglage individuel de ces hyperparamètres pour chaque modèle de prévision est très chronophage. Même l'optimisation de ces hyperparamètres pour l'ensemble de jeu de données de 720 ménages globalement n'est pas une tâche aisée. Pour cette raison, nous nous sommes basés généralement sur les recommandations de la littérature ainsi que notre propre expérience pour fixer certains hyperparamètres. D'autres hyperparamètres ont été sélectionnés manuellement après avoir testé plusieurs valeurs possibles sur un petit échantillon de données tiré aléatoirement de l'ensemble du jeu de données. Les valeurs utilisées pour les hyperparamètres de notre modèle sont alors représentées dans la table suivante 4.10.

Hyperparamètre	**Description**	**Valeur**
Couches cachées	Nombre de couches d'un réseau de neurones hormis les couches d'entrée et de sortie	2
Neurones cachées	Nombre de neurones dans chaque couche cachée	20 et 10
Fonction d'activation	Fonction d'activation utilisée à la sortie de chaque couche	sigmoïde et relu
Algorithme d'optimisation	Pour la descente de gradient	*Adam Optimizer*
Learning rate	Nombre qui détermine l'importance accordée à chaque mise à jour des poids	0,001
Fonction de perte/coût	Fonction qui estime la perte du modèle et intervient à la mise à jour des poids	erreur quadratique moyenne
Batch	Nombre d'échantillons traités avant la mise à jour des paramètres du modèle	1
Epochs	Nombre de passages de l'ensemble de données dans le réseau	150

TABLE 4.10 – Les hyperparamètres du modèle *RNN-LSTM*.

Le `package Keras`[19] du logiciel R a été utilisé pour mettre en œuvre le modèle *RNN-LSTM*. L'architecture de notre modèle de prévision *RNN-LSTM* est présentée sur la figure 4.46. La première couche du réseau est une couche d'entrée qui reçoit les données mises à l'échelle sous forme d'une matrice à trois dimensions comme expliqué précédemment. Cette couche est suivie d'une couche *LSTM* contenant 20 cellules *LSTM*, puis d'une couche *LSTM* contenant 10 cellules *LSTM*. Enfin, une couche dense (*fully connected*) avec 48 neurones permettra de fournir les prévisions correspondantes à 24 heures au pas demi-horaire de la charge électrique. La fonction d'activation sigmoïde a été utilisée pour transformer la sortie de chaque cellule *LSTM* de la première couche *LSTM* et la fonction d'activation relu pour la deuxième couche *LSTM* et la couche dense de sortie. Pour la fonction du coût, nous avons choisi l'erreur quadratique moyenne et pour l'algorithme d'optimisation *Adam optimizer* avec un *learning rate* égal à 0,001. Durant la phase d'entraînement du modèle, le fonctionnement de l'algorithme de rétropropagation (*Back-Propagation Through Time*) (voir la sous-section 4.5.1), nécessite l'ajustement de deux hyperparamètres qui sont la taille de *batch* (*batch size*) et le nombre d'*epochs*. La description de ces deux hyperparamètres est présentée dans la table 4.10. Dans notre modèle le nombre d'*epochs* est fixé à 150 et la taille de *batch* à 1. En effet, la taille de *batch* limite le nombre de valeurs d'entrée dans le réseau avant la mise à jour des poids. Cette limitation est imposée non seulement pendant la phase d'entraînement mais également pendant la phase de prévision. Par conséquent, la même taille doit être utilisée pour l'entraînement

19. https ://cran.r-project.org/web/packages/keras/index.html.

du modèle et la prévision. Comme nous prévoyons une journée entière de 24h à chaque sortie du modèle donc le nombre de *batch* que nous avons fixé à 1 correspond exactement à ce que nous voulons.

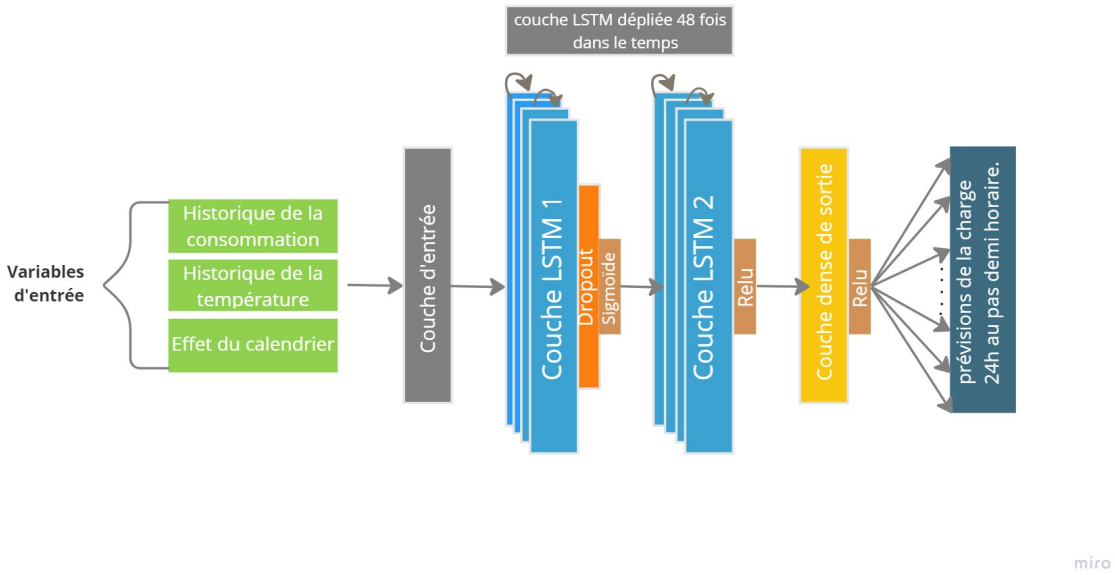


FIGURE 4.46 – L'architecture du modèle *RNN-LSTM*.

4.6 Comparaison des performances de tous les modèles de prévision

4.6.1 Précision

Dans cette sous-section, nous comparons la performance du modèle *RNN-LSTM* proposé dans la section précédente aux performances des modèles *KWF*, *GAM* et *MARS*. Les moyennes des erreurs de prévision NMAE, NRMSE, MASE et sMAPE à $(J+1)$ de la charge électrique des ménages dans le jeu de données sont présentées dans la table 4.11. Ces résultats révèlent que le modèle *KWF* avec *clustering* est le plus performant en moyenne en termes de précision de prévision pour les quatre métriques dans le cas des courbes de charge thermosensibles. Dans le cas des courbes de charge non-thermosensibles, aucun modèle s'est révélé le plus performant en moyenne pour les quatre métriques. Les résultats montrent qu'en moyenne, le modèle *GAM* est le plus performant en termes de l'erreur NMAE, tandis que le modèle *MARS* est le plus performant en termes de l'erreur sMAPE. Le modèle *KWF* s'est révélé le plus performant en termes de l'erreur NRMSE et MASE. Néanmoins, l'amélioration de l'erreur NRMSE du modèle *KWF* par rapport aux modèles *GAM* et *MARS* est plus importante que l'amélioration des erreurs NMAE et sMAPE par ces derniers. Le modèle *RNN-LSTM* et le modèle *MARS* ont la même précision de

prévision mesurée en NRMSE.

La figure 4.47 montre les boîtes à moustaches des erreurs de prévisions à $(J+1)$ MASE, NMAE et NRMSE des quatre modèles.

Modèle	**Thermosensibilité**	NMAE	NRMSE	MASE	sMAPE
GAM	Non-thermosensible	0,46	0,82	0,84	45,26
	Thermosensible	0,44	0,80	0,82	52,51
KWF	Non-thermosensible	0,47	0,72	0,81	44,57
KWF* avec *clustering	Thermosensible	0,41	0,69	0,78	52,05
MARS	Non-thermosensible	0,47	0,83	0,85	43,87
	Thermosensible	0,49	0,89	0,82	55,62
RNN-LSTM	Non-thermosensible	0,48	0,83	0,87	45,02
	Thermosensible	0,51	0,88	0,86	56,73

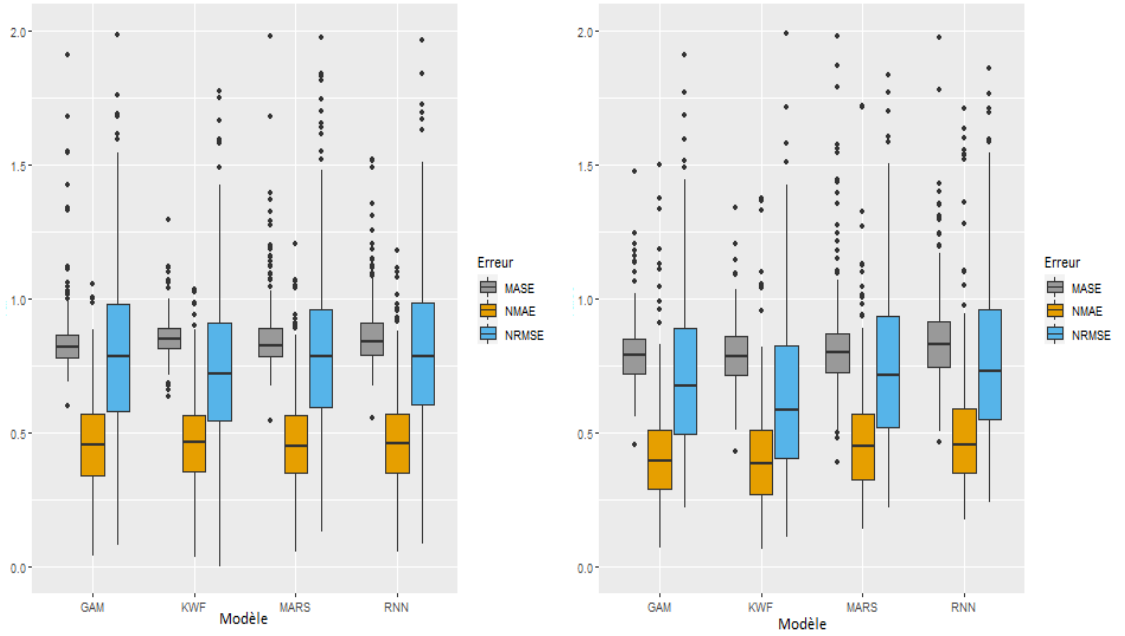
TABLE 4.11 – La performance moyenne des modèles de prévision *KWF*, *GAM*, *MARS* et *RNN-LSTM* à $(J+1)$ selon les quatre erreurs NMAE, NRMSE, MASE et sMAPE. Les meilleurs résultats sont affichés en bleu pour les courbes de charge non-thermosensibles, tandis que les meilleurs résultats pour les courbes de charge thermosensibles sont affichés en orange.

Nous examinons tout d'abord la performance des modèles de prévision des courbes de charge non-thermosensibles. En ce qui concerne l'erreur NRMSE, la comparaison visuelle des boîtes à moustaches des erreurs NRMSE des modèles testés montre une performance supérieure en distribution du modèle *KWF* par rapport aux autres modèles. Par contre, le classement par niveau de performance des modèles *GAM*, *MARS* et *RNN-LSTM* peut difficilement être établie par une comparaison visuelle des boîtes à moustaches des erreurs NRMSE. Par contre, l'hypothèse selon laquelle nous supposons que les trois modèles ont des performances similaires par rapport à l'erreur NRMSE est rejetée par le test statistique de Friedman [20] (p-valeur $< 1,10^{-5}$) indiquant ainsi qu'il existe au moins un modèle ayant une performance différente des deux autres modèles. Les tests non paramétriques de Wilcoxon effectués aux erreurs de prévision NRMSE produites par les modèles *GAM*, *MARS* et *RNN-LSTM* deux à deux rejettent également tous l'hypothèse nulle des performances similaires (des valeurs de p-valeur inférieures à $0,001$). Nous classons donc les performances des modèles de prévision dans l'ordre décroissant par rapport à la distribution de l'erreur NRMSE (la dispersion et le niveau de la médiane) de la manière suivante : le modèle *KWF*, le modèle *MARS*, le modèle *GAM* et le modèle *RNN-LSTM*.

En ce qui concerne l'erreur NMAE, visuellement les modèles semblent avoir des performances similaires avec des niveaux de médianes légèrement différents. Cette hypothèse de similarité de la performance par rapport à l'erreur NMAE est rejetée par le test statistique

20. Le test de Friedman est un test statistique utilisé pour évaluer s'il existe des différences statistiquement significatives entre les distributions de trois groupes appariés ou plus.

de Friedman (p-valeur = $2,10^{-5}$). Cependant, le test de Wilcoxon ne rejette pas l'hypothèse nulle de performances similaires par rapport à cette erreur entre le modèle *GAM* et *MARS* (p-valeur = $0,37$) d'une part, et le modèle *KWF* et *RNN-LSTM* (p-valeur = $0,29$) d'une autre part. Par conséquent, nous pouvons conclure que les deux modèles *GAM* et *MARS* d'une part ainsi que les deux modèles *KWF* et *RNN-LSTM* d'une autre part ont des performances similaires par rapport à l'erreur NMAE.



(a) Courbes de charge non-thermosensibles. (b) Courbes de charge thermosensibles.

FIGURE 4.47 – Les boîtes à moustaches des erreurs NMAE, NRMSE et MASE par les modèles *GAM*, *KWF*, *MARS* et *RNN-LSTM*.

Par rapport aux résultats de la prévision des courbes de charge thermosensibles, nous pouvons d'après les distributions des erreurs présentées par les boîtes à moustaches de classer les performances des modèles de prévision par rapport aux métriques NRMSE et NMAE par ordre décroissant de la manière suivante : le modèle *KWF*, le modèle *GAM*, le modèle *MARS* et le modèle *RNN-LSTM*.

4.6.2 Temps de calcul

Le temps de calcul est l'un des critères de performance à tenir en compte lors du choix entre les modèles de prévision ayant des performances similaires surtout quand ces derniers sont utilisés à des fins industrielles. Dans le contexte de la prévision de la charge électrique à l'échelle des ménages, le temps de calcul joue un rôle crucial. En effet, le modèle de prévision doit être entraîné pour des milliers de ménages, pour ensuite, générer des prévisions quotidiennes. Le temps de calcul dépend alors du nombre de ménages, de

l'horizon de prévision et de la technique de prévision sélectionnée. Dans notre étude, les calculs parallèles ont considérablement réduit le temps de calcul. Les modèles de prévision de la charge électrique des ménages que nous avons proposés sont appliqués aux données propres à chaque ménage indépendamment des autres et d'une manière automatisée qui nécessite aucune intervention humaine. Cela rend les étapes d'entraînement indépendantes les unes des autres et permet la répartition de la tâche de prévision des ménages entre les cœurs de calcul. Le temps de calcul est alors réduit proportionnellement au nombre des cœurs de calcul disponibles. Le calcul parallèle a été utilisé en deux étapes dans notre approche. La première étape consiste à répartir les tâches chronophages indépendantes dans les modèles de prévision comme dans le cas de l'algorithme de calcul et de sélection des *features* de la méthode de *clustering* intégrée dans le modèle *KWF* (voir la sous-section 4.3.5) ou dans le calcul de chaque équation correspondant à un pas de temps du modèle *GAM* (voir la sous-section 4.4.1). Les modèles ont été testés dans un premier temps sur un ordinateur portable ayant la configuration et l'environnement suivant : Intel(R) Core(TM) i5-7200U CPU @ 2,50GHz avec 15,9 Go de RAM. Ensuite, nous avons mis en place des scripts automatisés qui permettent de lancer les modèles de prévision pour tous les ménages en parallèle sur un serveur de calcul et de sauvegarder les résultats en exploitant tous les cœurs de calcul disponibles. Les temps d'entraînement et de prévision des modèles sont présentés dans la table 4.12. Le temps d'entraînement correspond au temps nécessaire pour entraîner le modèle de prévision sur toute la période d'entraînement et pour un seul ménage alors que le temps de prévision correspond au temps nécessaire pour calculer la prévision d'une journée.

Modèle	Temps d'entraînement (en s)	Temps de prévision (en s)
GAM	23,31	2,65
KWF	0,9	0,12
KWF* avec *clustering	34,9	0,12
MARS	8,50	0,0059
RNN-LSTM	300	1,25

TABLE 4.12 – Le temps d'entraînement et de prévision en secondes (s) des quatre modèles de prévision de la charge électrique à l'échelle des ménages.

Le temps d'entraînement du modèle *KWF* avec *clustering* correspond à la somme du temps requis pour l'étape de *clustering* (34,0 s) et du temps requis pour l'étape d'entraînement du modèle *KWF* (0,9 s). En pratique, cette étape de *clustering* se fait une seule fois pour l'ensemble des données d'entraînement. Ensuite, le modèle de *clustering* ainsi que les *clusters* des courbes de charge journalières sont sauvegardés et au fur et à mesure de l'arrivée des nouvelles données. Elles sont attribuées aux *clusters* déjà déterminés dans l'étape de *clustering* et intégrées dans le modèle de prévision *KWF*. Le temps d'entraînement élevé du modèle *RNN-LSTM* revient au fait que nous avons fait le choix d'ajuster le modèle pour un seul *batch*. L'augmentation du nombre de *batchs* permet d'accélérer

l'étape d'entraînement mais diminue la précision de la prévision pour certaines courbes de charge.

Les résultats de la table 4.12 montrent que, à l'exception du modèle *RNN-LSTM*, le temps d'entraînement et le temps de prévision des modèles que nous avons proposés sont raisonnables. Ils ne font pas obstacle à l'utilisation de ces modèles quotidiennement pour la prévision de la charge électrique à l'échelle des ménages.

4.7 Cas d'étude

L'analyse des courbes de charge des ménages dans le jeu de données a révélé que la majorité des courbes ayant une très faible précision de prévision présentaient un changement brutal entre la période d'entraînement et la période de test. Par conséquent, les modèles de prévision ont été entraînés sur une période différente de la période test, ce qui explique la faible précision de la prévision de ces ménages. Ces changements brutaux peuvent être causés par l'installation ou l'utilisation de nouveaux appareils électriques dont la consommation serait suffisamment importante pour générer un changement dans la courbe de charge électrique (comme des chauffages d'appoint, des ventilateurs, un aquarium, un congélateur, ...) ou par des modifications des habitudes de consommation (comme le changement de la température de chauffage du logement, l'augmentation du nombre des occupants, ...).

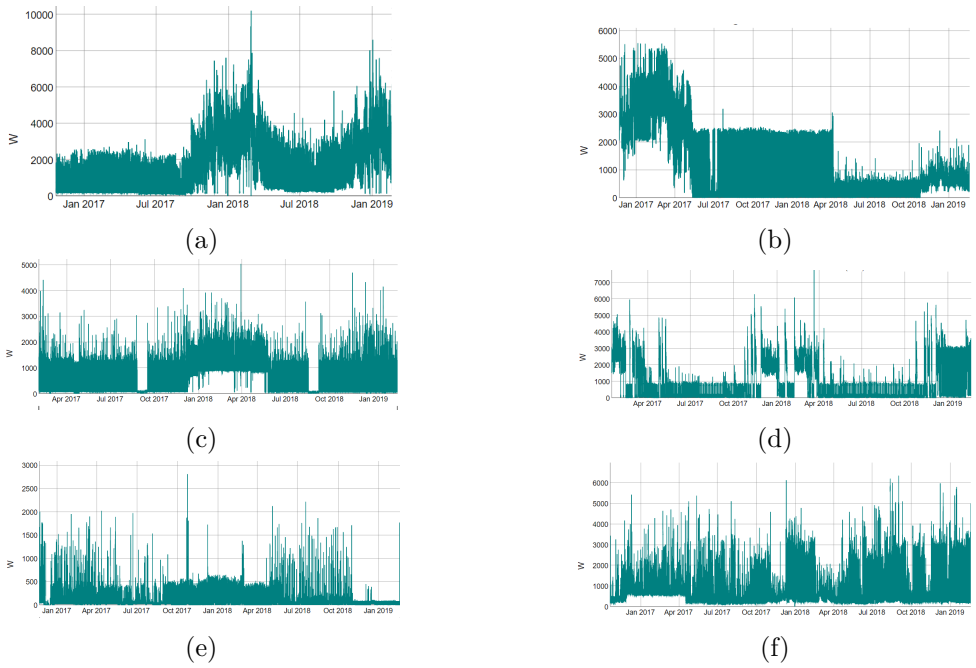


FIGURE 4.48 – Exemples des courbes représentant des changements brutaux de la charge électrique.

Des exemples de courbes de charge dans le jeu de données qui illustrent cette situation sont présentés sur la figure 4.48. De plus, nous avons remarqué que nous avons parmi les courbes de charge ayant des résultats de prévision médiocres, 12 courbes de charge électrique des logements étudiants (voir la figure 4.49). Ces courbes sont très volatiles, très bruitées et irrégulières avec des longues périodes de vacances. Elles reflètent le mode de vie dynamique et irrégulier des étudiants. L'ensemble de ces logements étudiants ainsi que les ménages dont les courbes de charge présentent des changements brutaux (noté Σ) représente environ 14,8% de l'ensemble des ménages dans le jeu de données (noté D).

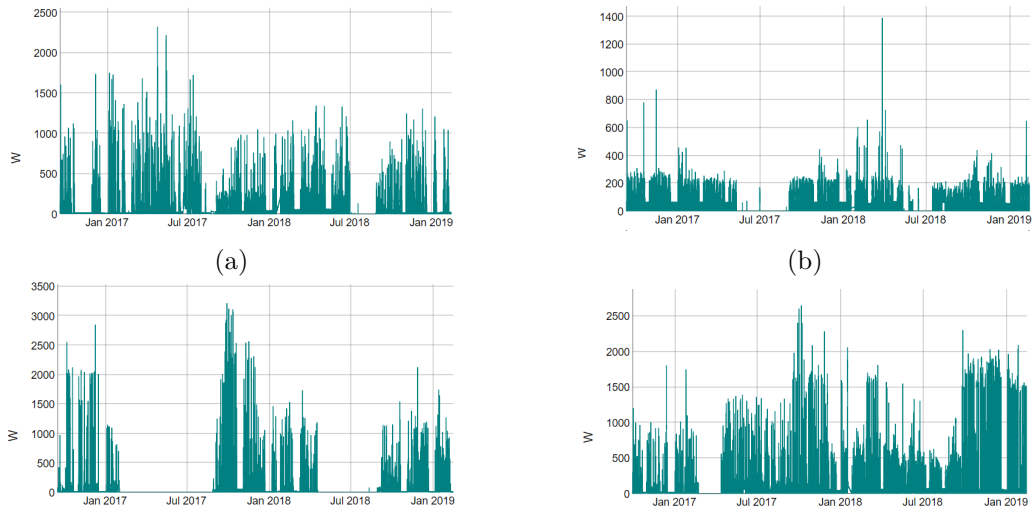


FIGURE 4.49 – Exemple de quatre courbes de charge électrique des logements étudiants.

Le tableau 4.13 montre les moyennes des erreurs de prévision par le modèle KWF de l'ensemble des ménages dans le jeu de données (D) ainsi que les deux sous-ensembles (Σ) et $(D - \Sigma)$ (le sous-ensemble des ménages restant de (D) après la privation de l'ensemble (Σ)).

Données	**Modèle**	**Thermosensibilité**	NMAE	NRMSE	sMAPE
(D)	KWF	Non-thermosensible	0,47	0,72	44,57
	KWF avec *clust*	Thermosensible	0,41	0,69	52,05
$(D - \Sigma)$	KWF	Non-thermosensible	0,41	0,64	40,88
	KWF avec *clust*	Thermosensible	0,35	0,56	47,55
(Σ)	KWF	Non-thermosensible	0,74	1,30	67,57
	KWF avec *clust*	Thermosensible	0,76	1,57	84,45

TABLE 4.13 – Les moyennes des erreurs de prévision par le modèle KWF (KWF avec *clustering* pour les courbes de charge thermosensibles) à $(J + 1)$ de l'ensemble du jeu de données (D) et des sous-ensembles $(D - \Sigma)$ et (Σ).

Ces résultats montrent un écart important entre les moyennes des trois erreurs de

prévision (NMAE, NRMSE et sMAPE) des deux ensembles des données (D) et ($D - \Sigma$) et celles de l'ensemble (Σ). Celles de l'ensemble (Σ) sont très élevées par rapport aux moyennes des erreurs des ménages dans l'ensemble du jeu de données (D) et encore plus importantes que celles dans l'ensemble ($D - \Sigma$) qui sont les plus faibles. Cela signifie que les ménages de l'ensemble (Σ) ont un impact négatif sur la précision des prévisions pour l'ensemble des données.

Certes, l'exclusion de ces ménages de l'ensemble du jeu de données comme c'est le cas dans la majorité des études portant sur la prévision de la charge électrique à l'échelle des ménages permettra de mettre en avance la performance des modèles de prévision et de réduire les erreurs de prévision. Par contre, cette pratique ne permet pas d'évaluer l'adaptabilité de ces modèles à la prévision des données réelles de différents profils de courbes de charge électrique. Afin d'améliorer la prévision de la charge électrique de ce type de ménage, la technique d'entraînement récursif du modèle de prévision est recommandée. Elle consiste à mettre à jour le modèle régulièrement en l'entraînant sur des nouvelles données, ce qui permettra d'intégrer avec le temps tout changement brutal dans la courbe de charge. La fréquence de l'entraînement du modèle peut être fixée régulièrement tous les mois par exemple ou bien lorsque un grand écart inhabituel est constaté entre les prévisions et les nouvelles données de consommation électrique du ménage.

4.8 Approche d'énergie pour les courbes de charge les plus volatiles

Comme nous l'avons expliqué précédemment, la difficulté de la prévision de la charge électrique à l'échelle des ménages réside dans le fait que cette charge est très bruitée, irrégulière et très volatile. Dans cette section, nous proposons une nouvelle approche qui consiste à prédire l'énergie électrique consommée pendant la journée au lieu de prédire la courbe de puissance. Cette approche permet d'une part de réduire l'irrégularité dans ces courbes de charge, et d'autre part de rendre les prévisions plus faciles à interpréter pour les consommateurs.

En réalisant la prévision de l'énergie électrique consommée pendant la journée, nous essayons d'éliminer le bruit et l'irrégularité dans les courbes de puissance liés au décalage des habitudes de consommation dans la journée qui est souvent identifié dans les courbes de charge électrique des ménages. En effet, le consommateur ne se soucie pas de décaler d'une heure, par exemple, la mise en route de sa machine à laver pour des raisons de commodité personnelle en particulier lorsque le prix de l'énergie électrique est fixe pendant la journée ou pendant des plages horaires fixes. Cependant, quelle que soit la raison pour laquelle un consommateur choisit de décaler sa consommation d'électricité, cela peut entraîner une irrégularité dans les courbes de charge et affecter la qualité des prévisions

de sa consommation électrique. En revanche, ce décalage ne signifie pas forcément une surconsommation de l'électricité dans la journée. Pour cette raison, nous estimons que le consommateur s'intéresse davantage à recevoir des prévisions de la totalité de sa consommation d'électricité à un instant donné t de la journée plutôt qu'à des prévisions de la puissance.

L'énergie électrique consommée pendant la journée est calculée au pas demi-horaire à partir de la courbe de puissance journalière en transformant tout d'abord la puissance (W) en énergie (Wh) puis en additionnant toutes les valeurs d'énergie obtenues à chaque pas de temps de la journée pour obtenir le total cumulé de l'énergie de 00h00 jusqu'à 23h30 (voir la figure 4.50). Nous appelons l'approche de prévision de la courbe de puissance l'approche de puissance et l'approche de prévision de la série temporelle de l'énergie électrique consommée pendant la journée par l'approche de l'énergie.

Afin d'évaluer l'impact de cette approche sur la volatilité des données de consommation d'électricité, nous avons calculé les indices de volatilité définis dans la partie 4.2.2.1 de l'ensemble des ménages dans le jeu de données pour les courbes de puissance et les séries temporelles de l'énergie journalière consommée. Les distributions des indices de volatilité dans les deux cas sont présentées sur la figure 4.51. Cette figure montre que les séries temporelles de l'énergie journalière consommée sont moins volatiles que les courbes de puissance des ménages dans le jeu de données.

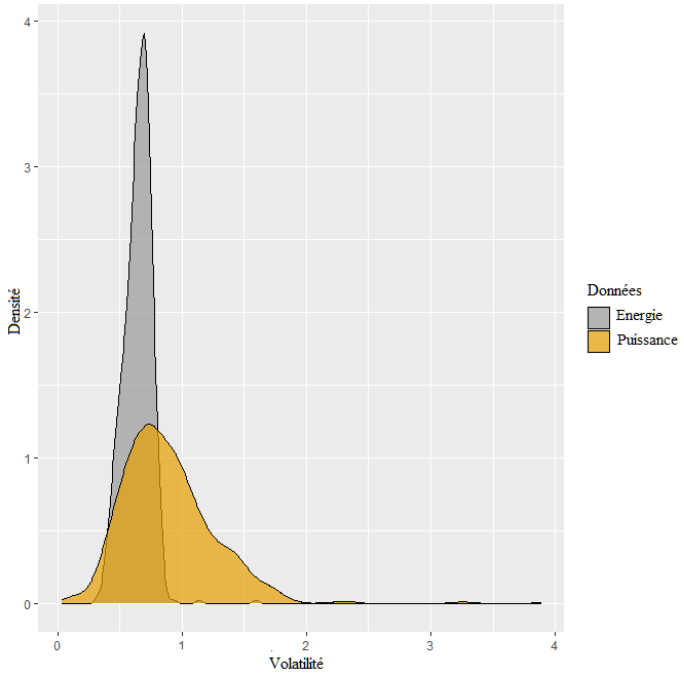
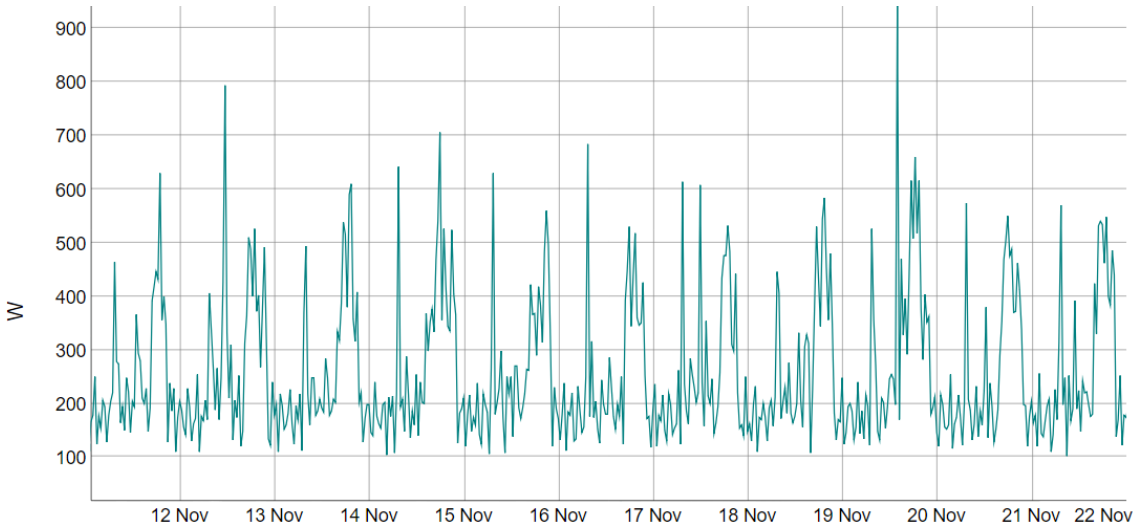
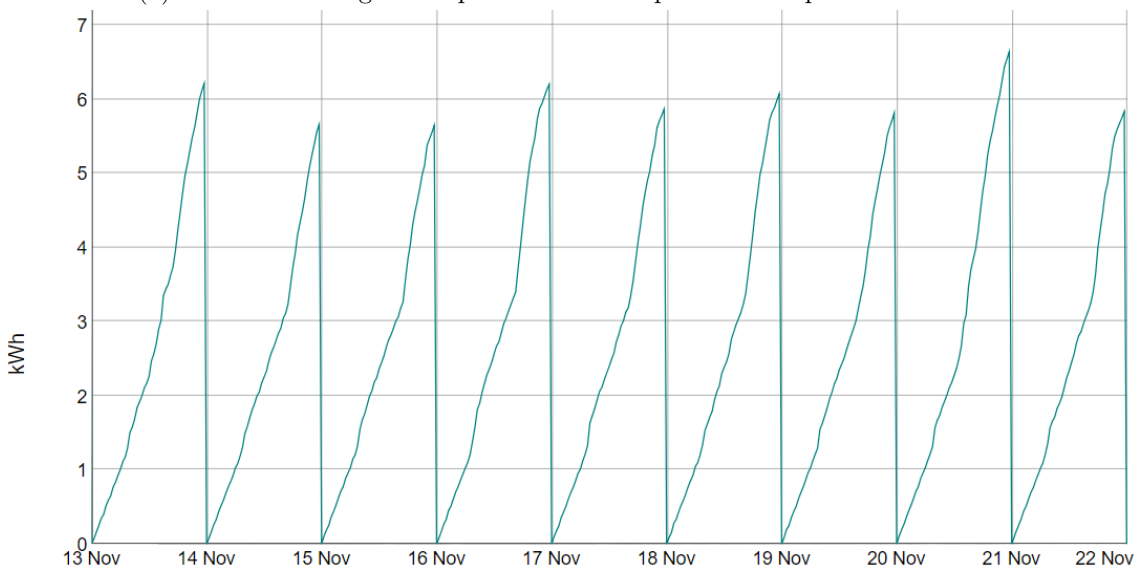


FIGURE 4.51 – Les distributions des indices de volatilité des courbes de charge électrique en puissance (en orange) et en énergie cumulée (en gris) de l'ensemble des ménages dans le jeu de données.



(a) Courbe de charge de la puissance électrique en W au pas demi-horaire.



(b) La série temporelle de l'énergie électrique journalière consommée en kWh au pas demi-horaire.

FIGURE 4.50 – Exemple de la série temporelle de l'énergie électrique consommée pendant une journée calculée au pas demi-horaire 4.50b à partir des courbes journalières de puissance 4.50a d'un ménage M_8 de la période allant de 13 à 22 novembre 2018.

En plus de l'analyse de l'impact de cette approche sur la volatilité des données, nous avons également examiné son impact sur la périodicité de ces dernières. Pour ce faire, nous avons prédit les courbes de puissance et les séries temporelles de l'énergie journalière consommée de tous les ménages dans le jeu de données par un modèle de persistance saisonnier. Les modèles de persistance saisonnier sont largement utilisés pour montrer l'existence d'une périodicité dans les séries temporelles. Ce modèle attribue à la prévision au temps t l'observation du même pas de temps du jour précédent. Les moyennes des erreurs NMAE et NRMSE de prévision à $(J+1)$ par ce modèle selon les deux approches sont présentées dans la table 4.14. Comme nous pouvions nous y attendre, les résultats montrent qu'en moyenne la prévision par le modèle de persistance des séries temporelles de l'énergie journalière consommée est plus précise que la prévision des courbes de puissance pour les deux erreurs NMAE et NRMSE pour l'ensemble de tous les ménages dans le jeu de données. Cela indique que les données de l'énergie journalière consommée sont plus périodiques (période égale à 48) que les données des courbes de charge de puissance. Les résultats de la prévision par le modèle de persistance sont présentés dans la table 4.14. Il est important de noter que dans cette partie, nous avons décidé de ne pas utiliser la métrique sMAPE pour comparer la précision des modèles de prévision des deux approches. Cela s'explique par le fait que les données de l'énergie électrique contiennent souvent des valeurs nulles ou très proches de zéro, ce qui rend l'erreur sMAPE très sensible et peut biaiser les résultats. Pour éviter cette confusion, nous avons opté pour les métriques NMAE et NRMSE qui sont moins sensibles aux valeurs nulles ou très proches de zéro. En utilisant ces métriques, nous pouvons obtenir une évaluation plus précise et juste de la performance des modèles de prévision, sans discriminer l'approche qui se base sur l'énergie par rapport à l'approche qui se base sur la puissance en raison de la nature des données.

Approche	**NMAE**	**NRMSE**
Puissance	0,53	1,05
Énergie	0,25	0,44

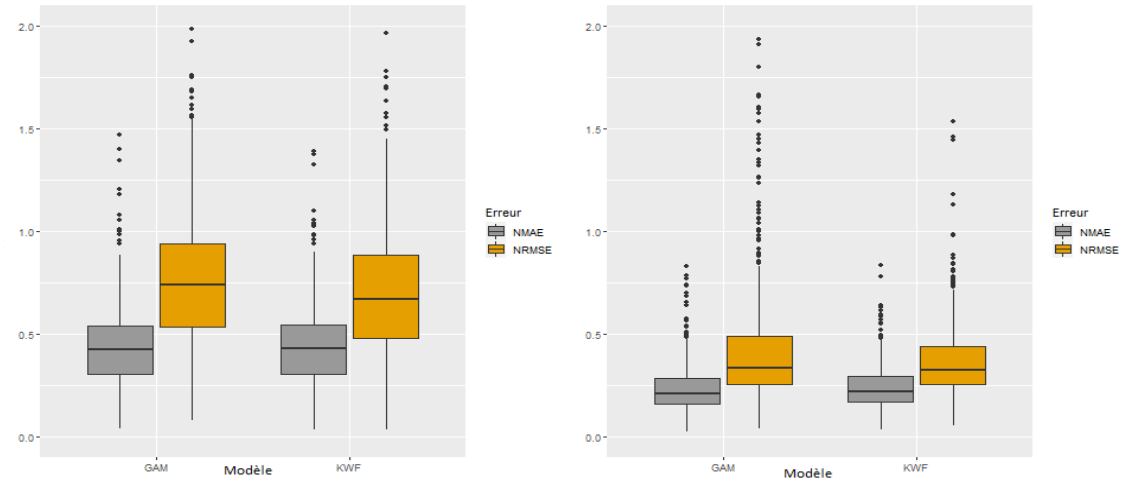
TABLE 4.14 – Les moyennes des erreurs NMAE et NRMSE du modèle de persistance saisonnier à $(J+1)$ pour les deux approches de puissance et d'énergie.

À ce stade, nous pouvons conclure que l'approche d'énergie a produit des résultats positifs en réduisant la volatilité des données et en rétablissant la périodicité. Nous nous attendons alors à ce que ces résultats aient un impact positif sur la précision de la prévision de la consommation d'électricité à l'échelle des ménages. Dans la suite de notre analyse, nous allons tester les deux modèles de prévision *KWF* et *GAM* définis dans les sections précédentes (voir la section 4.3 et la sous-section 4.4.1) sur les données de puissance et d'énergie journalière consommée pour tous les ménages du jeu de données afin de confirmer l'efficacité de cette approche. Les moyennes des erreurs NMAE et NRMSE de prévision de deux modèles sont présentées dans la table 4.15.

Approche	Modèle	NMAE	NRMSE
Puissance	*KWF*	0,44	0,70
	GAM	0,45	0,81
Énergie	*KWF*	0,24	0,36
	GAM	0,23	0,50

TABLE 4.15 – Les moyennes des erreurs NMAE et NRMSE par les modèles de prévision *KWF* et le modèle *GAM* à $(J+1)$ pour tous les ménages du jeu de données.

Les boîtes à moustaches de ces erreurs de prévision sont également présentées sur la figure 4.52 pour les deux approches. Les résultats présentés dans la table 4.15 montrent une amélioration significative des moyennes des erreurs de prévision NMAE et NRMSE pour les deux modèles de prévision *KWF* et *GAM* de l'énergie journalière consommée par rapport à la prévision de la puissance. Cette amélioration est également visible dans les distributions des erreurs de prévision NMAE et NRMSE présentées sur la figure 4.52 de ces deux modèles. En effet, en comparant les boîtes à moustaches des erreurs de prévision pour les deux approches (puissance et énergie), nous pouvons voir que les erreurs sont moins dispersées et ont des médianes inférieures lorsque les séries temporelles de l'énergie journalière consommée sont utilisées. Ces résultats confirment l'hypothèse selon laquelle l'utilisation des données d'énergie permet d'améliorer la précision des prévisions des données de consommation électrique des ménages puisqu'elles sont moins volatiles et plus périodiques.



(a) Approche de puissance.

(b) Approche d'énergie.

FIGURE 4.52 – Les boîtes à moustaches des erreurs NMAE et NRMSE par les modèles de prévision *GAM* et le modèle *KWF* à $(J+1)$ pour les deux approches de puissance et d'énergie.

En conclusion, notre approche permet de transformer les données de consommation électrique des ménages en données moins volatiles et plus périodiques. Cette approche peut

être considérée comme industriellement viable, car elle permet de fournir des prévisions fiables et interprétables de la consommation électrique à l'échelle des ménages. Ces prévisions peuvent aider les consommateurs à mieux gérer leur consommation d'électricité et à éviter une surconsommation coûteuse, tout en offrant aux fournisseurs de l'énergie une meilleure visibilité sur les besoins en électricité de leurs clients. En résumé, notre approche permet de transformer les données brutes de consommation de l'énergie des ménages recueillis par les compteurs intelligents en services de pilotage qui répondent aux besoins des consommateurs.

Les deux figures 4.53 et 4.54 montrent un exemple des courbes prévisionnelles de la consommation d'électricité d'un ménage M_9 dans le jeu de données par le modèle *KWF*. La figure 4.53 montre la prévision de la courbe de charge électrique de puissance en W alors que la figure 4.54 montre la prévision des données d'énergie électrique journalière consommée en kWh.

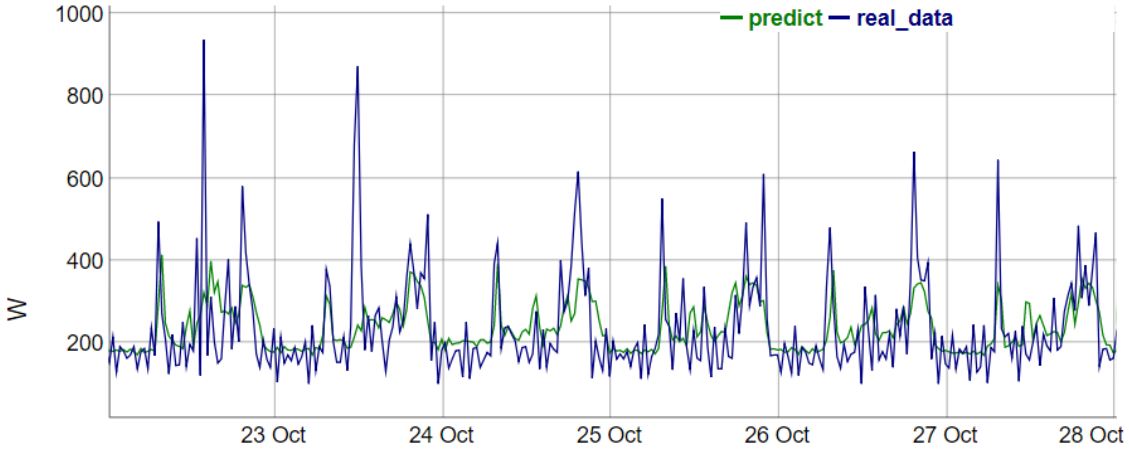


FIGURE 4.53 – Exemple de prévision à $(J + 1)$ par le modèle *KWF* d'une courbe de charge de puissance d'un ménage M_9 pour la période allant de 22 à 28 octobre 2022. La courbe en bleu représente la consommation réelle tandis que la courbe en vert représente les prévisions. Les erreurs de prévision NMAE et NRMSE des données de puissance de ce ménage sont respectivement 0,25 et 0,39.

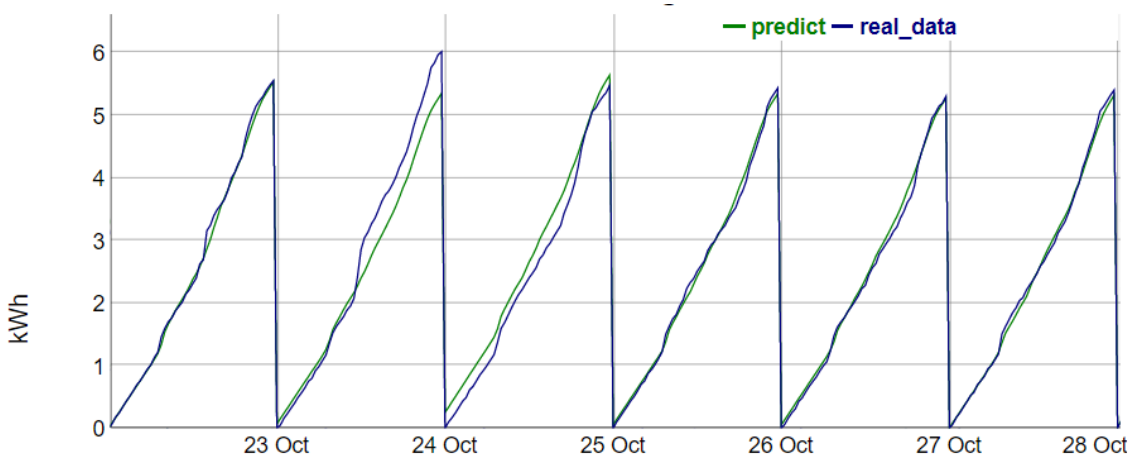


FIGURE 4.54 – Exemple de prévision à $(J+1)$ par le modèle KWF de la série temporelle de l'énergie journalière consommée du ménage M_9 pour la période allant de 22 à 28 octobre 2022. La courbe en bleu représente la consommation réelle tandis que la courbe en vert représente les prévisions. Les erreurs de prévision NMAE et NRMSE des données d'énergie journalière consommée de ce ménage sont respectivement 0,09 et 0,14.

4.9 Conclusion

Dans ce chapitre, nous avons abordé la problématique de la prévision de la charge électrique à l'échelle des ménages. Nous avons d'abord décrit les caractéristiques de la charge électrique des ménages, puis nous avons proposé six modèles de prévision de la charge électrique qui répondent aux différentes contraintes industrielles mentionnées dans le chapitre 3 et varient en termes de complexité. Chaque modèle proposé a des avantages qui le rendent intéressant pour la prévision de la charge électrique à l'échelle des ménages. Tous les modèles ont été testés sur un jeu de données privé de 720 ménages de profils de consommation disparates. La précision des modèles de prévision a été quantifiée par les quatre erreurs NMAE, NRMSE, MASE et sMAPE. Le premier modèle appelé KWF a été adapté et testé pour la prévision du lendemain au pas demi-horaire de la charge électrique des ménages. Nous avons comparé la performance de ce modèle à celle de deux modèles de référence : un modèle de forêt aléatoire, et un modèle climatologique. Les résultats montrent que le modèle KWF est plus performant pour la prévision de la charge électrique thermosensible et non-thermosensible. La précision de la prévision par le modèle KWF a été également étudiée en fonction de la thermosensibilité, la volatilité, la moyenne de consommation journalière ainsi que le type de tarification des ménages dans le jeu de données. Cette étude a montré que la précision des prévisions augmente avec l'augmentation de la moyenne de la consommation du ménage et son degré de thermosensibilité alors que cette précision diminue avec l'augmentation de l'indice de volatilité.

En revanche, ce modèle n'intègre aucune information sur les variables météorologiques. Pour cette raison et dans l'objectif d'améliorer la précision de la prévision des ménages

thermosensibles, nous avons intégré les résultats d'une méthode de *clustering* des données temporelles fonctionnelles dans le modèle *KWF*. Cette intégration permet de tenir compte de l'effet du calendrier et de l'impact de la température sur la consommation électrique. Cette méthode s'est révélée plus performante globalement pour la prévision de la charge électrique thermosensible. Les résultats montrent que le taux d'amélioration de la prévision augmente avec l'augmentation du degré de la thermosensibilité des ménages. La performance de cette méthode a été également testée sur les données agrégées des courbes de charge thermosensibles dans le jeu de données et a montré une précision supérieure à celle du modèle *KWF*.

D'autres modèles de régression non paramétrique comme le modèle *GAM* et *MARS* ont été également mis en œuvre et testés pour la prévision de la charge électrique des ménages. L'impact de l'intégration de la température extérieure sous différentes formes a été étudié. Les résultats ont montré que l'intégration des données décalées d'un jour de la température extérieure a améliorée la précision des prévisions des courbes de charge thermosensibles plus que les températures prédites ou lissées. L'interprétation des résultats de la prévision par ces deux modèles sont plus faciles à comprendre et à interpréter par rapport au modèle *KWF*. Un modèle *RNN-LSTM* est aussi mis en œuvre pour la prévision de la charge électrique des ménages dans l'objectif de rechercher des relations plus compliquées entre les variables exogènes et la charge électrique. Les résultats de prévision par tous les modèles sont comparés par rapport à la moyenne des trois erreurs NMAE, NRMSE et sMAPE ainsi que leurs distributions et analysés en fonction de la précision, l'interprétabilité et le temps de calcul. Les deux modèles *KWF* et *GAM* se sont révélés les plus performants globalement pour la prévision de la charge électrique par rapport aux autres modèles proposés.

Malgré la mise en place de différents modèles de prévision, l'amélioration de la précision des prévisions de certains ménages n'a pas été importante. Cette situation est due à la qualité limitée des données de consommation électrique de ces ménages, notamment leur irrégularité et leur volatilité. De plus, l'adoption de modèles plus sophistiqués ne garantit pas nécessairement une amélioration de la précision de la prévision des données si la prévisibilité de ces dernières n'est pas également améliorée. Dans ce contexte, une approche alternative pour améliorer la qualité des prévisions de consommation électrique à l'échelle des ménages a été proposée dans ce chapitre, à savoir la prévision de la quantité d'énergie électrique consommée pendant la journée plutôt que les puissances. Les résultats ont démontré que cette approche a amélioré la capacité à prédire des données de consommation électrique à l'échelle des ménages en réduisant leur volatilité et en les rendant plus périodiques. Par conséquent, cette approche permet d'améliorer la précision des prévisions même avec des données de consommation électrique volatiles, ce qui peut aider le fournisseur d'électricité à proposer des services adaptés à la consommation même pour les clients ayant des données de consommation moins régulières et plus difficiles à prévoir.

Chapitre 5

Application industrielle autour de la prévision de la charge électrique

Objectifs

L'objectif de ce chapitre est de répondre aux différents besoins du fournisseur d'énergie en termes de prévision de la charge électrique à différents niveaux d'agrégation, pour différents portefeuilles et différents types de données. Dans la première partie, nous nous concentrons sur la prévision de la charge électrique agrégée des ménages. Nous évaluons et comparons trois approches différentes de prévision de cette charge agrégée. Ensuite, nous adaptons et testons les modèles de prévision de la charge électrique à l'échelle des ménages présentés dans le chapitre 4 pour la prévision de la charge électrique des clients du secteur tertiaire à l'échelle individuelle. Nous calculons également des prévisions probabilistes qui permettent de tenir compte des incertitudes de la prévision de la charge électrique à l'échelle des clients dans le secteur résidentiel et tertiaire. Enfin, les deux dernières sections sont consacrées à la prévision des pertes et de la charge électrique du réseau de distribution.

Sommaire

5.1 Introduction

La prévision de la charge électrique est en effet un sujet vaste qui ne se limite pas à la prévision de la charge des ménages pour des applications de réponse à la demande (DRS) [1]. Chez un fournisseur d'électricité, la prévision de la charge électrique est cruciale pour différents types de charge électrique et de portefeuilles de clients, à différents niveaux d'agrégation et pour de nombreux enjeux, tels que la gestion du portefeuille de production et de stockage, la stratégie de vente et d'achat, la planification des investissements et des travaux sur le réseau de distribution, ...

Dans ce chapitre, nous explorons comment les modèles de prévision que nous avons mis en place pour la prévision de la charge électrique à l'échelle des ménages dans le chapitre 4 peuvent être utilisés pour répondre aux besoins du fournisseur d'énergie en termes de prévision. Plus précisément, nous avons étudié comment ces modèles peuvent être adaptés pour :

1. la prévision de la charge électrique agrégée de plusieurs ménages (voir la section 5.2).
2. la prévision de la charge électrique à l'échelle des clients dans le secteur tertiaire (voir la section 5.3).
3. la prévision des pertes électriques du réseau de distribution (voir la section 5.5).
4. la prévision de la charge électrique du réseau de distribution (voir la section 5.6).

Ces prévisions à différents niveaux et pour différents types de charge électrique permettent au fournisseur de prendre des décisions stratégiques et opérationnelles. Les enjeux et les avantages de chaque type de prévision seront détaillés dans les sections suivantes.

5.2 Prévision de la charge électrique agrégée des ménages

5.2.1 Objectif et intérêts

Il y a de nombreux avantages à prévoir la charge électrique agrégée des ménages. Tout d'abord, cela peut aider les fournisseurs d'énergie à planifier et gérer efficacement l'alimentation en électricité. En prédisant la charge électrique agrégée d'un ensemble de ménages, les fournisseurs d'électricité peuvent s'assurer qu'ils disposent de suffisamment de capacité de production pour répondre à la demande. Cela peut éviter les pannes de courant et les perturbations de l'alimentation en électricité.

Les fournisseurs d'électricité peuvent également identifier les horaires de pointe de la demande de ces ménages et prendre des mesures pour la réduire ou la déplacer. Ce qui a

1. https://en.wikipedia.org/wiki/Demand_response

pour effet de réduire le coût de la production d'électricité élevé pendant les périodes de pointe et d'offrir des tarifs plus compétitifs aux consommateurs.

Les tarifs adaptatifs qui reflètent mieux la demande de l'ensemble de ces ménages peuvent également être mis en place, ce qui peut encourager les consommateurs à réduire leur consommation pendant les heures de pointe, grâce aux services de réponse à la demande (DRS). Cela peut contribuer à une meilleure gestion des ressources et des moyens de production et à une réduction de la consommation d'énergie.

Il existe trois approches dans la littérature pour la prévision de la charge électrique agrégée (c'est-à-dire la consommation totale d'électricité d'un groupe de ménages ou de consommateurs) (Yildiz et al., 2017) : la première consiste à prévoir la charge électrique des ménages séparément, puis à additionner ces prévisions pour obtenir la prévision de la charge agrégée (voir la figure 5.1).

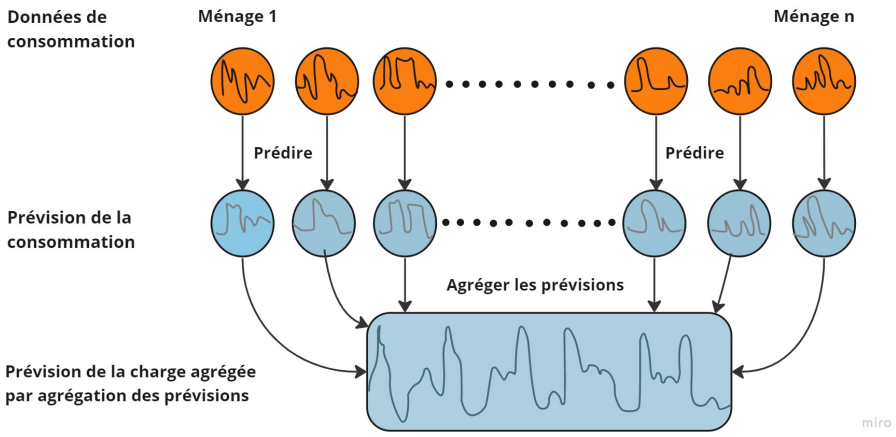


FIGURE 5.1 – La première approche de prévision de la charge électrique agrégée (approche 1).

La deuxième approche consiste à agréger en additionnant toutes les charges électriques des ménages, puis à prévoir la charge électrique agrégée (voir la figure 5.2). En effet, c'est l'approche la plus intuitive parmi les trois.

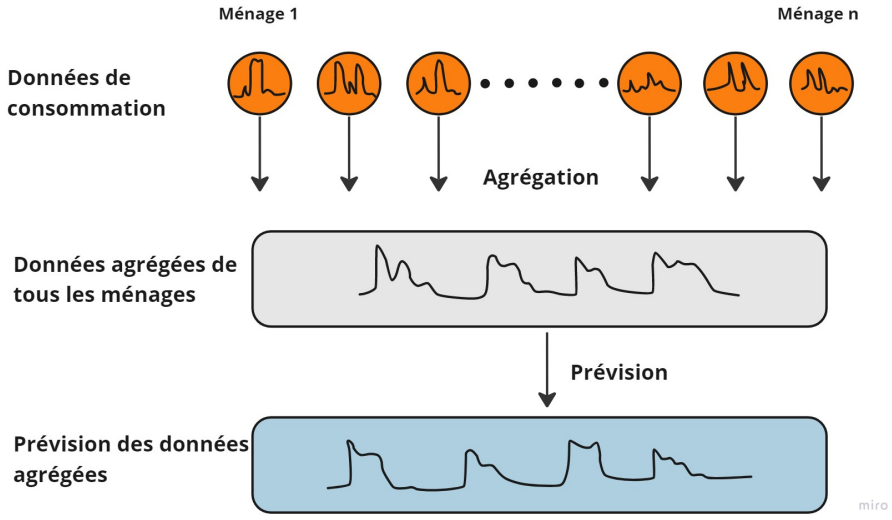


FIGURE 5.2 – La deuxième approche de prévision de la charge électrique agrégée (approche 2).

La troisième approche consiste à utiliser des méthodes de *clustering* en combinaison avec les méthodes de prévision afin d'améliorer la performance globale de la prévision. Cette dernière approche consiste à regrouper les ménages en fonction de la similitude de leurs habitudes de consommation, puis à additionner les charges électriques des ménages au sein de chaque *cluster* et à calculer une prévision pour chacun. Les prévisions obtenues pour chaque *cluster* sont ensuite à nouveau additionnées pour obtenir la prévision de charge agrégée finale (voir la figure 5.3). Cette approche a été suggérée par WIJAYA, SFRJ HUMEAU et al. (2014) et a été utilisée également dans des études comme celle de Franklin L QUILUMBA et al. (2014), qui ont proposé de regrouper les profils de la charge électrique en fonction de la moyenne de la consommation des ménages de chaque jour de la semaine pendant les périodes de pointe. Cette approche a été également utilisée dans SHAHZADEH et al. (2015) et les auteurs ont conclu que le *clustering* a amélioré la précision de la prévision de la charge agrégée pour toutes les méthodes de *clustering* utilisées. Les trois approches décrites précédemment sont référencées respectivement comme approche 1, approche 2 et approche 3 dans la suite de ce manuscrit.

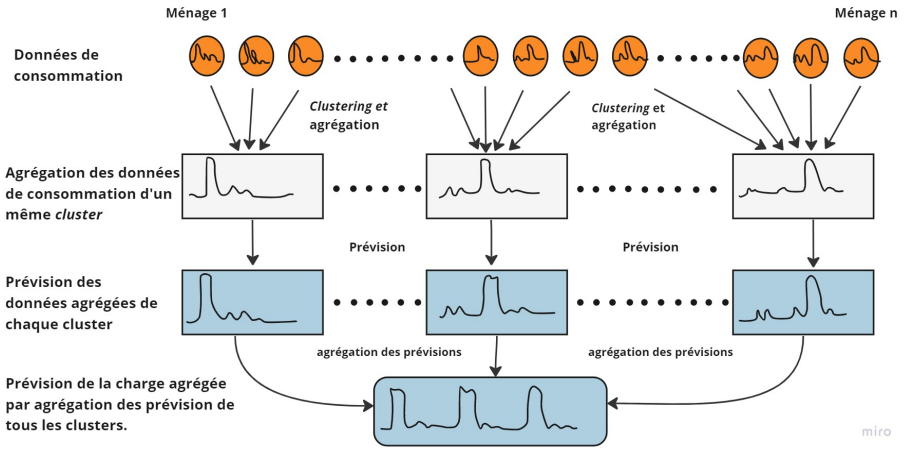


FIGURE 5.3 – La troisième approche de prévision de la charge électrique agrégée (approche 3).

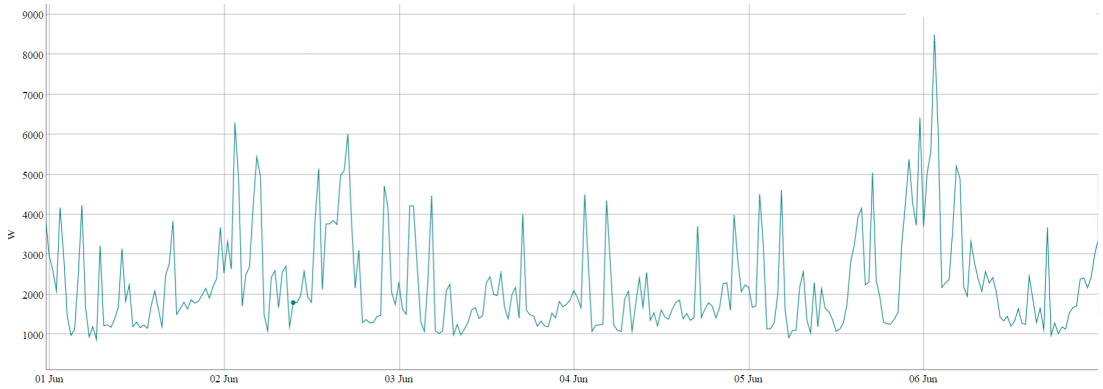
Dans les sous-sections suivantes, nous testons ces trois approches pour la prévision de la charge électrique agrégée des ménages dans notre jeu de données afin de déterminer laquelle est la plus appropriée. Nous comparons les performances de ces approches et choisissons celle qui se révèle la plus performante. Les deux modèles de prévision *KWF* et *GAM* qui se sont révélés les plus performants pour la prévision de la charge électrique à l'échelle des ménages sont adaptés et utilisés pour la prévision de la charge électrique agrégée dans les trois approches.

5.2.2 Effet de l'agrégation aléatoire

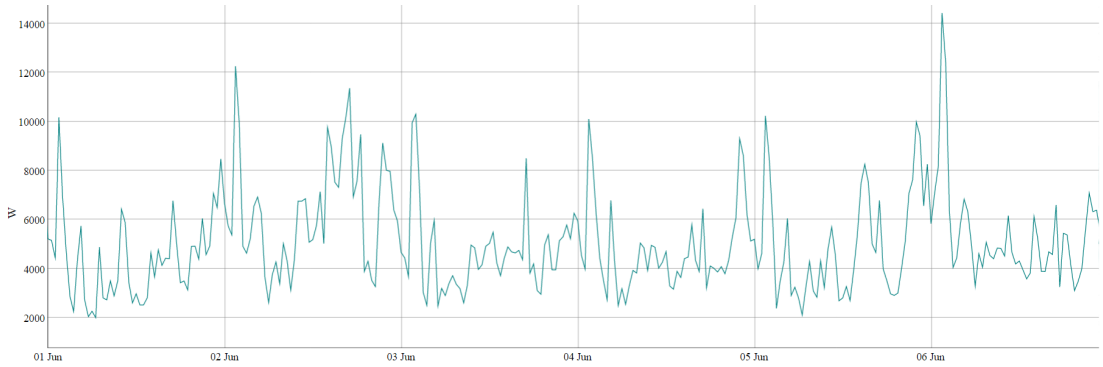
Il semble que les études comparant la prévision de la charge électrique à l'échelle des ménages et à différents niveaux d'agrégation sont encore anecdotiques et que les résultats obtenus ne permettent pas de trancher en faveur d'une approche plutôt qu'une autre. Par exemple, dans l'étude de KONG et al. (2017) l'approche 1 qui consiste à agréger les prévisions des ménages calculées individuellement est comparée à l'approche 2 qui consiste à prédire la charge électrique agrégée. Les résultats obtenus ont montré que l'approche d'agrégation des prévisions fournit des résultats de prévision plus précis. Une autre étude sur le même sujet dans PEÑALOZA et al. (2022) a montré que l'erreur de la prévision de la chargé agrégée est comparable à celle obtenue par agrégation des prévisions individuelles lorsque le nombre de consommateurs est inférieur à dix. Des comparaisons de la prévision de la charge électrique avec différents niveaux d'agrégation sont présentées dans SEVLIAN et al. (2014), mais les résultats ne sont pas comparés à l'approche de l'agrégation des prévisions des charges électriques des ménages individuellement.

Dans l'objectif de déterminer l'approche la plus appropriée pour la prévision de la charge électrique à différents niveaux d'agrégation dans le jeu de données, nous présentons

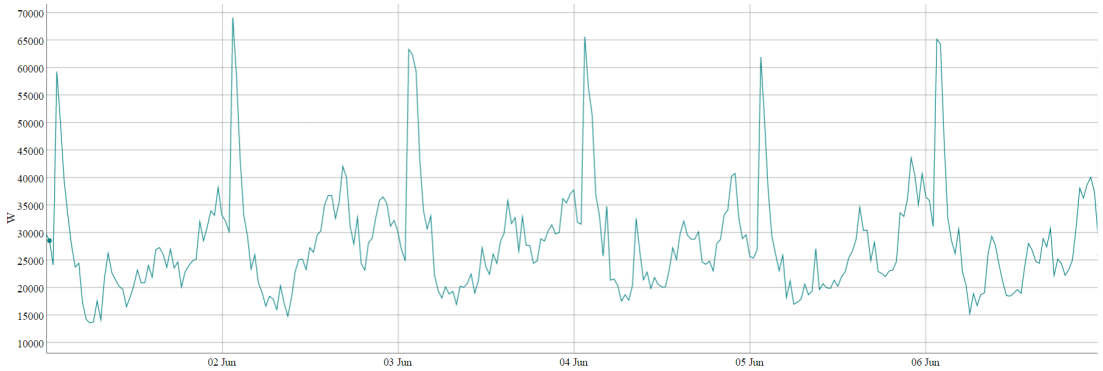
dans cette sous-section une comparaison de la prévision de la charge électrique à différents niveaux d'agrégation par les deux premières approches présentées dans la sous-section précédente. À partir de 720 ménages présents dans le jeu de données, nous avons sélectionné aléatoirement des ensembles de 5, 10, 50, 100, 200, 300, 400, 500 et 600 ménages. La figure 5.4 montre des exemples de données agrégées de la consommation électrique de plusieurs ménages. Nous remarquons que le caractère aléatoire domine toujours la charge électrique agrégée des ménages, même si la saisonnalité et la régularité de la charge sont plus marquées que pour les charges électriques individuelles.



(a) Données agrégées de 5 ménages.

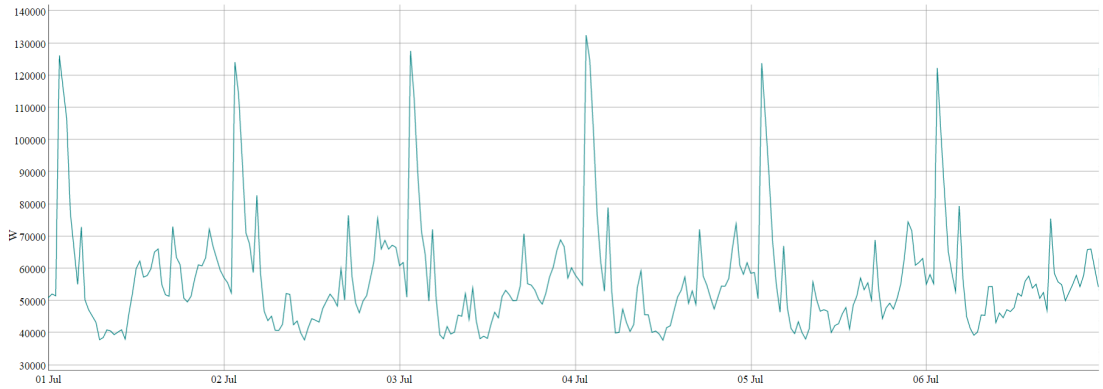


(b) Données agrégées de 10 ménages.

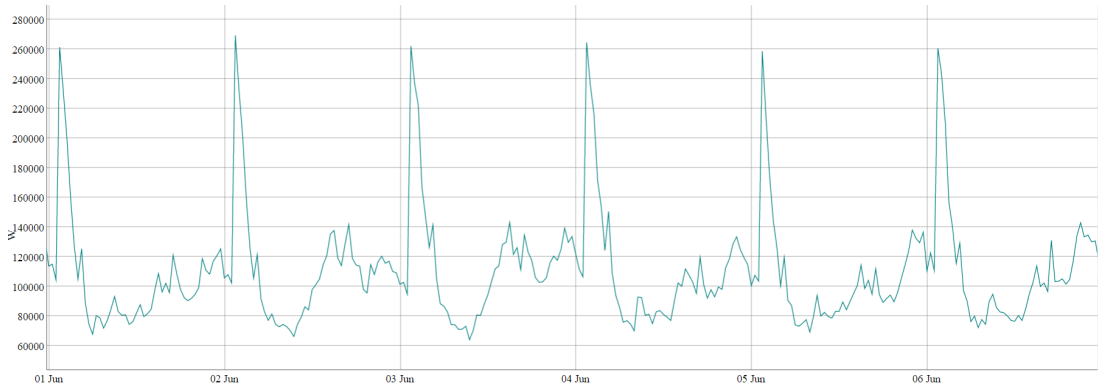


(c) Données agrégées de 50 ménages.

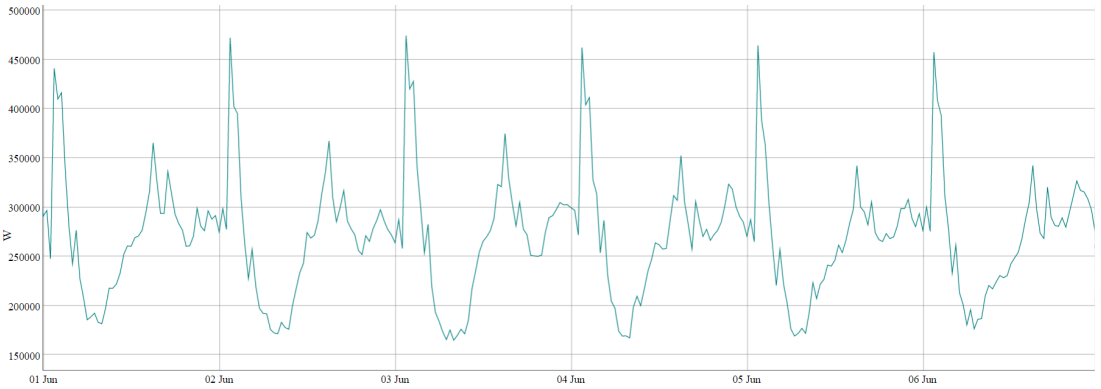
FIGURE 5.4 – Exemples de courbes de charge de la consommation électrique agrégée pour la période allant du 1 juin jusqu'au 6 juin 2018 en W.



(d) Données agrégées de 100 ménages.



(e) Données agrégées de 200 ménages.



(f) Données agrégées de 500 ménages.

FIGURE 5.4 – Exemples de courbes de charge de la consommation électrique agrégée pour la période allant du 1 juin jusqu'au 6 juin 2018 en W.

La prévision de la charge électrique a été effectuée par les deux modèles *KWF* et *GAM* qui ont été mis en œuvre pour la prévision de la charge électrique à l'échelle des ménages dans le chapitre 4 et qui se sont révélés les plus performants. Les données de chaque ménage sont séparées en 70% pour la période d'entraînement et 30% pour la période de test. Les prévisions de 24h au pas demi-horaire d'un ensemble de ménages ayant été calculées individuellement par les deux modèles de prévision et sur toute la période de test, elles sont agrégées. Elles sont ensuite comparées aux prévisions de la charge électrique agrégée

de tous les ménages dans le même ensemble de données obtenues par les mêmes modèles de prévision pour la même période de test. Cette expérience a été répétée 20 fois.

Les figures 5.5, 5.6 et 5.7 montrent les résultats de la prévision du jour $(J + 1)$ par les modèles *KWF* et *GAM* de la charge électrique des ménages à différents niveaux d'agrégation en se basant, pour chaque modèle de prévision, sur 20 expériences de prévision pour chaque ensemble de ménages. Les erreurs de prévision diminue avec l'augmentation des niveaux d'agrégation. Les moyennes des trois erreurs de prévision par les deux modèles sont beaucoup plus importantes lorsque le niveau d'agrégation est faible. Ceci s'explique par la haute volatilité et l'irrégularité des données de consommation électrique agrégées à des faibles niveaux. Les résultats montrent également que l'agrégation des données de la consommation électrique à partir de 50 ménages réduit considérablement les erreurs de la prévision de la charge agrégée pour les deux modèles testés. En effet, les données de consommation électrique agrégées au-delà de 50 ménages ont un caractère moins aléatoire, plus périodique et régulier que celles des ménages individuellement ou pour des niveaux d'agrégation plus faibles et par conséquent, elles sont plus prévisibles (voir les figures 5.4 et 5.4).

Les performances des deux modèles de prévision varient également en fonction du niveau d'agrégation. Pour un niveau d'agrégation inférieur à 50, le modèle *KWF* s'est avéré plus précis pour la prévision de la charge agrégée, tandis que pour un niveau d'agrégation supérieur à 50, c'est le modèle *GAM* qui a donné les meilleurs résultats. Cette différence de performance peut s'expliquer par le fait que lorsque le niveau d'agrégation augmente, la forme de la charge prévue pour le jour suivant devient de plus en plus similaire à celle du jour précédent, ce qui renforce la périodicité quotidienne à la base de la construction du modèle *GAM* et par conséquent, améliore sa performance.

De plus, les résultats montrent que l'approche 2 est globalement plus appropriée à la prévision de la charge agrégée que l'approche 1 pour le jeu de données et pour les différents niveaux d'agrégation testés. Ce résultat comme expliqué précédemment est dû au fait que la charge électrique agrégée est plus régulière et périodique que la charge électrique des ménages et par conséquent, la prévision de la charge agrégée est plus précise que la somme des prévisions de la charge électrique des ménages calculées individuellement. Plus précisément, pour un niveau d'agrégation inférieur à 50 ménages, le résultat des deux approches de prévision par le modèle *GAM* sont très comparables. Par contre, au-delà de 50 ménages, l'approche 2 est clairement plus performante. Concernant le modèle *KWF*, nous remarquons que pour tous les niveaux d'agrégation et pour les trois métriques, l'approche 2 est plus performante pour la prévision à $(J + 1)$ de la charge électrique agrégée.

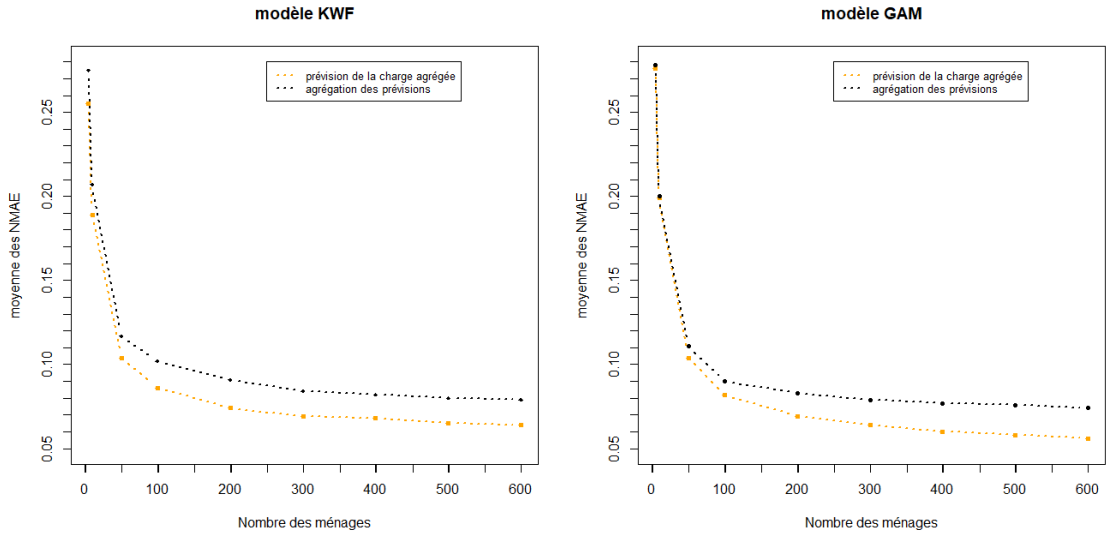


FIGURE 5.5 – Les moyennes des erreurs NMAE de prévision à $(J+1)$ par les deux modèles *KWF* et *GAM* en fonction du nombre des ménages dans l'ensemble de données et par les deux approches 1 et 2. La ligne en orange représente les résultats de la prévision par l'approche 2 (prévision de la charge agrégée). La ligne en noire représente les résultats de la prévision par l'approche 1 (agrégation des prévisions).

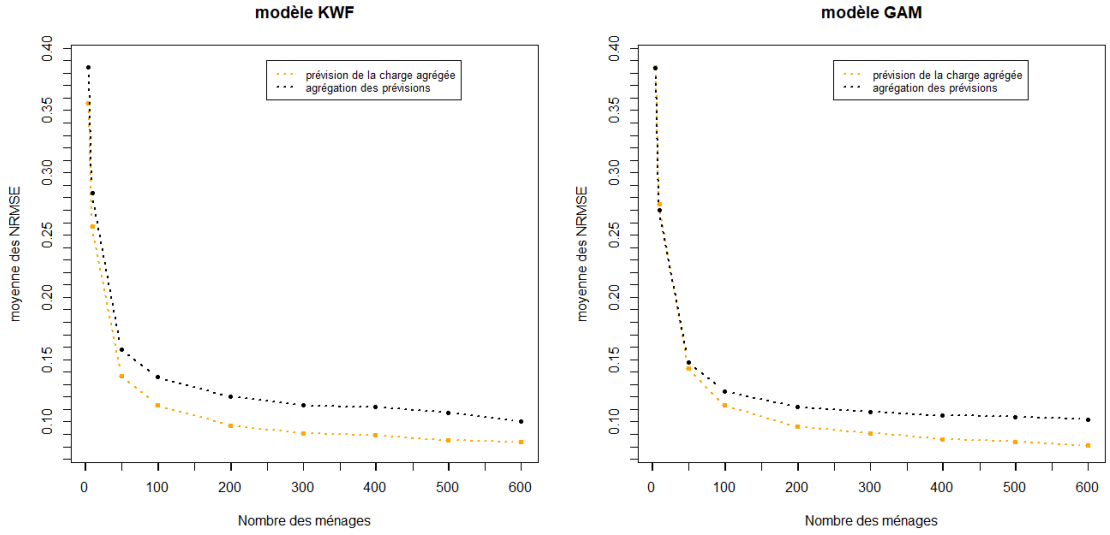


FIGURE 5.6 – Les moyennes des erreurs NRMSE de prévision à $(J+1)$ par les deux modèles *KWF* et *GAM* en fonction du nombre des ménages dans l'ensemble de données et par les deux approches 1 et 2. La ligne en orange représente les résultats de la prévision par l'approche 2 (prévision de la charge agrégée). La ligne en noire représente les résultats de la prévision par l'approche 1 (agrégation des prévisions).

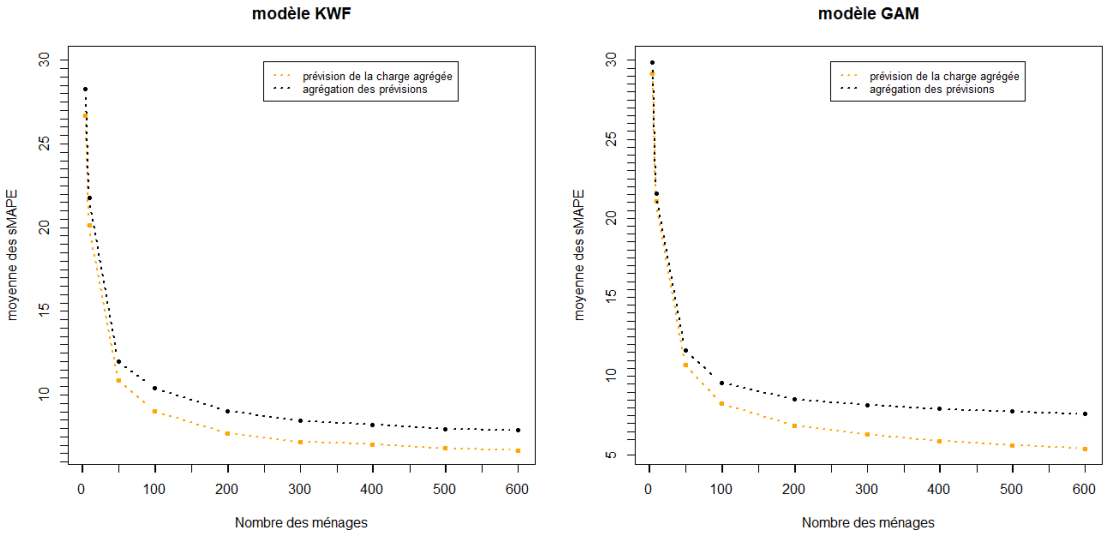


FIGURE 5.7 – Les moyennes des erreurs sMAPE de prévision à $(J+1)$ par les deux modèles *KWF* et *GAM* en fonction du nombre des ménages dans l'ensemble de données et par les deux approches 1 et 2. La ligne en orange représente les résultats de la prévision par l'approche 2 (prévision de la charge agrégée). La ligne en noire représente les résultats de la prévision par l'approche 1 (agrégation des prévisions).

En conclusion, l'agrégation des prévisions individuelles de la charge électrique à l'échelle des ménages n'a pas eu d'effet positif sur la précision de la prévision de la charge électrique agrégée pour les modèles et les données que nous avons examinés. Cependant, notre expérience a montré que l'agrégation de la charge électrique au-delà de 50 ménages permet d'améliorer considérablement la précision de la prévision de la charge électrique agrégée. Le modèle *KWF* est le plus performant pour la prévision de la charge agrégée à un niveau d'agrégation inférieur à 50 ménages, tandis que le modèle *GAM* est plus performant pour la prévision de la charge agrégée au-delà de 50 ménages.

Dans la sous-section suivante, nous explorons la possibilité d'améliorer la précision de la prévision de la charge électrique agrégée en regroupant les ménages ayant des caractéristiques de consommation électrique similaires en k *clusters*, pour ensuite prédire la charge électrique agrégée des ménages dans le même *cluster*. Le résultat de l'agrégation des k prévisions est comparé au résultat obtenu par l'approche 2. Nous supposons par cette approche que le *clustering* des courbes de charge selon des caractéristiques de consommation similaires permet d'obtenir des courbes de charge plus homogènes dans chaque *cluster* et par conséquent, des données agrégées plus lisses, plus périodiques, moins bruitées et plus faciles à prédire. En principe, cela devrait améliorer la précision de la prévision de la charge électrique agrégée de l'ensemble des données.

5.2.3 Effet du *Clustering*

Le *clustering* est une technique de traitement de données qui consiste à regrouper des données en groupes ou *clusters* ayant des caractéristiques similaires. Dans le contexte de l'analyse des données de compteurs intelligents de la consommation électrique de ménages (Wijaya, SFRJ Humeau et al., 2014 ; Haben, Singleton et al., 2015 ; Laurinec et al., 2017 ; Jeong et al., 2021), le *clustering* peut être utilisé pour mieux comprendre les différents profils de consommation électrique des ménages, tels que leur consommation à différents moments de la journée et leur efficacité énergétique globale (Haben, Singleton et al., 2015 ; Franklin L. Quilumba et al., 2015). Cette compréhension peut aider les fournisseurs d'électricité à identifier les ménages qui ont besoin de solutions d'économie d'énergie électrique, comme la réponse à la demande, et à mettre en place des programmes spécifiques pour répondre à leurs besoins, tels que des tarifs variables adaptés à leur consommation (Rhodes et al., 2014) ou des incitations pour encourager une réduction ou un déplacement de la consommation en dehors des périodes de pointe (Kwac et al., 2014).

Dans certains cas, le *clustering* des courbes de charge permet d'obtenir des profils de consommation qui peuvent aider les entreprises d'électricité à identifier lorsque certains groupes de clients connaissent des coupures de courant. En comparant le profil de consommation actuel d'un groupe de clients à leur profil habituel, les entreprises d'électricité peuvent détecter des écarts qui peuvent indiquer une coupure de courant. Cela peut les aider à réagir plus rapidement et de manière plus efficace pour rétablir le courant. Il peut également aider les entreprises d'électricité à mieux comprendre le fonctionnement du réseau électrique et à la façon dont il réagit face à différents modèles de demande. Cela peut être utile pour optimiser les opérations du réseau, par exemple en identifiant les meilleures façons de gérer la demande de pointe ou d'intégrer des sources d'énergie renouvelable.

Le *clustering* a également été utilisé pour la détection des anomalies dans les profils de consommation journalière d'électricité en comparant le profil de consommation réel à celui obtenu par *clustering* (Chicco et al., 2001). En plus de la détection des anomalies de la consommation d'électricité, les études (Abreu et al., 2012 ; Ozawa et al., 2016) ont comparé les profils de consommation réels de certains ménages avec des profils économes en énergie électrique de ces mêmes ménages ou de ménages voisins et ont utilisé ces comparaisons comme source d'information de retour pour aider les occupants de ces ménages à rationaliser leur consommation d'électricité et à réaliser des économies sur leurs factures d'électricité.

Finalement, le *clustering* peut contribuer à améliorer la prévision de la charge électrique des ménages à l'échelle individuelle comme dans le cas du modèle *KWF* avec *clustering* que nous avons présenté dans le chapitre 4 (Chaouch, 2013 ; Franklin L Quilumba

et al., 2014 ; Yildiz et al., 2017) (améliorer la précision des prévisions de charge intra-journalière d'un même consommateur) ou à des niveaux agrégés (Ilic et al., 2013 ; Viegas et al., 2015 ; Shahzadeh et al., 2015). En effet, en regroupant des profils de consommation similaires, les fournisseurs d'électricité peuvent mieux prévoir la demande en électricité de certains groupes de clients, ce qui peut améliorer la précision globale de leurs prévisions de la charge électrique.

5.2.3.1 *Clustering* des courbes de charge dans la littérature

Divers modèles sont utilisés pour le *clustering* des courbes de charge électrique, tels que le *clustering* hiérarchique (Abreu et al., 2012), les méthodes k-moyennes et k-médoïdes (*k-means* et *k-medoïds*) (Räsänen et al., 2009 ; Beniétez et al., 2014), les cartes auto-adaptatives (*self-organizing map (SOM)*) (Albert et al., 2013), l'analyse en composantes principales (Abreu et al., 2012), la modélisation en mélange fini (*finite mixture modelling* FMM) (Stephen et al., 2012 ; Haben, Singleton et al., 2015) et le partitionnement spectral (*spectral clustering*) (Abreu et al., 2012). Il n'y a pas de consensus dans la littérature quant à la meilleure approche de *clustering* à utiliser. En revanche, la méthode de *clustering* la plus largement utilisée est la méthode *k-means* en raison de sa faible complexité de calcul, sa convergence rapide, sa forte interprétabilité ainsi que son applicabilité à de grands ensembles de données.(Yildiz et al., 2017) ont fourni une revue détaillée sur les différentes méthodes de *clustering*. (T. Zhang et al., 2011) ont comparé plusieurs méthodes de *clustering* et ont conclu que la méthode de *k-means* était la méthode la plus cohérente.

Le choix de la forme des données a regroupé est important tout comme le choix de la méthode de *clustering*. En effet, en raison de la nature séquentielle des données de la consommation électrique nous pouvons distinguer deux approches principales, la première consiste à regrouper les données brutes de la charge électrique (Flath et al., 2012) et la deuxième consiste à regrouper des caractéristiques ou *features* spécifiques extraites de ces données (*feature-based clustering*) telles que la consommation annuelle d'électricité, les pointes de consommation et d'autres caractéristiques pour décrire les profils de consommation journalière (Räsänen et al., 2009 ; Haben, Singleton et al., 2015 ; Arco et al., 2017). (Räsänen et al., 2009) ont déterminé sept caractéristiques à partir de profils bruts hebdomadaires : la moyenne, l'écart-type, le coefficient d'asymétries (*skewness*), le coefficient d'aplatissement (*kurtosis*), le chaos, l'énergie et la périodicité, et ont utilisé la méthode *k-means* pour le *clustering*. (Haben, Singleton et al., 2015) ont analysé les profils de ménages en distinguant quatre périodes en fonction de l'heure de la journée (la nuit, le petit-déjeuner, la journée et le soir), le jour de la semaine/les week-ends et de la saison. Pour chacune de ces périodes, la puissance relative moyenne a été calculée et utilisée comme caractéristiques dans la méthode de *clustering*. Les auteurs dans (Arco et al., 2017) proposent une approche de *clustering* à deux niveaux qui utilise différentes

caractéristiques pour chaque niveau. Selon ces études, le *clustering* basé sur l'extraction des caractéristiques des données a entraîné une amélioration significative par rapport au *clustering* des données brutes.

En effet, l'utilisation des caractéristiques extraites des données permet de réduire la dimension des dernières, et de rendre le processus de *clustering* plus efficace et plus rapide. Elle permet également de réduire le bruit et la redondance dans les données, ce qui peut améliorer la qualité des *clusters* formés et les rendre plus interprétables. En même temps, l'utilisation d'un nombre réduit de caractéristiques dans la méthode de *clustering* peut entraîner une perte d'informations et affecter la qualité des *clusters* formés. Il est donc important de trouver un équilibre entre le nombre de *features* utilisées et la qualité des *clusters* obtenus.

L'extraction des *features* pour le *clustering* dépend souvent de l'application pour laquelle elle est utilisée (Räsänen et al., 2009). En effet, différentes méthodes d'extraction des *features* peuvent être plus ou moins adaptées en fonction de la nature des données d'origine et de l'objectif du *clustering*. Par exemple, certaines méthodes sont particulièrement adaptées aux données temporelles, tandis que d'autres sont mieux adaptées aux données textuelles ou aux données de type image. Il est donc important de choisir une méthode d'extraction de *features* qui convient à l'application en question afin d'obtenir des résultats satisfaisants (Liao, 2005).

Lors du processus de *clustering*, il est important également de définir une mesure de distance qui reflète de manière adéquate les similitudes ou les différences entre les données. Le choix de la mesure de distance dépend également de l'objectif du *clustering* et de la nature des données que ce soit pour trouver des similitudes dans le temps ou dans la forme des données. Il est donc important de sélectionner une mesure de distance appropriée afin de garantir la qualité et la fiabilité des résultats. En effet, il existe plusieurs mesures de distance couramment utilisées pour le *clustering* de séries temporelles. Voici quelques exemples :

1. la distance euclidienne : cette mesure de distance est calculée comme la racine carrée de la somme des carrés des différences entre chaque point de données des deux séries temporelles. Elle est souvent utilisée lorsque les données sont normalisées et que les dimensions ont une importance similaire (Chicco et al., 2001 ; Räsänen et al., 2009).
2. la distance de Manhattan : cette mesure de distance est calculée comme la somme des différences absolues entre chaque point de données des deux séries temporelles. Elle est souvent utilisée lorsque les dimensions ont des échelles différentes ou lorsque l'ordre des dimensions est important (Gang et al., 2022).
3. la distance de *Dynamic Time Warping* (DTW) : cette mesure de distance prend en compte la similitude temporelle entre deux séries temporelles en alignant les points

de données de chaque série temporelle de manière à minimiser la distance totale entre elles. Elle est souvent utilisée lorsque les séries temporelles ont des longueurs différentes ou lorsque la similitude temporelle est importante (Teeraratkul et al., 2017).

Dans l'approche des séries chronologiques brutes, le choix d'une distance appropriée influence significativement les résultats de *clustering*. Si les utilisateurs sont plus intéressés par la similarité entre les formes de profil de charge que par l'alignement temporel, une autre métrique de distance la métrique *Dynamic Time Warping* (DTW), est le meilleur choix (Teeraratkul et al., 2017). En effet, la distance DTW décale les profils de charge le long de l'axe du temps dans une certaine plage et trouve la distance minimale entre les profils, ce qui permet de regrouper des profils de formes similaires mais avec des décalages temporels, dans le même groupe (Yildiz et al., 2017). Dans l'approche basée sur le *clustering* des caractéristiques la distance euclidienne est souvent utilisée pour mesurer la similarité entre les caractéristiques des données surtout lorsque toutes les caractéristiques ont la même contribution dans la mesure de similarité.

5.2.3.2 Méthodologie proposée

Dans cette partie, nous évaluons la troisième approche présentée par la figure 5.3 dans l'objectif d'améliorer la précision de la prévision de la charge électrique agrégée de tous les ménages dans le jeu de données. Nous rappelons que cette approche consiste à regrouper les ménages dans notre jeu de données en k *clusters*, à agréger la consommation d'électricité de chaque ménage dans chaque *cluster* pour ensuite prévoir l'agrégation de la consommation électrique de chacun des k *clusters*. La prévision de la consommation électrique agrégée de tous les ménages est alors calculée en additionnant les prévisions des k *clusters*.

Nous nous attendons à ce que cette approche permettra de regrouper les ménages ayant des profils de consommation électrique similaires, ce qui nous permettra d'obtenir une courbe de charge électrique totale ou agrégée de chaque *cluster* moins bruyante, plus régulière et périodique. Cette homogénéité, en théorie, devrait améliorer la précision de la prévision de la charge électrique au sein de chaque *cluster* et, par conséquent, celle de la prévision de la charge électrique agrégée de tous les ménages.

Nous avons testé une méthode de *clustering* des données de consommation électrique des ménages qui s'inspire des travaux de Haben, Singleton et al. (2015). Dans cette étude, les auteurs présentent une méthode de *clustering* de données de consommation électrique de clients résidentiels basée sur un modèle de mélange fini (FMM) pour identifier les ménages qui pourraient être bénéfiques pour réduire la demande en électricité. En raison de la forte volatilité et de l'irrégularité de ces données, les auteurs ont utilisé une analyse

détaillée des données de compteurs intelligents à l'échelle des ménages pour déterminer les caractéristiques ou *features* qui ont permis de mieux comprendre les pointes de la demande de ces ménages et les principales sources de variabilité de leurs données de consommation.

Les résultats de leur analyse sur 2 700 clients ont montré que les pointes de la demande d'électricité de ces ménages se produisent généralement à des moments précis de la journée en fonction de la saison (effet de la température) et du type du jour (jour de la semaine ou week-end). Par exemple, pour le jeu de données utilisé, les demandes les plus faibles ont tendance à se produire en fin d'après-midi et en milieu de journée, tandis que les demandes les plus importantes ont tendance à se produire pendant la nuit et le matin. De plus, la demande d'électricité la plus importante de ces ménages est assez saisonnière et se produit à des moments spécifiques de la journée (en raison de l'utilisation du chauffage en hiver et de l'éclairage pendant les horaires de présence au domicile). Les effets du week-end sont également visibles sur les courbes de charge des ménages. Les auteurs décrivent que les pointes de la demande d'électricité sont décalées pendant les jours de week-end par rapport aux autres jours de la semaine en raison des soirées tardives et du réveil tardif des occupants.

D'après l'analyse mentionnée, les auteurs ont défini sept caractéristiques pour regrouper les données de consommation d'électricité des ménages en utilisant la technique de *clustering*. Quatre de ces caractéristiques permettent de regrouper les données des consommateurs en fonction de la répartition de leur consommation d'électricité au cours de quatre périodes prédéterminées de la journée qui correspondent à des périodes typiques d'activité des ménages. Deux autres caractéristiques ont été définies pour prendre en compte l'effet des jours de week-end et de la saisonnalité sur la consommation d'électricité. Enfin, la dernière caractéristique a été attribuée à la mesure générale de la variabilité et de l'irrégularité des données de consommation de chaque client. Nous avons fait le choix de diviser la journée dans notre cas en cinq périodes différentes de celles proposées dans l'article de Haben, Singleton et al. (2015). Ce choix prend en compte à la fois le mode de vie des Français et l'impact du cycle économique sur leurs habitudes de consommation. Les cinq périodes que nous avons choisies sont alors les suivantes :

1. **la période matinale** : c'est la période que nous avons fixée entre **5h et 9h** du matin. Elle correspond à la première partie de la journée, elle est caractérisée par une augmentation de la consommation d'électricité en raison de l'utilisation de différents appareils électriques pour les activités quotidiennes. Cela peut inclure les activités suivantes :
   (a) préparation du petit déjeuner : utilisation de la cafetière, du grille-pain, du four à micro-ondes, ...
   (b) préparation de la maison : allumage de l'éclairage, utilisation de l'aspirateur, du fer à repasser, ...
   (c) hygiène personnelle : utilisation du chauffe eau pour la douche, du sèche-

cheveux, ...

(d) déplacements : utilisation de la voiture car si elle est électrique, cela peut entraîner une augmentation de la consommation d'électricité ;

(e) travail et études : utilisation de l'ordinateur, de l'imprimante, ...

2. **la période diurne de la journée** : c'est la période entre **9h30 à 11h30** et de **14h30 à 17h30**. Elle correspond généralement aux heures de travail ou d'études pour de nombreux ménages. La consommation d'électricité durant cette période peut varier en fonction des activités menées par chaque ménage mais généralement elle est moins importante par rapport aux autres périodes de la journée en raison de l'absence des occupants du domicile. Pour les ménages occupés par des retraités ou des personnes en télétravail, la consommation d'électricité durant la période diurne peut ne pas suivre le profil décrit précédemment des ménages occupés par des personnes travaillant ou étudiant à l'extérieur du domicile. Ces ménages n'ont pas souvent les mêmes habitudes de consommation puisqu'ils ne sont pas soumis aux contraintes de l'activité professionnelle et de l'absence du domicile.
3. **la période du milieu de la journée** : c'est la période entre **12h et 14h**. La période entre 12h et 14h correspond généralement au milieu de la journée et peut être caractérisée par une variation de la consommation d'électricité en fonction des activités menées par chaque ménage. Dans certains ménages, la consommation d'électricité peut augmenter pendant cette période en raison de l'utilisation de l'électroménager pour la préparation du repas de midi comme le four, la plaque de cuisson, le lave-vaisselle, ... Dans d'autres ménages, la consommation d'électricité peut rester stable si les occupants ne sont pas présents à la maison ou s'ils ne font pas d'activités particulières nécessitant l'utilisation d'appareils électriques.
4. **la période de la fin de la journée** : c'est la période entre **18h à 22h**. Elle correspond souvent à la période de retour au domicile après le travail ou les études. La consommation d'électricité pendant cette période est généralement plus importante par rapport aux autres périodes de la journée en raison de la présence des occupants au domicile et de l'utilisation de différents appareils électriques comme le four, le micro-ondes, le lave-vaisselle, la télévision et les ordinateurs. Cette période peut également être caractérisée par l'utilisation de l'éclairage intérieur et extérieur, ainsi que par l'utilisation de l'appareil de chauffage/climatisation ou de ventilation.
5. **la période de la nuit** : c'est la période entre le **22h30 à 4h30**. Elle correspond souvent à la période du repos. La consommation d'électricité durant cette période peut être relativement faible par rapport à d'autres périodes de la journée en raison de l'absence d'activités nécessitant l'utilisation d'appareils électriques. Il est possible que la consommation d'électricité augmente légèrement en fonction de l'utilisation de certains appareils tels que l'appareil de chauffage/climatisation ou l'appareil de ventilation. Cependant, dans les ménages ayant des contrats d'électricité de tarif HC/HP, il est possible que la consommation d'électricité soit élevée pendant la nuit, si cette

période tombe pendant les heures creuses. Les clients dans ce cas là, programment automatiquement le déclenchement de leurs appareils électriques énergivores pendant cette période afin de réduire les coûts de leur consommation d'électricité.

Pour définir les caractéristiques utilisées dans la méthode de *clustering*, nous adoptons la notation utilisée dans (Haben, Singleton et al., 2015). Pour une courbe de charge électrique d'un ménage particulier, P_i est définie comme étant la puissance moyenne pour chaque période de la journée $i = 1, 2, 3, 4, 5$ calculée sur une année entière de données, avec l'écart type correspondant σ_i. $\hat{P}$ est la puissance moyenne journalière, P_i^E et P_i^H sont les moyennes de puissance pendant les saisons d'été et les saisons d'hiver respectivement pour chaque période i de la journée. Finalement, les puissances moyennes des jours de week-end et des jours de la semaine pour chaque période de la journée i sont définies respectivement par P_i^{WE} et P_i^{WD}. En utilisant la notation ci-dessus, les caractéristiques suivantes pour chaque courbe de charge sont calculées de la façon suivante :

1. **la puissance moyenne relative** pour chaque période de la journée i calculée sur une année entière de données

$$P_i^R = \frac{P_i}{\hat{P}}, \quad i = 1, 2, 3, 4, 5,$$

2. **l'écart-type relatif moyen** sur une année entière de données

$$\hat{\sigma} = \frac{1}{5} \sum_{i=1}^{5} \frac{\sigma_i}{P_i},$$

3. **un score saisonnier** défini par

$$S = \sum_{i=1}^{5} \frac{|P_i^H - P_i^E|}{P_i},$$

4. **un score de différence entre le week-end et le jour de la semaine** défini par

$$W = \sum_{i=1}^{5} \frac{|P_i^{WE} - P_i^{WD}|}{P_i}.$$

Pour chaque ménage dans le jeu de données, nous avons calculé les huit caractéristiques décrites ci-dessus pour une année entière de données. Concernant le choix d'un modèle de *clustering*, nous avons opté pour l'algorithme *k-means* plutôt que le modèle *FMM* décrit dans (Haben, Singleton et al., 2015). En effet, l'algorithme *k-means* offre une grande facilité d'interprétation à l'inverse, du modèle *FMM* qui n'est pas aussi facilement interprétable car ses paramètres sont latents et les relations entre les caractéristiques sont représentées dans un espace à haute dimension. De plus, le modèle *FMM* est principalement utilisé dans les cas où les caractéristiques représentent des interactions complexes entre elles, ce qui n'est pas le cas pour les huit caractéristiques calculées à partir des don-

nées. Dans l'algorithme *k-means*, le nombre de *clusters* doit être prédéfini. Généralement, le modèle est appliqué à un nombre différent k de *clusters*. Ensuite, la meilleure solution parmi elles est sélectionnée à l'aide d'un indice de validité comme l'indice de silhouette (Rousseeuw, 1987), l'indice de Davies-Bouldin (Xiao et al., 2017) (Rousseeuw, 1987) ou l'indice de Calinski-Harabasz (X. Wang et al., 2019). Étant donné que notre objectif est d'améliorer la prévision de la charge en regroupant les clients en fonction de leurs profils de consommation, nous avons décidé d'évaluer le nombre de *clusters* en fonction de la performance de la prévision. Par conséquent, plutôt que d'utiliser un indice de sélection de la valeur optimale du nombre de *clusters* k, nous avons fait le choix de sélectionner la valeur de k qui permet d'améliorer la prévision de la charge en regroupant les données de consommation électrique des ménages dans le même *cluster*. Le nombre de *clusters* k sélectionné est alors celui qui minimise l'erreur de prévision.

Notre méthodologie de *clustering* et de prévision se résume alors par les étapes suivantes.

1. Calculer les huit caractéristiques décrites ci-dessus pour chaque ménage présent dans le jeu de données décrit dans la sous-section 4.2.1.
2. Appliquer l'algorithme de *clustering k-means* aux caractéristiques calculées dans l'étape précédente, pour k =1, 3, 5, 10, 20, 30 et 40 avec 1000 répétitions pour des initialisations différentes dans l'objectif de surmonter le problème des minima locaux (Jain, 2010).
3. Agréger les données de consommation électrique de tous les ménages dans chaque *cluster*.
4. Prédire à $(J+1)$ les k séries chronologiques de consommation électrique agrégées de chaque *cluster* par les deux modèles de prévision *KWF* et *GAM* proposés dans le chapitre 4.
5. Agréger les k prévisions obtenues dans l'étape précédente et comparer le résultat aux données agrégées de tous les ménages dans le jeu de données (cas où $k = 1$). Lorsque $k = 1$, tous les ménages sont regroupés dans un seul *cluster* et la prévision est effectuée pour les données de consommation électrique agrégées de tous les ménages dans le jeu de données. Ceci correspond à la deuxième approche (approche 2) de prévision présentée sur la figure 5.2.

Les résultats de la prévision de la charge agrégée par les deux modèles *KWF* et *GAM* à $(J+1)$ obtenus pour chaque valeur de k sont présentés dans la partie suivante.

5.2.3.3 Résultats

La table 5.1 montre les moyennes de chacune de huit caractéristiques de chaque *cluster* obtenu par la méthode de *clustering* proposée pour $k = 3$.

Caractéristiques								
cluster	P_1^R	P_2^R	P_3^R	P_4^R	P_5^R	$\hat{\sigma}$	S	W
1	0,95	0,88	0,94	0,95	1,19	1,11	5,33	0,57
2	0,78	0,99	1,13	1,21	0,94	0,85	0,86	0,79
3	0,86	0,90	0,99	1,08	1,11	1,00	3,16	0,76

TABLE 5.1 – Les moyennes de chacune des huit caractéristiques pour chaque *cluster*. Cas où le nombre de *cluster* est égal à 3 ($k = 3$).

Comme le montrent les résultats dans la table 5.1, les *clusters* obtenus peuvent être distingués par les valeurs moyennes de puissance relative durant la période du milieu de la journée, la fin de la journée et la nuit. Les valeurs moyennes de puissance relative durant la période matinale et diurne semblent moins contribuantes dans la détermination des *clusters*. La figure 5.8 montre le profil moyen journalier normalisé de la charge électrique des trois *clusters*. Le profil moyen du premier *cluster* présente une moyenne de consommation électrique importante au milieu et à la fin de la journée avec trois pointes aux alentours de 12h, 20h et minuit. Le troisième *cluster* a un profil moyen similaire au profil moyen du premier *cluster*.

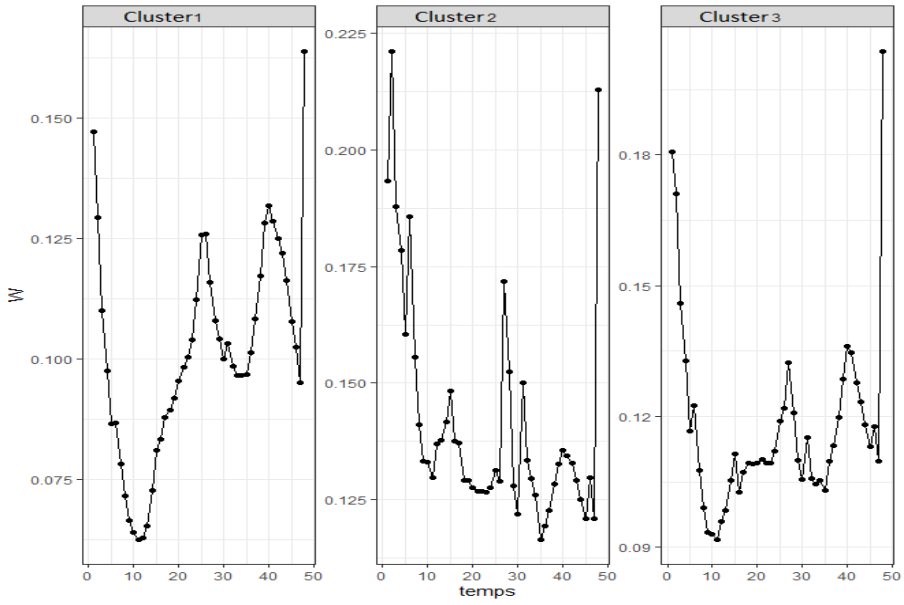


FIGURE 5.8 – Les profils moyens journaliers normalisés de la charge électrique de chaque *cluster* (cas où $k = 3$).

Cependant, ces deux *clusters* sont distinguables l'un de l'autre par leurs scores de saisonnalité (voir la figure 5.9).

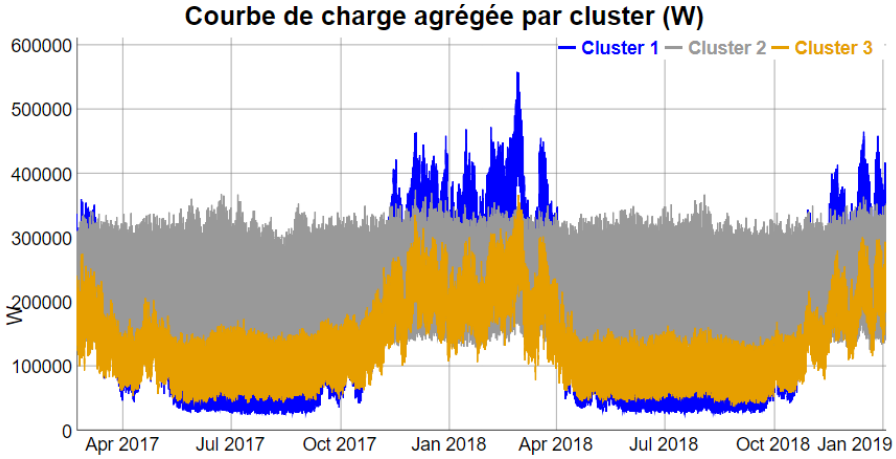


FIGURE 5.9 – Les courbes de charge électrique agrégées par *cluster* (cas où $k = 3$) de la période allant du 1 avril 2017 jusqu'au 30 janvier 2019.

La figure 5.9 présente les courbes de charge électrique agrégées par *cluster*. Nous observons une saisonnalité annuelle dans les courbes de charge agrégées des *clusters* 1 et 3, caractérisée par une augmentation de la moyenne de la consommation en hiver, contrairement à la courbe de charge agrégée du *cluster* 2, qui présente une moyenne de consommation presque constante tout au long de l'année. Ce résultat permet d'identifier deux groupes de ménages thermosensibles ayant des profils de consommation distincts. Les *clusters* 1 et 3 regroupent également les ménages dont les données de consommation électrique présentent une grande variabilité, caractérisée par un écart-type relatif moyen élevé par rapport aux ménages du *cluster* 2. Les *clusters* 2 et 3 regroupent les ménages ayant une différence dans la consommation électrique entre les jours de la semaine et les week-ends.

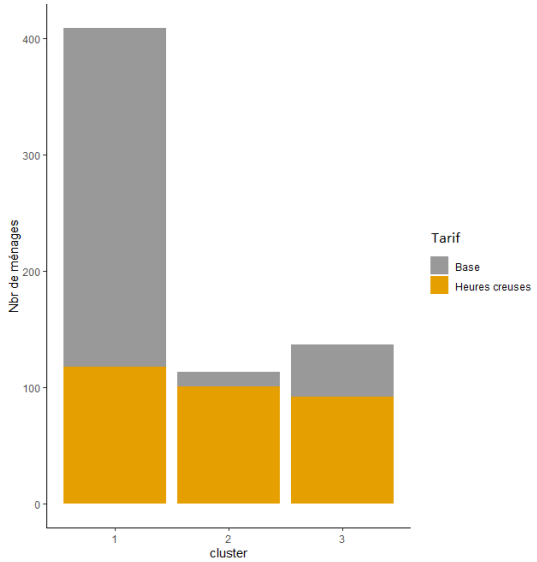


FIGURE 5.10 – Nombre de ménages par *cluster* en fonction de l'option tarifaire. Cas où le nombre de *cluster* est égal à 3 ($k = 3$).

L'analyse des résultats de la méthode de *clustering* pour $k = 3$ en fonction de l'option tarifaire des ménages a permis de constater que le *cluster* 1 est principalement composé de ménages ayant une option tarifaire de base (voir la figure 5.10). Les ménages ayant opté pour l'option tarifaire HC/HP sont quant à eux répartis de manière équitable dans les trois *clusters*.

Dans ce qui suit nous étudions l'application de la méthode de *clustering* proposée sur notre jeu de données décrit dans la sous-section 4.2.1 dans l'objectif d'améliorer la précision de la prévision de la charge agrégée. Comme expliqué dans la section précédente, nous avons utilisé la méthode de *clustering* pour regrouper les ménages ayant des caractéristiques de consommation électrique similaires. Une fois les *clusters* sont obtenus, les données de consommation électrique des ménages appartenant au même *cluster* sont agrégées (voir la figure 5.11).

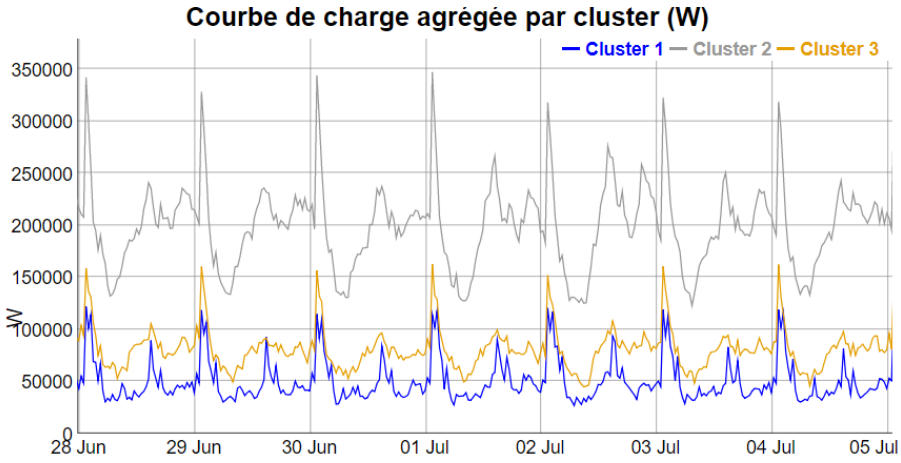
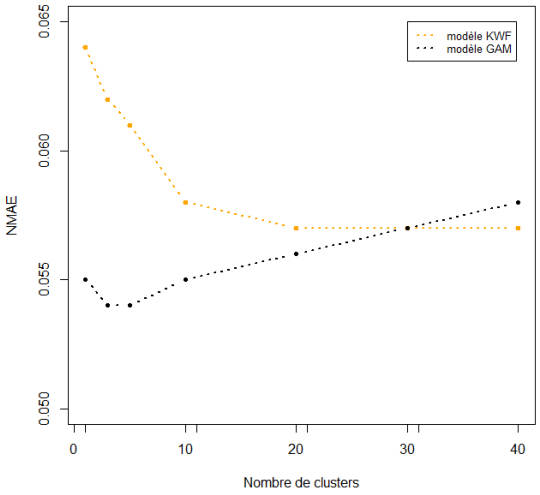


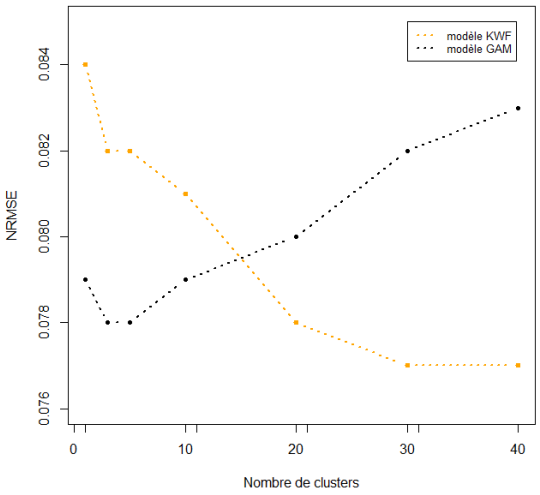
FIGURE 5.11 – Les courbes de charge électrique agrégées par *cluster* (cas où $k = 3$) d'une semaine du 28 juin au 4 juillet 2018.

Pour toute valeur de k dans l'ensemble $\{1, 3, 5, 10, 20, 30, 40\}$ les k courbes de charge obtenues de chaque *cluster* sont prédites par les modèles *KWF* et *GAM*. Les données sont divisées en 70% pour l'entraînement des modèles et 30% pour le test. Les prévisions obtenues sont ensuite agrégées et les résultats sont présentés sur la figure 5.12.

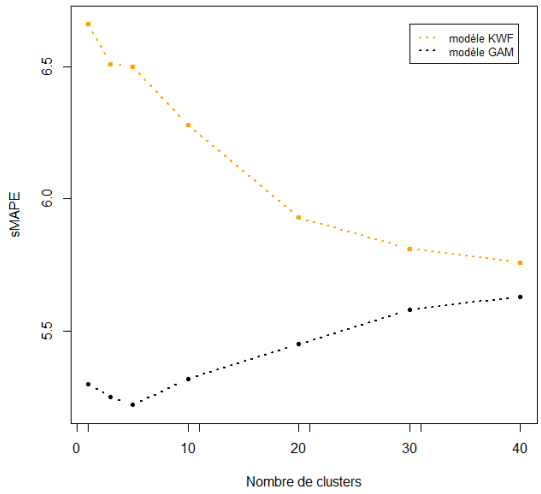
La figure 5.12 montrent les erreurs de prévision NRMSE, NMAE et sMAPE des modèles *KWF* et *GAM* pour différentes valeurs du nombre de *clusters* k. Nous rappelons que lorsque $k = 1$, tous les ménages sont regroupés dans un seul *cluster* et une seule prévision est effectuée par chaque modèle.



(a) NMAE



(b) NRMSE



(c) sMAPE

FIGURE 5.12 – Les erreurs NRMSE, NMAE et sMAPE de la prévision de la charge électrique agrégée par les deux modèles *KWF* et *GAM* à $(J+1)$ et pour différents nombres k de *clusters*.

Les résultats indiquent que les performances des deux modèles varient en fonction du nombre de *clusters* k. Pour $k = 1$ (c'est-à-dire en absence de *clustering*), le modèle *GAM* fournit une prévision de la charge électrique agrégée plus précise que le modèle *KWF* pour les trois métriques d'erreur NMAE, NRMSE et sMAPE. Pour $k = 3$, les trois erreurs de prévision du modèle *GAM* atteignent leurs valeurs minimales, indiquant ainsi une amélioration de la précision de la prévision de la charge agrégée après l'utilisation de la méthode de *clustering*. Pour $k \geqslant 5$ les trois erreurs de prévision du modèle *GAM* augmentent jusqu'à ce que le *clustering* ne soit plus bénéfique, puisque les erreurs deviennent

plus grandes que celles de la prévision de la charge agrégée sans *clustering* (cas où $k = 1$). En revanche, pour le modèle *KWF*, les trois métriques diminuent considérablement avec l'augmentation de k, atteignant une certaine stabilité à partir de $k = 30$. Par conséquent, nous pouvons conclure que la méthode de *clustering* a permis aux deux modèles, *KWF* et *GAM*, d'améliorer la précision de la prévision de la charge électrique pour $k = 30$ et $k = 3$, respectivement, par rapport à l'absence de *clustering*.

Bien que la méthode de *clustering* soit plus avantageuse pour le modèle *KWF*, le modèle *GAM* obtient des prévisions plus précises pour $k = 3$ en termes d'erreurs NMAE et sMAPE. En revanche, le modèle *KWF* pour $k = 30$ fournit la prévision la plus précise en termes d'erreur NRMSE. La différence significative entre le nombre optimal de *clusters* k pour le modèle *KWF* et le modèle *GAM* peut s'expliquer par les résultats obtenus dans la partie précédente. En effet, comme nous l'avons montré précédemment, le modèle *KWF* est plus performant pour la prévision de la charge électrique à des faibles niveaux d'agrégation, tandis que le modèle *GAM* est plus performant pour des niveaux d'agrégation supérieurs. Par conséquent, lorsque le nombre de *clusters* augmente, le nombre de ménages dans chaque *cluster* diminue, ce qui va à l'encontre de la capacité du modèle *GAM* à fournir des prévisions précises, tandis que le modèle *KWF* est plus à l'aise avec des *clusters* plus petits et peut donc continuer à améliorer ses prévisions.

Il est en effet intéressant de souligner que la taille de l'échantillon de clients peut avoir un impact significatif sur l'efficacité du *clustering* pour prévoir la charge électrique agrégée. Cela est dû au fait que la taille de l'échantillon peut influencer la représentativité des données dans chaque *cluster* et la précision des résultats de *clustering*. En générale, une taille d'échantillon plus importante peut entraîner une meilleure représentativité des données dans chaque *cluster*, ce qui peut permettre d'obtenir des *clusters* plus homogènes avec une quantité plus importante de données, et ainsi des données agrégées plus régulières. Cela peut conduire à une meilleure prévision de la charge électrique agrégée. Par conséquent, il serait intéressant d'étudier l'amélioration en fonction non seulement du nombre de *clusters*, mais également de la taille de l'échantillon de clients, afin d'optimiser la précision de la prévision de la charge électrique agrégée. Cette analyse pourrait aider à déterminer la taille optimale de l'échantillon pour obtenir les résultats les plus précis possibles, tout en limitant la complexité et le temps de calcul associés au processus de *clustering*.

5.3 Prévision des courbes de charge dans le secteur tertiaire

5.3.1 Objectif et intérêts

Les clients d'électricité dans le secteur tertiaire sont des entreprises et des organisations qui offrent des services plutôt que des biens matériels comme les hôpitaux, les écoles, les banques, les restaurants, les hôtels, les centres commerciaux, les bureaux, ... Ces clients ont des besoins électriques variables en fonction de leurs activités et de leurs équipements électriques.

La prévision de la charge électrique à l'échelle des clients dans le secteur tertiaire revêt une importance croissante dans le secteur de l'électricité. En effet cette pratique est essentielle pour les deux parties prenantes majeures, à savoir les clients et les fournisseurs d'électricité. Les clients du secteur tertiaire accordent une grande importance à la prévision de leur charge électrique car elle leur permet de gérer leur consommation d'électricité de manière proactive. En se basant sur ces prévisions, ils peuvent adapter leur consommation en fonction des heures de pointe ou des tarifs variables. Ainsi, ils évitent les surcharges et les pénalités financières, ce qui réduit considérablement leurs dépenses énergétiques. Les prévisions de charge électrique peuvent aider les clients tertiaires à surveiller le bon fonctionnement de leurs équipements électriques et à détecter rapidement les pannes ou les défaillances, ce qui permet une maintenance préventive et une réduction des temps d'arrêt. La gestion de la consommation électrique peut être un enjeu crucial pour les clients tertiaires qui cherchent à atteindre des objectifs environnementaux. Dans ce contexte, les prévisions de charge électrique peuvent être très utiles pour aider les clients à planifier leur consommation d'énergie en fonction des sources d'énergie renouvelable disponibles ou pour éviter la consommation électrique lors des périodes de pointe. En effet, ces prévisions leur permettent de prendre des décisions éclairées et de réduire leur empreinte environnementale [2], ce qui est de plus en plus important pour les entreprises et les organisations qui cherchent à améliorer leur image de marque et leur réputation.

La prévision de la charge électrique à l'échelle des clients dans le secteur tertiaire offre également de nombreux avantages pour les fournisseurs d'électricité. Tout d'abord, elle permet d'optimiser la production et la distribution de l'électricité en anticipant les fluctuations de la demande et en évitant les pics de consommation qui pourraient entraîner des perturbations sur le réseau électrique. En outre, la prévision de la charge électrique permet aux fournisseurs d'électricité de proposer des tarifs adaptés aux besoins de chaque client, de personnaliser leurs offres et ainsi de fidéliser leur clientèle. Elle permet également à l'entreprise de proposer des services qui encouragent les clients à l'efficacité énergétique, ce qui peut se traduire par une réduction de la consommation d'énergie et des coûts associés, ainsi qu'une réduction de l'impact environnemental de l'entreprise.

2. impact de l'activité humaine sur l'environnement.

Contrairement à la prévision de la charge électrique à l'échelle des ménages, la prévision de la charge électrique dans le secteur tertiaire peut être plus facile et plus précise. Cela est dû au fait que les données dans le secteur tertiaire sont souvent plus régulières et périodiques, avec des horaires d'ouverture et de fermeture qui restent généralement constants. De plus, les activités économiques des entreprises et des organisations ont tendance à être moins volatiles que celles des ménages, ce qui rend la charge électrique plus prévisible. Toutefois, la complexité réside dans la variété des activités et des équipements électriques des clients, ainsi que les variations saisonnières qui peuvent rendre la prévision de la charge électrique plus complexe. Il convient également de prendre en compte l'évolution de la consommation électrique au fil du temps pour chaque entreprise, qui peut être influencée par divers facteurs tels que les changements dans les activités commerciales, l'introduction de nouveaux équipements ou technologies économes en énergie, ou même les changements dans les politiques environnementales. Ces facteurs peuvent rendre la prévision de la charge électrique plus complexe.

Notre objectif dans cette section est d'adapter et de tester les modèles que nous avons mis en œuvre pour la prévision de la charge électrique à l'échelle des ménages afin de les appliquer à la prévision de la charge à l'échelle des clients dans le secteur tertiaire à $(J+1)$. Actuellement, chez le fournisseur, la méthode utilisée pour la prévision de la charge électrique des clients tertiaires est très simpliste, se limitant à une simple copier-coller en fonction des jours de la semaine et des jours fériés. Nous cherchons donc à améliorer cette méthode en utilisant des modèles plus sophistiqués et adaptés à la nature de la charge électrique des clients du secteur tertiaire.

5.3.2 Description des courbes de charge tertiaires

Les données utilisées dans cette étude sont des données privées et anonymes sur la consommation d'électricité de 290 clients tertiaires du fournisseur d'énergie en France. Ces données comprennent des mesures de la consommation d'électricité (en kW) toutes les demi-heures pour la période allant de janvier 2017 à octobre 2019. Cependant, aucune information supplémentaire sur la nature ou le type d'entreprise de ces clients n'est disponible dans le jeu de données.

Les caractéristiques des courbes de charge électrique à l'échelle individuelle des clients tertiaires peuvent varier considérablement en fonction de plusieurs facteurs, notamment le type d'entreprise, les heures d'opération, les jours de la semaine, les saisons, les périodes de vacances, ...

Cependant, en général, les courbes de charge électrique des clients tertiaires peuvent être décrites comme suit :

1. **La succession des pics et des creux** : les courbes de charge des entreprises

tertiaires sont généralement caractérisées par une succession de pics et de creux. Les pics de charge importants apparaissent pendant les heures de travail, telles que le matin ou le début de l'après-midi, lorsque les équipements électriques sont en utilisation, tandis que les creux se produisent pendant les heures de fermeture ou les week-ends, lorsque les équipements électriques sont éteints ou en mode veille (voir la figure 5.13d). Certaines entreprises peuvent avoir un usage électrique stable tout au long de la journée, sans pics ni creux importants. Cela peut être le cas pour les entreprises qui n'ont pas de cycles de travail définis, telles que les bureaux administratifs, les centres de données, les bibliothèques ou les centres d'appels.

2. **La saisonnalité** : elle se caractérise par la variation de la consommation d'électricité selon les saisons, les jours de la semaine et les heures de la journée, et est influencée par divers facteurs tels que la température, les habitudes de consommation, les activités économiques et les comportements des clients.

   Au niveau annuel, la consommation d'électricité des clients tertiaires peut varier en fonction des températures extérieures dues à la climatisation ou au chauffage. Elle peut varier également en fonction des événements saisonniers, des fêtes et des vacances (voir la figure 5.13c). Par exemple, les centres commerciaux peuvent connaître une augmentation de leur consommation d'électricité pendant les périodes de soldes, de Noël ou de la rentrée scolaire.

   Au niveau hebdomadaire, la consommation d'électricité peut varier en fonction des jours de la semaine. Les entreprises tertiaires peuvent avoir des schémas de consommation électrique différents les jours de la semaine, selon leur type d'activité. Par exemple, les restaurants peuvent connaître une augmentation de la consommation électrique le week-end, alors que les bureaux peuvent avoir une consommation électrique plus importante en semaine (voir la figure 5.13a).

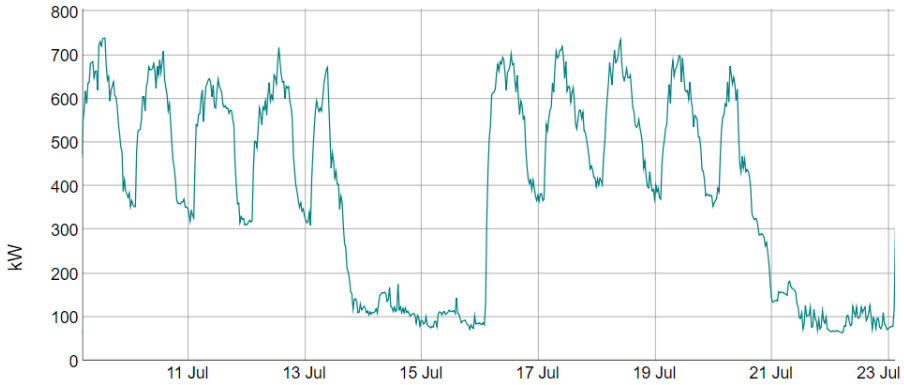
   Au niveau journalier, la consommation électrique peut varier en fonction des habitudes de consommation des clients tertiaires au cours de la journée. Par exemple, la consommation d'électricité peut être plus élevée pendant les heures de travail que pendant les pauses ou après le travail.

   Certaines entreprises tertiaires peuvent avoir des courbes de charge électrique qui dépendent de la saison. Par exemple, les entreprises du secteur du tourisme peuvent avoir une consommation plus élevée en été et une consommation plus faible en hiver.

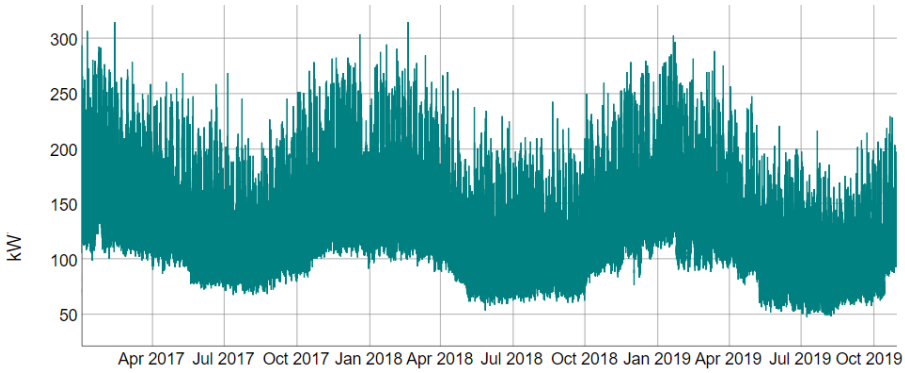
3. **La thermosensibilité** : les courbes de charge électrique des entreprises tertiaires sont sensibles aux conditions thermiques, ce qui signifie que la consommation électrique de ces entreprises dépend de la température extérieure. Cette sensibilité thermique est souvent observée dans les entreprises ayant une forte consommation en climatisation et chauffage, comme les hôtels, les centres commerciaux, ... (voir la figure 5.13b)

4. **Le cycle économique** : la consommation électrique est influencée par l'état de l'économie. Pendant les périodes de croissance économique, la consommation électrique

peut augmenter en raison de l'accroissement des activités économiques et commerciales. Au contraire, pendant les périodes de récession économique, la consommation électrique peut diminuer en raison de la réduction de l'activité économique. Les entreprises tertiaires peuvent fermer temporairement ou réduire leur activité, ce qui entraîne une baisse de la consommation d'électricité. Ces facteurs ont un rôle essentiel dans l'évolution des tendances dans les courbes de charge tertiaire. Par exemple, pendant la pandémie de Covid-19, les mesures de confinement et la fermeture de nombreux commerces et entreprises ont entraîné une baisse de la consommation d'électricité des clients tertiaires, tels que les locaux professionnels.



(a) Courbe de charge électrique d'un client tertiaire présentant une dichotomie entre les jours de la semaine et le week-end.



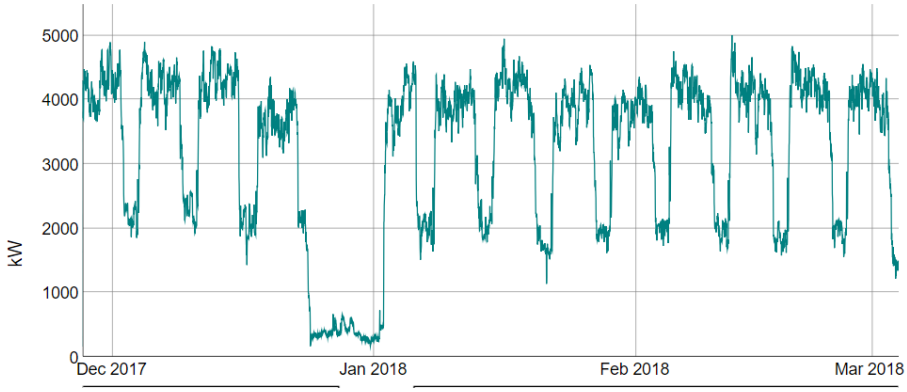
(b) Courbe de charge électrique d'un client tertiaire montrant une saisonnalité annuelle liée à la thermosensibilité.

5.3.3 Modèles et approche de prévision

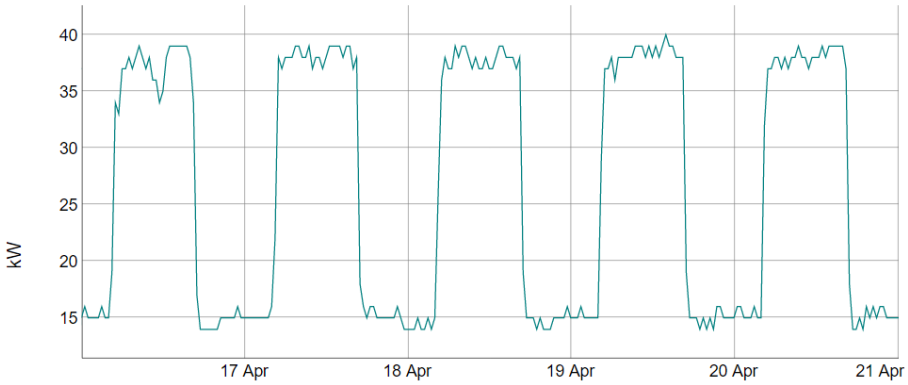
Nous avons exploré la possibilité d'utiliser les modèles que nous avons mis en œuvre pour la prévision de la charge électrique à l'échelle des ménages afin de prédire la charge électrique des clients tertiaires à $(J+1)$. Notre objectif était de trouver comment nous pouvions exploiter ces modèles pour améliorer la méthode actuellement utilisée chez le fournisseur, qui consiste en une simple copie-coller en fonction des jours de la semaine. En adaptant nos modèles et en les appliquant aux données tertiaires, nous cherchions à proposer une méthode plus sophistiquée et mieux adaptée à la nature et les caractéristiques

décrites dans la sous-section 5.3.2 précédente de la charge électrique des clients du secteur tertiaire.

Nous avons choisi d'utiliser les modèles *KWF*, *GAM* et *MARS* pour la prévision de la charge électrique à l'échelle des clients tertiaires, en raison de leur efficacité prouvée dans la prévision de la charge électrique des ménages de notre jeu de données en termes de précision, d'interprétabilité et de temps de calcul. Les courbes de charge tertiaires sont également réparties en deux ensembles, avec 70% des données utilisées pour l'entraînement et 30% pour le test.



(c) Courbe de charge électrique d'un client tertiaire révélant l'impact spécifique du calendrier, en particulier la semaine de Noël.



(d) Courbe de charge électrique d'un client tertiaire illustrant l'alternance entre des périodes de pics et de creux de consommation.

FIGURE 5.13 – Extrait des courbes de charge électrique des clients dans le secteur tertiaire.

Pour le modèle *KWF*, les groupes déterministes définit dans la partie 4.3.1.1 ont été modifiés pour devenir : les lundis, les mardis-mercredis-jeudis, les vendredis, les samedis, les dimanches, les jours fériés et les jours précédant les jours fériés. La décomposition des jours en différents groupes dans le modèle *KWF* est basée sur les caractéristiques de la consommation électrique des clients dans le secteur tertiaire notamment sa variation en fonction des jours de la semaine et des jours fériés. En effet, la charge électrique des clients du secteur tertiaire peut avoir des caractéristiques différentes de celle des ménages en

raison des différences dans les habitudes de consommation, les équipements électriques et les horaires d'activité. Par conséquent, en adaptant les groupes déterministes du modèle *KWF* aux données tertiaires, nous avons pris en compte ces différences et nous avons identifié les groupes de jours qui ont des caractéristiques similaires en termes de charge électrique des clients tertiaires. Cette décomposition des jours a été choisie de manière à refléter les particularités de la charge électrique des clients du secteur tertiaire et à permettre une meilleure modélisation de la variabilité de la charge électrique.

Il est courant de regrouper les mardis, mercredis et jeudis ensemble en raison de leur similitude dans les habitudes de consommation (Cugliari, 2011 ; Goude et al., 2013 ; Gaillard et al., 2016). Ces jours sont généralement considérés comme des jours de travail réguliers, où les activités économiques sont plus stables et prévisibles, ce qui se traduit par des profils de consommation électrique similaires dans le secteur tertiaire. Ainsi, en regroupant les mardis, mercredis et jeudis ensemble, nous capturons mieux cette similarité de consommation que si nous les séparions en groupes distincts. Cette approche permet également de réduire le nombre de groupes déterministes, ce qui peut faciliter la mise en œuvre du modèle et peut améliorer sa capacité de prévision.

En plus, les lundis sont souvent considérés comme des jours de forte consommation électrique dans le secteur tertiaire, car les entreprises et les institutions reprennent leur activité après le week-end et peut être ont besoin d'une quantité d'énergie électrique plus importante pour fonctionner. De même, le vendredi peut également être un jour de consommation énergétique plus élevée que les autres jours de la semaine, car certaines entreprises peuvent prolonger leurs horaires de travail jusqu'au soir ou organiser des événements après le travail. C'est pourquoi il peut être judicieux de considérer le lundi et le vendredi comme des jours différents des autres jours de la semaine.

Les samedis peuvent être différents des dimanches en termes de consommation d'électricité dans le secteur tertiaire en fonction du type d'activités exercées. En effet, certains entreprises dans le secteur tertiaire travaillent les samedis mais pas les dimanches, ou vice versa. Par exemple, les magasins ont souvent une plus grande affluence de clients les samedis, tandis que les bureaux sont fermés les deux jours.

En incluant le groupe des jours précédant les jours fériés dans le modèle *KWF*, cela permet d'indiquer que les jours qui suivent sont des jours fériés. Cette information est importante pour la prévision de la charge électrique des clients du secteur tertiaire, car la consommation électrique peut varier considérablement pendant les jours fériés en raison de la fermeture de certains établissements ou de la diminution du nombre de personnes présentes sur les lieux de travail.

Le modèle *KWF* avec *clustering* décrit dans la sous-section 4.3.5 est utilisé pour la prévision des courbes de charge thermosensibles. Les mêmes structures de modèles *GAM* et *MARS* décrites dans les sous-sections 4.4.1 et 4.4.2 sont utilisées. Les variables d'entrée

comprennent la charge décalée d'un jour, la charge décalée d'une semaine, le décalage d'un jour de la température extérieure (pour les clients thermosensibles), les jours de la semaine et les jours fériés. Comme dans le modèle *KWF*, les trois jours de la semaine, à savoir mardi, mercredi et jeudi, sont considérés comme un seul jour dans ces modèles. Les modèles sont comparés entre eux ainsi qu'à la méthode utilisée chez le fournisseur, et les résultats sont présentés dans la sous-section suivante.

5.3.4 Résultats

La table 5.2 présente les résultats de l'évaluation de performance moyenne de différents modèles de prévision de la consommation électricité à l'aide de quatre métriques différentes, à savoir NMAE, NRMSE, MASE et sMAPE. Les modèles évalués comprennent *KWF*, *GAM*, *MARS* et la méthode du fournisseur décrite dans la sous-section précédente. Les meilleurs résultats pour chaque métrique sont surlignés en orange.

En examinant les résultats, nous pouvons voir que le modèle *GAM* est celui qui obtient les meilleurs résultats en termes de NMAE, NRMSE, sMAPE et MASE, ce qui indique qu'il est le modèle le plus précis pour la prévision de la consommation d'électricité à $(J+1)$ pour notre jeu de données. Le modèle *KWF* obtient également de bons résultats, mais il est légèrement moins précis que le modèle *GAM*. Le modèle *MARS*, quant à lui, est le moins précis des trois modèles de prévision, obtenant des résultats plus élevés pour les trois métriques à l'exception de l'erreur sMAPE.

En ce qui concerne la méthode utilisée actuellement chez le fournisseur, les résultats sont significativement moins précis que les modèles de prévision que nous avons proposés.

Si ces modèles sont adoptés à la place de la méthode utilisée actuellement chez le fournisseur, il est possible d'obtenir des gains significatifs en termes de précision de prévision. Par exemple, si le modèle *GAM* est adopté cela leur permettrait d'obtenir des améliorations significatives de **18,78%** en termes de NMAE, **28,76%** en termes de NRMSE, **7,80%** en termes de sMAPE et **20,87%** en termes de MASE sur notre jeu de données. Il pourrait donc être intéressant pour le fournisseur d'énergie de considérer ces modèles comme des alternatives potentielles à leur méthode actuelle.

Modèle	NMAE	NRMSE	MASE	sMAPE
KWF	0,242	0,377	0,698	29,59
GAM	0,238	0,374	0,694	27,50
MARS	0,246	0,392	0,728	29,54
méthode fournisseur	0,293	0,525	0,876	29,83

TABLE 5.2 – Performance moyenne des modèles *KWF*, *GAM* et *MARS* à $(J+1)$ ainsi que la méthode utilisée actuellement chez le fournisseur d'énergie selon les métriques NMAE, NRMSE, MASE et sMAPE. Meilleurs résultats en orange.

En complément des performances moyennes, nous avons également tracé les boîtes à moustaches pour chaque modèle de prévision afin de visualiser la distribution des erreurs (voir la figure 5.14). Les boîtes à moustaches des erreurs de prévision des trois modèles, *KWF*, *GAM* et *MARS*, révèlent des niveaux de médianes distincts, ce qui a été confirmé par les tests statistiques de *Freidmann* et *Wilcoxon*.

Les métriques du modèle *GAM* ont présenté les médianes les plus basses parmi les trois modèles. Les boîtes à moustaches des erreurs NMAE et NRMSE du modèle *GAM* étaient moins étalées que celles des modèles *KWF* et *MARS*. Ces résultats corroborent les performances moyennes, montrant que le modèle *GAM* est le plus performant pour la prévision de la charge à l'échelle des clients tertiaires.

Les médianes les plus basses après celles du modèle *GAM* sont associées au modèle *MARS* pour les erreurs NMAE et NRMSE. Cependant, contrairement aux performances moyennes qui suggèrent que le modèle *KWF* est plus performant que le modèle *MARS*, les boîtes à moustaches montrent que le modèle *KWF* a des niveaux de médianes plus élevées que le modèle *MARS*, mais avec moins de valeurs extrêmes. Cela suggère que le modèle *KWF* est plus robuste que le modèle *MARS*, car ses erreurs de prévision sont moins susceptibles de varier de manière significative puisque les distributions sont moins étalées. Cependant, si la précision de la prévision est plus importante que la robustesse, le modèle *MARS* peut être plus approprié car il est capable de faire des prévisions plus précises malgré ses valeurs extrêmes. Concernant la méthode utilisée par le fournisseur, les boîtes à moustaches montrent une médiane plus grande et une dispersion plus importante que les autres modèles, ce qui indique que les erreurs de prévision sont plus importantes pour cette méthode. Cela confirme donc les résultats de performance moyenne qui ont montré que la méthode de fournisseur est moins précise que les autres modèles.

En résumé, le modèle *GAM* est le plus performant pour la prévision de la consommation électrique à l'échelle des clients dans le secteur tertiaire, suivis par le modèle *MARS*, tandis que la méthode utilisée chez le fournisseur est moins précise que les modèles statistiques. En moyenne, les résultats du modèle *KWF* sont meilleurs que ceux du modèle *MARS*, mais en ce qui concerne la distribution des erreurs, le modèle *MARS* est légère-

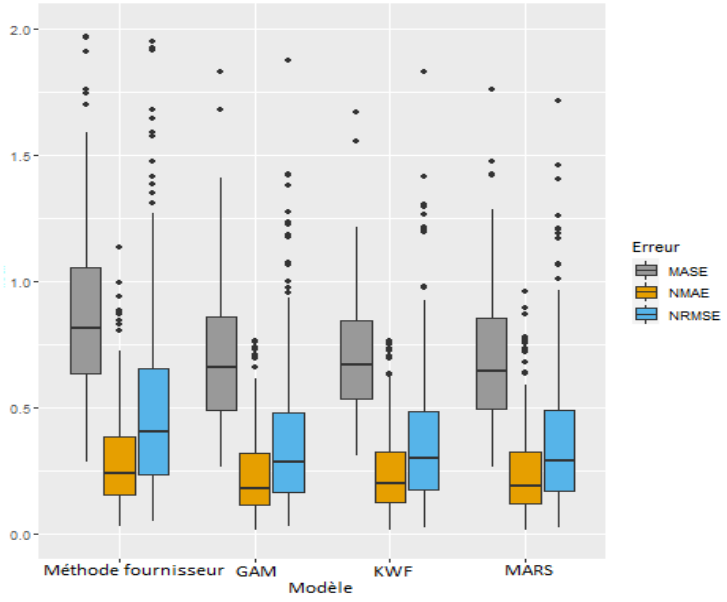


FIGURE 5.14 – Les boîtes à moustaches des erreurs de prévision à $(J+1)$ par les trois modèles *GAM*, *KWF* et *MARS* ainsi que la méthode utilisée chez le fournisseur d'énergie.

ment meilleur. Bien que les résultats pour le modèle *KWF* et *MARS* soient légèrement inférieurs à ceux de *GAM*, ils restent néanmoins supérieurs à ceux de la méthode utilisée actuellement chez le fournisseur d'énergie pour toutes les métriques.

La prévision de la charge électrique pendant les jours fériés est essentielle étant donné la variation significative de la consommation électrique qui peut avoir des répercussions importantes sur la gestion de l'énergie électrique. Cette variation peut être observée d'un jour férié à un autre. Par exemple, pendant Noël, la fermeture des restaurants et des lieux de divertissement entraîne une baisse importante de leur consommation électrique. En revanche, lors d'autres jours fériés comme la fête nationale ou le Nouvel An, la consommation électrique dans ces établissements peut augmenter en raison de l'augmentation de la fréquentation. Cette variation rend la prévision de ces jours par les modèles que nous avons proposés difficile, car ces modèles considèrent tous les jours fériés de manière uniforme. Cependant, certaines études de la littérature ont divisé ces jours fériés en fonction de leurs caractéristiques économiques et sociales (DORDONNAT et al., 2008). Malheureusement, en raison du nombre limité de jours fériés présents dans les historiques des données, cette approche ne peut pas être mise en œuvre dans notre étude car cela impliquerait de travailler avec des groupes de jours ayant un nombre de données insuffisant, ce qui n'est pas adapté à la prévision par des modèles de régression non paramétriques. Ainsi, la méthode du fournisseur d'énergie peut être plus efficace pour la prévision de ces jours, car elle se base sur une méthode simple consistant à attribuer les valeurs de consommation d'un jour férié au même jour de l'année précédente.

Nous avons comparé les résultats de la prévision de ces jours avec les modèles *GAM*,

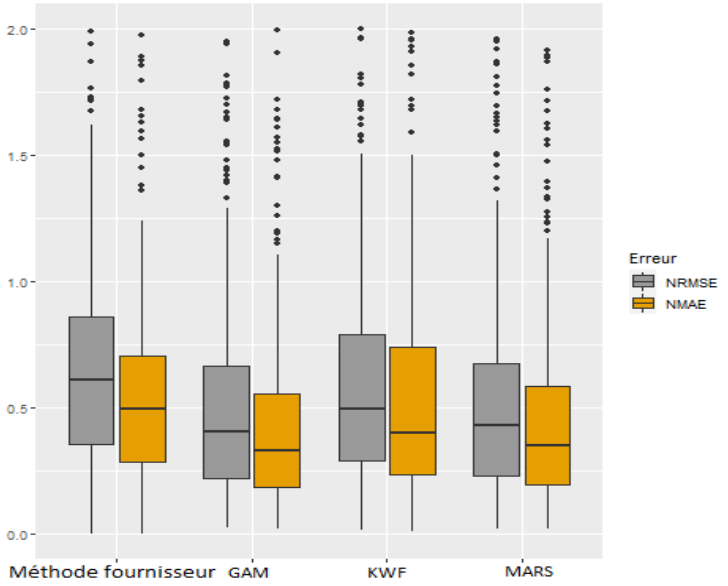


FIGURE 5.15 – Les boîtes à moustaches des erreurs de prévision à $(J+1)$ pour les modèles *GAM*, *KWF* et *MARS*, comparées à la méthode utilisée chez le fournisseur pour les jours fériés.

MARS, *KWF* et la méthode utilisée chez le fournisseur d'énergie. La figure 5.15 présente les erreurs de prévision NMAE et NRMSE des jours fériés de chaque courbe de charge dans le jeu de données pour toute la période test, sous forme de boîtes à moustaches. Nous remarquons que pour les jours fériés, tous les modèles présentent des médianes d'erreurs très élevées par rapport aux médianes d'erreurs globales de la période de test (voir la figure 5.14). En outre, la dispersion importante des résultats et la présence de nombreuses valeurs extrêmes pour chaque modèle témoignent de la difficulté des modèles à prédire avec précision la consommation électrique pendant les jours fériés. Cela met en évidence les limites de ces modèles de prévision. Contrairement à ce que nous avions supposé, la méthode du fournisseur n'a pas donné les meilleurs résultats pour la prévision des jours fériés. Les modèles *GAM* et *MARS* ont montré des performances similaires et supérieures à la fois au modèle *KWF* et à la méthode du fournisseur.

Les résultats obtenus peuvent être expliqués par le fait que le jeu de données de consommation électrique contient des jours fériés dont la consommation varie d'une année à l'autre en raison de facteurs variables tels que la météo (la température et les précipitations), les jours de la semaine et les événements spéciaux comme les festivals ou les matchs de football importants. Par conséquent, la méthode de prévision consistant à simplement attribuer la même valeur que l'année précédente pour ces jours fériés s'est avérée inefficace. En revanche, les modèles statistiques qui sont capables de détecter des relations plus complexes entre les différentes variables ont donné de meilleurs résultats.

5.4 Prévisions probabilistes

5.4.1 Objectif et intérêts

Jusqu'à présent, les prévisions de la consommation électrique étaient réalisées de manière déterministe, communément appelée prévisions ponctuelles, qui ne prenaient pas en compte les incertitudes associées à la prévision et ne fournissaient qu'une seule valeur de prévision. Cependant, pour la prévision de la charge électrique à l'échelle des clients, les données de consommation peuvent être très volatiles et influencées par de nombreux facteurs imprévisibles tels que les conditions météorologiques, les jours fériés, les vacances, ... Ainsi, l'utilisation de prévisions déterministes peut comporter des risques, notamment dans les applications industrielles telles que la gestion de la consommation d'électricité des clients. Par exemple, supposons qu'un fournisseur d'énergie électrique utilise des prévisions déterministes pour envoyer des alertes aux clients en cas d'anomalies de consommation, c'est-à-dire lorsqu'il y a une consommation anormalement élevée ou basse. Si une telle alerte est envoyée à un client sur la base d'une prévision déterministe qui ne prend pas en compte les incertitudes, il est possible qu'une fausse alerte soit envoyée. Les prévisions probabilistes peuvent aider à éviter les fausses alertes en fournissant une plage de valeurs possibles pour la consommation électrique future plutôt qu'une seule estimation ponctuelle. Avec cette plage de valeurs, il est possible de définir un seuil d'alerte qui prend en compte la variabilité de la consommation plutôt que de se baser uniquement sur une valeur ponctuelle.

Il est essentiel de noter que les prévisions probabilistes et déterministes sont complémentaires et non pas exclusives. En effet, les prévisions déterministes sont plus simples à calculer et sont pertinentes dans les situations où l'incertitude est minime et où une seule valeur prévue est suffisante pour la prise de décision. En revanche, les prévisions probabilistes sont plus adaptées pour tenir compte de l'incertitude dans les prévisions, mais elles nécessitent davantage de calculs. Par exemple, pour les clients résidentiels, les prévisions déterministes peuvent être employées pour prédire la consommation électrique d'une maison avec des habitudes de vie régulières, telle qu'une maison de vacances ou une résidence secondaire. Toutefois, pour une maison occupée par une famille avec des enfants en bas âge, dont les habitudes de vie varient considérablement, les prévisions probabilistes peuvent être plus adaptées pour tenir compte des changements imprévus dans les habitudes de vie. Pour les clients tertiaires, les prévisions déterministes peuvent être utiles pour anticiper la consommation électrique d'un grand bâtiment commercial avec des activités régulières et prévisibles, par exemple un centre commercial. En revanche, pour les bâtiments où les activités sont variables, comme les hôtels, les prévisions probabilistes peuvent être plus appropriées pour tenir compte de l'incertitude liée à la variabilité des activités et des événements.

Les prévisions probabilistes (Gneiting et Katzfuss, 2014) peuvent être calculées à partir des intervalles de prévision des modèles de prévision. Les intervalles de prévision paramétriques sont basés sur des hypothèses sur la distribution de l'erreur, tandis que les méthodes de bootstrap (Stine, 1985) et de quantiles sont des méthodes non paramétriques qui permettent de calculer des intervalles de prévision mais n'imposent pas d'hypothèses spécifiques sur la distribution de l'erreur.

La méthode de scénario (Mohammadi et al., 2014) est une autre méthode qui peut être utilisée pour générer des prévisions probabilistes. Elle consiste à définir différents scénarios possibles pour les variables d'entrée du modèle et à utiliser ces scénarios pour générer une gamme de résultats possibles pour la variable à prédire. Chaque scénario est généralement associé à une probabilité d'occurrence, qui peut être déterminée à partir d'informations historiques ou d'expertise du domaine. Les résultats de chaque scénario sont ensuite combinés pour fournir une prévision probabiliste de la variable à prédire.

Les mesures d'évaluation les plus couramment utilisées pour les prévisions probabilistes sont le taux de couverture empirique (ECR) (Antoniadis, Brossat et al., 2016), l'erreur quadratique moyenne des prévisions probabilistes (CRPS) (Gneiting et Raftery, 2007) et la divergence de Kullback-Leibler (KL) (Kullback et al., 1951). Le taux de couverture empirique évalue la capacité du modèle à capturer la variation réelle des données, tandis que l'erreur quadratique moyenne des prévisions probabilistes mesurent la distance entre la distribution prédictive et la distribution observée. Le choix de la mesure dépendent de l'objectif de la prévision et du contexte d'application.

Dans cette étude, les prévisions probabilistes ont été générées en utilisant les deux modèles les plus performants pour la prévision de chaque type de client. Le modèle KWF a été retenu pour les clients résidentiels et les méthodes de calcul des intervalles de prévision décrites dans (Antoniadis, Brossat et al., 2016) a été utilisée pour générer des prévisions probabilistes. Pour les clients du secteur tertiaire, le modèle GAM a été choisi et la méthode de régression de quantile a été utilisée pour générer des prévisions probabilistes. La qualité des prévisions a été évaluée en utilisant le taux de couverture empirique, qui mesure la capacité du modèle à capturer la variation réelle des données. Cette mesure correspond à la proportion d'observations qui se trouvent à l'intérieur des intervalles de prévision générés par un modèle probabiliste. Elle permet de déterminer si les intervalles de prévision sont bien calibrés, c'est-à-dire si le niveau de confiance choisi pour les intervalles correspond à la proportion d'observations observées à l'intérieur de ces intervalles. Un taux de couverture empirique proche du niveau de confiance choisi indique que les intervalles de prévision sont fiables et que le modèle est bien calibré. Cette mesure est simple à calculer et à interpréter, ce qui la rend adaptée à des applications dans des contextes industriels. Le taux de couverture empirique à un niveau de confiance α peut être représenté par la formule suivante :

$$ECR_y(\alpha) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \in [y_{i,\alpha}, y_{i,1-\alpha}]),$$

où y_i est la valeur observée, $y_{i,\alpha}$ est le quantile d'ordre α de la distribution des prévisions pour cette observation, et I est la fonction indicatrice (égale à 1 si l'événement entre parenthèses est vrai et 0 sinon). La moyenne de ces indicateurs pour l'ensemble des observations donne le taux de couverture empirique.

5.4.2 Intervalles de prévision pour les clients résidentiels

5.4.2.1 Description des méthodes

Nous allons présenter les trois méthodes de calcul des intervalles de prévision pour le modèle *KWF*, à savoir *NS-KWF*, *S-KWF* et *k-FWE*, décrites dans Antoniadis, Brossat et al. (2016). Ces méthodes visent à calculer les intervalles de prévision du modèle *KWF* à l'aide de la méthode de bootstrap (Stine, 1985). Pour ce faire, une prévision ponctuelle $\widehat{Z}_{n+1}$ est d'abord obtenue à partir du modèle *KWF* pour le segment Z_{n+1}. Ensuite, des pseudo-réalisations de bootstrap sont générées en fonction des poids associés aux observations du modèle. Deux cas ont été distingués : le cas stationnaire et le cas non-stationnaire. Pour le cas stationnaire, la méthode consiste à générer des pseudo-réalisations de bootstrap en deux étapes : tout d'abord, calculer la prévision ponctuelle $\widehat{Z}_{n+1}$ pour le segment Z_{n+1} en utilisant l'historique des segments $Z_1, \ldots, Z_n$, puis générer B pseudo-réalisations $Z_{n+1}^{(b)}$ en fonction des poids $\tilde{w}_{m,n}$ (voir l'équation (3.31)) qui sont utilisés pour construire la distribution de telle sorte que $\mathbb{P}(Z_{n+1}^{(b)} = Z_{m+1}|Z_n) = \tilde{w}_{m,n}$. Dans le cas non-stationnaire, la méthode est similaire, mais en prenant en compte les parties d'approximation du segment $\widehat{Z}_{n+1}$, ce qui conduit à la décomposition du processus Z en termes d'approximations et détails. Enfin, les pseudo-réalisations sont utilisées pour construire l'intervalle de prévision pour la fonction prédite $\widehat{Z}_{n+1}$.

La méthode *NS-KWF* implique d'obtenir les pseudo-résidus, de calculer les quantiles empiriques correspondants, puis de construire l'intervalle de prévision en utilisant ces quantiles. **Dans le cas stationnaire**, la procédure consiste à :

1. Calculer les pseudo-résidus par $\widehat{R}_{n+1}^{(b)}(t_i) = Z_{n+1}^{(b)}(t_i) - \hat{Z}_{n+1}(t_i)$, pour $b \in \{1, \ldots, B\}$;
2. Calculer les quantiles α et $1-\alpha$ notés respectivement $\hat{R}_{n+1,\alpha}(t_i)$ et $\hat{R}_{n+1,1-\alpha}(t_i)$ pour chaque t_i ;
3. Construire l'intervalle de prévision en utilisant les quantiles obtenus : $U_{n+1,\alpha}(t_i) = \hat{Z}_{n+1}(t_i) + \hat{R}_{n+1,1-\alpha}(t_i)$ et $L_{n+1,\alpha}(t_i) = \hat{Z}_{n+1}(t_i) + \hat{R}_{n+1,\alpha}(t_i)$.

Dans le cas non stationnaire, la procédure de construction d'intervalle de prévision est modifiée pour inclure à la fois les parties d'approximation et de détail du segment. Pour

cela, des pseudo-résidus sont obtenus à la fois pour les détails et les approximations, et les quantiles empiriques pour les deux sont calculés séparément. Ensuite, l'intervalle de prévision est calculé en utilisant les deux ensembles de quantiles et l'écart-type des résidus bootstrap. Ainsi, la nouvelle procédure sera la suivante :

1. Obtenir les pseudo-résidus des détails et des approximations de la manière suivante :

   Pour $\quad b \in \{1, \ldots, B\}, \widehat{R}_{n+1}^{(b)}(t_i) = D_{m+1}^{(b)}(t_i) - \widehat{D}_{n+1}(t_i), \quad \widehat{Q}_{n+1}^{(b)}(t_i) = S_{m+1}^{(b)}(t_i) - \widehat{S}_{n+1}(t_i);$

2. Calculer pour chaque t_i les quantiles empiriques α et $1-\alpha$ pour les résidus de détails $\widehat{R}_{n+1,\alpha}(t_i)$ et $\widehat{R}_{n+1,1-\alpha}(t_i)$ et sélectionner les parties d'approximation correspondantes $\widehat{Q}_{n+1,\alpha}(t_i)$ et $\widehat{Q}_{n+1,1-\alpha}(t_i)$;
3. Sélectionner les parties approximation correspondantes. Pour chaque t_i du maillage d'échantillonnage, l'intervalle de prévision est donné par :

$$L_{n+1,\alpha}(t_i) = \widehat{Q}_{n+1,\alpha}(t_i) + \widehat{R}_{n+1,\alpha}(t_i) + \widehat{Z}_{n+1}(t_i) \quad \text{et}$$
$$U_{n+1,\alpha}(t_i) = \widehat{Q}_{n+1,1-\alpha}(t_i) + \widehat{R}_{n+1,1-\alpha}(t_i) + \widehat{Z}_{n+1}(t_i).$$

La méthode *S-KWF* construit un intervalle symétrique dont la longueur dépend des écarts-types des résidus de bootstrap et d'un quantile théorique dépendant de $1-\alpha$. Cette méthode est typiquement de la forme $\widehat{Z}_{n+1}(t_i) \pm 1,96\hat{\sigma}_{t_i}$, pour le quantile gaussien symétrique au niveau 0,95 avec $\hat{\sigma}_{t_i}$ étant l'écart-type empirique de $\widehat{Z}_{t_i}^{(b)}$.

La méthode *k-FWE* consiste à calculer les résidus bootstrap standardisés et à obtenir les k-maximums de ces résidus. Ensuite, le quantile $1-\alpha$ est calculé à partir de ces k-maximums pour obtenir d-max. Enfin, l'intervalle de prévision pour le segment de prévision est calculé en utilisant d-max et l'écart-type des résidus bootstrap. Soit $\widehat{Z}_{n+1}(t)$ la prévision du segment $Z_{n+1}(t)$ et B les prévisions bootstrap. Les étapes de calcul des intervalles de prévision par cette méthode sont résumées dans Antoniadis, Brossat et al. (2016) de la façon suivante :

1. Calculer les résidus bootstrap standardisés $\hat{s}_b^* \in \mathbb{R}^H$ avec $b = 1, \ldots, B$;
2. Obtenir les k-max$_b$, les k plus grandes valeurs de $|\hat{s}_b|$;
3. Calculer le quantile $1-\alpha$ de k-max$_b^*$ noté d-max ;
4. Obtenir l'intervalle de prévision pour $\widehat{Z}_{n+1}(t)$ par : $\widehat{Z}_{n+1}(t_i) \pm d_{\max} \cdot \hat{\sigma}_{t_i}$.

5.4.2.2 Expérience numérique

Nous avons implémenté en R les trois méthodes de calcul des intervalles de prévision décrites précédemment. Nous avons ensuite testé ces méthodes sur les données de consommation électrique des ménages. Nous avons choisi de travailler avec les premiers

cent ménages dans le jeu de données décrit dans la sous-section 4.2.1 pour lesquels le modèle *KWF* présente les meilleurs résultats en termes de précision des prévisions ponctuelles, évaluées à l'aide des métriques NMAE, NRMSE et sMAPE. Cette sélection nous a permis de travailler dans un scénario où le modèle de prévision *KWF* est performant. Il est probable que les ménages ayant donné de bons résultats de prévision ponctuelle ont des données de consommation électrique régulières et suivent des schémas de consommation prévisibles, ce qui peut rendre les prévisions probabilistes de la consommation pour ces ménages plus fiables et significatifs. Nous avons ensuite divisé les données de ces ménages en un ensemble d'entraînement (70%) et un ensemble de test (30%). En appliquant la méthodologie décrite dans la section 4.3, nous avons obtenu les prévisions à $(J+1)$ pour le modèle *KWF* appliqué à l'ensemble de ménages sélectionnés. Nous avons suivi la procédure décrite dans le paragraphe précèdent pour produire 100 échantillons bootstrap. Ensuite, chacune des méthodes présentées dans le paragraphe précédent a été utilisée pour calculer les intervalles de prévision à $(J+1)$ pour les deux niveaux de confiance (90% et 95%) recommandés par les experts du fournisseur d'énergie. Nous avons fixé à 2 le paramètre k indiquant le nombre de points en dehors de la construction des intervalles de prévision dans la méthode *k-FWE*. Enfin, l'évaluation de la qualité des prévisions probabilistes a été effectuée en calculant le taux de couverture empirique pour chaque ménage. Les résultats de cette expérience sont présentés dans la partie suivante.

5.4.2.3 Résultats

La table 5.3 montre la distribution des taux de couverture empirique exprimés en pourcentage des intervalles de prévision calculés par les trois méthodes *NS-KWF*, *S-KWF* et *k-FWE* pour un niveau de confiance de 90%. Les valeurs statistiques présentées dans la table sont les suivantes : minimum (Min.), premier quartile (Q1.), médiane, moyenne, troisième quartile (Q3.), maximum (Max.).

	Min.	Q1.	Médiane	Moyenne	Q3.	Max.
NS-KWF	48,47	63,91	68,17	68,11	72,46	83,05
S-KWF	54,10	73,56	81,19	79,35	86,25	99,74
k-FWE	96,41	97,62	98,25	98,25	98,84	99,99

TABLE 5.3 – Distribution des taux de couverture empirique exprimés en pourcentage pour les méthodes *NS-KWF*, *S-KWF* et *k-FWE* pour un niveau de confiance de 90%.

Les résultats montrent que la méthode *NS-KWF* a des taux de couverture empirique relativement faibles pour tous les quantiles testés, avec une moyenne de 68,11% pour un niveau de confiance de 90%. Cela suggère que les intervalles de prévision obtenus avec cette méthode ont une probabilité relativement faible de couvrir les valeurs réelles des courbes de charge électrique des ménages testés dans le jeu de données. En revanche, la méthode *S-KWF* présente des taux de couverture empirique plus élevés que la méthode *NS-KWF* pour tous les quantiles, avec une moyenne de 79,35%. Cette différence suggère

que les intervalles de prévision générés par la méthode *S-KWF* sont de meilleure qualité que ceux générés par la méthode *NS-KWF* en termes de taux de couverture empirique. Les indicateurs statistiques des deux méthodes *S-KWF* et *NS-KWF* sont globalement inférieurs au niveau de confiance théorique de 90%. Cela signifie que les intervalles de prévision générés par ces deux méthodes ont une probabilité relativement faible de couvrir les valeurs réelles des courbes de charge électrique des ménages testés avec un niveau de confiance de 90%.

La méthode *k-FWE*, quant à elle, affiche des taux de couverture empirique élevés pour tous les quantiles testés, avec une moyenne de 98,25% pour un niveau de confiance de 90%. De plus, tous les indicateurs statistiques dépassent le niveau de confiance fixé. Ces résultats suggèrent que la méthode *k-FWE* est capable de couvrir avec précision les valeurs réelles des courbes de charge électrique des ménages testés, ce qui la positionne comme une méthode supérieure aux deux autres méthodes testées.

Nous avons également effectué des tests sur les méthodes de calcul des intervalles de prévision à un niveau de confiance de 95% afin d'évaluer leurs performances à ce niveau. Cette analyse permet de vérifier si ces méthodes maintiennent des niveaux de couverture empirique similaires à ceux obtenus à un niveau de confiance de 90%, ce qui peut être plus important pour le fournisseur d'énergie. En conséquence, cela permet de sélectionner la méthode la plus adaptée pour chaque niveau de confiance requis. Les résultats obtenus sont présentés dans la table 5.4.

	Min.	Q1.	Médiane	Moyenne	Q3.	Max.
NS-*KWF*	61,91	74,42	78,27	78,04	82,00	91,36
S-*KWF*	68,42	81,93	87,82	86,70	90,61	99,93
k-FWE	96,85	98,03	98,60	98,57	99,15	100,00

TABLE 5.4 – Distribution des taux de couverture empirique exprimés en pourcentage pour les méthodes *NS-KWF*, *S-KWF* et *k-FWE* pour un niveau de confiance de 95%.

En comparant les résultats des deux tables 5.3 et 5.4, nous pouvons constater que le passage d'un niveau de confiance de 90% à 95% conduit à une augmentation des taux de couverture empirique pour toutes les méthodes testées. Cela est cohérent avec l'idée que plus le niveau de confiance est élevé, plus les intervalles de prévision seront larges, ce qui augmente la probabilité que les valeurs réelles de consommation tombent dans les intervalles de prévisions. En particulier, pour la méthode *NS-KWF*, nous pouvons observer une amélioration significative de ses performances en termes de taux de couverture empirique, passant d'une moyenne de 68,11% à 78,04%, soit une augmentation de 9,93%. De même, la méthode *S-KWF* a également connu une amélioration de ses performances, avec une augmentation de sa moyenne de 79,35% à 86,70%, soit une augmentation de 7,35%. Cependant, malgré cette amélioration, la méthode *S-KWF* reste toujours en deuxième position par rapport à la méthode *k-FWE* qui maintient sa performance à un niveau élevé et stable, avec une moyenne des taux de couverture autour de 98% pour les deux niveaux

de confiance.

Il est important de souligner que pour le niveau de confiance 95%, les méthodes *NS-KWF* et *S-KWF* ont toujours des taux de couverture empirique inférieurs au niveau de confiance théorique, ce qui signifie que ces méthodes ont tendance à sous-estimer la variabilité des données et à produire des intervalles de prévision plus étroits que ce qui est nécessaire pour atteindre le niveau de confiance souhaité.

En résumé, les résultats obtenus permettent de conclure que la méthode *k-FWE* est la plus performante des trois méthodes testées en termes de taux de couverture empirique, et ce pour les deux niveaux de confiance 90% et 95%.

Il est à souligner que bien que la méthode *k-FWE* ait montré une performance supérieure en termes de taux de couverture empirique, l'utilisation de la méthode *S-KWF* peut être plus appropriée dans certains contextes. En effet, le choix entre ces deux méthodes dépend des préférences et des besoins spécifiques du fournisseur d'énergie, ainsi que du contexte d'application. En général, un taux de couverture empirique élevé est souhaitable pour les intervalles de prévision car cela indique que les intervalles sont plus fiables et que les valeurs réelles sont plus susceptibles de tomber dans ces intervalles. Toutefois, il est important de noter que le taux de couverture élevé ne garantit pas toujours une bonne qualité de prévision, car cela peut résulter d'intervalles de prévision très larges et peu informatifs pour la prise de décision. Par conséquent, l'interprétation des résultats des taux de couverture doit être faite en fonction des objectifs d'utilisation des prévisions. Par exemple, si les prévisions seront utilisées pour générer des alertes de sur ou sous-consommation, il serait souhaitable d'utiliser une méthode avec un taux de couverture empirique élevé comme la méthode *k-FWK* pour minimiser les risques de fausses alertes. Cependant, si les prévisions sont utilisées pour inciter les clients à réduire leur consommation électrique pendant les périodes de pointe, une méthode de prévision avec des intervalles de prévision ayant un taux de couverture moins élevé comme la méthode *S-KWF* peut être plus appropriée pour générer des alertes précises et ciblées. Il serait pertinent de mener une analyse de performance de ces deux méthodes dans le contexte spécifique de chaque application envisagée par le fournisseur d'énergie. Cette approche permettra de mieux comprendre les forces et les limites de chacune et d'identifier la plus adaptée aux besoins spécifiques de l'application en question. D'autres critères d'évaluation de la qualité des prévisions probabilistes tels que la symétrie des intervalles de prévision, l'exactitude de la densité de probabilité prédite ou encore la rapidité de la mise à jour des prévisions peuvent également être testés selon le contexte d'application.

Après avoir comparé les performances de trois méthodes de calcul des intervalles de prévision du modèle *KWF*, nous avons étudié comment la méthode la plus performante, à savoir la méthode *k-FWK*, varie en termes de performance en fonction des différentes tranches horaires de la journée. Cette analyse peut aider à déterminer si les fluctuations

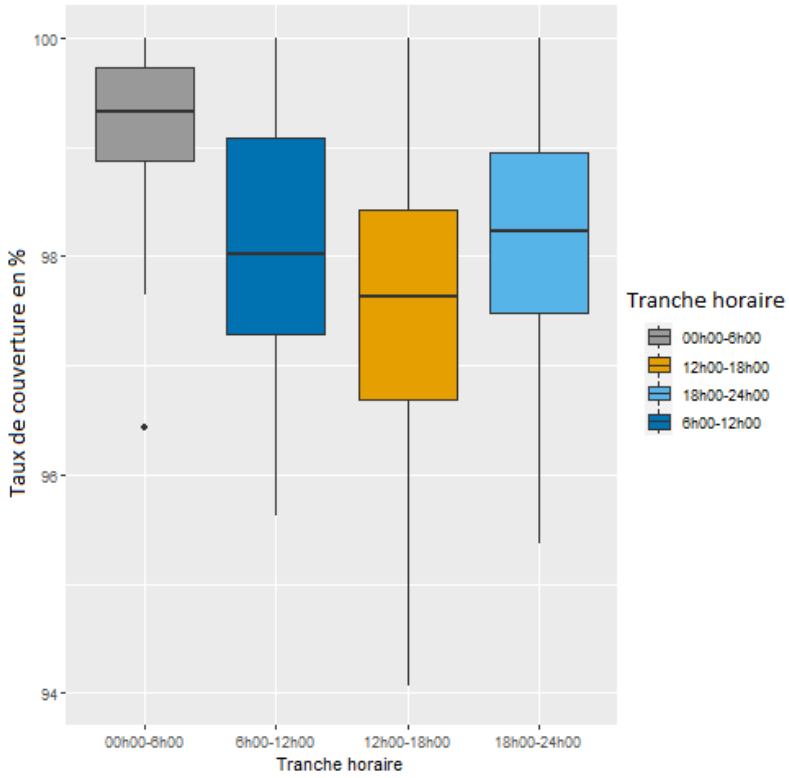


FIGURE 5.16 – Les boîtes à moustaches des taux de couverture empirique en pourcentage des intervalles de prévision du modèle KWF calculés par la méthode k-FWK à un niveau de confiance de 90% en fonction des tranches horaires de la journée.

des taux de couverture des intervalles de prévision sont liées aux comportements des consommateurs et à leur mode de vie.

Nous avons donc divisé la journée en quatre périodes de six heures chacune : de minuit à 6 heures, de 6 heures à midi, de midi à 18 heures et de 18 heures à minuit, afin d'étudier la variation des taux de couverture empirique en fonction des tranches horaires de la journée. Pour chaque ménage parmi les cent ménages sélectionnés, nous avons calculé le taux de couverture pour chaque tranche horaire. Les résultats sont présentés sous forme de boîtes à moustaches dans la figure 5.16.

L'analyse des boîtes à moustaches des taux de couverture empirique montre que les périodes de minuit à 6 heures du matin et de 18 heures à minuit ont des médianes de taux de couverture empirique plus élevées, tandis que la période de 12 heures à 18 heures a la médiane la plus basse. Cela signifie que les prévisions probabilistes de consommation d'électricité des ménages pour les périodes de minuit à 6 heures du matin et de 18 heures à minuit sont plus précises que pour les autres périodes de la journée en termes de taux de couverture empirique.

La boîte la moins étalée pour la tranche horaire de minuit à 6 heures du matin

indique que les taux de couverture empirique sont plus homogènes pendant cette période. En revanche, les boîtes de 6 heures à 12 heures et de 18 heures à minuit sont les plus étalées, ce qui signifie que les taux de couverture empirique sont plus variables pendant ces deux périodes de la journée.

Les résultats obtenus sont cohérents avec les habitudes de consommation d'électricité des ménages. En effet, la période de minuit à 6 heures du matin correspond généralement aux heures de sommeil et de repos, où les activités énergivores sont réduites, à l'exception des appareils programmés pour fonctionner automatiquement pendant les heures creuses de la nuit. Cela conduit à une consommation d'électricité plus stable et prévisible pendant ces périodes chez la majorité des ménages, ce qui se reflète dans les taux de couverture empirique plus élevés.

La variabilité importante des habitudes de consommation d'électricité durant les périodes de 6 heures à 12 heures et de 12 heures à 18 heures pourrait expliquer la baisse des valeurs médianes des taux de couverture empirique observée. En effet, durant ces heures, les ménages ont tendance à vaquer à leurs activités quotidiennes telles que la cuisine, la lessive et la charge de véhicules électriques, ce qui peut provoquer des pics de consommation d'électricité difficiles à prévoir. Cette variabilité peut donc affecter la capacité des intervalles de prévision calculés par la méthode *k-FWE* à couvrir les données de manière précise, ce qui se reflète dans la baisse des taux de couverture empirique.

Les périodes entre 18h et minuit présentent des taux de couverture empirique généralement meilleurs que les périodes de 6h à 12h et de 12h à 18h, en raison de la domination de la consommation d'électricité par des appareils électroménagers tels que la télévision, l'éclairage et les appareils électroniques. Ces derniers ont tendance à avoir une consommation d'électricité plus stable et moins variable que les appareils utilisés pendant les périodes de haute activité domestique de 6h à 12h et de 12h à 18h. Cela explique les médianes des taux de couverture empirique plus élevés pour cette période. Cependant, ces taux de couverture restent inférieurs à ceux de la période de minuit à 6h en raison de la faible consommation électrique pendant la nuit, qui présente une stabilité de consommation électrique accrue et une moindre variabilité.

La relation entre les taux de couverture des intervalles de prévision et les habitudes de consommation est une observation cruciale car elle implique des applications pratiques significatives pour la gestion de la consommation d'électricité. Ces intervalles de prévision peuvent également être utilisés pour générer des alertes aux consommateurs lorsque leur consommation réelle dépasse les bornes supérieures de leurs intervalles de prévision. Ces alertes peuvent être classées en intensité en fonction de leur position dans la journée et de leur distance par rapport à la période de pointe. En prenant des mesures en réponse à ces alertes, les consommateurs peuvent réduire leur consommation d'électricité pendant les périodes de pointe, contribuant ainsi à réduire la demande d'électricité et à réaliser

des économies d'énergie. En encourageant les comportements économes en énergie, tels que la réduction de l'utilisation d'appareils énergivores pendant les périodes de pointe ou la participation à des programmes de tarification dynamique, les consommateurs peuvent prendre des mesures pour améliorer leur efficacité énergétique.

5.4.3 Intervalles de prévision pour les clients tertiaires

Comme pour les clients résidentiels, nous avons calculé les intervalles de prévision pour les clients tertiaires en utilisant le modèle *GAM* qui s'est avéré le plus performant pour la prévision de la charge électrique à cette échelle. Pour cela, nous avons sélectionné un échantillon test de cent clients tertiaires tirés aléatoirement du jeu de données décrit dans la sous-section 5.3.2, et calculé les intervalles de prévision par la méthode de quantile de régression. Contrairement aux clients résidentiels, la sélection aléatoire des clients tertiaires était justifiée par une qualité globalement satisfaisante de la prévision de la charge électrique pour ce secteur. Nous nous attendons ainsi à ce que les prévisions probabilistes soient également précises pour la majorité des clients dans le secteur tertiaire. Les données de chaque client ont été divisées en 70% pour l'entraînement et 30% pour le test. Les intervalles de prévision ont été calculés et évalués pour toute la période de test.

5.4.3.1 Description de la méthode

Nous avons utilisé la méthode de régression quantile pour calculer les intervalles de prévision du modèle *GAM* décrit dans la sous-section 4.4.1 et adapté pour la prévision des courbes de charge des clients dans le secteur tertiaire (section 5.3). La méthode de régression quantile a été introduite par Koenker et al. (1978) pour estimer la fonction quantile conditionnelle de la variable à expliquer y, étant donné les variables explicatives X. La fonction quantile conditionnelle de y à un quantile donné τ est notée $q_\tau(y|X)$.

Les modèles *GAM* peuvent être utilisés pour modéliser la relation entre les variables explicatives et la fonction quantile conditionnelle de la variable à expliquer. Dans un modèle *GAM*, on suppose que la fonction quantile conditionnelle de la variable à expliquer y peut être écrite comme une somme de fonctions lisses des variables explicatives :

$$q_\tau(y|X) = \sum_{i=1}^{p} f_i(x_i),$$

où x_i est la i-ème variable explicative, f_i est une fonction lisse associée à la i-ème variable explicative et p est le nombre total de variables explicatives.

La fonction de perte utilisée pour la régression quantile est définie comme suit :

$$L_\tau(y, q_\tau(y|X)) = \rho_\tau(y - q_\tau(y|X)) = \begin{cases} \tau(y - q_\tau(y|X)) & \text{si } y - q_\tau(y|X) > 0 \\ (\tau - 1)(y - q_\tau(y|X)) & \text{si } y - q_\tau(y|X) \leq 0, \end{cases}$$

La fonction de perte totale est définie comme la somme des fonctions de perte pour tous les points de données :

$$Q_\tau = \sum_{i=1}^{n} L_\tau(y_i, q_\tau(y_i|X_i)) + \sum_{i=1}^{p} \lambda_i J(f_i)$$

où n est le nombre de points de données, y_i est la variable à expliquer associée à l'i-ème observation, X_i est le vecteur des variables explicatives associées à l'i-ème observation, λ_i est le paramètre de régularisation pour la i-ème fonction lisse f_i et $J(f_i)$ est la mesure de régularisation pour la i-ème fonction lisse f_i. Le paramètre de régularisation λ_i contrôle la quantité de régularisation appliquée à la fonction lisse f_i, ce qui permet d'éviter le surajustement. La mesure de régularisation $J(f_i)$ quantifie la rugosité de la fonction lisse f_i. Les méthodes de régularisation les plus couramment utilisées sont la pénalisation L_1 également appelée régularisation de Lasso (Shrinkage, 1996) et la pénalisation L_2 également appelée régularisation de Ridge (Hoerl et al., 1970).

Une fois le modèle *GAM* ajusté et les quantiles de régression estimés, nous pouvons déterminer les intervalles de prévision pour les niveaux de confiance choisis. Les intervalles de prévision sont définis par l'intervalle $[q_{\tau_{1-\gamma/2}}(X), q_{\tau_{\gamma/2}}(X)]$, où $q_{\tau_{1-\gamma/2}}(X)$ et $q_{\tau_{\gamma/2}}(X)$ représentent respectivement le $\tau_{1-\gamma/2}$-ème quantile conditionnel et le $\tau_{\gamma/2}$-ème quantile conditionnel de la variable à expliquer y sachant les variables explicatives X. Les valeurs de $\tau_{1-\gamma/2}$ et $\tau_{\gamma/2}$ dépendent du niveau de confiance γ choisi. Par exemple, si $\gamma = 0,9$, alors $\tau_{1-\gamma/2} = 0,05$ et $\tau_{\gamma/2} = 0,95$.

Nous avons appliqué la méthode de quantile de régression en utilisant le `package qgam` du logiciel R (Fasiolo et al., 2020) pour calculer les intervalles de prévision à $(J + 1)$ du modèle *GAM* pour les cent clients tirés aléatoirement du jeu de données décrit dans la section 5.3. Les deux niveaux de confiance de 90% et 95% ont été ciblés. Nous avons calculé les quantiles de régression correspondants aux niveaux de confiance de 90% pour les intervalles de prévision, avec des quantiles de 0,05 et 0,95, ainsi que pour les intervalles de prévision à 95% avec des quantiles de 0,025 et 0,975. La performance de cette méthode a été évaluée en calculant le taux de couverture empirique, de la même manière que pour les données résidentielles. Les résultats obtenus sont présentés dans la partie suivante.

5.4.3.2 Résultats

Les résultats des taux de couverture empirique des intervalles de prévision du modèle *GAM* calculés par la méthode de quantile de régression pour les courbes de charge des cent clients à deux niveaux de confiance (90% et 95%) sont présentés dans la table suivante.

Niveau de confiance	Min.	Q1	Médiane	Moyenne	Q3	Max.
90%	81,39	90,14	92,29	91,92	93,69	98,90
95%	88,97	95,26	96,68	96,37	97,75	99,82

TABLE 5.5 – Distribution des taux de couverture empirique exprimés en pourcentage des intervalles de prévision du modèle *GAM* à $(J+1)$ calculés par la méthode de quantile de régression pour les deux niveaux de confiance (90% et 95%).

Nous pouvons voir que pour le niveau de confiance de 90%, les taux de couverture empirique varient entre 81,39% et 98,90%, avec une moyenne de 91,92% supérieure au niveau de confiance. Pour le niveau de confiance de 95%, les taux de couverture empirique varient entre 88,97% et 99,82%, avec une moyenne de 96,37% également supérieure au niveau de confiance. Nous pouvons remarquer que les taux de couverture empirique sont généralement plus élevés pour le niveau de confiance de 95% que pour le niveau de confiance de 90% pour tous les indicateurs statistiques, ce qui est attendu étant donné que les intervalles de prévision associés à un niveau de confiance plus élevé sont plus larges et par conséquent, sont capables d'inclure les valeurs réelles des courbes de charge des clients.

De plus, nous pouvons remarquer que la médiane et le Q1 sont très proches de la moyenne pour les deux niveaux de confiance, ce qui suggère que la distribution des taux de couverture empirique est relativement symétrique et ne présente pas de biais important. Le fait que les valeurs de Q1 pour les deux niveaux de confiance soient supérieures aux valeurs de niveau de confiance suggère que la méthode de calcul des intervalles de prévision utilisée est globalement performante en termes de taux de couverture. En effet, cela signifie également que 75% des clients ont un taux de couverture empirique équivalent ou supérieur au niveau de confiance qui leur est théoriquement proposée dans leurs intervalles de prévisions.

Cependant, il est important de souligner que les 25% des clients restants présentent un taux de couverture empirique inférieur au niveau de confiance correspondant, ce qui pourrait indiquer une certaine imprécision dans la prévision probabiliste de leur consommation. Il est donc essentiel d'analyser plus en détail ces cas afin d'identifier les causes de cette imprécision.

L'analyse plus détaillée de ces 25% de clients a révélé que la majorité d'entre eux (44% pour un niveau de confiance de 90% et 64% pour un niveau de confiance de 95%) ont un taux de couverture compris dans une fourchette proche des niveaux de confiance théoriques, ce qui est considéré comme acceptable. Pour les autres clients qui présentent

un taux de couverture empirique plus inférieur au niveau de confiance, nous avons identifié que leurs données de consommation évoluent avec le temps, ce qui peut entraîner des imprécisions dans la prévision probabiliste de leur consommation et ainsi affecter leur taux de couverture empirique. Ces changements peuvent se produire brusquement ou graduellement (appelés couramment des "drifts"), comme le montrent respectivement les figures 5.17 et 5.18. Le même problème a été identifier dans la prévision ponctuelle des données de consommation des ménages (voir la section 4.7). Il est important de souligner que nous désignons par "changements brusques de la consommation électrique" tous les changements qui se produisent soudainement et persistent dans le temps, tandis que les "drifts" sont ceux qui se produisent graduellement et peuvent évoluer dans différentes directions. Par exemple, un changement brusque pourrait être une entreprise qui installe de nouveaux équipements énergivores, tandis qu'un "drift" pourrait être causé par une entreprise qui adopte progressivement des pratiques plus éco-responsables, réduisant ainsi sa consommation d'énergie au fil du temps.

La détection automatique de ces changements dans les données de consommation électrique est alors essentielle pour améliorer le taux de couverture des intervalles de prévision des clients tertiaires. En les identifiant rapidement, il serait possible d'ajuster les modèles de prévision et les méthodes de calcul des intervalles de prévision en conséquence.

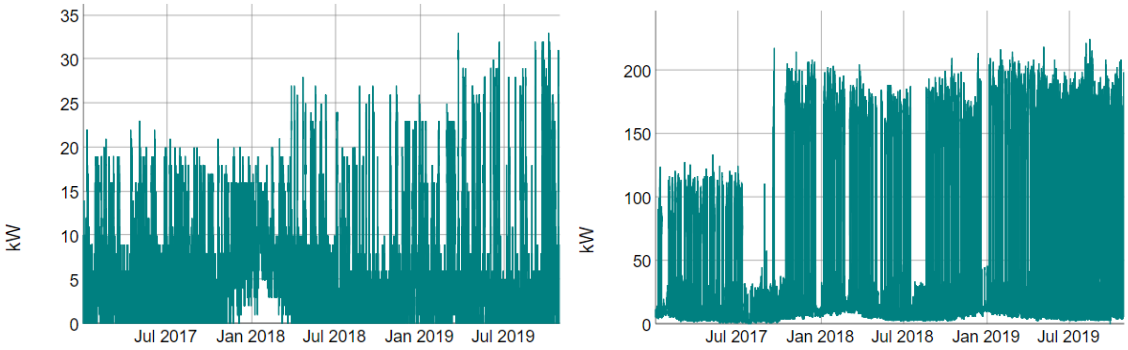


FIGURE 5.17 – Exemples des courbes de charge montrant des changements brusques dans les données de consommation électrique de deux clients tertiaire.

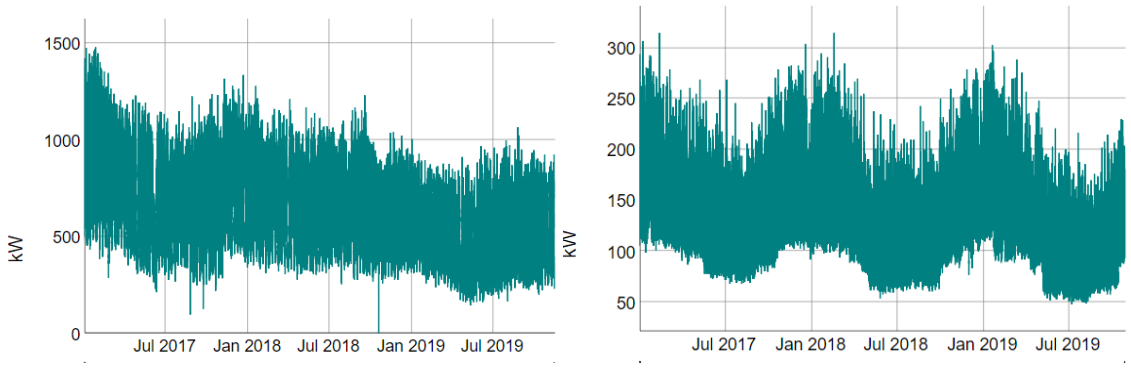


FIGURE 5.18 – Exemples des courbes de charge montrant des *drifts* dans les données de consommation électrique de deux clients tertiaire.

En conclusion, l'analyse des taux de couverture empirique pour les deux niveaux de confiance a montré que les intervalles de prévision du modèle GAM calculés à $(J+1)$ par la méthode de quantile de régression ont une forte probabilité de couvrir les valeurs réelles des courbes de charge électrique et donnent des résultats satisfaisants pour la plupart des clients, avec des taux de couverture proches ou supérieurs aux niveaux de confiance théoriques. Cependant, pour certains clients, les résultats montrent un taux de couverture empirique inférieur au niveau de confiance proposé, ce qui peut être attribué à des changements brusques ou graduels dans leurs habitudes de consommation. Afin d'améliorer le taux de couverture des intervalles de prévision pour les clients dont les taux de couverture empiriques sont inférieurs au niveau de confiance, il est crucial de surveiller régulièrement leurs données de consommation et d'adapter la méthode de quantile de régression en conséquence. C'est pourquoi la mise en place de méthodes de détection automatique de ces changements est essentielle pour détecter et traiter rapidement les changements significatifs dans les données de consommation en ajustant les modèles de prévision ponctuelle et probabiliste.

5.5 Prévision des pertes

5.5.1 Objectif et intérêts

La compensation des pertes est un sujet très intéressant pour les fournisseurs d'électricité. En effet, lors du transport de l'électricité tout au long du réseau de transport d'électricité, elle subit des pertes qui dépendent de plusieurs facteurs notamment de la quantité d'électricité injectée, les conditions météorologiques ainsi que la distance de transport. Ces pertes se quantifient par l'écart entre la quantité d'énergie injectée dans les lignes de transport et l'énergie livrée au compteur et facturée aux clients. Selon RTE, 78% des pertes sont

causées par effet Joules [3] qui se traduit par un dégagement de la chaleur suite au passage du courant électrique dans les lignes de transport formées des matériaux conducteurs. De plus, les conditions météorologiques ont un fort impact également sur les pertes dont 8% [4] sont liées au phénomène de décharge électrique connu par l'effet couronne [5] entre l'air et les câbles transportant du courant sous haute tension. Il se manifeste par l'apparition d'une gaine lumineuse bleuâtre qui se forme autour de ces câbles [6]. L'autoconsommation des transformateurs et des auxiliaires des postes constitue également une source de perte qui représente environ 3% de la quantité d'énergie.

FIGURE 5.19 – Phénomène de couronne observé tout au long des câbles de transport d'électricité [Source [7]].

Les fournisseurs d'électricité doivent alors veiller à la compensation des pertes électriques résultant du transit sur leurs réseaux pour pouvoir assurer l'équilibrage entre l'achat et la vente de l'énergie électrique. L'objectif est alors de prévoir ces pertes avec le plus de précision possible pour pouvoir les anticiper puisque tout écart entre l'énergie achetée et l'énergie livrée effective aura un impact sur le prix de l'achat et de vente de l'énergie.

5.5.2 Méthode de prévision actuelle dans l'entreprise et besoin

Actuellement chez le fournisseur, les prévisions des pertes sont effectuées de manière manuelle en utilisant une moyenne pondérée des jours équivalents passés, en fonction du calendrier et de la température extérieure. Cette méthode n'est pas automatisée et nécessite une intervention manuelle de la part d'un membre de l'équipe de prévision chaque matin. La qualité des prévisions dépend donc de l'expertise de l'opérateur chargé de réaliser ces prévisions. Il est donc nécessaire de mettre en place un modèle de prévision automatisé et précis pour les prévisions de pertes. L'automatisation dans ce contexte fait référence à

3. shorturl.at/ADUX4
4. shorturl.at/bjzOX
5. https://fr.wikipedia.org/wiki/Effet_corona
6. shorturl.at/ILXZ8
7. https://betiac.ma/effet-corona/

la capacité des modèles à générer des prévisions de manière autonome sans l'intervention humaine constante. L'automatisation est un atout car elle permet d'accélérer le processus de prévision et de réduire les variations dans les prévisions manuelles qui sont souvent liées aux différences d'expérience et d'expertise des opérateurs chargés de réaliser les prévisions. En outre, ce modèle doit satisfaire les exigences industrielles en matière de rapidité, de précision et d'interprétabilité.

5.5.3 Description des données

Les données utilisées dans cette étude comprennent l'historique des pertes du réseau de distribution du fournisseur, enregistré toutes les demi-heures, pour la période allant de janvier 2015 à janvier 2020, ainsi que l'historique des températures extérieures pour la même période.

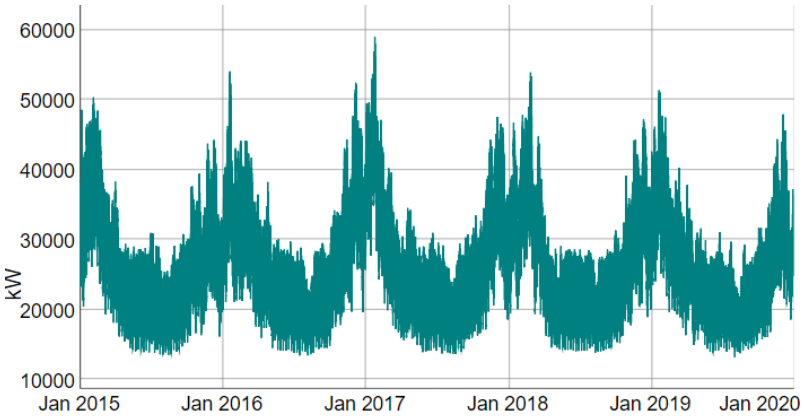
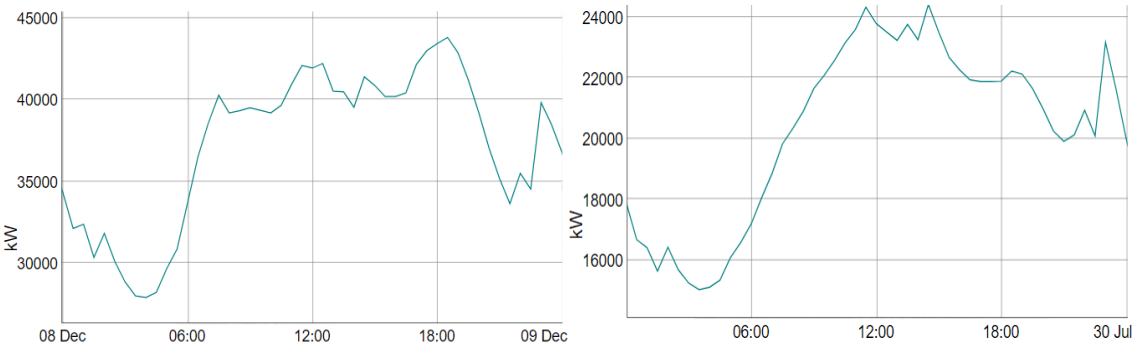


FIGURE 5.20 – L'historique des données de pertes.

Ces données ont des caractéristiques qui ressemblent à certains égards aux caractéristiques de la demande d'électricité à l'échelle nationale (voir figure 4.4) et à l'échelle des données agrégées de la demande (voir figure 4.6). La figure 5.20 montre une forte saisonnalité annuelle dépendante des variations saisonnières de la température. Cette thermosensibilité est plus prononcée aux températures froides et se manifeste par une forte augmentation des pertes en hiver donnant ainsi des formes des pointes et des creux consécutifs entre l'hiver et l'été et menant à un écart important entre les pointes des pertes d'hiver et d'été. Les données montrent également une forte saisonnalité journalière qui se manifeste par des heures de pointes et des creux au cours de la journée (voir la figure 5.21). La figure 5.21 montre deux exemples de cycle journalier des pertes en hiver (voir la figure 5.21a) et en été (voir la figure 5.21b). Les deux figures montrent un creux des pertes la nuit de 00h00 à 6h30 qui correspond au minimum des pertes sur les 24 heures de la journée. La pointe du matin en hiver est autour de 7h (voir la figure 5.21a) alors qu'en été est autour de 11h30 (voir la figure 5.21b). Le maximum des pertes est atteint à la pointe vers 14h30 en été

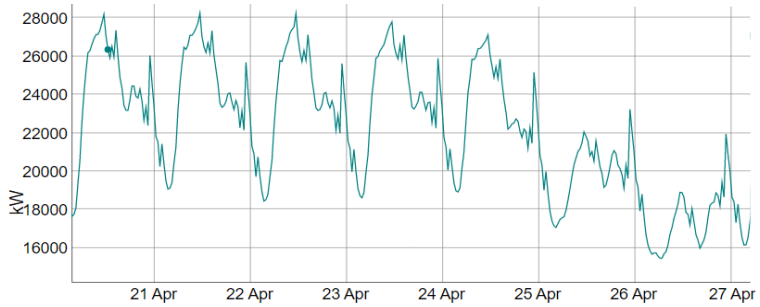
(voir la figure 5.21b) et vers 18h30 en hiver (voir la figure 5.21a).



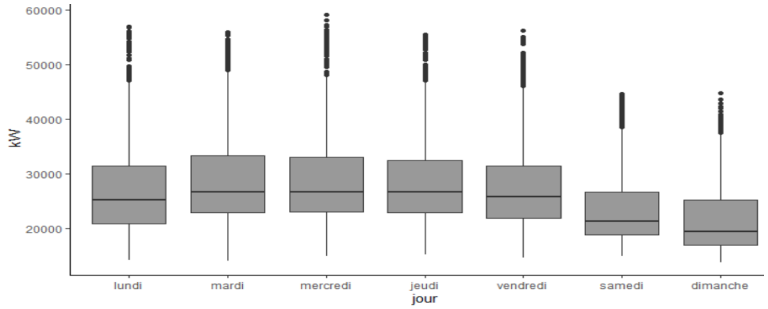
(a) En hiver (journée de 8 décembre 2015). (b) En été (journée de 29 juillet 2015).

FIGURE 5.21 – Exemple de cycle journalier des pertes.

La saisonnalité hebdomadaire est également présente dans les données des pertes (voir la figure 5.22). La figure 5.22b montre la différence de la distribution des données des pertes entre les jours de week-end et les autres jours de la semaine. Une diminution importante des pertes est observée les week-end par rapport aux autres jours de la semaine où les pertes suivent des profils très similaires et réguliers (voir la figure 5.22a). Ceci est dû au fait que l'augmentation de la demande d'électricité liée au cycle économique entraîne l'augmentation des pertes également.



(a) Exemple de cycle hebdomadaire des pertes.

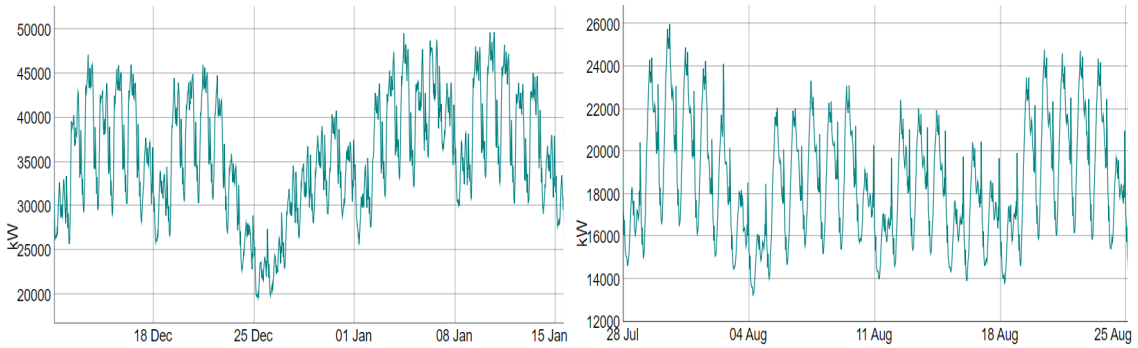


(b) Les distributions des pertes selon les jours de la semaine.

FIGURE 5.22 – Exemples de la saisonnalité hebdomadaire des données de pertes.

Le cycle économique a un fort impact sur les pertes (voir la figure 5.23). En effet,

la diminution de la consommation les jours de week-end, la diminution du niveau moyen des pertes pendant les périodes des vacances d'été (voir la figure 5.23b) et les vacances de Noël (voir la figure 5.23a) ainsi que pendant les jours fériés reflètent cet impact.



(a) Semaine des vacances de Noël en 2015. (b) Semaine des vacances d'été en août 2015.

FIGURE 5.23 – Exemples de la diminution des pertes pendant les périodes des vacances.

5.5.4 Modèles et approche de prévision

Notre objectif est alors de prédire les pertes d'électricité à $(J+1)$ en utilisant l'historique des pertes, tout en considérant les caractéristiques de la courbe de charge des pertes qui ont été présentées précédemment. Comme les données à notre disposition ne comportent pas de valeurs manquantes, la phase de préparation des données se limite à la division du jeu de données en un ensemble d'entraînement et un ensemble de test ainsi que la détermination des variables d'entrée de chaque modèle. Cette dernière étape consiste à sélectionner les variables pertinentes à inclure dans le modèle, en fonction de leur impact sur les pertes d'électricité et de leur capacité à fournir des informations utiles pour la prévision en tenant compte des fondements théoriques de chaque modèle et la façon dont il traite les données.

Nous avons donc divisé le jeu de données en 80% de données d'entraînement et 20% de données test. Dans cette partie, nous utilisons l'erreur MAPE (définition dans la sous-section 3.4.1) pour évaluer la précision des différents modèles de prévision et les comparer à la méthode utilisée chez le fournisseur. Étant donné que l'erreur MAPE est également utilisée chez le fournisseur pour évaluer la précision de leurs prévisions, cela nous permettra de faire une comparaison pertinente entre les performances de nos modèles et celles de la méthode existante. En fonction des caractéristiques des données décrites dans la sous-section 5.5.3, nous avons identifié les variables d'entrée suivantes :

1. **Les pertes décalées** : nous avons utilisé les pertes décalées d'un jour (P_{t-48}) et d'une semaine (P_{t-336}). Ces deux variables correspondent aux pertes d'électricité enregistrées à la même pas de temps, mais le jour précédent et la semaine précédente respectivement. L'inclusion de ces deux variables peut permettre de prendre

en compte la saisonnalité journalière et hebdomadaire des données de pertes dans la prévision ;

2. **Les températures extérieures** : pour modéliser la relation entre les pertes et les températures extérieures, nous avons utilisé plusieurs variables liées à la température, notamment les températures extérieures réelles (T_t), les températures décalées d'un jour (T_{t-48}), la moyenne journalière de la température du jour à prédire (T_{mean}), la température maximale (T_{max}) et minimale (T_{min}) du jour précédent le jour à prédire ainsi que les températures lissées ($T_t^{(\gamma)}$) de T_t avec un paramètre de lissage exponentiel $\gamma \in [0, 1]$. Elle est définie à l'instant t par : $T_t^{(\gamma)} = \gamma T_{t-1}^{(\gamma)} + (1 - \gamma)T_t$.
   En utilisant ces variables, nous pouvons capturer les tendances et les changements dans les températures extérieures qui peuvent affecter les pertes électriques ;
3. **Les variables calendaires** : ces variables permettent de saisir les effets des jours fériés, des jours de la semaine et des mois sur les données. Nous avons utilisé des variables binaires pour les jours fériés (*Holidays*), des variables factorielles pour les jours de la semaine avec 5 niveaux correspondant à différents types de jour (*Weekday*) : lundi, mardi-mercredi-jeudi, vendredi, samedi, dimanche, les variables factorielles pour les mois (*Month*) ainsi que la position d'un jour donné dans l'année en termes de jour numéroté (*Posan*). Elles sont utiles pour capter les tendances saisonnières dans les données et ajuster les modèles de prévision en conséquence ;
4. **La variable de temps** (*Time*) : cette variable est utilisée pour modéliser les différents pas de temps. Dans le cas des données mesurées à intervalles de demi-heure, elle prend des valeurs entre 1 et 48. En utilisant cette variable, il est possible d'identifier des motifs récurrents dans des plages temporelles spécifiques, ce qui peut améliorer la performance des modèles de prévision ;
5. **Les variables saisonnières de Fourier** (VarFourier) : les coefficients de Fourier sont utilisés comme variables d'entrée dans le modèle de prévision pour modéliser la saisonnalité annuelle dans les données. Ces coefficients sont calculés à partir de fonctions trigonométriques et permettent de décomposer la variation saisonnière annuelle en termes de fréquences ;
6. **La variable modélisant l'heure d'été et l'heure d'hiver** (*Daylight*) : nous avons introduit une variable binaire pour représenter le changement d'heure entre l'été et l'hiver. Cette variable prend une valeur de 0 pour l'hiver et une valeur de 1 pour l'été. En ajoutant cette variable dans les modèles de prévision, nous espérons qu'ils pourront mieux capturer les variations saisonnières liées aux changements d'heure ;

En utilisant les variables d'entrée présentées précédemment, nous avons adapté plusieurs modèles pour prévoir les pertes d'électricité. Il convient de noter que les variables présentées précédemment ne sont pas nécessairement toutes utilisées dans tous les modèles. Pour chaque modèle, nous avons choisi les variables qui maximisent sa performance. Les modèles testés sont les suivants :

1. Le modèle de ***Hong*** : ce modèle est couramment utilisé comme modèle de référence pour la prévision de la charge électrique. Comme les données de pertes électriques partagent des caractéristiques similaires avec les données de la charge électrique, le modèle de prévision de la charge électrique de *Hong*, décrit dans la partie 3.2.1.1, peut également être utilisé pour prédire les pertes électriques. Toutefois, afin d'améliorer sa capacité à prédire les pertes à court terme, nous l'avons modifié en y ajoutant les variables de pertes décalées d'un jour et d'une semaine (P_{t-48} et P_{t-336} respectivement) et les températures décalées d'un jour et d'une semaine (T_{t-48} et T_{t-336} respectivement). L'équation dans (3.6) devient alors :

$$\begin{aligned}P_t = {} & \beta_0 + \beta_1 t + \beta_2 Time_t \times T_t + \beta_3 Time_t \times T_t^2 + \beta_4 Time_t \times T_t^3 \\ & + \beta_5 Weekday_t \times Time_t + \beta_6 Month_t + \beta_7 Month_t \times T_t + \beta_8 Month_t \times T_t^2 \\ & + \beta_9 Month_t \times T_t^3 + \beta_{10} P_{t-48} + \beta_{11} T_{t-48} + \beta_{12} Time_t \times T_{t-48} + \beta_{13} Time_t \times T_{t-48}^2 \\ & + \beta_{14} Time_t \times T_{t-48}^3 + \beta_{15} Month_t \times T_{t-48} + \beta_{16} Month_t \times T_{t-48}^2 + \beta_{17} Month_t \times T_{t-48}^3 \\ & + \beta_{18} P_{t-336} + \beta_{19} T_{t-336} + \beta_{20} Time_t \times T_{t-336} + \beta_{21} Time_t \times T_{t-336}^2 + \beta_{22} Time_t \times T_{t-336}^3 \\ & + \beta_{23} Month_t \times T_{t-336} + \beta_{24} Month_t \times T_{t-336}^2 + \beta_{25} Month_t \times T_{t-336}^3 + \epsilon_t.\end{aligned}$$

2. Le modèle ***KWF*** : l'hypothèse selon laquelle des conditions similaires dans le passé entraînent des conditions similaires dans le futur peut également s'appliquer aux données de pertes, nous avons donc utilisé le modèle *KWF* pour prédire les pertes. Le modèle intègre les groupes déterministes proposés dans (Cugliari, 2011), qui prennent en compte diverses informations du calendrier telles que les jours de la semaine, les jours fériés et les jours de transition. Ces groupes peuvent être résumés comme suit : les jours de la semaine de lundi à dimanche, avec les mardis, mercredis et jeudis considérés comme un seul groupe. De plus, nous prenons également en compte les jours précédant et suivant les jours fériés pour permettre au modèle de tenir compte de l'effet spécifique du type de jour férié sur les données. En effet, l'impact d'un jour férié peut varier en fonction du jour de la semaine où il tombe, ce qui peut influencer les résultats de la prévision. Par exemple, un jour férié tombant un jeudi peut avoir un effet différent de celui tombant un lundi ou un autre jour.
3. Un modèle ***GAM*** : nous avons utilisé le `package mgcv` du logiciel R pour implémenter plusieurs modèles *GAM* visant à prévoir les pertes électriques en manipulant les variables d'entrée et en ajustant la relation entre ces variables. Nous présentons ici le modèle le plus performant sélectionné parmi les différents modèles conçus dans ce but. Nous avons ajusté un modèle pour chaque tranche de demi-heure de la journée (soit 48 modèles au total), comme dans les études (Pierrot et al., 2011) et (Gaillard et al., 2016). Ce modèle a été traité en parallèle, avec un modèle par cœur de calcul, ce qui a considérablement réduit le temps de calcul. Bien que nous ayons tenté de créer un modèle unique pour capturer la structure temporelle des données et la corrélation entre les tranches de temps de la journée, nous avons constaté de meilleurs résultats de précision en utilisant un modèle distinct pour chaque tranche

de demi-heure. Nous avons également essayé d'utiliser une transformation logarithmique des données, mais n'avons pas constaté d'amélioration significative, donc nous avons opté pour l'utilisation des données brutes. Si nous considérons la série temporelle des pertes à l'instant t, notée P_t, l'équation du modèle peut être exprimée comme suit pour chaque tranche de demi-heure de la journée dh dans $1, \ldots, 48$:

$$\begin{aligned} P_t^{dh} &= m_1^{dh}(Weekday) + m_2^{dh}(T_t^{(0,95)}) + m_3^{dh}(T_{min}, T_{max}) \\ &+ m_4^{dh}(T_t^{(0,99)}) + m_5^{dh}(P_{t-48}, P_{t-336}) \\ &+ m_6^{dh}(Posan)\mathbb{1}_{Daylight} + m_7^{dh}(T_t) + \text{Holidays}. \end{aligned}$$

Les m_i^{dh} sont les fonctions de lissage des *splines* de régression à estimer par le modèle *GAM*. Les températures lissées avec des paramètres de lissage exponentiel de $0,95$ et $0,99$ sont respectivement $T_t^{(0,95)}$ et $T_t^{(0,99)}$.

4. Le modèle de **forêt aléatoire** : nous avons utilisé le `package randomForest` du logiciel R pour ajuster un modèle de forêt aléatoire pour chaque tranche de demi-heure, en utilisant les variables d'entrée suivantes : P_{t-48}, P_{t-336}, T_t, T_{mean}, *Time*, *Weekday*, *Month*, *VarFourier*, *Holidays*. Ces modèles ont également été traités en parallèle pour diminuer le temps de calcul.
5. Le modèle ***MARS*** : nous avons utilisé le `package earth` du logiciel R pour ajuster le modèle en utilisant toutes les variables présentées précédemment. La recherche des hyperparamètres optimaux, tels que le degré maximal des interactions entre les variables dans le modèle *degree* et le nombre de variables à élaguer (*nprune*), a été réalisée en effectuant une recherche de grille (*grid search*). Cette méthode a testé toutes les combinaisons de paramètres spécifiées dans la grille afin de trouver les hyperparamètres qui donnent les meilleures performances du modèle en termes de précision. La procédure de sélection automatique du modèle *MARS* (voir 3.2.1.3) a permis de sélectionner automatiquement les variables les plus pertinentes parmi celles proposées. Les variables ont été sélectionnées selon leur ordre d'importance, et voici la liste des variables retenues : P_{t-48}, *Weekday*, T_{mean}, *Holidays*, *Time*, T_t, T_{t-48} et *VarFourier*.

5.5.5 Résultats

Dans la table 5.6, les performances de différents modèles de prévision pour l'année 2019 sont comparées en termes d'erreur MAPE pour la prévision des pertes électriques à $(J+1)$. Les résultats indiquent que le modèle *GAM* présente la meilleure performance en termes de MAPE, suivi de près par le modèle *MARS*.

	GAM	*MARS*	*RF*	*KWF*	*Hong*
MAPE%	2,75	3,85	3,99	4,59	5,70

TABLE 5.6 – Les performances des modèles de prévision des pertes à $(J+1)$: *GAM*, *MARS*, forêt aléatoire (*RF*), *KWF* et modèle de *Hong*.

La figure 5.24 illustre les courbes prévisionnelles des pertes électriques par le modèle *GAM* pour la semaine de 2 à 9 septembre 2019. Nous pouvons remarquer que le modèle est capable de prendre en compte les variations entre les jours de la semaine et les *weekends*, et les différences entre les prévisions et les données réelles de pertes sont relativement faibles, avec une erreur MAPE de 1,7% pour cette semaine.

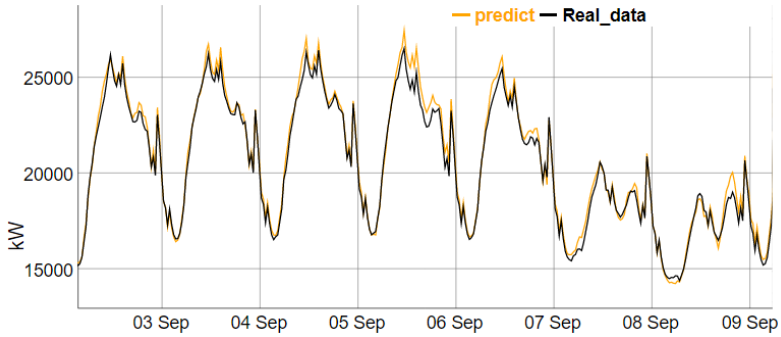


FIGURE 5.24 – Exemple des courbes prévisionnelles par le modèle *GAM* pour la semaine de 2 à 9 septembre 2019, avec les prévisions en orange et les données réelles de pertes en noir.

Nous avons reçu des retours positifs de la part des experts du fournisseur d'énergie concernant les modèles de prévision que nous avons proposés, en particulier les deux modèles les plus performants *GAM* et *MARS*. Ils ont notamment souligné la précision, la rapidité et l'interprétabilité de ces modèles ainsi que leur capacité à modéliser efficacement l'impact de la température (voir la figure 5.25) et des jours fériés (voir la figure 5.26) sur les pertes électriques.

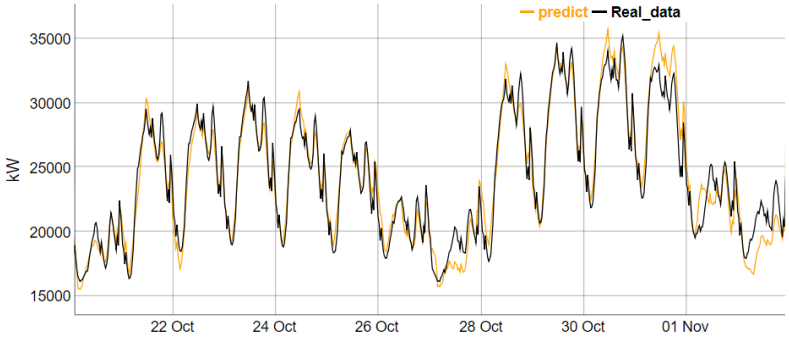


FIGURE 5.25 – Extrait des courbes prévisionnelles montrant la capacité d'adaptation du modèle *GAM* aux variations de tendance liées à la baisse de température : la courbe orange représente les prévisions tandis que la courbe noire correspond aux données réelles de pertes électriques.

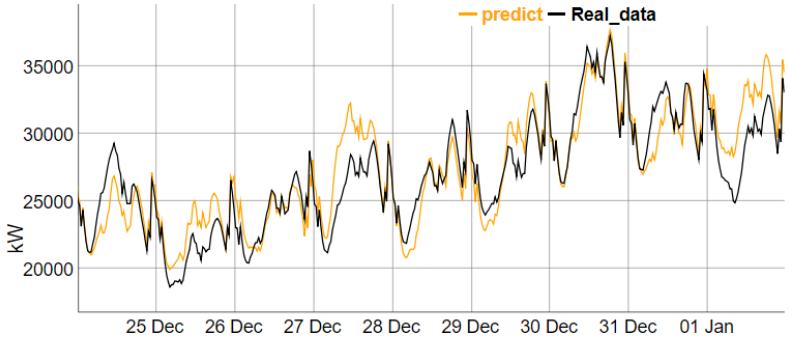
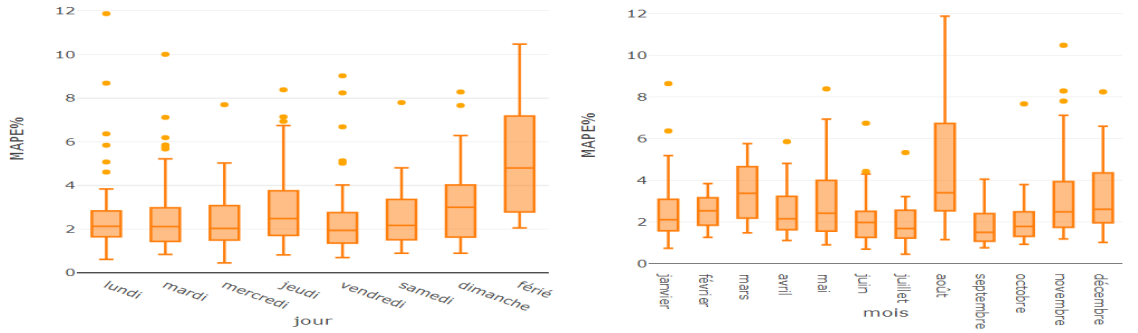


FIGURE 5.26 – Extrait des courbes prévisionnelles de la semaine de Noël en 2019 par le modèle *GAM* montrant l'adaptabilité du modèle *GAM* aux fluctuations des pertes durant les jours fériés : la courbe orange représente les prévisions tandis que la courbe noire correspond aux données réelles de pertes électriques.

Bien que le modèle *GAM* permette de détecter et de prédire les effets du calendrier, la figure 5.27 montre que les erreurs de prévision peuvent varier en fonction du type de jour et du mois. Les distributions des erreurs MAPE de prévision à $(J+1)$ du modèle *GAM* sont représentées pour différents types de jours (jours de la semaine, week-ends et jours fériés) et différents mois sur la figure 5.27. La figure 5.27a montre que le modèle de prévision est sensible à certains jours de la semaine, en particulier les dimanches et les jours fériés, pour lesquels les erreurs de prévision sont plus importantes. En effet, les erreurs de prévision peuvent rester élevées même si les jours fériés et les dimanches ont été intégrés dans les modèles de prévision en raison de la variabilité accrue de ces jours par rapport aux autres jours, ce qui peut rendre la prévision plus difficile. En analysant la distribution de l'erreur de prévision en fonction des mois, il a été observé que les erreurs sont plus importantes pendant le mois d'août. Cette observation s'explique par le fait que le mois d'août est traditionnellement un mois de vacances en France, ce qui peut entraîner des changements significatifs dans les habitudes de consommation d'électricité des ménages et des entreprises. Par exemple, les entreprises peuvent fermer pour quelques semaines,

réduisant ainsi leur consommation d'électricité. De même, de nombreuses familles partent en vacances et réduisent leur consommation d'électricité à leur domicile. Cette variabilité dans les habitudes de consommation d'électricité peut donc entraîner une grande variabilité dans les pertes, rendant leur prévision pendant cette période plus difficile.



(a) Distribution des erreurs MAPE en fonction du type de jour.

(b) Distribution des erreurs MAPE en fonction des mois.

FIGURE 5.27 – Les distributions des erreurs de prévisions à $(J+1)$ du modèle *GAM* classées par type de jour et par mois.

Après avoir comparé les performances de nos modèles de prévision *GAM*, *MARS*, et *RF* à celles de la méthode manuelle utilisée chez le fournisseur d'énergie pour l'année 2019, les résultats ont montré que nos trois modèles étaient plus précis pour la prévision à $(J+1)$ des pertes électriques, avec des erreurs MAPE respectives de 2,75%, 3,85% et 3,99%, comparé à l'erreur MAPE de la méthode manuelle qui était de 4,13%. En plus d'améliorer la précision de la prévision, l'utilisation des modèles de prévision permet également de gagner du temps car leur processus d'ajustement et de prévision ne prend pas plus d'une minute, contrairement à la méthode manuelle qui nécessite 15 minutes pour l'ajustement et le calcul des prévisions.

Cependant, il est important de noter que les modèles de prévision ont leurs limites, en particulier lorsqu'il s'agit d'événements imprévus et exceptionnels tels que le confinement lié à la pandémie de Covid-19. Ainsi, pour l'année 2020, la méthode manuelle s'est avérée plus performante, en particulier pour la prévision des périodes de confinement (15 mars - 11 mai) et (29 octobre - 15 décembre). Le tableau 5.7 compare les performances des modèles *GAM* et *MARS* ainsi que la méthode manuelle utilisée chez le fournisseur d'énergie pour l'année 2020.

	méthode manuelle	*GAM*	*MARS*
2020	3,81%	4,90%	4,99%
Périodes de confinement 2020	4,53%	9,82%	7,04%

TABLE 5.7 – Les résultats de la prévision des pertes électriques à $(J+1)$ pour l'année 2020, y compris les périodes de confinement (15 mars - 11 mai) et (29 octobre - 15 décembre) en termes d'erreur MAPE pour les modèles de prévision *GAM*, *MARS*, et la méthode manuelle du fournisseur.

Les résultats dans la table 5.7 montrent que la méthode manuelle est la plus performante pour l'année 2020, avec un MAPE de 3,81%. En effet, pour les périodes de confinement, les modèles de prévision ont des MAPE beaucoup plus élevés que la méthode manuelle, en particulier pour le modèle *GAM* avec un MAPE de 9,82%. La période de confinement a eu un impact significatif sur les performances des modèles de prévision. La figure 5.28 illustre cet effet en montrant les erreurs de prévision MAPE du modèle *MARS* et de la méthode manuelle pour l'année 2020. Nous pouvons constater que, pour la période de début d'année et avant le confinement, le modèle *MARS* était plus performant que la méthode manuelle. Cependant, à partir du confinement, la performance du modèle *MARS* s'est dégradée car il ne pouvait pas tenir compte de ces changements soudains dans les données contrairement à la méthode manuelle utilisée chez le fournisseur qui se base sur l'expertise humaine pour ajuster les prévisions en fonction des événements passés. Cette approche permet de modifier rapidement les prévisions en conséquence. C'est pourquoi, dans le contexte de la pandémie de Covid-19, la méthode manuelle s'est avérée plus performante que les modèles de prévision basés sur les données historiques uniquement.

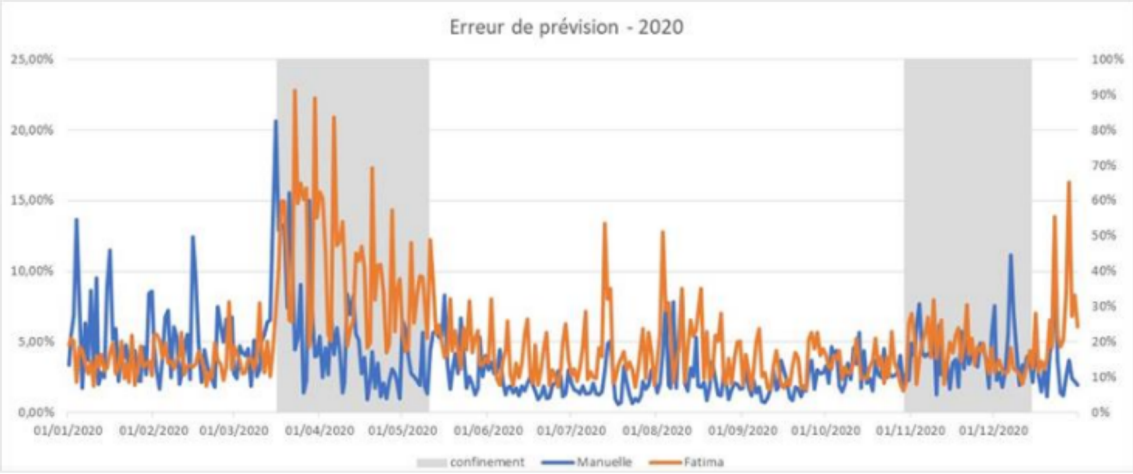
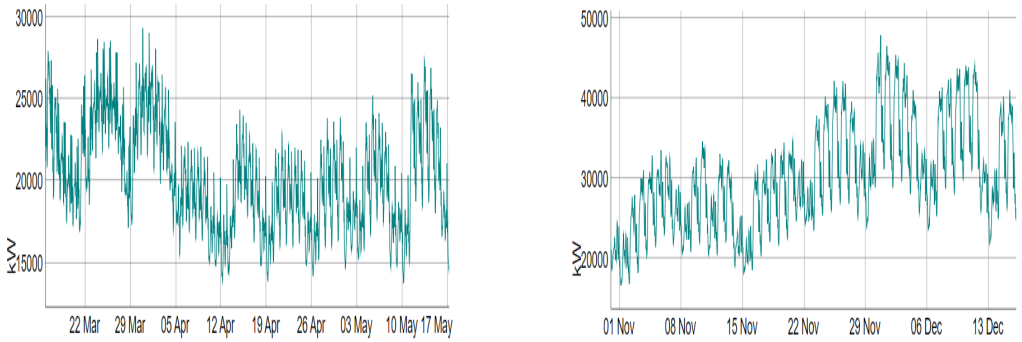


FIGURE 5.28 – Les erreurs de prévision des pertes en 2020 : la courbe orange représente les erreurs MAPE de prévision par notre modèle *MARS* tandis que la courbe bleue correspond aux erreurs de la méthode manuelle chez le fournisseur. Les intervalles grisés indiquent les périodes de confinement en France (17 mars - 11 mai) et (30 octobre - 15 décembre) [Source ingénieure d'achat chez le fournisseur].

Dans le but d'améliorer les performances des modèles de prévision en période de confinement, différentes approches ont été explorées. Tout d'abord, l'utilisation de variables fictives a été mise en place pour indiquer si une période donnée correspondait ou non à une période de confinement dans le modèle *MARS*. Ensuite, pour tenir compte des informations récentes, la médiane des pertes de la semaine précédente a été ajoutée dans le modèle *GAM*. D'autres méthodes de prévision ont également été testées, comme le modèle Prophet. Ce modèle développé par "Facebook" en 2017 (S. J. Taylor et al., 2018) permet de capturer les tendances non linéaires, les effets saisonniers et les événements exceptionnels. Il permet également de modéliser des points de changement de la tendance grâce à l'ajout de covariables appelées *changepoints*. Ces covariables permettent de spécifier des moments dans le temps où la tendance de la série peut changer de manière significative. Le modèle Prophet utilise ensuite une approche bayésienne pour estimer les paramètres du modèle en prenant en compte les points de changement de la tendance. Les utilisateurs peuvent fixer manuellement les dates des *changepoints* ou laisser le modèle les estimer automatiquement. La caractéristique de fixer des points de changement de tendance dans le modèle Prophet peut être utile pour la prévision de la période de confinement. En effet, nous pouvons fixer un point de changement de tendance à la date de début du confinement pour que le modèle prenne en compte cette information dans sa prévision. L'utilisation de ce modèle a permis d'améliorer légèrement la prévision de la période de confinement en réduisant l'erreur MAPE à 6,78%. Toutefois, cette amélioration est relativement modeste par rapport à celle obtenue avec la méthode manuelle.

Cependant, il n'a pas été possible de valider l'amélioration de la précision envisagée par l'intégration des variables fictives et des médianes de la semaine précédente dans les modèles *MARS* et *GAM*. En effet, ces approches sont difficiles à évaluer en raison de l'impact variable du confinement sur les pertes. Le premier confinement, qui a eu lieu du 17 mars au 11 mai (voir la figure 5.29a), a eu un impact plus fort que celui de la période entre le 30 novembre et le 15 décembre (voir la figure 5.29b) sur les pertes. De plus, cet impact variait d'une semaine à une autre, pour la même période (voir la figure 5.29a). Par conséquent, l'intégration de variables fictives, de médianes de la semaine précédente ou de points de changement dans le modèle Prophet ne permettra probablement pas de surpasser les performances des prévisions manuelles à court terme. Ces modèles statistiques ont besoin d'une période d'entraînement suffisante avant de pouvoir s'adapter à prédire une situation inhabituelle, ce qui rend difficile l'évaluation de l'efficacité de ces approches. Une possibilité serait d'utiliser ces approches pour améliorer l'entraînement du modèle en vue de prévoir des situations similaires qui pourraient se produire dans le futur. Une intervention manuelle d'experts occasionnelle peut être nécessaire pour ajuster les modèles statistiques pendant des périodes anormales telles que le confinement.



(a) Période de confinement du 17 mars au 11 mai 2020.

(b) Période de confinement du 30 octobre au 15 décembre 2020.

FIGURE 5.29 – L'impact des périodes de confinement sur les pertes d'électricité.

5.5.6 Intégration dans un logiciel de prévision

Nous avons intégré le modèle de prévision *MARS* dans un logiciel de prévision utilisé chez le fournisseur. Ce logiciel permet à l'utilisateur d'importer les données de la source et de les visualiser. Il permet également de sélectionner la période de temps à prédire et de choisir l'historique de données à utiliser pour entraîner le modèle de prévision. Une fonctionnalité d'évaluation de la qualité des prévisions est également disponible, en comparant les résultats du modèle avec les données réelles sur une période définie (voir la figure 5.30).

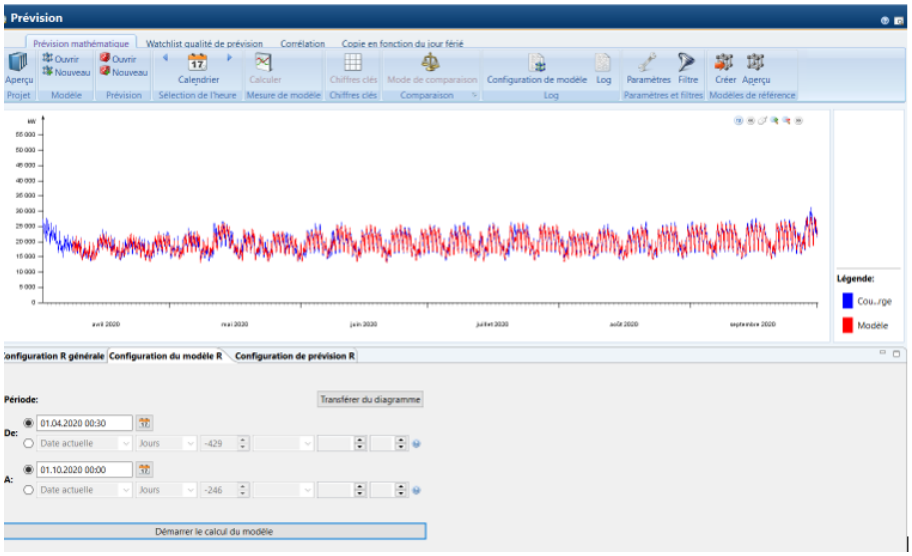


FIGURE 5.30 – Capture d'écran du logiciel de prévision utilisé par le fournisseur d'énergie montrant les courbes prévisionnelles générées par notre modèle de prévision *MARS*, ainsi que les données réelles des pertes d'électricité.

Cette intégration a été réalisée à travers deux scripts R : un pour l'entraînement du modèle et l'autre pour la génération de prévisions. L'utilisateur doit fournir la période et

l'horizon de prévision souhaités. Le modèle est entraîné avec les données sélectionnées et sauvegardé sous forme d'objet R. Pour la prévision, l'utilisateur peut choisir entre deux options : prédire à partir du modèle pré-entraîné sauvegardé ou effectuer une prévision en temps réel en entraînant le modèle avec les données d'entrée fournies. Le modèle est actuellement en phase de test chez le fournisseur sur une période prolongée afin de pouvoir évaluer sa fiabilité et sa rentabilité sur une durée plus étendue.

5.6 Prévision de la courbe de charge du réseau de distribution

5.6.1 Objectif et intérêts

La prévision de la courbe de charge du réseau de distribution d'un fournisseur électrique est une tâche importante dans la planification et l'optimisation de la production d'énergie électrique. Cette prévision permet aux fournisseurs d'électricité d'ajuster leur production en conséquence et de répondre aux besoins de leurs clients de manière plus efficace.

Dans le but de satisfaire les exigences du fournisseur d'énergie et en restant dans le domaine de la prévision de la charge électrique et ses applications, nous avons également abordé la prévision de la charge électrique de distribution chez le fournisseur d'énergie. Nous avons envisagé la possibilité d'adapter les modèles de prévision que nous avons mis en œuvre pour la prévision de la charge électrique à d'autres applications et portefeuilles de clients, afin de prédire la charge électrique du réseau de distribution.

Comme pour la prévision des pertes, la méthode actuelle de prévision de la charge électrique du réseau de distribution utilisée chez le fournisseur d'énergie implique des ajustements manuels effectués par les opérateurs en charge de la gestion de la demande d'électricité, qui modifient la charge électrique du jour précédent en fonction de la température et du calendrier. Ces ajustements manuels visent à tenir compte des variations de la charge électrique causées par des facteurs tels que la température extérieure et les jours fériés mais se font d'une manière imprécise. Par conséquent, les outils de prévision plus avancés basés sur des modèles statistiques ou d'apprentissage automatique peuvent améliorer la précision des prévisions. Cela permettrait de mieux anticiper les variations de la demande d'électricité, mieux gérer la production et la distribution d'électricité et d'automatiser le processus de prévision, réduisant ainsi le temps et les ressources nécessaires à l'ajustement manuel.

5.6.2 Description des données

Les données mises à notre disposition pour cette étude comprennent l'historique de la charge du réseau de distribution du fournisseur, enregistré toutes les demi-heures, sur une période s'étendant de janvier 2017 à janvier 2020, ainsi que les données historiques des températures extérieures pour la même période et la même fréquence (voir la figure 5.31).

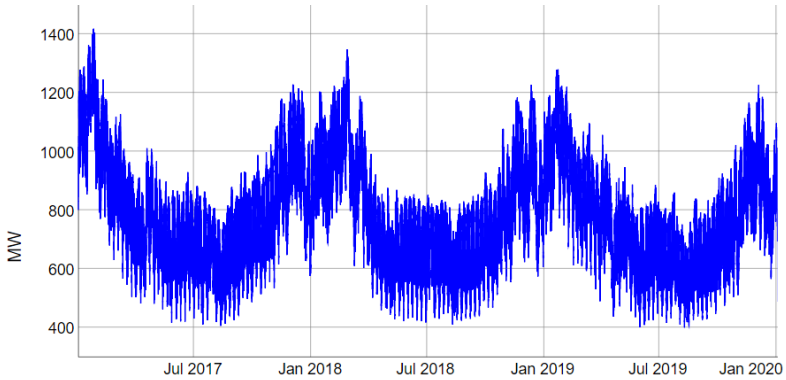


FIGURE 5.31 – Courbe de charge électrique du réseau de distribution pour la période du 1 janvier 2017 au 1 janvier 2020.

Les données fournies présentaient un taux de 1,85% de données manquantes (voir la figure 5.32). Ces données manquantes de la charge électrique du réseau de distribution peuvent être causées par divers facteurs, tels que des problèmes de communication et des erreurs de transmission de données.

Étant donnée que les valeurs manquantes dans les séries temporelles peuvent affecter la qualité de l'analyse et de la prévision des données. Leur présence peut entraîner une baisse de la précision des prévisions, et affecter la performance des modèles de prévision. Il est donc important de traiter ces valeurs avant de procéder à la prévision.

Le traitement des données manquantes est un sujet vaste et complexe qui nécessite une analyse approfondie. Cependant, dans le cadre de cette étude, notre objectif était de mettre en place des modèles de prévision de la charge électrique et non pas de déterminer la meilleure méthode de traitement des données manquantes. Nous avons donc proposé une méthode d'imputation en fonction des caractéristiques des données et de notre expérience. Il serait judicieux de réaliser une étude plus rigoureuse en comparant différentes méthodes d'imputation de données manquantes à l'avenir.

Les méthodes d'imputation des données manquantes pour les séries temporelles peuvent être divisées en deux catégories principales : les méthodes de substitution et les méthodes de modélisation. Les méthodes de substitution consistent à remplacer les valeurs manquantes par des valeurs observées comme par exemple la méthode de la moyenne mobile qui consiste à remplacer chaque valeur manquante par la moyenne de la valeur précédente

et de la valeur suivante. Cette méthode est simple à mettre en œuvre, mais peut être moins précise que les méthodes de modélisation qui consistent à utiliser un modèle de la série temporelle comme les modèles *ARIMA*, les modèles de régression pour estimer les valeurs manquantes. Ces méthodes sont plus complexes à mettre en œuvre, mais peuvent fournir des estimations plus précises des valeurs manquantes.

Nous avons choisi d'utiliser une méthode qui se base sur la décomposition saisonnière et convient particulièrement aux séries temporelles avec une saisonnalité marquée et un faible taux de données manquantes, ce qui est le cas dans notre étude. Elle implique de supprimer la saisonnalité de la série temporelle, d'appliquer une méthode d'imputation classique telle que l'interpolation ou la moyenne sur la série désaisonnalisée, puis de réintégrer la saisonnalité pour obtenir une série complète. Cette méthode est implémentée dans la fonction `na_seadec`() du `package imputeTS` du logiciel R (voir la figure 5.32).

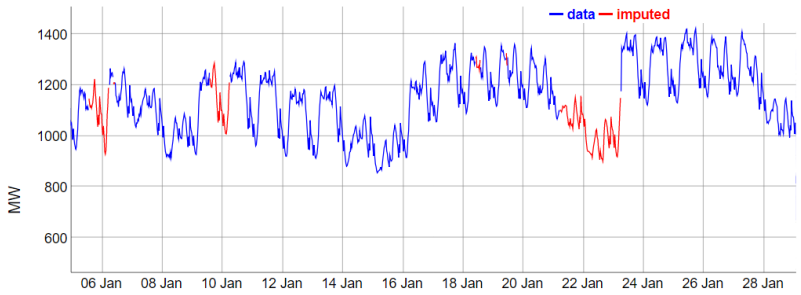


FIGURE 5.32 – Courbe de charge électrique du réseau de distribution montrant l'imputation des données manquantes par `na_seadec`(). Les valeurs de charge électrique sont représentées en bleu, tandis que les parties en rouge correspondent aux données manquantes qui ont été imputées.

Les caractéristiques des données de la charge électrique du réseau de distribution sont similaires à celles des pertes décrites dans la sous-section 5.5.3, telles que la saisonnalité journalière, hebdomadaire et annuelle, ainsi que la dépendance à la température ou thermosensibilité. Il est donc raisonnable de supposer que les modèles qui ont bien prédit les pertes pourraient également être efficaces pour prédire la charge électrique. C'est pourquoi nous avons testé les modèles qui ont obtenu les meilleurs résultats pour la prévision des pertes afin de prédire la charge électrique du réseau de distribution du fournisseur. Cependant, étant donné l'absence d'un historique des prévisions calculées par la méthode du fournisseur, nous avons sollicité les retours des experts du domaine pour évaluer la qualité de nos prévisions.

5.6.3 Résultats

Les deux premières années, 2017 et 2018, ont été utilisées pour entraîner les modèles, tandis que l'année 2019 a été utilisée pour le test. Les résultats de la prévision pour les

horizons $(J+1)$ et $(J+7)$ ont été obtenus à l'aide de trois modèles : *GAM*, *MARS* et forêt aléatoire *RF*. Ces modèles ont été construits de la même manière que les modèles de prévision des pertes décrits dans la sous-section 5.5.4. Pour la prévision à $(J+7)$, seules les variables ne dépendant pas du jour précédant la prévision ont été incluses, car ces données ne seraient pas disponibles pour prédire la charge à $(J+7)$. La table 5.8 montre les performances des modèles de prévision de la charge électrique du réseau à $(J+1)$ et $(J+7)$ pour l'année 2019, mesurées par la métrique MAPE.

	GAM	*MARS*	*RF*
$(J+1)$	2.97	3,65	4,02
$(J+7)$	3,95	4,70	4,92

TABLE 5.8 – Performance des modèles de prévision de la charge électrique du réseau à $(J+1)$ et $(J+7)$ mesurées par la métrique MAPE : *GAM*, *MARS* et forêt aléatoire (*RF*) pour l'année 2019.

Les résultats montrent que le modèle *GAM* a la meilleure performance pour la prévision à $(J+1)$ avec un taux d'erreur de seulement 2,97%, suivi du modèle *MARS* avec 3,65% et de la forêt aléatoire avec 4,02%. Pour la prévision à $(J+7)$, le modèle *GAM* est encore le plus performant avec une erreur de 3,95%, suivi de *MARS* avec 4,70% et de la forêt aléatoire avec 4,92%. En général, les trois modèles ont des performances assez bonnes avec des taux d'erreur relativement faibles. Nous pouvons remarquer que les performances des modèles de prévision de la charge électrique du réseau sont légèrement meilleures à $(J+1)$ qu'à $(J+7)$ pour chaque modèle. Cela peut s'expliquer par le fait que plus on s'éloigne de la date de prévision, plus il y a d'incertitudes et de variations possibles dans les données d'entrée, ce qui peut rendre la prévision plus difficile. Les changements saisonniers, les événements météorologiques extrêmes, les vacances, les jours fériés, les grèves et autres perturbations imprévues dans le système électrique sont autant de facteurs qui peuvent expliquer ces incertitudes. De plus, la prévision à $(J+7)$ nécessite de supprimer les variables relatives au jour précédent, ce qui peut réduire la précision des modèles. En effet, cette variable peut intégrer les changements récents qui peuvent avoir un impact significatif sur la charge électrique. Par conséquent, en retirant cette variable, les modèles peuvent perdre de l'information pertinente pour la prévision de la charge électrique, ce qui peut expliquer en partie la dégradation des performances observées pour la prévision à $(J+7)$. Cependant, les différences de performances entre les trois modèles ne sont pas très significatives et ils ont tous démontré leur capacité à prédire la charge électrique du réseau avec une erreur MAPE de moins de 5% à $(J+1)$ et $(J+7)$.

Nous avons également calculé des intervalles de prévision pour le modèle de prévision de la charge électrique du réseau, notamment pour le modèle *GAM* qui a affiché les meilleures performances. Ces intervalles permettent d'évaluer la marge d'erreur possible des prévisions et sont donc un outil précieux pour le fournisseur d'énergie. En effet, ils

permettent d'aider les opérateurs à prendre des décisions plus éclairées en prenant en compte les incertitudes associées aux prévisions. Les intervalles de prévision ont été calculés par la méthode de régression de quantile décrites dans la section 5.4 avec un niveau de confiance de 95%.

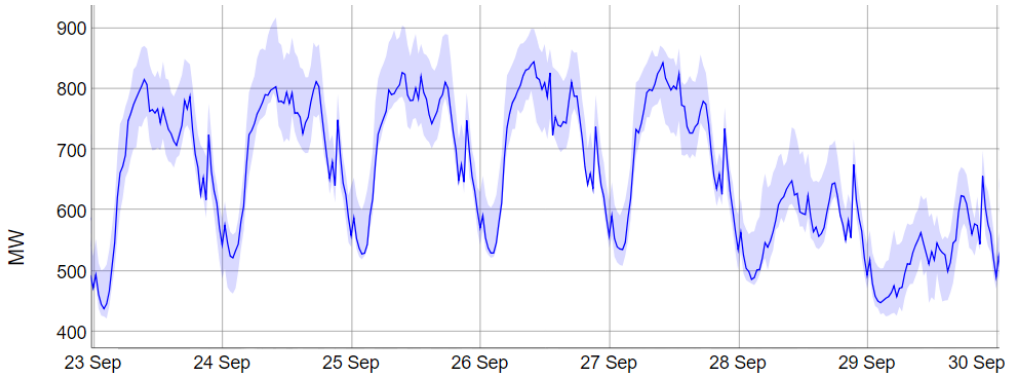


FIGURE 5.33 – Intervalles de prévision à $(J+1)$ à 95% de niveau de confiance pour la courbe de charge du réseau de distribution du 23 septembre au 30 septembre 2019, calculés par la méthode de régression du quantile pour le modèle *GAM*. La courbe en bleu correspond aux données réelles, tandis que l'effet d'ombre représente le tube de l'intervalle de prévision.

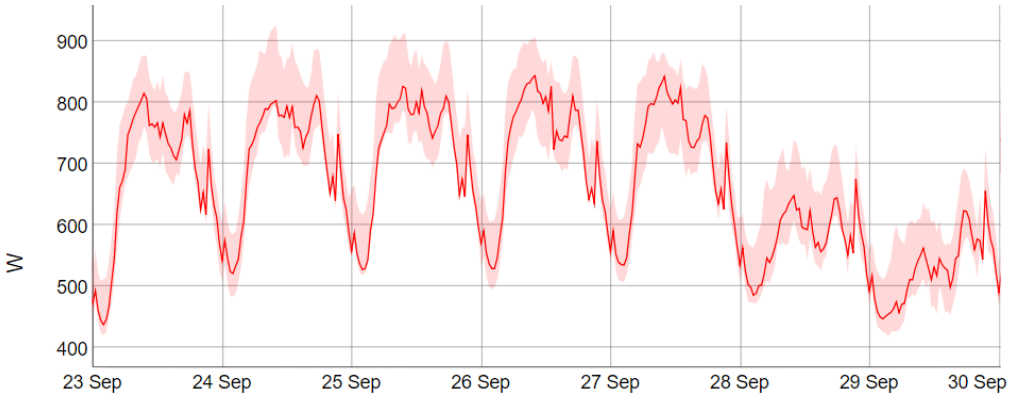


FIGURE 5.34 – Intervalles de prévision à $(J+7)$ à 95% de niveau de confiance pour la courbe de charge du réseau de distribution du 23 septembre au 30 septembre 2019, calculés par la méthode de régression du quantile pour le modèle *GAM*. La courbe en rouge correspond aux données réelles, tandis que l'effet d'ombre représente le tube de l'intervalle de prévision.

Le taux de couverture définit dans la section 5.4 a été utilisé pour évaluer la qualité et la fiabilité des intervalles de prévision. Le résultat obtenu montre un taux de couverture empirique égal à 96,4% pour les intervalles de prévision à $(J+1)$. Cela signifie que dans 96,4% des cas, les valeurs réelles se trouvent effectivement à l'intérieur de l'intervalle de prévision calculé. C'est une indication importante de la qualité de notre modèle de prévision de charge électrique et de l'efficacité de la méthode de régression du quantile pour calculer les intervalles de prévision. La figure 5.33 illustre un exemple des intervalles

de prévision calculés à $(J + 1)$.

Concernant le taux de couverture de la prévision à $(J + 7)$, les résultats ont montré un taux de 93,5%, légèrement inférieur au niveau de confiance théorique et inférieur à celui à $(J + 1)$. Ceci s'explique par le fait que les modèles de prévision peuvent devenir moins précis à mesure que l'horizon de prévision s'éloigne dans le temps (voir les résultats dans la table 5.8). Il est donc courant d'observer une diminution du taux de couverture à mesure que l'horizon de prévision est étendu, car l'incertitude augmente avec l'horizon de prévision. Cependant, un taux de couverture de 93,5% reste raisonnablement élevé et proche de la valeur théorique de 95%, indiquant ainsi un taux élevé de fiabilité des prévisions probabilistes calculées. La figure 5.34 illustre un exemple des intervalles de prévision calculés à $(J + 7)$, qui apparaissent plus larges que ceux calculés à $(J + 1)$ et présentés dans la figure 5.33.

En conclusion, les résultats ont montré que la qualité des prévisions probabilistes, représentées par les intervalles de prévision, était satisfaisante. Cela donne une confiance accrue au fournisseur d'énergie et aux opérateurs de réseau lorsqu'ils prennent des décisions en utilisant ces prévisions. Cependant, il est important de noter que le taux de couverture peut varier en fonction des périodes et des conditions météorologiques, et il est donc crucial de continuer à surveiller la performance des intervalles de prévision et de les mettre à jour si nécessaire pour maintenir leur qualité et fiabilité.

5.6.4 Application d'aide à la décision et applications pour d'autres portefeuilles

Suite au retour positif des experts chez le fournisseur sur la qualité et l'intérêt des résultats de nos prévisions ponctuelles et probabilistes, nous avons apporté notre aide pour le développement d'une application permettant de faciliter l'utilisation des modèles de prévision par les opérateurs du fournisseur d'énergie. Cette application a été conçue pour être une interface utilisateur conviviale permettant de charger facilement les données, de sélectionner un modèle de prévision, d'effectuer des prévisions et de calculer les intervalles de prévision. Elle permet également de tracer les courbes de prévisions, de calculer les métriques d'erreur de prévision pour une période donnée et de faire des statistiques sur les erreurs.

L'application a été conçue pour répondre aux besoins spécifiques des opérateurs de réseau, qui doivent souvent prendre des décisions en fonction des prévisions de la charge électrique. Grâce à cette application, les opérateurs peuvent accéder rapidement aux prévisions de charge électrique pour une période donnée, ainsi qu'aux intervalles de prévision associés.

Nous avons également adapté avec succès les modèles que nous avons mis en œuvre

pour la prévision de la charge électrique à différents niveaux pour la prévision des données de consommation de gaz à l'échelle des clients, ainsi que pour d'autres périmètres tels que les transformateurs de distribution. Bien que ces résultats ne soient pas présentés dans ce manuscrit, les modèles que nous avons mis en place ont montré une performance similaire pour la prévision de la consommation de gaz, ce qui souligne la robustesse de notre approche de modélisation et de prévision. Ces résultats ouvrent des perspectives intéressantes pour l'application de ces modèles de prévision à d'autres niveaux et pour d'autres types de données dans le secteur d'énergie.

5.7 Conclusion

Dans l'ensemble, ce chapitre a permis de répondre aux différents besoins du fournisseur d'énergie en termes de prévisions de charge électrique à différents niveaux d'agrégation, pour différents portefeuilles et types de données. Nous avons pu évaluer l'utilité des modèles de prévision que nous avons mis en œuvre pour les ménages en testant trois approches différentes pour la prévision de la consommation agrégée des ménages.

La première approche consiste à agréger les prévisions de consommation électrique à l'échelle individuelle pour prédire la consommation électrique agrégée. La deuxième approche consiste à agréger les données de consommation électrique des ménages, puis à appliquer les modèles de prévision pour prédire la consommation électrique agrégée. La troisième approche consiste à regrouper les courbes de consommation des ménages en *clusters* en fonction de caractéristiques calculées à partir des données de consommation électrique. Les données de chaque *cluster* ont ensuite été agrégées pour prédire la consommation électrique agrégée de chaque *cluster*, et ces prévisions ont été agrégées pour représenter les prévisions de la charge électrique agrégée de tous les ménages. Les deux modèles qui se sont révélés les plus performants pour la prévision de la consommation électrique des ménages ont été utilisés dans cette expérience à savoir le modèle *KWF* et *GAM*. Les résultats obtenus ont montré que l'agrégation des données de consommation électrique au-delà de 50 ménages permettait une amélioration significative des prévisions pour les deux modèles testés. Le modèle *KWF* s'est avéré plus performant pour la prévision de la charge électrique agrégée à un niveau d'agrégation faible (moins de 50 ménages) tandis que le modèle *GAM* a montré de meilleures performances pour des niveaux d'agrégation élevés (plus de 50 ménages). En revanche, l'agrégation des prévisions individuelles des ménages n'a pas contribué à améliorer la précision des prévisions de la consommation électrique agrégée par les deux modèles. La troisième approche s'est révélée la plus performante parmi les trois pour la prévision de la consommation électrique agrégée par les deux modèles de prévisions testés. Il a été démontré que le choix optimal du nombre de *clusters* pour chaque modèle de prévision est crucial pour obtenir de bonnes performances de cette approche.

Ensuite, les modèles de prévision ont été adaptés pour la prévision de la charge électrique à l'échelle des clients dans le secteur tertiaire, en prenant en compte les caractéristiques spécifiques de ces données. Les prévisions effectuées par les modèles proposés ont été comparées à celles effectuées par la méthode actuelle du fournisseur, et les résultats ont montré que les trois modèles proposés étaient plus performants que la méthode du fournisseur.

Des méthodes de calcul des intervalles de prévision à $(J+1)$ ont été implémentées pour tenir compte des incertitudes liées à la prévision de la charge électrique à l'échelle des clients résidentiels et tertiaires. Pour les clients résidentiels, trois méthodes de calcul des intervalles de prévision ont été testées sur les données de consommation de cent clients résidentiels pour le modèle *KWF*, qui avait été préalablement identifié comme le plus performant pour la prévision ponctuelle de la charge électrique des ménages. Les méthodes ont été comparées en termes de taux de couverture empirique, et les résultats ont montré qu'une des méthodes avait un taux de couverture globalement satisfaisant pour les deux niveaux de confiance testés. Les intervalles de prévision à $(J+1)$ pour le modèle *GAM* ont été calculés à l'aide de la méthode de quantile de régression pour les clients tertiaires. Ce choix de modèle s'explique par sa performance avérée pour la prévision ponctuelle de la charge électrique pour les clients de ce secteur. Les résultats ont démontré que la méthode était performante pour la majorité des clients testés, et que les taux de couverture empirique des intervalles de prévision calculées étaient proches du niveau de confiance théorique proposé. Il a été identifié que le problème de l'existence des changements brusques ou *drifts* dans les données peut affecter la qualité des prévisions probabilistes en termes de taux de couverture des intervalles de prévision.

Des modèles ont été adaptés et testés pour prévoir les pertes électriques du réseau et la charge électrique du réseau de distribution. Les performances des modèles de prévision ont été évaluées en comparant leurs prévisions de pertes à celles obtenues par une méthode manuelle actuellement utilisée par le fournisseur d'énergie. Les résultats ont démontré que les modèles avaient une performance supérieure à la méthode manuelle pour la prévision des pertes en 2019. Toutefois, ces modèles ont présenté des limites pour la prévision des pertes pendant la période de confinement. Ainsi, plusieurs approches ont été proposées pour tenir compte des changements liés à la période de la crise du "Covid-19". Pour la prévision des données de réseau, les résultats obtenus ont été validés par des experts du domaine de l'énergie qui ont exprimé leur satisfaction quant à la performance et à l'efficacité des modèles proposés. Les experts ont souligné que les résultats obtenus étaient prometteurs pour améliorer la gestion du réseau électrique et permettre une meilleure planification de l'achat et de la vente d'énergie.

Conclusion et perspectives

À la fin de ce manuscrit, une synthèse des travaux réalisés est présentée, suivie d'une discussion sur les limites de la recherche et des propositions de perspectives majeures pour les futures études.

6.1 Synthèse des travaux effectués

Cette thèse se concentre principalement sur la prévision de la consommation électrique à l'échelle des ménages, motivée par l'intérêt des applications de gestion de la consommation et de réponse à la demande. Notre approche consiste à transformer les données de consommation collectées auprès des clients via les compteurs intelligents en informations significatives qui permettent de mieux comprendre et de gérer leur consommation électrique. Pour ce faire, nous avons exploré la faisabilité de fournir des prévisions fiables et facilement interprétables de la consommation de chaque ménage, basées sur ses propres données de consommation. Ces prévisions peuvent être communiquées aux clients via des applications mobiles, afin de les informer et de les alerter sur leur comportement énergétique par rapport à leurs habitudes de consommation électrique.

Par la suite, nous avons élargi la problématique pour inclure la prévision de la charge électrique à d'autres niveaux et types de données, telles que les données agrégées d'un groupe de ménages, les données de consommation des clients du secteur tertiaire, les pertes et la charge électrique du réseau de distribution.

Dans le chapitre 1, nous avons abordé le cadre général de cette thèse, justifiant la nécessité de la recherche sur la prévision à court terme à l'échelle des ménages pour les fournisseurs et les consommateurs d'électricité. Nous avons exposé la problématique industrielle que cette thèse tente de résoudre, ainsi que les objectifs spécifiques qui visent à répondre aux besoins des consommateurs et des fournisseurs d'électricité.

Dans le chapitre 2, nous avons présenté une revue de la littérature portant sur la prévision de la charge électrique à court terme $(J + 1)$, à l'échelle nationale et locale.

Cette revue a mis en évidence une abondance d'études portant sur la prévision à l'échelle nationale, mais elles sont plus rares à l'échelle locale, et plus particulièrement à l'échelle des ménages. Par ailleurs, il a été mis en évidence que la prévision de la charge électrique à l'échelle des ménages est difficile, principalement en raison de la caractéristique irrégulière et hautement volatile des données, entraînant une erreur de prévision de la charge des ménages qui peut largement dépasser les 30% pour seulement 1% à l'échelle nationale. Après avoir comparé et analysé les différentes approches proposées dans la littérature pour la prévision de la consommation électrique des ménages, nous avons conclu qu'il est difficile de déterminer quelle approche est la plus fiable car chacune présente des avantages et des inconvénients en fonction des données disponibles et des objectifs de la prévision. Le choix de la technique de prévision doit donc dépendre de la qualité et de la nature des données utilisées. Par ailleurs, l'utilisation de la prévision à l'échelle des ménages dans un environnement industriel pour le jour $(J+1)$ implique des contraintes supplémentaires en plus de la précision des modèles de prévision. Les contraintes telles que les ressources de calcul disponibles, le temps de calcul requis pour générer les prévisions, ainsi que l'interprétabilité doivent également être prises en compte, pouvant restreindre les choix des techniques de prévision à utiliser et nécessitant des compromis entre la précision et la faisabilité. En résumé, ce chapitre souligne que la prévision de la charge électrique à court terme à l'échelle des ménages est un sujet complexe qui nécessite une attention particulière à la qualité des données, à la sélection de la technique de prévision appropriée, ainsi qu'à la prise en compte des contraintes opérationnelles.

Dans le chapitre 3, nous avons abordé les bases théoriques de certains modèles de prévision pour la charge électrique à court terme, mettant en avant leurs atouts et leurs limites. Nous avons également présenté les critères de sélection utilisés pour choisir les modèles les mieux adaptés à nos besoins. Ces éléments, combinés avec des expériences antérieures qui ne sont pas incluses dans cette thèse, nous ont guidés vers la sélection de plusieurs modèles de prévision, notamment le modèle *KWF*, *GAM*, *MARS*, *RNN* et les forêts aléatoires. Nous avons choisi ces modèles pour leurs capacités à gérer des données non stationnaires et des interactions non linéaires entre les différentes variables, comme c'est souvent le cas avec la charge électrique. De plus, ces modèles offrent une grande flexibilité pour la modélisation des tendances saisonnières et journalières, qui sont des éléments importants pour la prévision de la charge électrique. Nous avons présenté les caractéristiques de chacun de ces modèles, ainsi que les résultats de leurs performances dans le contexte de la prévision de la charge électrique à court terme. Enfin, nous avons exposé les métriques d'erreur généralement utilisées dans la littérature pour évaluer la précision des modèles de prévision, ainsi que celles spécifiques au contexte de la charge électrique.

Dans le chapitre 4, nous avons traité les objectifs de la thèse en matière de prévision de la consommation électrique à l'échelle des ménages. Ce chapitre présente un plan chronologique détaillant les différentes étapes que nous avons suivies pour mettre en œuvre

des modèles de prévision adaptés aux besoins du fournisseur électrique. Ces étapes vont du traitement des données, à la sélection des variables d'entrée des modèles à partir des caractéristiques des données de consommation électrique à l'échelle des ménages, en passant par la sélection et l'adaptation des modèles de prévision, jusqu'à l'évaluation de la performance des modèles et la comparaison de leurs précisions. Les caractéristiques des données de consommation électrique à l'échelle des ménages sont analysées et comparées à celles à l'échelle nationale. Cela a permis de justifier la difficulté de la prévision à cette échelle par rapport à l'échelle nationale. Ensuite, six modèles de prévision de la charge électrique à l'échelle des ménages répondant aux différentes contraintes industrielles et variant en termes de complexité ont été mis en œuvre. Les modèles ont été testés sur un jeu de données privé de 720 ménages de profils de consommation divers. Dans notre étude, nous avons pris en compte le fait que tous les modèles que nous avons adaptés pour la prévision de la charge à l'échelle des ménages ne sont pas construits de la même manière. Nous avons donc exploré chaque modèle en détail, en examinant la façon dont il fonctionne et en sélectionnant les meilleures variables pour maximiser sa performance. Pour ce faire, nous avons mené des expériences sur un petit échantillon de clients, en testant différentes configurations pour chaque modèle. Nous avons ensuite sélectionné les meilleures configurations parmi celles testées en termes de précision de la prévision pour chaque modèle. Cependant, il est important de noter que toutes les expériences menées ne sont pas décrites dans ce manuscrit. Nous avons inclus les résultats les plus significatifs pour chaque modèle. Nous avons apporté une attention particulière à l'intégration de la température dans les modèles de prévision de la charge électrique. Pour le modèle *KWF*, nous avons opté pour une adaptation de son fonctionnement en utilisant des *clusters*, afin d'intégrer l'impact de la température sur les courbes de charge journalières, contrairement aux autres modèles pour lesquels nous avons directement utilisé les données brutes de température. Les différentes approches d'intégration de la température ont montré une amélioration significative de la précision des modèles de prévision, par rapport aux modèles ne tenant pas compte de la température. Le modèle *KWF* suivi du modèle *GAM* ont présenté les meilleures performances en termes de moyennes d'erreurs (NMAE, NRMSE, MASE et sMAPE) ainsi que de distribution des erreurs, comparativement aux autres modèles de prévision. Les résultats des prévisions ont montré une grande hétérogénéité, avec des performances satisfaisantes pour certains ménages, mais pas pour d'autres, en raison des différences dans les caractéristiques des données de consommation d'électricité à l'échelle des ménages. Nous avons identifié que la qualité de la prévision dépendait de la volatilité et de la thermosensibilité des ménages. Nous avons constaté que plus la courbe de charge est thermosensible, plus la prévision est précise. Toutefois, pour les ménages ayant des habitudes de consommation irrégulières et volatiles, la qualité de la prévision était médiocre. De plus, nous avons constaté que la qualité de la prévision se dégradait lorsque les données présentent un changement brutal entre la période d'entraînement et la période de test. Une approche alternative de prévision pour les courbes de charge les plus volatiles est également proposée, qui consiste à prédire l'énergie électrique consommée pendant la journée plutôt que les

puissances. Les résultats obtenus montrent que cette approche permet de transformer les données de consommation en données plus régulières et moins volatiles, et donc plus prédictibles. Cette méthode présente un intérêt commercial, car elle permet aux fournisseurs d'électricité de proposer des services de gestion de la consommation électrique même pour les clients ayant des données de consommation volatiles. Les résultats préliminaires de ce chapitre ont été présentés lors la Conférence Internationale Francophone sur la Science des Données (CIFSD) et des résultats plus avancés sont présentés lors de la Conférence Internationale sur les Statistiques Computationnelles (COMPSTAT2022).

Dans le chapitre 5, nous avons exposé les travaux que nous avons menés concernant les applications liées à la prévision. Le fournisseur d'énergie a souhaité exploiter les modèles de prévision mis en place à l'échelle des ménages afin de répondre à ses besoins à plusieurs niveaux et pour différents types de données. Cette démarche a débuté par la prévision des données agrégées des ménages, puis a été étendue aux données des clients du secteur tertiaire, aux données de pertes électriques ainsi qu'à la charge électrique du réseau de distribution. Pour les données agrégées des ménages, nous avons testé trois approches de prévision différentes : l'agrégation de la charge électrique de tous les ménages suivie de la prévision de la charge agrégée, la prévision de tous les données des ménages individuellement suivie de l'agrégation des prévisions, ainsi que le *clustering* des données de consommation des ménages suivi de la prévision des données agrégées dans chaque *cluster* et de l'agrégation des prévisions. La troisième approche s'est avérée la plus appropriée pour prévoir les données agrégées. Les résultats obtenus ont démontré que l'agrégation des prévisions des données de consommation électrique des ménages réalisées au niveau individuel n'a pas entraîné une amélioration de la précision de la prévision de la charge électrique agrégée. En revanche, l'utilisation de la méthode de *clustering* a permis d'améliorer cette précision. En d'autres termes, le *clustering* a permis de regrouper des courbes de charge de ménages ayant des profils de consommation similaires, ce qui a conduit à une amélioration de la prédictibilité des agrégations, d'abord au niveau des *clusters* puis à l'échelle globale de la charge agrégée de tous les ménages. Les modèles de prévision ont été adaptés également pour la prévision des données à l'échelle des clients dans le secteur tertiaire. Dans un échantillon de test comprenant 290 clients, les trois modèles proposés (*GAM*, *MARS* et *KWF*) ont donné de meilleurs résultats pour la prévision à $(J+1)$ que la méthode utilisée par le fournisseur d'énergie. Des modèles de prévision ont été adaptés et comparés pour la prévision des pertes électriques à $(J+1)$. Les modèles *GAM* et *MARS* ont montré une précision supérieure à celle de la méthode utilisée par le fournisseur d'énergie en 2019. Cependant, ces modèles ont montré des limites dans la prévision des périodes anormales qui perturbent la régularité des données, telles que la crise du « Covid-19 », par rapport à la méthode manuelle utilisée par le fournisseur. Des essais visant à améliorer les performances des modèles n'ont pas abouti principalement en raison du temps d'adaptation nécessaire pour les modèles afin de réagir face à une situation anormale, qui est supérieur à celui de l'adaptation manuelle. Ces résultats permettent au fournisseur d'élec-

tricité d'utiliser les modèles de prévision les plus performants pour prédire les pertes, ce qui lui permet de bénéficier d'une meilleure précision et d'une automatisation accrue. Par exemple, l'utilisation du modèle *GAM* a permis une amélioration significative du taux de précision de 33% par rapport à la méthode manuelle pour l'année 2019. Les opérateurs peuvent continuer à ajuster manuellement les prévisions lors des périodes anormales pour le moment. Nous avons intégré l'un des modèles les plus performants dans un logiciel de prévision chez le fournisseur d'énergie et il est actuellement en phase de test. Nous avons utilisé les modèles les plus performants pour la prévision des pertes, à savoir le modèle *GAM* et *MARS*, et le modèle de forêt aléatoire pour prédire la charge électrique du réseau de distribution du fournisseur à $(J+1)$ et $(J+7)$, car ces deux types de données présentent des caractéristiques similaires. Nous avons également élargi notre analyse en utilisant des prévisions probabilistes qui intègrent les incertitudes associées à la prévision de la charge électrique par la méthode du quantile de régression pour le modèle *GAM*, qui s'est avéré le plus performant pour les prévisions ponctuelles. Les résultats ont montré que ces prévisions probabilistes étaient fiables et apportaient des informations complémentaires sur la distribution de la charge électrique prévue. Les experts ont donné des retours positifs sur l'utilité et la performance des modèles de prévision que nous avons mis en place et les prévisions probabilistes dans leur prise de décision quotidienne. C'est ainsi qu'une application a été mise en place pour faciliter l'utilisation de ces modèles de prévision par les opérateurs chez le fournisseur d'énergie. Cette application permet de charger des données, de calculer des prévisions pour les périodes $(J+1)$ et $(J+7)$ à l'aide de nos modèles, de calculer les intervalles de prévision correspondants, de visualiser les données, les prévisions et les intervalles de prévision, et d'analyser l'historique des erreurs de prévision commises par le modèle.

6.2 Ouverture et perspectives

La prévision de la consommation électrique à l'échelle des ménages est un défi important en raison des limites liées à la disponibilité des données historiques, à la diversité des profils de consommation, ainsi qu'à la difficulté d'intégrer des variables exogènes telles que les conditions météorologiques. Il existe donc de nombreuses perspectives à explorer pour améliorer l'intégration de ces modèles de prévision dans un environnement opérationnel.

Les modèles que nous avons mis en place pour la prévision de la consommation électrique à l'échelle des clients des secteurs résidentiels et tertiaires seront utilisés par le fournisseur d'énergie pour offrir des services de gestion de la consommation. Cette approche de prévision offre un grand potentiel commercial, car elle permet de prédire la consommation électrique de chaque client en utilisant ses propres données, offrant ainsi une approche personnalisée et adaptée aux besoins de chaque client. De cette façon, les clients peuvent mieux comprendre leur consommation électrique et prendre des décisions

éclairées pour réduire leur consommation et leur facture d'électricité.

Cependant, il convient de souligner que notre approche présente des limites. L'une des principales limites est que nos modèles reposent sur l'utilisation de données historiques de consommation électrique pour effectuer les prévisions, ce qui peut limiter leur applicabilité dans les situations où ces données ne sont pas disponibles ou ne sont pas suffisantes, ce que l'on appelle communément le « problème de démarrage à froid ».

Afin de pallier ces limites, nous proposons deux approches alternatives de prévision. La première utilise la segmentation de la clientèle, une méthode courante pour prédire la consommation des clients en l'absence de données historiques. Cette approche consiste à regrouper les clients ayant des caractéristiques similaires, telles que la localisation géographique, le type de logement, la taille de la famille, le type de chauffage, et ainsi de suite. Ensuite, la consommation moyenne du groupe est utilisée pour prédire la consommation du client n'ayant pas encore d'historique de consommation. Une fois qu'un historique de consommation de quelques semaines est disponible pour le client, des modèles de prévision nécessitant peu de données, tels que le modèle climatologique présenté dans le chapitre 4 ou des modèles saisonniers naïfs, peuvent être utilisés pour surmonter le problème de démarrage à froid. Bien que ces modèles ne fournissent pas des prévisions très précises, ils permettent d'obtenir une première estimation de la consommation du client. Ensuite, une fois que suffisamment de données sont collectées, les modèles présentés dans le chapitre 4 peuvent être utilisés pour affiner les prévisions.

La deuxième approche consiste à utiliser un modèle de réseau de neurones unique pour prédire la consommation électrique de chaque ménage, plutôt que d'avoir un réseau séparé pour chaque ménage comme présenté dans le chapitre 4. Bien que la méthode précédente fournisse des prévisions plus précises et plus personnalisées pour chaque ménage, la nouvelle approche présente plusieurs avantages.

Tout d'abord, cette méthode permet de surmonter le problème de manque ou de limitation d'historique de données pour certains ménages grâce au transfert d'apprentissage. Ce transfert consiste à utiliser l'apprentissage acquis par le réseau de neurones pré-entraîné sur un ensemble de données de ménages ayant un historique long et complet de consommation électrique pour améliorer les performances sur les données des ménages ayant un historique plus court. Le modèle apprend des caractéristiques utiles pour la prévision de la consommation d'électricité, telles que la saisonnalité, les jours fériés et l'impact de la température, qui peuvent être utilisées pour prédire la consommation d'électricité pour chaque ménage, même avec peu d'historique de données.

De plus, il est possible d'adapter ce modèle pour prédire d'autres types de données de consommation, telles que les données de consommation électrique des bâtiments ou des quartiers, ainsi que les données de consommation de gaz. Cette approche nécessite également moins de temps et de ressources pour entraîner le modèle, ce qui facilite sa mise

à l'échelle pour un grand nombre de ménages. Cependant, elle nécessite une plus grande quantité de données de ménages pour l'entraînement du modèle. Par conséquent, si des données à l'échelle des ménages sont disponibles, cette approche offre une piste intéressante à explorer en raison de ses nombreux avantages.

Les résultats de notre étude portant sur les modèles de prévision de la charge électrique à l'échelle des ménages, présentés dans le chapitre 4, ont démontré que la présence de changements brusques dans les séries temporelles de consommation électrique peut entraîner une grande disparité entre la période d'entraînement et la période de test, ce qui a pour effet de réduire la précision des prévisions. Afin de remédier à cette problématique, il serait judicieux d'explorer des méthodes de détection des changements pour identifier les périodes de temps où ces changements ont eu lieu et ainsi adapter les modèles de prévision en conséquence. Par exemple, le modèle pourrait être ré-entraîné en utilisant les données les plus récentes afin de tenir compte des ces changements dans les données de consommation d'électricité. En combinant les modèles de prévision avec des algorithmes de détection des changements, il serait possible de détecter les changements brusques dans les données de consommation d'électricité et de régler automatiquement le modèle de prévision en conséquence. Enfin, pour évaluer leur robustesse face aux changements brutaux dans les données, une étude comparative de la performance des différents modèles que nous avons mis en œuvre pour la prévision de la charge à l'échelle des ménages serait pertinente et permettrait de déterminer quel modèle est le plus adapté pour une utilisation par le fournisseur d'électricité dans ce contexte. De plus, il convient de noter que les perspectives que nous avons proposées peuvent être généralisées aux clients tertiaires, puisque ces données présentent également des périodes de changement.

L'approche de prévision par *clustering* présentée dans le chapitre 5 pour l'agrégation des données de charge électrique des ménages est très prometteuse puisqu'elle a montré que le *clustering* des données de consommation des ménages avant la prévision permet d'améliorer la précision de la prévision des données agrégées. L'étude de Wijaya, SFRJ Humeau et al. (2014) a également montré que cette amélioration augmente avec l'augmentation du nombre de ménages dans le jeu de données. Il serait intéressant de compléter notre étude par une analyse de la relation entre la taille de l'ensemble de données et le taux d'amélioration de la prévision, ainsi que l'impact de la sélection a priori des limites temporelles pour les périodes de la journée utilisées pour la définition de certaines caractéristiques.

Les modèles mis en œuvre pour la prévision des pertes et de la charge électrique du réseau, comme décrits dans le chapitre 5, ont été intégrés dans des applications afin de faciliter leur utilisation par les opérateurs chez le fournisseur. Le périmètre d'application de ces modèles a été étendu pour inclure également les prévisions de consommation de gaz. Cependant, ces modèles s'appuient sur des données calendaires et météorologiques historiques, ce qui les rend incapables de capturer les effets significatifs de la crise du

« Covid-19 » sur la consommation d'énergie. En conséquence, ces modèles ont montré des performances médiocres depuis le début de la pandémie. Pour améliorer la précision des prévisions, il serait donc intéressant d'explorer des méthodes qui permettent d'ajuster les modèles de prévision en fonction des nouvelles données de perte et de réseau. Des travaux intéressants qui peuvent offrir des pistes pour remédier à ce problème sont présentés dans l'étude menée par Obst et al. (2021).

Liste des communications

Communications orales

1. F. Fahs, F. Bertrand, M. Maumy, Forecasting electricity consumption at the household level, 24th International Conference on Computational Statistics (COMPSTAT2022), du 23 août à 26 août 2022.
2. F. Fahs, F. Bertrand, M. Maumy, Prévision de la consommation d'électricité à l'échelle individuelle dans les secteurs résidentiel et tertiaire, Congrès des Jeunes Chercheuses et Jeunes Chercheurs en Mathématiques Appliquées à Paris (France), 27 au 29 octobre 2021.
3. F. Fahs, F. Bertrand, M. Maumy, Prévision de la consommation d'électricité à l'échelle individuelle dans les secteurs résidentiel et tertiaire, p187-199 in Conférence Internationale Francophone sur la Science des Données (CIFSD). Actes de la 9e édition, à Marseille (France), du 9 au 11 juin 2021.
4. J.-B. Wahl, F. Fahs, Maumy-Bertrand, F. Bertrand, Energy consumption analysis, International Congress on Industrial and Applied Mathematics, à Valence (Espagne), du 15 au 19 juillet 2019.
5. C. Caldini-Queiros, Z. Belhachmi, F. Bertrand, V. Chabannes, F. Fahs, Ph. Helluy, R. Hild, M. Maumy-Bertrand, C. Prud'homme, J.-B. Wahl, Center of Modeling and Simulation of Strasbourg, International Congress on Industrial and Applied Mathematics, à Valence (Espagne), du 15 au 19 juillet 2019.
6. F. Fahs, M. Maumy-Bertrand, C. Caldini-Queiros, F. Bertrand, Prévision de la consommation d'électricité à l'échelle des ménages, Actes 51èmes Journées de Statistique de la SFDS, à Nancy, du 3 au 7 juin 2019.

Communications écrites

1. F. Fahs, F. Bertrand, M. Maumy, Prévision de la consommation d'électricité à l'échelle individuelle dans les secteurs résidentiel et tertiaire, p187-199 in Conférence Internationale Francophone sur la Science des Données (CIFSD). Actes de la 9e édition, à Marseille (France), du 9 au 11 juin 2021.
2. F. Fahs, M. Maumy-Bertrand, C. Caldini-Queiros, F. Bertrand, Prévision de la consommation d'électricité à l'échelle des ménages, Actes 51èmes Journées de Statistique de la SFDS, à Nancy, du 3 au 7 juin 2019.

Bibliographie

Abd Jalil, Nur Adilah, Maizah Hura Ahmad et Norizan Mohamed (2013). « Electricity load demand forecasting using exponential smoothing methods ». *World Applied Sciences Journal* 22.11, p. 1540-1543 (cf. p. 13, 14).

Abreu, Joana M, Francisco Câmara Pereira et Paulo Ferrão (2012). « Using pattern recognition to identify habitual behavior in residential electricity consumption ». *Energy and buildings* 49, p. 479-487 (cf. p. 20, 154, 155).

Albert, Adrian et Ram Rajagopal (2013). « Smart meter driven segmentation : What your consumption says about you ». *IEEE Transactions on power systems* 28.4, p. 4019-4030 (cf. p. 155).

Alduailij, Mona A, Ioan Petri, Omer Rana, Mai A Alduailij et Abdulrahman S Aldawood (2021). « Forecasting peak energy demand for smart buildings ». *The Journal of Supercomputing* 77.6, p. 6356-6380 (cf. p. 52).

Alonso, Andrés M, Francisco J Nogales et Carlos Ruiz (2020). « A single scalable LSTM model for short-term forecasting of massive electricity time series ». *Energies* 13.20, p. 5328 (cf. p. 124, 127).

Amato, Anthony D, Matthias Ruth, Paul Kirshen et James Horwitz (2005). « Regional energy demand responses to climate change : methodology and application to the commonwealth of Massachusetts ». *Climatic Change* 71.1, p. 175-201 (cf. p. 70).

Antoniadis, Anestis, Xavier Brossat, Jairo Cugliari et Jean-Michel Poggi (2014). « Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité ». *Journal de la Société Française de Statistique* 155.2, p. 202-219 (cf. p. 15, 41, 42, 46, 81).

— (2016). « A prediction interval for a function-valued forecast model : Application to load forecasting ». *International Journal of Forecasting* 32.3, p. 939-947 (cf. p. 178-180).

Antoniadis, Anestis, Efstathios Paparoditis et Theofanis Sapatinas (2006). « A functional wavelet–kernel approach for time series prediction ». *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 68.5, p. 837-857 (cf. p. 41, 43, 83).

Aragon, Yves (2011). *Séries temporelles*. Ed. Techniques Ingénieur (cf. p. 33).

Arco, Leticia, Gladys Casas et Ann Nowé (2017). « Clustering methodology for smart metering data based on local and global features ». *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, p. 1-13 (cf. p. 155).

Atan, Rodziah, Hasimah Abdul Rahman, Afiqah Abu Bakar et Ruzanna Ab Ghazali (2014). « Electricity load forecasting using statistical linear regression models ». *Renewable and Sustainable Energy Reviews* 31, p. 607-617 (cf. p. 29).

Bacher, Raphael, Reinhard Madlener et Tim Stoffel (2017). « Combining a top-down and bottom-up approach for forecasting electricity consumption in local grids ». *Applied Energy* 194, p. 476-491 (cf. p. 235).

Barbier, Thibaut (2017). « Modélisation de la consommation électrique à partir de grandes masses de données pour la simulation des alternatives énergétiques du futur ». Thèse de doct. Paris Sciences et Lettres (ComUE) (cf. p. 235).

Bashir, ZA et ME El-Hawary (2009). « Applying wavelets to short-term load forecasting using PSO-based neural networks ». *IEEE transactions on power systems* 24.1, p. 20-27 (cf. p. 15).

Beccali, M, M Cellura, V Lo Brano et Antonino Marvuglia (2008). « Short-term prediction of household electricity consumption : Assessing weather sensitivity in a Mediterranean area ». *Renewable and Sustainable Energy Reviews* 12.8, p. 2040-2065 (cf. p. 22).

Bedi, Jatin et Durga Toshniwal (2019). « Deep learning framework to forecast electricity demand ». *Applied energy* 238, p. 1312-1326 (cf. p. 16).

Bengio, Yoshua, Aaron Courville et Pascal Vincent (2013). « Representation learning : A review and new perspectives ». *IEEE transactions on pattern analysis and machine intelligence* 35.8, p. 1798-1828 (cf. p. 16).

Beniétez, Ignacio, Alfredo Quijano, José-Luis Diéez et Ignacio Delgado (2014). « Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers ». *International Journal of Electrical Power & Energy Systems* 55, p. 437-448 (cf. p. 155).

Bessec, Marie et Julien Fouquau (2008). « The non-linear link between electricity consumption and temperature in Europe : A threshold panel approach ». *Energy Economics* 30.5, p. 2705-2721 (cf. p. 70).

Bianco, Vincenzo, Oronzio Manca et Sergio Nardini (2009). « Electricity consumption forecasting in Italy using linear regression models ». *Energy* 34.9, p. 1413-1421 (cf. p. 13, 14).

Biswas, MA Rafe, Melvin D Robinson et Nelson Fumo (2016). « Prediction of residential building energy consumption : A neural network approach ». *Energy* 117, p. 84-92 (cf. p. 49).

Bouktif, Salah, Ali Fiaz, Ali Ouni et Mohamed Adel Serhani (2019). « Single and multi-sequence deep learning models for short and medium term electric load forecasting ». *Energies* 12.1, p. 149 (cf. p. 16).

Bowerman, Bruce L, Richard T O'Connell et Anne B Koehler (2005). *Forecasting, time series, and regression : an applied approach*. T. 4. South-Western Pub (cf. p. 53).

Box, George E. P., Gwilym M. Jenkins et Gregory C. Reinsel (1970). *Time series analysis : Forecasting and control*. San Francisco, CA : Holden-Day (cf. p. 32, 33).

Breiman, Leo (1996). « Bagging predictors ». *Machine learning* 24, p. 123-140 (cf. p. 40).

— (2000). « Randomizing outputs to increase prediction accuracy ». *Machine Learning* 40, p. 229-242 (cf. p. 40).

— (2001). « Random forests ». *Machine learning* 45.1, p. 5-32 (cf. p. 39).

Breiman, Leo et Adele Cutler (2004). « Random forest-manual ». *Online : http ://www. stat. berkeley. edu/˜ breiman/RandomForests/cc_manual. htm* (cf. p. 40).

Breiman, Leo, Jerome H Friedman, Richard A Olshen et Charles J Stone (2017). *Classification and regression trees*. Routledge (cf. p. 40).

Brooks, Harold E et Charles A Doswell III (1996). « A comparison of measures-oriented and distributions-oriented approaches to forecast verification ». *Weather and forecasting* 11.3, p. 288-303 (cf. p. 54).

Capasso, Alfonso, W Grattieri, R Lamedica et A Prudenzi (1994). « A bottom-up approach to residential load modeling ». *IEEE transactions on power systems* 9.2, p. 957-964 (cf. p. 12).

Cerquitelli, Tania, Gianfranco Chicco, Evelina Di Corso, Francesco Ventura, Giuseppe Montesano, Anita Del Pizzo et al. (2018). « Discovering electricity consumption over time for residential consumers through cluster analysis ». *2018 International Conference on Development and Application Systems (DAS)*. IEEE, p. 164-169 (cf. p. 20).

Chaouch, Mohamed (2013). « Clustering-based improvement of nonparametric functional time series forecasting : Application to intra-day household-level load curves ». *IEEE Transactions on Smart Grid* 5.1, p. 411-419 (cf. p. 154).

Chapagain, Kamal, Somsak Kittipiyakul et Pisut Kulthanavit (2020). « Short-term electricity demand forecasting : Impact analysis of temperature for Thailand ». *Energies* 13.10, p. 2498 (cf. p. 16).

Charles, Catherine, Gervais Leclerc, Jean-Jacques Pireaux et Jean-Paul Rasson (2004). « Introduction to wavelet applications in surface spectroscopies ». *Surface and Interface Analysis : An International Journal devoted to the development and application of techniques for the analysis of surfaces, interfaces and thin films* 36.1, p. 49-60 (cf. p. xiii, 245).

Chicco, Gianfranco, Roberto Napoli et Federico Piglione (2001). « Load pattern clustering for short-term load forecasting of anomalous days ». *2001 IEEE Porto Power Tech Proceedings (Cat. No. 01EX502)*. T. 2. IEEE, 6-pp (cf. p. 154, 156).

Chin, Hiroshi (2016). « An analytical evaluation of top-down versus bottom-up forecast in the electricity demand ». *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, p. 1-2 (cf. p. 12).

Cleveland, William S (1979). « Robust locally weighted regression and smoothing scatterplots ». *Journal of the American statistical association* 74.368, p. 829-836 (cf. p. 237).

Cooley, Thomas F et Edward C Prescott (1973). « An adaptive regression model ». *International Economic Review*, p. 364-371 (cf. p. 243).

Cugliari, Jairo (2011). « Prévision non paramétrique de processus à valeurs fonctionnelles : Application à la consommation d'électricité ». Thèse de doct. Université Paris Sud-Paris XI (cf. p. 15, 17, 41, 42, 44-46, 81, 83, 84, 94-100, 109, 110, 172, 196, 248).

Dang-Ha, The-Hien, Filippo Maria Bianchi et Roland Olsson (2017). « Local short term electricity load forecasting : Automatic approaches ». *2017 international joint conference on neural networks (ijcnn)*. IEEE, p. 4267-4274 (cf. p. 19).

De Boor, Carl et Carl De Boor (1978). *A practical guide to splines*. T. 27. springer-verlag New York (cf. p. 242).

Dietterich, Thomas G (2000). « Ensemble methods in machine learning ». *Multiple Classifier Systems : First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*. Springer, p. 1-15 (cf. p. 40).

Ding, Yong, Julio Borges, Martin A Neumann et Michael Beigl (2015). « Sequential pattern mining—A study to understand daily activity patterns for load forecasting enhancement ». *2015 IEEE First International Smart Cities Conference (ISC2)*. IEEE, p. 1-6 (cf. p. 21).

Dordonnat, V, Siem Jan Koopman, Marius Ooms, A Dessertaine et J Collet (2008). « An hourly periodic state space model for modelling French national electricity load ». *International Journal of Forecasting* 24.4, p. 566-587 (cf. p. ix, 13, 14, 16, 69-71, 116, 175).

Dryar, Henry A (1944). « The effect of weather on the system load ». *Electrical Engineering* 63.12, p. 1006-1013 (cf. p. 16).

Dudek, Grzegorz (2015). « Short-Term Load Forecasting Using Random Forests ». *2015* 323, p. 821-828 (cf. p. 41).

Eneyew, Dagimawi D, Miriam AM Capretz, Girma T Bitsuamlak et Syed Mir (2020). « Predicting Residential Energy Consumption Using Wavelet Decomposition with Deep Neural Network ». *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, p. 895-900 (cf. p. 19, 56).

Engle, Robert F, Clive WJ Granger, John Rice et Andrew Weiss (1986). « Semiparametric estimates of the relation between weather and electricity sales ». *Journal of the American statistical Association* 81.394, p. 310-320 (cf. p. 70, 77).

Erişen, Esra, Cem Iyigun et Fehmi Tanrisever (2017). « Short-term electricity load forecasting with special days : an analysis on parametric and non-parametric methods ». *Annals of Operations Research*, p. 1-34 (cf. p. 13).

Esteves, Gheisa RT, Bruno Q Bastos, Fernando L Cyrino, Rodrigo F Calili et Reinaldo C Souza (2015). « Long term electricity forecast : a systematic review ». *Procedia Computer Science* 55, p. 549-558 (cf. p. 12).

Eubank, Randall L (1999). *Nonparametric regression and spline smoothing*. CRC press (cf. p. 241, 243, 244).

Fan, Shu et Rob J Hyndman (2011). « Short-term load forecasting based on a semi-parametric additive model ». *IEEE transactions on power systems* 27.1, p. 134-141 (cf. p. 14, 111).

— (2012). « Short-term load forecasting based on a semi-parametric additive model ». *IEEE Transactions on Power Systems* 27.1, p. 134-141 (cf. p. 36).

Fasiolo, Matteo, Simon N Wood, Margaux Zaffran, Raphaël Nedellec et Yannig Goude (2020). « qgam : Bayesian non-parametric quantile regression modelling in R ». *arXiv preprint arXiv :2007.03303* (cf. p. 187).

Fathi, Oussama (2019). « Time series forecasting using a hybrid ARIMA and LSTM model ». *Velvet Consulting*, p. 1-7 (cf. p. 33).

Flath, Christoph, David Nicolay, Tobias Conte, Clemens van Dinther et Lilia Filipova-Neumann (2012). « Cluster analysis of smart metering data ». *Business & Information Systems Engineering* 4.1, p. 31-39 (cf. p. 155).

Fox, John (2015). *Applied regression analysis and generalized linear models*. 3e éd. Thousand Oaks, CA : Sage Publications (cf. p. 29).

Freund, Yoav, Robert E Schapire et al. (1996). « Experiments with a new boosting algorithm ». *icml*. T. 96. Citeseer, p. 148-156 (cf. p. 40).

Friedman, Jerome H (1991). « Multivariate adaptive regression splines ». *The annals of statistics* 19.1, p. 1-67 (cf. p. 36).

Friedman, Jerome H et Charles B Roosen (1995). *An introduction to multivariate adaptive regression splines* (cf. p. 37, 39).

Fung, WY, Ka Se Lam, WT Hung, SW Pang et YL Lee (2006). « Impact of urban temperature on energy consumption of Hong Kong ». *Energy* 31.14, p. 2623-2637 (cf. p. 16).

Gaillard, Pierre, Yannig Goude et Raphaël Nedellec (2016). « Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting ». *International Journal of forecasting* 32.3, p. 1038-1050 (cf. p. 172, 196).

Gajowniczek, Krzysztof et Tomasz Ząbkowski (2017). « Electricity forecasting on the individual household level enhanced based on activity patterns ». *PloS one* 12.4, e0174098 (cf. p. 21, 113).

Gang, Wu, Zhang Dongdong et Fan Shengrong (2022). « Application of an improved k-means clustering algorithm in power user grouping ». *International Journal of Numerical Modelling : Electronic Networks, Devices and Fields* 35.4, e2990 (cf. p. 156).

Genuer, Robin et Jean-Michel Poggi (2016). « Arbres CART et Forêts aléatoires, Importance et sélection de variables » (cf. p. 40).

Gerossier, Alexis (2019). « Short-term forecasting of electricity demand of smart homes and distribution grids ». Thèse de doct. Paris Sciences et Lettres (ComUE) (cf. p. 67, 73, 85, 114, 116).

Gerossier, Alexis, Robin Girard, George Kariniotakis et Andrea Michiorri (2017). « Probabilistic day-ahead forecasting of household electricity demand ». *CIRED-Open Access Proceedings Journal* 2017.1, p. 2500-2504 (cf. p. 22, 25, 36, 56, 111, 113, 114).

Gers, Felix A, Nicol N Schraudolph et Jürgen Schmidhuber (2002). « Learning precise timing with LSTM recurrent networks ». *Journal of machine learning research* 3.Aug, p. 115-143 (cf. p. 122).

Ghofrani, Mahmoud, Mohammad Hassanzadeh, Mehdi Etezadi-Amoli et M Sami Fadali (2011). « Smart meter based short-term load forecasting for residential customers ». *2011 North American Power Symposium*. IEEE, p. 1-5 (cf. p. 18).

Gijbels, Irène et Ilaria Prosdocimi (2010). « Loess ». *Wiley Interdisciplinary Reviews : Computational Statistics* 2.5, p. 590-599 (cf. p. 239).

Gneiting, Tilmann et Matthias Katzfuss (2014). « Probabilistic forecasting ». *Annual Review of Statistics and Its Application* 1, p. 125-151 (cf. p. 178).

Gneiting, Tilmann et Adrian E Raftery (2007). « Strictly proper scoring rules, prediction, and estimation ». *Journal of the American statistical Association* 102.477, p. 359-378 (cf. p. 178).

Goodfellow, Ian, Yoshua Bengio et Aaron Courville (2016). *Deep Learning*. Cambridge, MA : MIT Press (cf. p. 46, 47, 49).

Goude, Yannig, Raphael Nedellec et Nicolas Kong (2013). « Local short and middle term electricity load forecasting with semi-parametric additive models ». *IEEE transactions on smart grid* 5.1, p. 440-446 (cf. p. 71, 172).

Gouty, Félix (2022). *Un quart des ménages français disposaient d'un système de climatisation en 2020*. URL : https://www.actu-environnement.com/ae/news/menages-francais-systeme-climatisation-chaleur-ademe-2020-37991.php4 (visité le 02/08/2021) (cf. p. 66).

Grafe, Rosmarie, José Ramón Cancelo et Antoni Espasa (2007). *Forecasting from one day to one week ahead for the Spanish system operator*. Rapp. tech. Universidad Carlos III de Madrid. Departamento de Estadıéstica (cf. p. 70).

Gross, George et Francisco D Galiana (1987). « Short-term load forecasting ». *Proceedings of the IEEE* 75.12, p. 1558-1573 (cf. p. 49, 77).

Guisan, Antoine, Thomas C Edwards Jr et Trevor Hastie (2002). « Generalized linear and generalized additive models in studies of species distributions : setting the scene ». *Ecological modelling* 157.2-3, p. 89-100 (cf. p. 240).

Guo, Ying-Chun, Dong-Xiao Niu et Yan-Xu Chen (2006). « Support vector machine model in electricity load forecasting ». *2006 International Conference on Machine Learning and Cybernetics*. IEEE, p. 2892-2896 (cf. p. 15).

Gupta, Pradeep C et Keigo Yamada (1972). « Adaptive short-term forecasting of hourly loads using weather information ». *IEEE Transactions on Power Apparatus and Systems* 5, p. 2085-2094 (cf. p. 16).

Haben, Stephen, Georgios Giasemidis, Florian Ziel et Siddharth Arora (2019). « Short term load forecasting and the effect of temperature at the low voltage level ». *International Journal of Forecasting* 35.4, p. 1469-1484 (cf. p. 21, 77).

Haben, Stephen, Colin Singleton et Peter Grindrod (2015). « Analysis and clustering of residential customers energy behavioral demand using smart meter data ». *IEEE transactions on smart grid* 7.1, p. 136-144 (cf. p. 154, 155, 157, 158, 160).

Haben, Stephen, Jonathan Ward, Danica Vukadinovic Greetham, Colin Singleton et Peter Grindrod (2014). « A new error measure for forecasts of household-level, high resolution electrical energy consumption ». *International Journal of Forecasting* 30.2, p. 246-256 (cf. p. viii, 54, 55).

Hanke, John E et Dean W Wichern (2005). *Business forecasting*. Pearson Educación (cf. p. 53).

Hastie, TJ, William A BARNETT, James POWELL et George E TAUCHEN (1988). « Eubank, R. L.(1988), Spline Smoothing and Nonparametric Regression ». *Wolfgang Hardle Paul Speckman*, p. 1468 (cf. p. 241).

Hastie, Trevor, Robert Tibshirani et Jerome Friedman (2009). *The elements of statistical learning : data mining, inference, and prediction*. 2nd. Springer. Chap. 3 (cf. p. 31, 41, 49, 126).

Hastie, Trevor J (2017). « Generalized additive models ». *Statistical models in S*. Routledge, p. 249-307 (cf. p. 34).

Hastie, Trevor J et Robert J Tibshirani (1990). *Generalized additive models*. T. 43. CRC press (cf. p. 34, 35).

Hayes, Barry, Jorn Gruber et Milan Prodanovic (2015). « Short-term load forecasting at the local level using smart meter data ». *2015 IEEE Eindhoven PowerTech*. IEEE, p. 1-6 (cf. p. 22).

Henley, Andrew et John Peirson (1997). « Non-linearities in electricity demand and temperature : parametric versus non-parametric methods ». *Oxford bulletin of economics and statistics* 59.1, p. 149-162 (cf. p. 71).

Hinojosa, VH et A Hoese (2010). « Short-term load forecasting using fuzzy inductive reasoning and evolutionary algorithms ». *IEEE Transactions on power systems* 25.1, p. 565-574 (cf. p. 14).

Hippert, Henrique Steinherz, Carlos Eduardo Pedreira et Reinaldo Castro Souza (2001). « Neural networks for short-term load forecasting : A review and evaluation ». *IEEE Transactions on power systems* 16.1, p. 44-55 (cf. p. 77).

Hippert, HS, DW Bunn et RC Souza (2005). « Large neural networks for electricity load forecasting : Are they overfitted ? » *International Journal of forecasting* 21.3, p. 425-434 (cf. p. 15).

Ho, Tin Kam (1998). « The random subspace method for constructing decision forests ». *IEEE transactions on pattern analysis and machine intelligence* 20.8, p. 832-844 (cf. p. 40).

Hochreiter, Sepp, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber et al. (2001). *Gradient flow in recurrent nets : the difficulty of learning long-term dependencies* (cf. p. 122).

Hochreiter, Sepp et Jürgen Schmidhuber (1997). « Long short-term memory ». *Neural computation* 9.8, p. 1735-1780 (cf. p. 122).

Hoerl, Arthur E et Robert W Kennard (1970). « Ridge regression : Biased estimation for nonorthogonal problems ». *Technometrics* 12.1, p. 55-67 (cf. p. 187).

Hoffman, Ross N, Zheng Liu, Jean-Francois Louis et Christopher Grassoti (1995). « Distortion representation of forecast errors ». *Monthly Weather Review* 123.9, p. 2758-2770 (cf. p. 54).

Hong, Tao (2010). *Short term electric load forecasting*. North Carolina State University (cf. p. 29).

Hong, Tao et Shu Fan (2016). « Probabilistic electric load forecasting : A tutorial review ». *International Journal of Forecasting* 32.3, p. 914-938 (cf. p. 12, 13, 23).

Hong, Tao, Pierre Pinson et Shu Fan (2014). « Global energy forecasting competition 2012 ». *International Journal of Forecasting* 30.2, p. 357-363 (cf. p. 32).

Hong, Tao, Peng Wang et Henry L Willis (2011). « A naïve multiple linear regression benchmark for short term load forecasting ». *Power and Energy Society General Meeting, 2011 IEEE*. IEEE, p. 1-6 (cf. p. 31).

Hong, Tao, Pu Wang et Laura White (2015). « Weather station selection for electric load forecasting ». *International Journal of Forecasting* 31.2, p. 286-295 (cf. p. 29).

Hou, Tingting, Rengcun Fang, Jinrui Tang, Ganheng Ge, Dongjun Yang, Jianchao Liu et al. (2021). « A novel short-term residential electric load forecasting method based on adaptive load aggregation and deep learning algorithms ». *Energies* 14.22, p. 7820 (cf. p. 68, 69, 92, 120, 127).

Huang, Ning, Guoyin Lu et Dongmei Xu (2016). « A permutation importance-based feature selection method for short-term electricity load forecasting using random forest ». *Energies* 9.10, p. 767 (cf. p. 41).

Huang, Qian, Yonghua Li, Shaohui Liu et et al. (2017). « Short term load forecasting based on wavelet decomposition and random forest ». *Proceedings of the Workshop on Smart Internet of Things*. ACM, p. 2 (cf. p. 41).

Huck, GE, AA Mahmoud, RB Comerford, J Adams et E Dawson (1980). « Load forecast bibliography phase I ». *IEEE Transactions on Power Apparatus and Systems* 99.1, p. 53-58 (cf. p. 13).

Humeau, Samuel, Tri Kurniawan Wijaya, Matteo Vasirani et Karl Aberer (2013). « Electricity load forecasting for residential customers : Exploiting aggregation and correlation between households ». *2013 Sustainable internet and ICT for sustainability (SustainIT)*. IEEE, p. 1-6 (cf. p. 22, 25).

Hurvich, Clifford M, Jeffrey S Simonoff et Chih-Ling Tsai (1998). « Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion ». *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 60.2, p. 271-293 (cf. p. 238).

Hyndman, Rob J et George Athanasopoulos (2018). *Forecasting : principles and practice*. OTexts (cf. p. 27).

Hyndman, Rob J et Anne B Koehler (2006). « Another look at measures of forecast accuracy ». *International Journal of forecasting* 22.4, p. 679-688 (cf. p. 27, 52).

Ilic, Dejan, Stamatis Karnouskos et P Goncalves Da Silva (2013). « Improving load forecast in prosumer clusters by varying energy storage size ». *IEEE Grenoble PowerTech* (cf. p. 155).

ISSDA (2020). *Commission for Energy Regulation (CER)*. url : https://www.ucd.ie/issda/data/commissionforenergyregulationcer/ (cf. p. 25).

Jahan, Ibrahim Salem, Vaclav Snasel et Stanislav Misak (2020). « Intelligent systems for power load forecasting : A study review ». *Energies* 13.22, p. 6105 (cf. p. 12).

Jain, Anil K (2010). « Data clustering : 50 years beyond K-means ». *Pattern recognition letters* 31.8, p. 651-666 (cf. p. 161).

James, Gareth, Daniela Witten, Trevor Hastie et Robert Tibshirani (2013). *An Introduction to Statistical Learning*. Springer (cf. p. 28).

Javed, Fahad, Naveed Arshad, Fredrik Wallin, Iana Vassileva et Erik Dahlquist (2012). « Forecasting for demand response in smart grids : An analysis on use of anthropologic and structural data and short term multiple loads forecasting ». *Applied Energy* 96, p. 150-160 (cf. p. 21).

Jeong, Hyun Cheol, Minseok Jang, Taegon Kim et Sung-Kwan Joo (2021). « Clustering of load profiles of residential customers using extreme points and demographic characteristics ». *Electronics* 10.3, p. 290 (cf. p. 154).

Kaytez, Fazil, M Cengiz Taplamacioglu, Ertugrul Cam et Firat Hardalac (2015). « Forecasting electricity consumption : A comparison of regression analysis, neural networks and least squares support vector machines ». *International Journal of Electrical Power & Energy Systems* 67, p. 431-438 (cf. p. 14, 15).

Keil, Christian et George C Craig (2009). « A displacement and amplitude score employing an optical flow technique ». *Weather and Forecasting* 24.5, p. 1297-1308 (cf. p. 54).

Kelly, Jack et William Knottenbelt (2015). « The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes ». *Scientific Data* 2.150007 (cf. p. 25).

Khuntia, Swasti R, José L Rueda et Mart AMM van Der Meijden (2016). « Forecasting the load of electrical power systems in mid-and long-term horizons : A review ». *IET Generation, Transmission & Distribution* 10.16, p. 3971-3977 (cf. p. 12).

Kim, Sungil et Heeyoung Kim (2016). « A new metric of absolute percentage error for intermittent demand forecasts ». *International Journal of Forecasting* 32.3, p. 669-679 (cf. p. 53).

Kodogiannis, Vassilis S, Mahdi Amina et Ilias Petrounias (2013). « A clustering-based fuzzy wavelet neural network model for short-term load forecasting ». *International journal of neural systems* 23.05, p. 1350024 (cf. p. 15).

Koenker, Roger et Gilbert Bassett Jr (1978). « Regression quantiles ». *Econometrica : journal of the Econometric Society*, p. 33-50 (cf. p. 186).

Kong, Weicong, Zhao Yang Dong, Youwei Jia, David J Hill, Yan Xu et Yuan Zhang (2017). « Short-term residential load forecasting based on LSTM recurrent neural network ». *IEEE Transactions on Smart Grid* 10.1, p. 841-851 (cf. p. 120, 124, 127, 148).

Kreuwel, Frank PM, Wouter H Knap, Lennard R Visser, Wilfried GJHM van Sark, Jordi Vilà-Guerau de Arellano et Chiel C van Heerwaarden (2020). « Analysis of high frequency photovoltaic solar energy fluctuations ». *Solar Energy* 206, p. 381-389 (cf. p. 85).

Kullback, Solomon et Richard A Leibler (1951). « On information and sufficiency ». *The annals of mathematical statistics* 22.1, p. 79-86 (cf. p. 178).

Kwac, Jungsuk, June Flora et Ram Rajagopal (2014). « Household energy consumption segmentation using hourly data ». *IEEE Transactions on Smart Grid* 5.1, p. 420-430 (cf. p. 154).

Lahouar, Abdelkader et Jean Bernard Hayet Slama (2015). « Day-ahead load forecast using random forest and expert input selection ». *Energy Conversion and Management* 103, p. 1040-1051 (cf. p. 41).

Laurinec, Peter et Mária Lucká (2017). « New clustering-based forecasting method for disaggregated end-consumer electricity load using smart grid data ». *2017 IEEE 14th international scientific conference on informatics*. IEEE, p. 210-215 (cf. p. 20, 154).

Le Comte, Douglas M et Henry E Warren (1981). « Modeling the impact of summer temperatures on national electricity consumption ». *Journal of Applied Meteorology and Climatology* 20.12, p. 1415-1419 (cf. p. 70).

Le Cun, Yann, Lawrence D Jackel, Brian Boser, John S Denker, Hans Peter Graf, Isabelle Guyon et al. (1989). « Handwritten digit recognition : Applications of neural network chips and automatic learning ». *IEEE Communications Magazine* 27.11, p. 41-46 (cf. p. 47).

Leite Coelho da Silva, Felipe, Kleyton da Costa, Paulo Canas Rodrigues, Rodrigo Salas et Javier Linkolk López-Gonzales (2022). « Statistical and Artificial Neural Networks Models for Electricity Consumption Forecasting in the Brazilian Industrial Sector ». *Energies* 15.2, p. 588 (cf. p. 33).

LeLynx.fr, Logo (2022). *Heures creuses/Heures pleines : les horaires et les prix*. URL : https://www.lelynx.fr/energie/comparateur-electricite/prix-electricite/heures-pleines-heures-creuses-electricite/ (cf. p. 61).

Liao, T Warren (2005). « Clustering of time series data—a survey ». *Pattern recognition* 38.11, p. 1857-1874 (cf. p. 156).

Lipton, Zachary C (2016). « The mythos of model interpretability ». *ICML Workshop on Human Interpretability in Machine Learning* (cf. p. 28).

Loader, Clive R (1999). « Bandwidth selection : classical or plug-in ? » *The Annals of Statistics* 27.2, p. 415-438 (cf. p. 240).

Lusis, Peter, Kaveh Rajab Khalilpour, Lachlan Andrew et Ariel Liebman (2017). « Short-term residential load forecasting : Impact of calendar effects and forecast granularity ». *Applied energy* 205, p. 654-669 (cf. p. 25, 66, 83, 113, 114).

Madhulatha, T Soni (2012). « An overview on clustering methods ». *arXiv preprint arXiv :1205.1117* (cf. p. 235).

Makridakis, Spyros (1993). « Accuracy measures : theoretical and practical concerns ». *International journal of forecasting* 9.4, p. 527-529 (cf. p. 53).

Makridakis, Spyros, Allan Andersen, Robert Carbone, Robert Fildes, Michele Hibon, Rudolf Lewandowski et al. (1982). « The accuracy of extrapolation (time series) methods : Results of a forecasting competition ». *Journal of forecasting* 1.2, p. 111-153 (cf. p. 53).

Makridakis, Spyros, Steven C Wheelwright et Rob J Hyndman (1998). *Forecasting : methods and applications*. T. 3. John Wiley & Sons (cf. p. 27, 49).

Mallat, Stephane (2009). *A Wavelet Tour of Signal Processing*. Academic Press (cf. p. 245).

Mallat, Stephane G (1989). « A theory for multiresolution signal decomposition : the wavelet representation ». *IEEE Transactions on pattern analysis and machine intelligence* 11.7, p. 674-693 (cf. p. 246).

Mallat, Stephane Georges (1988). *Multiresolution representations and wavelets*. University of Pennsylvania (cf. p. 41, 82).

MaPetiteEnergie (2021). *LA LIBÉRALISATION DU MARCHÉ DE L'ÉNERGIE EN EUROPE ET EN FRANCE*. https://www.monpetitforfait.com/energie/aides/liberalisation-marche-energie (cf. p. 2, 5).

Marsh, Lawrence C et David R Cormier (2001). *Spline regression models*. 137. Sage (cf. p. 243).

Metaxiotis, K, A Kagiannas, D Askounis et J Psarras (2003). « Artificial intelligence in short term electric load forecasting : a state-of-the-art survey for the researcher ». *Energy conversion and Management* 44.9, p. 1525-1534 (cf. p. 13).

Mohamed, Norizan, Maizah Hura Ahmad, Zuhaimy Ismail et S Suhartono (2010). « Short term load forecasting using double seasonal ARIMA model ». *Proceedings of the regional conference on statistical sciences*. T. 10, p. 57-73 (cf. p. 13, 14).

Mohammadi, Sirus et Ali Mohammadi (2014). « Stochastic scenario-based model and investigating size of battery energy storage and thermal energy storage for micro-grid ». *International Journal of Electrical Power & Energy Systems* 61, p. 531-546 (cf. p. 178).

Molnar, Christoph (2021). *Interpretable Machine Learning* (cf. p. 28).

Molnar, Christoph, Giuseppe Casalicchio et Bernd Bischl (2018). « iml : An R package for interpretable machine learning ». *Journal of Open Source Software* 3.26, p. 786 (cf. p. 28).

Moral-Carcedo, Julian et José Vicéns-Otero (2005). « Modelling the non-linear response of Spanish electricity demand to temperature variations ». *Energy economics* 27.3, p. 477-494 (cf. p. viii, 70).

Murphy, Allan H (1977). « The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation ». *Monthly Weather Review* 105.7, p. 803-816 (cf. p. 85).

Al-Musaylh, Mohanad S, Ravinesh C Deo, Jan F Adamowski et Yan Li (2018). « Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia ». *Advanced Engineering Informatics* 35, p. 1-16 (cf. p. 39, 117).

Nadaraya, Elizbar A (1964). « On estimating regression ». *Theory of Probability & Its Applications* 9.1, p. 141-142 (cf. p. 239).

Nason, Guy P et Bernard W Silverman (1995). « The stationary wavelet transform and some statistical applications ». *Wavelets and statistics*. Springer, p. 281-299 (cf. p. 235).

Nelder, John Ashworth et Robert WM Wedderburn (1972). « Generalized linear models ». *Journal of the Royal Statistical Society : Series A (General)* 135.3, p. 370-384 (cf. p. 34).

Nielsen, Michael A (2015). *Neural Networks and Deep Learning*. determination press (cf. p. 47).

Niu, Dong-Xiao, Qiang Wanq et Jin-Chao Li (2005). « Short term load forecasting model using support vector machine based on artificial neural network ». *2005 International Conference on Machine Learning and Cybernetics*. T. 7, 4260-4265 Vol. 7 (cf. p. 49).

Noémie (2018). *Compteur électrique – Relevés, index, PDL et consommation*. URL : https://www.capitaine-energie.com/fournisseur-electricite/consommation-electrique/compteur-electrique/ (visité le 08/11/2018) (cf. p. 80).

Nowicka-Zagrajek, Joanna et Rafal Weron (2002). « Modeling electricity loads in California : ARMA models with hyperbolic noise ». *Signal Processing* 82.12, p. 1903-1915 (cf. p. 13, 14).

Obst, David, Joseph De Vilmarest et Yannig Goude (2021). « Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France ». *IEEE transactions on power systems* 36.5, p. 4754-4763 (cf. p. 219).

Ozawa, Akito, Ryota Furusato et Yoshikuni Yoshida (2016). « Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles ». *Energy and Buildings* 119, p. 200-210 (cf. p. 154).

Özkizilkaya, Özlem (2014). « Thermosensibilité de la demande électrique : identification de la part non linéaire par couplage d'une modélisation bottom-up et de l'approche bayésienne ». Thèse de doct. Paris, ENMP (cf. p. 69, 71, 115).

Papadakis, SE, JB Theocharis, SJ Kiartzis et AG Bakirtzis (1998). « A novel approach to short-term load forecasting using fuzzy neural networks ». *IEEE Transactions on power systems* 13.2, p. 480-492 (cf. p. 15).

Papalexopoulos, Alex D et Timothy C Hesterberg (1990). « A regression-based approach to short-term system load forecasting ». *IEEE Transactions on power systems* 5.4, p. 1535-1547 (cf. p. 29, 71).

Peirson, John et Andrew Henley (1994). « Electricity load and temperature : Issues in dynamic specification ». *Energy Economics* 16.4, p. 235-243 (cf. p. 16).

Peñaloza, Ana Apolo, Roberto Chouhy Leborgne et Alexandre Balbinot (2022). « Comparative Analysis of Residential Load Forecasting with Different Levels of Aggregation ». *Engineering Proceedings* 18.1, p. 29 (cf. p. 120, 148).

Pierrot, Amandine et Yannig Goude (2011). « Short-term electricity load forecasting with generalized additive models ». *Proceedings of ISAP power* 2011 (cf. p. viii, 13, 15, 17, 36, 65, 112, 196).

Pinon, Xavier (2022). *Quel est le tarif en heures creuses pour l'électricité ?* URL : https://www.kelwatt.fr/prix/heures-creuses (visité le 28/06/2022) (cf. p. 62).

Poggi, Jean-Michel (1994). « Prévision non paramétrique de la consommation électrique ». *Revue de Statistique Appliquée* 42.4, p. 83-98 (cf. p. 41, 45).

Quilumba, Franklin L, Wei-Jen Lee, Heng Huang, David Y Wang et Robert L Szabados (2014). « Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities ». *IEEE Transactions on Smart Grid* 6.2, p. 911-918 (cf. p. 20, 113, 147, 154).

— (2015). « Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities ». *IEEE Transactions on Smart Grid* 6.2, p. 911-918 (cf. p. 154).

Rajbhandari, Yaju, Anup Marahatta, Bishal Ghimire, Ashish Shrestha, Anand Gachhadar, Anup Thapa et al. (2021). « Impact study of temperature on the time series electricity demand of urban nepal for short-term load forecasting ». *Applied System Innovation* 4.3, p. 43 (cf. p. 16).

Ramanathan, Ramu, Robert Engle, Clive WJ Granger, Farshid Vahid-Araghi et Casey Brace (1997). « Short-run forecasts of electricity loads and peaks ». *International journal of forecasting* 13.2, p. 161-174 (cf. p. 29).

— (2001). « Short-run forecasts of electricity loads and peaks ». *Essays in econometrics : Collected Papers of Clive WJ Granger Volume 1, Spectral Analysis, Seasonality, Nonlinearity, Methodology, and Forecasting*, p. 497-516 (cf. p. 16).

Räsänen, Teemu et Mikko Kolehmainen (2009). « Feature-based clustering for electricity use time series data ». *International conference on adaptive and natural computing algorithms*. Springer, p. 401-412 (cf. p. 155, 156).

Rhodes, Joshua D, Wesley J Cole, Charles R Upshaw, Thomas F Edgar et Michael E Webber (2014). « Clustering analysis of residential electricity demand profiles ». *Applied Energy* 135, p. 461-471 (cf. p. 154).

Rice, John et Murray Rosenblatt (1983). « Smoothing splines : regression, derivatives and deconvolution ». *The annals of Statistics*, p. 141-156 (cf. p. 244).

RJ, Hastie TJ Tibshirani (1990). « Generalized additive models ». *CRC Monographs on Statistics & Applied Probability. New York : Chapman & Hall* (cf. p. 35).

Rodrigues, Filipe, Carlos Cardeira et João Manuel Ferreira Calado (2014). « The daily and hourly energy consumption and load forecasting using artificial neural network method : a case study using a set of 93 households in Portugal ». *Energy Procedia* 62, p. 220-229 (cf. p. 21, 56).

Rossi, Maurizio et Davide Brunelli (2013). « Electricity demand forecasting of single residential units ». *2013 IEEE Workshop on Environmental Energy and Structural Monitoring Systems*. IEEE, p. 1-6 (cf. p. 56).

Rousseeuw, Peter J (1987). « Silhouettes : a graphical aid to the interpretation and validation of cluster analysis ». *Journal of computational and applied mathematics* 20, p. 53-65 (cf. p. 161).

Rumelhart, David E, Geoffrey E Hinton et Ronald J Williams (1986). « Learning representations by back-propagating errors ». *nature* 323.6088, p. 533-536 (cf. p. 48).

Sadaei, Hossein Javedani, Petrônio Cândido de Lima e Silva, Frederico Gadelha Guimaraes et Muhammad Hisyam Lee (2019). « Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series ». *Energy* 175, p. 365-377 (cf. p. 49).

Selectra (2021). *Compteurs communicants : quelle situation ailleurs en Europe ?* https://selectra.info/energie/actualites/expert/compteurs-communicants-europe (cf. p. 3).

Sevlian, Raffi et Ram Rajagopal (2014). « Short term electricity load forecasting on varying levels of aggregation ». *arXiv preprint arXiv :1404.0058* (cf. p. 22, 91, 148).

Shahzadeh, Abbas, Abbas Khosravi et Saeid Nahavandi (2015). « Improving load forecast accuracy by clustering consumers using smart meter data ». *2015 international joint conference on neural networks (IJCNN)*. IEEE, p. 1-7 (cf. p. 147, 155).

Shao, Zhen, Chengshan Wang, Jin Lin, Hui Li et Yingtao Liu (2019). « Nonparametric Kernel Regression for Hourly Peak Load Forecasting ». *IEEE Access* 7, p. 85316-85328 (cf. p. 29).

Shcherbakov, Maxim Vladimirovich, Adriaan Brebels, Nataliya Lvovna Shcherbakova, Anton Pavlovich Tyukov, Timur Alexandrovich Janovsky, Valeriy Anatol'evich Kamaev et al. (2013). « A survey of forecast error measures ». *World applied sciences journal* 24.24, p. 171-176 (cf. p. 53).

Shi, Heng, Minghao Xu et Ran Li (2017). « Deep learning for household load forecasting—A novel pooling deep RNN ». *IEEE Transactions on Smart Grid* 9.5, p. 5271-5280 (cf. p. 25, 49, 124).

Shrinkage, Robert Tibshirani Regression (1996). « Selection via the LASSO Journal of the Royal Statistical Society ». *Series B (Methodological)* 58.1, p. 267-288 (cf. p. 187).

Sigauke, Caston et Delson Chikobvu (2010). « Daily peak electricity load forecasting in South Africa using a multivariate non-parametric regression approach ». *ORiON* 26.2 (cf. p. 117).

Singh, Rayman Preet, Peter Xiang Gao et Daniel J Lizotte (2012). « On hourly home peak load prediction ». *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, p. 163-168 (cf. p. 56).

Steinley, Douglas et Michael J Brusco (2008). « A new variable weighting and selection procedure for K-means cluster analysis ». *Multivariate Behavioral Research* 43.1, p. 77-108 (cf. p. 98, 99).

Stephen, Bruce et Stuart J Galloway (2012). « Domestic load characterization through smart meter advance stratification ». *IEEE Transactions on Smart Grid* 3.3, p. 1571-1572 (cf. p. 155).

Stine, Robert A (1985). « Bootstrap prediction intervals for regression ». *Journal of the American Statistical Association* 80.392, p. 1026-1031 (cf. p. 178, 179).

Strang, Gilbert et Truong Nguyen (1996). *Wavelets and Filter Banks*. Wellesley-Cambridge Press (cf. p. 41).

Sugar, Catherine A et Gareth M James (2003). « Finding the number of clusters in a dataset : An information-theoretic approach ». *Journal of the American Statistical Association* 98.463, p. 750-763 (cf. p. 98, 99).

Taieb, Souhaib Ben, James W Taylor et Rob J Hyndman (2021). « Hierarchical probabilistic forecasting of electricity demand with smart meter data ». *Journal of the American Statistical Association* 116.533, p. 27-43 (cf. p. 22).

Tascikaraoglu, Akin, AR Boynuegri et Mehmet Uzunoglu (2014). « A demand side management strategy based on forecasting of residential renewable sources : A smart home system in Turkey ». *Energy and Buildings* 80, p. 309-320 (cf. p. 64).

Taylor, James W (2003). « Short-term electricity demand forecasting using double seasonal exponential smoothing ». *Journal of the Operational Research Society* 54.8, p. 799-805 (cf. p. 13, 14, 17).

Taylor, James W et Roberto Buizza (2003). « Using weather ensemble predictions in electricity demand forecasting ». *International Journal of Forecasting* 19.1, p. 57-70 (cf. p. 17, 21).

Taylor, James W, Lilian M De Menezes et Patrick E McSharry (2006). « A comparison of univariate methods for forecasting electricity demand up to a day ahead ». *International journal of forecasting* 22.1, p. 1-16 (cf. p. 14).

Taylor, James W et Patrick E McSharry (2007). « Short-term load forecasting methods : An evaluation based on european data ». *IEEE Transactions on Power Systems* 22.4, p. 2213-2219 (cf. p. 17).

Taylor, Sean J et Benjamin Letham (2018). « Forecasting at scale ». *The American Statistician* 72.1, p. 37-45 (cf. p. 202).

Teeraratkul, Thanchanok, Daniel O'Neill et Sanjay Lall (2017). « Shape-based approach to household electric load curve clustering and prediction ». *IEEE Transactions on Smart Grid* 9.5, p. 5196-5206 (cf. p. 157).

Veit, Andreas, Christoph Goebel, Rohit Tidke, Christoph Doblander et Hans-Arno Jacobsen (2014). « Household electricity demand forecasting : benchmarking state-of-the-art methods ». *Proceedings of the 5th international conference on Future energy systems*, p. 233-234 (cf. p. 18).

Viegas, Joaquim L, Susana M Vieira et João MC Sousa (2015). « Fuzzy clustering and prediction of electricity demand based on household characteristics ». *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*. Atlantis Press, p. 1040-1046 (cf. p. 155).

Voss, Marcus, Christian Bender-Saebelkampf et Sahin Albayrak (2018). « Residential short-term load forecasting using convolutional neural networks ». *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, p. 1-6 (cf. p. 25, 49).

Wand, Matt P et M Chris Jones (1994). *Kernel smoothing*. CRC press (cf. p. 239, 241).

Wang, Xu et Yusheng Xu (2019). « An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index ». *IOP Conference Series : Materials Science and Engineering*. T. 569. 5. IOP Publishing, p. 052024 (cf. p. 161).

Wang, Yizhen, Ningqing Zhang et Xiong Chen (2021). « A short-term residential load forecasting model based on lstm recurrent neural network considering weather features ». *Energies* 14.10, p. 2737 (cf. p. 72).

Wang, Yuedong (2011). *Smoothing splines : methods and applications*. CRC press (cf. p. 244).

Wangpattarapong, Kiattiporn, Somchai Maneewan, Nipon Ketjoy et Wattanapong Rakwichian (2008). « The impacts of climatic and economic factors on residential electricity consumption of Bangkok Metropolis ». *Energy and Buildings* 40.8, p. 1419-1425 (cf. p. 16).

Werbos, Paul J (1990). « Backpropagation through time : what it does and how to do it ». *Proceedings of the IEEE* 78.10, p. 1550-1560 (cf. p. 121).

Wijaya, Tri Kurniawan, SFRJ Humeau, Matteo Vasirani et Karl Aberer (2014). « Residential electricity load forecasting : evaluation of individual and aggregate forecasts ». *EPFL : Lausanne, Switzerland*, p. 1-22 (cf. p. 18, 20, 21, 25, 56, 91, 147, 154, 218).

Wijaya, Tri Kurniawan, Matteo Vasirani, Samuel Humeau et Karl Aberer (2015). « Cluster-based aggregate forecasting for residential electricity demand using smart meter data ». *2015 IEEE international conference on Big data (Big data)*. IEEE, p. 879-887 (cf. p. 56).

Wood, Simon N (2001). « mgcv : GAMs and generalized ridge regression for R ». *R news* 1.2, p. 20-25 (cf. p. 36).

— (2006). « Generalized additive models : an introduction with R. Chapman and Hall/CRC ». *Texts Stat. Sci.* 67, p. 391 (cf. p. 36).

Wu, Jianxin (2017). « Introduction to convolutional neural networks ». *National Key Lab for Novel Software Technology. Nanjing University. China* 5.23, p. 495 (cf. p. 235).

Xiao, Junwei, Jianfeng Lu et Xiangyu Li (2017). « Davies Bouldin Index based hierarchical initialization K-means ». *Intelligent Data Analysis* 21.6, p. 1327-1338 (cf. p. 161).

Yan, Ke, Wei Li, Zhiwei Ji, Meng Qi et Yang Du (2019). « A hybrid LSTM neural network for energy consumption forecasting of individual households ». *Ieee Access* 7, p. 157633-157642 (cf. p. 19, 25).

Yildiz, Baran, Jose I Bilbao, Jonathon Dore et Alistair B Sproul (2017). « Recent advances in the analysis of residential electricity consumption and applications of smart meter data ». *Applied Energy* 208, p. 402-427 (cf. p. 23, 64, 69, 78, 92, 146, 155, 157).

— (2018). « Short-term forecasting of individual household electricity loads with investigating impact of data resolution and forecast horizon ». *Renewable Energy and Environmental Sustainability* 3, p. 3 (cf. p. 20, 25, 56).

Zareipour, Hamidreza, Kankar Bhattacharya et Claudio A Cañizares (2007). « Electricity market price volatility : The case of Ontario ». *Energy policy* 35.9, p. 4739-4748 (cf. p. 68).

Al-Zayer, Jamal et Abdulla A Al-Ibrahim (1996). « Modelling the impact of temperature on electricity consumption in the eastern province of Saudi Arabia ». *Journal of Forecasting* 15.2, p. 97-106 (cf. p. 16, 70).

Zhang, Jie, Xiao Liu, Yikang Cheng et Hongqiang Li (2017). « Hourly electrical load forecasting using a novel hybrid model based on random forest and fuzzy time series ». *Applied Energy* 206, p. 52-63 (cf. p. 29).

Zhang, Tiefeng, Guangquan Zhang, Jie Lu, Xiaopu Feng et Wanchun Yang (2011). « A new index and classification approach for load pattern analysis of large electricity customers ». *IEEE Transactions on Power Systems* 27.1, p. 153-160 (cf. p. 155).

Zhang, X, Katarina Grolinger et Miriam AM Capretz (2018). « Forecasting Residential Energy Consumption Using Support Vector Regressions ». *Proceedings of the IEEE International Conference on Machine Learning and Applications, Orlando, FL, USA*, p. 17-18 (cf. p. 56).

Annexes

Annexes du chapitre 2

Quelques définitions

Méthode de *clustering* : c'est une méthode d'apprentissage automatique qui vise à regrouper les données présentant des propriétés de similitudes. La notion de similarité dans cette méthode est exprimée par le biais d'une mesure de distance. Cette méthode contrairement à la méthode de classification utilise des données non labellisées et par conséquent, les groupes sont définis à partir des caractéristiques présentes dans les données (Madhulatha, 2012).

Étape de pré-clustering : nous désignons par cette expression une étape de prétraitement qui consiste à faire du *clustering* de données avant de les utiliser. Par conséquent, le modèle de prévision ne sera pas appliqué à l'ensemble de jeu de données mais aux données de chaque *cluster* séparément.

Transformée en ondelettes stationnaire : elle est similaire à la transformée en ondelettes discrète sauf que le signal est sur-échantillonné dans chaque processus de subdivision dans le but d'éviter la perte d'informations (Nason et al., 1995).

Réseau de neurones convolutifs : c'est une sous-catégorie de réseau de neurones, largement utilisée dans le domaine de reconnaissance d'images et des données audio (Wu, 2017).

L'approche *bottom-up* : c'est une méthode de prévision ascendante qui consiste à agréger des prévisions à l'échelle individuelle (ménages, entreprises, ...) dans l'objectif d'estimer la demande électrique globale. Cette approche est souvent utilisée pour les prévisions à court terme de la demande d'électricité (Barbier, 2017 ; Bacher et al., 2017).

L'approche *top-down* : c'est une méthode de prévision descendante qui consiste à estimer la demande électrique globale d'une région ou d'un pays, puis à la décomposer en demandes plus détaillées pour chaque secteur d'activité, chaque zone géographique, groupes de clients ou encore chaque période horaire. Cette approche est souvent utilisée dans les prévisions macro-économiques et dans la planification énergétique à long terme (Barbier, 2017 ; Bacher et al., 2017).

Annexes du chapitre 3

Rappel de définitions sur les processus

Un **processus stochastique** ou encore appelé **processus aléatoire**, noté souvent par $X = (X(t))_{t\in T}$, représente une évolution, discrète ou à temps continu, d'une variable aléatoire notée $X(t)$ définie sur un espace de probabilité. Cette notion se généralise à plusieurs dimensions. Cette notion de **processus aléatoire** généralise la notion de variable aléatoire introduite en probabilités. On le définit alors comme une famille de variables aléatoires $X(t)$ associées à toutes les valeurs $t \in T$, où l'ensemble T représente souvent le temps.

On distingue généralement les **processus en temps discret** et les **processus en temps continu**, c'est-à-dire ceux à valeurs discrètes et ceux à valeurs continues. Si l'ensemble T est dénombrable nous parlons de **processus discret** ou de **série temporelle**. Si l'ensemble T est indénombrable nous parlons alors de **processus continu**. La différence n'a rien de fondamental : la stationnarité, constance en fonction du temps des propriétés statistiques, se définit de la même façon. Il ne s'agit même pas d'une différence pratique car les calculs sur un processus continu s'effectuent à partir de l'échantillonnage d'une réalisation du processus. La différence porte plutôt sur l'attitude adoptée face à l'utilisation d'une seule réalisation.

Un **processus stochastique** est dit **stationnaire** si pour toute valeur de $h > 0$, les deux processus stochastiques $(X(t+h))_{t\geq 0}$ et $(X(t))_{t\geq 0}$ ont la même loi, c'est-à-dire, si nous avons :

$$\forall h > 0 \quad \text{et} \quad t_1, \ldots, t_k \geqslant 0, \quad (X(t_1+h), \ldots, X(t_p+h)) \overset{\mathscr{L}}{=} (X_(t_1), \ldots, X(t_p)).$$

Autrement dit, si les propriétés mathématiques caractérisant ce processus sont indépendantes du temps. Il faut noter que la majorité des méthodes de modélisation des processus stochastiques exigent la vérification de l'hypothèse de stationnarité. Cependant cette hypothèse est difficilement vérifiable sur des données réelles.

Un **processus de Markov** est un processus stochastique qui vérifie la propriété de Markov, c'est-à-dire si la distribution conditionnelle de probabilité des états futurs, étant

donnés les états passés et l'état présent, ne dépend en fait que de l'état présent et non pas des états passés.

Ce processus en temps discret est une suite de variables aléatoires réelles $X_1, X_2, X_3 \ldots$. L'ensemble de leurs valeurs possibles est appelé **espace d'états**. Selon les auteurs, l'expression « chaîne de Markov » désigne les processus de Markov à temps discret et à espace d'états discret, c'est-à-dire les processus de Markov à temps discret dont l'espace d'états est fini ou dénombrable.

Méthodes de régression non paramétrique univariée

La régression locale (*Loess*)

Cette méthode a été proposée par Bill Cleveland et ses collègues à la fin des années 70 au début des années 90 (Cleveland, 1979), elle fait partie de la famille des méthodes de régression polynomiale locale. Elle est très répandue pour l'estimation de la fonction de régression non paramétrique en raison de sa simplicité et sa rapidité par rapport aux autres méthodes concurrentes. Elle combine une grande partie de la simplicité de la régression linéaire des moindres carrés avec la flexibilité de la régression non linéaire.

Dans cette partie X représente la variable indépendante et $\mathbf{x}$ est un point de X. L'idée de base du modèle est d'utiliser les k plus proches voisins de $\mathbf{x}$ pour estimer la fonction de régression $m(\mathbf{x})$ dans l'équation (3.1). Ces k plus proches voisins de $\mathbf{x}$ sont sélectionnés parmi les n observations de X dans l'échantillon. Soit $V(\mathbf{x})$ l'ensemble de k plus proches voisins de $\mathbf{x}$. La méthode de *Loess* peut alors se résumer par les étapes suivantes :

1. Identifier les k plus proches voisins de $\mathbf{x}$ (revient à identifier $V(\mathbf{x})$), pour tous les points $\mathbf{x}$ du domaine de X.
2. Calculer la distance $d(\mathbf{x})$ qui sépare $\mathbf{x}$ des points les plus éloignés de son voisinage $V(\mathbf{x})$. Cette distance est donnée par la définition suivante :

$$d(\mathbf{x}) = \max_{\mathbf{x}_i \in V(\mathbf{x})} |\mathbf{x} - \mathbf{x}_i|. \tag{7.1}$$

3. Calculer un poids pour chaque point dans $V(\mathbf{x})$ de la façon suivante :

$$w_i(\mathbf{x}) = K\left(\frac{|\mathbf{x} - \mathbf{x}_i|}{d(\mathbf{x})}\right). \tag{7.2}$$

où K est une fonction vérifiant certaines caractéristiques comme :

(a) $K(u) \geq 0 \quad \forall u$.

(b) $K(u) = 0 \quad \forall u > 1.$

La fonction K utilisée d'habitude dans la littérature est une fonction cubique pondérée la forme :

$$K(u) = \begin{cases} (1-u^3)^3 & \text{pour } 0 \leq u < 1 \\ 0 & \text{sinon.} \end{cases}$$

Cette fonction permet d'accorder un poids plus élevé aux points du voisinage $V(\mathbf{x})$ les plus proches de $\mathbf{x}$.

4. Calculer l'estimateur $\hat{m}$ au point $\mathbf{x}$ par un polynôme de degré à déterminer (souvent degré 1 ou 2) à l'aide de la méthode de moindres carrés pondérés appliquée à l'ensemble de voisinage $V(\mathbf{x})$. Soit $P(\mathbf{x}) = \sum_{i=0}^{d} \beta_i \mathbf{x}^i$ le polynôme de degré d qui va servir à calculer l'estimateur $\hat{m}$ au point $\mathbf{x}$ et $\hat{\beta} = (\hat{\beta_0}, \ldots, \hat{\beta_d})$ le vecteur des coefficients à estimer. Le vecteur $\hat{\beta}$ est alors estimé en minimisant la quantité suivante :

$$\sum_{i=1}^{k} w_i(\mathbf{x})(\mathbf{y}_i - \beta_0 - \cdots - \beta_d \mathbf{x}_i^d)^2 \quad \forall \mathbf{x}_i \in V(\mathbf{x}). \tag{7.3}$$

Le point lissé en $\mathbf{x}$, en utilisant une régression localement pondérée de degré d est alors $(\mathbf{x}, \hat{\mathbf{y}})$, où $\hat{\mathbf{y}}$ est la valeur ajustée de la régression au point $\mathbf{x}$. Par suite nous pouvons écrire :

$$\hat{\mathbf{y}} = \sum_{j=0}^{d} \hat{\beta}_j(\mathbf{x})\mathbf{x}^j = \sum_{i=1}^{k} l_i(\mathbf{x})\mathbf{y}_i. \tag{7.4}$$

L'équation (7.4) peut s'écrire sous une forme matricielle de la façon suivante : $\hat{\mathbf{Y}} = \mathbf{LY}$ où $\mathbf{L}$ est dite la matrice de lissage, $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ est le vecteur des valeurs observées et $\hat{Y} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n)$ le vecteur qui correspond aux valeurs prédites de la variable X.

En pratique, la méthode n'est pas appliquée pour tous les points $\mathbf{x}$ de l'échantillon de X. Un ensemble de points de X est sélectionné pour lequel nous calculons la valeur de l'estimateur. Ensuite, une méthode d'interpolation est utilisée pour obtenir l'estimateur de X en tout point $\mathbf{x}$.

Cette approche dépend du nombre k des points dans le voisinage $V(\mathbf{x})$. Plus le nombre k de points est grand plus l'estimateur est lisse. Le paramètre k est souvent exprimé en pourcentage par rapport à la taille de l'échantillon et est considéré comme paramètre de lissage. La détermination de la valeur optimale de k peut se faire en minimisant le critère d'information d'Akaike corrigé (AIC_{C1}) (Hurvich et al., 1998) donné par :

$$AIC_{C1} = n \log(\hat{\sigma}^2) + n \frac{\delta_1/\delta_2(n + \nu_1)}{\delta_1^2/\delta_2 - 2}. \tag{7.5}$$

avec :
n = nombre des observations dans l'échantillon.
$\delta_1 = \mathrm{Tr}(I - L)^T(I - L)$.

$\delta_2 = \mathrm{Tr}((I-L)^T(I-L))^2$.
$\nu_1 = \mathrm{Tr}(L)$.
$\hat{\sigma}^2 = \frac{1}{\delta_1}\sum_{i=1}^{n} \hat{\epsilon_i}^2$ où $\hat{\epsilon}$ est le vecteur des résidus.

Pour plus d'informations sur la méthode de *Loess*, l'auteur peut se référer à (Gijbels et al., 2010).

La régression à noyau

L'estimation par régression à noyau (Wand et al., 1994) consiste à estimer la fonction de régression (3.2) pour une valeur donnée $\mathbf{x}$ de X par une moyenne pondérée des observations de la variable Y par l'intermédiaire d'un polynôme de degré d et d'une fonction noyau qui attribue un coefficient de pondération ou poids élevé au point $\mathbf{x}$ et un poids proportionnel aux autres valeurs de X en fonction de la distance qui les sépare de $\mathbf{x}$. L'une des méthodes les plus populaires de la régression à noyau a été proposée par (Nadaraya, 1964) et est connue sous le nom d'estimateur de Nadaraya-Watson. Le noyau noté souvent K est une fonction qui vérifie les trois conditions suivantes :

1. $K(t) > 0$,
2. $K(-t) = K(t) \quad \forall t$,
3. $\int_{\mathbb{R}} K(t)\,dt = 1$.

Parmi les noyaux les plus utilisés nous citons :

Noyau	Expression
Uniforme	$K(t) = \frac{1}{2}\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$
Epanechnikov	$K(t) = \frac{3}{4}(1-t^2)\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$
Triangle	$K(t) = (1-\lvert t\rvert)\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$
Guassien	$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

Table 7.9 – Les quatre noyaux les plus utilisés pour l'estimation de la fonction de régression par la méthode de noyau.

Pour estimer la fonction de régression $m(\mathbf{x})$ (3.2) au point $\mathbf{x}$ à l'aide de la méthode de Nadaraya-Watson, la moyenne pondérée des observations $(\mathbf{y}_i)_{i\in 1,\ldots,n}$ de Y est calculée pour chaque point $\mathbf{x}$ de X de la façon suivante :

$$\hat{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})\mathbf{y}_i}{\sum_{i=1}^{n} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})}. \tag{7.6}$$

où h est un paramètre nommé fenêtre, qui contrôle le degré de lissage de l'estimation. En

pratique, l'estimateur est calculé sur une grille des points plutôt que sur l'ensemble du domaine de X, ensuite une méthode d'interpolation est utilisée pour former un estimateur continu.

Cette méthode peut être vue comme un cas particulier de l'estimation polynomiale locale. En effet, Ceci n'est pas réalisable directement, car aucune connaissance sur m n'est disponible. Cependant, d'après la formule de Taylor d'ordre d, il est possible d'écrire m pour tout point $\mathbf{x}$ au voisinage de $\mathbf{x}_i$ par :

$$m(\mathbf{x}_i) \approx m(\mathbf{x}) + m(\mathbf{x})'(\mathbf{x}_i - \mathbf{x}) + \cdots + \frac{m^{(d)}(\mathbf{x})}{d!}(\mathbf{x}_i - \mathbf{x})^d. \tag{7.7}$$

Ensuite, l'estimation par polynôme local consiste à minimiser la somme résiduelle des carrés alors le problème revient à minimiser la quantité suivante :

$$\sum_{i=1}^{n} (\mathbf{y}_i - m(\mathbf{x}_i))^2 \tag{7.8}$$

En remplaçant $m(\mathbf{x}_i)$ dans 7.8 par son expression dans la formule de Taylor (7.7) nous obtenons :

$$\sum_{i=1}^{n} \left(\mathbf{y}_i - \sum_{j=0}^{d} \frac{m^{(j)}(\mathbf{x})}{j!}(\mathbf{x}_i - \mathbf{x})^j \right)^2. \tag{7.9}$$

En posant $\beta_j := \frac{m^{(j)}(\mathbf{x})}{j!}$ le problème se transforme en un problème de régression linéaire où les paramètres inconnus sont $\beta = (\beta_0, \ldots, \beta_d)$:

$$\sum_{i=1}^{n} \left(\mathbf{y}_i - \sum_{j=0}^{d} \beta_j (\mathbf{x}_i - \mathbf{x})^j \right)^2. \tag{7.10}$$

Enfin, en intégrant les poids ou les coefficients de pondération dans l'équation (7.10) le problème de minimisation prend la forme suivante :

$$\underset{\beta \in \mathbb{R}^{d+1}}{\arg\min} \sum_{i=1}^{n} \left(\mathbf{y}_i - \sum_{j=0}^{d} \beta_j (\mathbf{x}_i - \mathbf{x})^j \right)^2 K_h(\mathbf{x} - \mathbf{x}_i). \tag{7.11}$$

En général, la méthode du noyau est assez robuste par rapport au choix du noyau comparativement au choix de la valeur du paramètre de lissage h (Guisan et al., 2002). Loader (1999) a proposé les deux critères CV (*cross validation*) et GCV (*generalized cross validation*) pour la sélection du paramètre de lissage h.

Pour plus d'informations sur la régression par la méthode du noyau, le lecteur peut se référer à (Wand et al., 1994).

Les *splines* de régression

La méthode des *splines* de régression (Hastie et al., 1988 ; Eubank, 1999) est une technique de régression non-linéaire largement reconnue en analyse numérique pour sa simplicité et sa facilité d'interprétation. Elle offre une alternative à la méthode de régression polynomiale, et est souvent considérée comme supérieure. D'après Eubank (1999), en utilisant le théorème de Taylor, il est possible de réécrire le modèle de régression univariée présenté dans l'équation (3.1)) pour tout $\mathbf{x} \in X$ comme :

$$\mathbf{y} = \sum_{i=1}^{m} \alpha_i \mathbf{x}^{i-1} + [(m-1)!]^{-1} \int m^{(m)}(\mathbf{x})(\mathbf{x} - \xi)_+^{m-1} d\xi + \epsilon, \tag{7.12}$$

où $m^{(m)}(x)$ est la m ème dérivée de $m(x)$ et

$$(x - \xi)_+ = \begin{cases} x - \xi & \text{si } x \geq \xi \\ 0 & \text{si } x < \xi \end{cases} \tag{7.13}$$

Grâce à Hastie et al. (1988), l'intégrale dans (7.12) peut être représentée par une somme de puissances tronquées et par suite l'estimateur de $m_\Gamma(\mathbf{x})$ est réécrit comme suivant :

$$\hat{m}_\Gamma(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \mathbf{x}^{i-1} + \sum_{i=1}^{k} \beta_i (\mathbf{x} - \xi_i)_+^{m-1}, \tag{7.14}$$

et $\Gamma = (\xi_1, \ldots, \xi_k)$ un vecteur de points dit vecteur des nœuds. Toute fonction de l'ensemble $\{x^0, \ldots, x^{m-1}, (x-\xi_1)_+^{m-1}, \ldots, (x-\xi_k)_+^{m-1}\}$ fait partie de la famille des fonctions *splines*. Ces fonctions sont des fonctions continues, polynomiales par morceaux de degré $m-1$. L'équation dans (7.14) peut s'écrire également de la façon suivante :

$$\hat{m}_\Gamma(\mathbf{x}) = \sum_{i=1}^{m+k} \theta_i B_i(\mathbf{x}), \tag{7.15}$$

où $B_i(x) = x^{i-1}, i = 1, \ldots, m$ et $B_{m+i}(x) = (x - \xi_i)_+^{m-1}, i = 1, \ldots, k$ forment ensemble la base des puissances tronquées et $\Theta = (\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_k)$ le vecteur de coefficients de régression.

Pour un ensemble fixe de nœuds Γ, une spline de régression est linéaire et peut être estimée paramétriquement en utilisant les techniques des moindres carrés. Les coefficients Θ du modèle sont alors calculés par :

$$\hat{\Theta}_\Gamma = \arg\min_\theta \sum_{i=1}^{n} (\mathbf{y}_i - \sum_{j=1}^{m+k} \theta_j B_j(\mathbf{x}_i))^2. \tag{7.16}$$

Si la matrice $\mathbf{X}_\Gamma = \{B_j(\mathbf{x}_i)\}_{i\in\{1,\ldots,n\},j\in\{1,\ldots,m+k\}}$ alors l'estimateur de Θ_Γ est donné par :

$$\hat{\Theta}_\Gamma = ({\mathbf{X}_\Gamma}^T \mathbf{X}_\Gamma)^{-1} {\mathbf{X}_\Gamma}^T \mathbf{Y}. \tag{7.17}$$

Bien que la base de puissance tronquée soit simple, elle représente certains problèmes de stabilité numérique qui réduisent sa porté. En effet, les supports de certaines fonctions de la base sont définis sur tout le domaine de définition des données ce qui peut entraîner des corrélations entre certaines *splines* de base et causer des problèmes d'instabilités numériques. Les B-splines fournissent une représentation alternative pratique et plus stables numériquement de la base de puissance tronquée. La représentation des B-*splines* se fait par une série des fonctions de base polynomiales qui ont un support compact et local.

Puisqu'ils sont définis localement leur construction nécessite alors une augmentation de la séquence des nœuds ainsi qu'une relation récursive pour construire les fonctions de la base. Afin de construire ces bases *splines*, nous devons ajouter $2m$ nœuds au vecteur $\Gamma = \{\xi_1, \ldots, \xi_k\}$ de la façon suivante : soient $\xi_0 < \xi_1$ et $\xi_k < \xi_{k+1}$ deux nœuds frontières, le vecteur de nœuds augmenté $\Gamma' = \{\tau_1, \ldots, \tau_{k+2m}\}$ est alors défini par :

1. $$\tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_m \leqslant \xi_0.$$
2. $$\tau_{j+m} = \xi_j, \quad \forall j = 1, \ldots, k.$$
3. $$\xi_{k+1} \leqslant \tau_{k+m+1} \leqslant \tau_{k+m+2} \leqslant \ldots \leqslant \tau_{k+2m}.$$

En effet, le choix des m premiers nœuds et des m derniers nœuds n'est pas spécifié. De Boor et al. (1978) a proposé de choisir :

$$\tau_1 = \tau_2 = \cdots = \tau_m = \xi_0 \quad \text{et} \quad \xi_{k+1} = \cdots = \tau_{k+m+1} = \tau_{k+m+2} = \cdots \tau_{k+2m}.$$

Les B-splines sont alors définies de manière récursive par :

$$N_i^m(x) = \left(\frac{x - \tau_i}{\tau_{i+m-1} - \tau_i}\right) N_i^{m-1}(x) + \left(\frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}}\right) N_{i+1}^{m-1}(x),$$

où

$$N_i^1(x) = \begin{cases} 1 & \text{si } x \in [\tau_i, \tau_{i+1}) \\ 0 & \text{sinon} \end{cases}$$

pour $i = 1, \ldots, k + m + 1$, avec par convention $(N_i^1(x) = 0 \quad \text{si} \quad \tau_i = \tau_{i+1})$. Dans ce cas,

l'estimateur du vecteur de coefficients dans (7.18) prend la forme suivante :

$$\hat{\Theta}_\Gamma = (N_\Gamma{}^T N_\Gamma)^{-1} N_\Gamma{}^T \mathbf{Y}, \tag{7.18}$$

où la matrice $N_\Gamma = \{N_j^m(\mathbf{x_i})\}_{i\in\{1,\ldots,n\},j\in\{1,\ldots,m+k\}}$.

En pratique, l'ensemble Γ n'est pas connu, l'estimation alors de $\hat{m_\Gamma}$ par les *splines* de régression nécessitent l'estimation de Γ, Θ et m (ordre des *splines*). En effet, il existe plusieurs méthodes d'estimation des paramètres Γ, Θ contrairement au paramètre m qui est considéré fixe au départ pour toutes les méthodes d'estimation des *splines* de régression.

Le choix du nombre k des nœuds et de leurs emplacements joue un rôle très important puisque un mauvais choix peut biaiser l'estimateur et conduire à des résultats erronés. L'une des méthodes propose de fixer un nombre k de nœuds et de sélectionner leurs emplacements d'une façon uniforme sur le domaine. Bien que cette méthode prenne en considération la dispersion des données, elle manque d'adaptabilité.

Il existe également des méthodes plus automatiques qui estiment le vecteur des nœuds et le vecteur de coefficients de régression Θ simultanément mais exigent la fixation du nombre des nœuds au départ ce qui n'est pas assez pratique.

D'autres méthodes dites adaptatives permettent d'estimer les splines de régression sans fixer le nombre et les emplacements de nœuds (Cooley et al., 1973). Ils sont calculés par ajustement à partir des données.

Pour plus d'informations sur les *splines* de régression, le lecteur peut se référer à (Eubank, 1999) ou (Marsh et al., 2001).

Les *splines* de lissage pour la régression

L'estimateur *spline* de lissage (Eubank, 1999) est une extension importante de l'estimateur de spline de régression. Ces estimateurs contournent le problème de la sélection des nœuds en plaçant des nœuds à tous les points de l'échantillon de la variable représentant la variable explicative X. Le vecteur des nœuds est alors donné par $(\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \}$ avec $\mathbf{x}_i$ une valeur de $X)$. Ils utilisent simultanément un terme de pénalité pour contrôler la complexité. L'estimation de la fonction de régression m par cette méthode se fait via la minimisation de la somme des carrés résiduelle pénalisée suivante :

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - m(\mathbf{x}_i))^2 + \lambda \int m^{(m)}(t)^2\, dt, \tag{7.19}$$

où $\lambda \in [0, +\infty)$ est un paramètre qui contrôle le degré de lissage de l'estimateur. Il contrôle la complexité en établissant un compromis entre le premier terme qui mesure l'erreur de l'estimation, et le second terme qui pénalise la courbure de la fonction. Plus la valeur

de λ est petite plus l'estimateur est flexible. Par contre, lorsque la valeur de λ augmente l'estimateur devient plus lisse. En pratique, il existe de nombreuses méthodes pour choisir le paramètre de lissage λ, telles que la méthode CV (*Cross Validation*) et la méthode GCV (*Generalized Cross-Validation*) (Yuedong Wang, 2011).

Eubank (1999) a démontré que la solution unique qui minimise le critère dans (7.19) parmi toutes les fonctions m ayant des dérivées continues jusqu'à l'ordre $(m-1)$ et la m$^{\text{ème}}$ dérivée est de carré intégrable est une spline naturelle (une fonction polynomiale par morceaux de degré $2m-1$) ayant pour nœuds les valeurs distinctes de la variable $X : \{\mathbf{x}_i, i = 1, \ldots, n, \quad \text{avec} \quad \mathbf{x}_i < \mathbf{x}_j, \forall i < j\}$ et prend la forme d'un polynôme de degré $(m-1)$ sur les intervalles $(-\infty, \mathbf{x}_1]$ et $[\mathbf{x}_n, +\infty)$ (en dehors des valeurs prises par la variable aléatoire X). Le nombre de paramètres à estimer dans ce cas est égal à n le nombre de points dans l'échantillon (nœuds) uniquement puisque les autres coefficients sont pénalisés par les contraintes imposées aux extrémités de l'estimateur. L'estimateur par *splines* de lissage prend alors la forme suivante :

$$\hat{m}_\lambda(\mathbf{x}) = \sum_{i=1}^{n} \phi_i N_i(\mathbf{x}), \tag{7.20}$$

où N_i est la i$^{\text{ème}}$ base de spline naturelle. Par suite l'estimateur du vecteur de coefficients $\Phi = (\phi_1, \ldots, \phi_n)$ s'écrit comme suivant :

$$\hat{\Phi}_\lambda = (S^T S + \lambda \Omega_n)^{-1} S^T \mathbf{Y}, \tag{7.21}$$

où la matrice $S = \{N_j(\mathbf{x}_i)\}_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,n\}}$ et $\Omega_n = \{\int N_i^{(m)}(t) N_j^{(m)}(t)\, dt\}$.

Les *splines* de lissage fournissent souvent des ajustements similaires à ceux de la régression par noyau. Le paramètre dit fenêtre h de la régression par noyau et le paramètre de lissage λ pour les *splines* de lissage sont tous les deux généralement déterminés par la méthode CV (*Cross Validation*). Cependant, la méthode de *splines* de lissage reste plus simple à utiliser en pratique puisque la méthode de régression par noyau nécessite en plus le choix d'un noyau convenable.

Finalement, malgré leurs avantages et leur utilité les *splines* de lissage ne sont pas facilement généralisables au cas de régression multivarié. Une raison pour laquelle il est intéressant d'avoir recours à d'autres méthodes comme le modèle *GAM* présentée dans la partie 3.2.1.2.

Le lecteur intéressé par plus d'informations sur les *splines* de lissage peut se référer à (Rice et al., 1983) ou à (Yuedong Wang, 2011).

La transformée en ondelettes

La transformée en ondelettes (S. Mallat, 2009) et la décomposition en série de Fourier sont les deux techniques les plus célèbres dans le domaine de l'analyse du signal. Contrairement à l'analyse de Fourier la transformée en ondelettes offre la possibilité d'analyser un signal simultanément dans le domaine du temps et des fréquences. Cette technique utilise des fonctions d'ondelettes, qui sont des signaux oscillants qui varient en fréquence et en amplitude pour décomposer la fonction en une somme de deux composantes : l'approximation et les détails. L'approximation modélise la variation de la tendance globale (évolution du niveau moyen de la fonction), alors que la composante de détails modélise les changements localisés en temps et en fréquences à différentes résolutions. En 1909, Alfred Haar a défini la première ondelette (Ondelette de Haar) comme une fonction composée d'une courte impulsion négative suivie d'une courte impulsion positive [8].

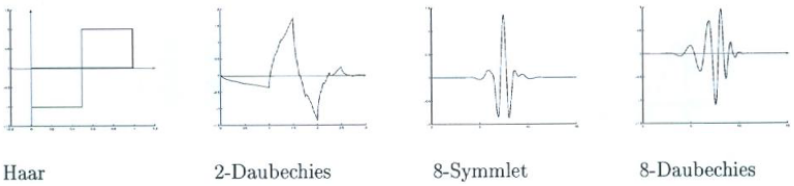


figure 7.35 – Exemples d'ondelettes [Source (Charles et al., 2004)].

En mathématiques, une ondelette ψ est une fonction de carré sommable de l'espace d'Hilbert $L^2(\mathbb{R})$, le plus souvent oscillante et de moyenne nulle. La figure 7.35 montre plusieurs exemples d'ondelettes. Une famille d'ondelettes est généralement constituée par des ondelettes analysantes issues de la translation et la dilatation d'une seule fonction dite ondelette mère $\psi(t)$. Elles sont notées par $\psi_{a,b}$ et localisées en temps autour d'un paramètre b et oscillant à une fréquence égale à $1/a$. La famille d'ondelettes est alors générée par la modification des valeurs de deux paramètres a et b dans l'espace de temps et de fréquence. Elle prend alors la forme suivante :

$$\left\{\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}), \quad a \in \mathbb{R}_+^*, b \in \mathbb{R}\right\},$$

où b est le paramètre du temps (délai) ou de la position et a le paramètre d'échelle. Il existe de nombreuses ondelettes mères ψ dans la littérature, parmi lesquelles nous citons :

1. L'ondelette mère de Morlet donnée par l'équation suivante :

$$\psi(t) = \cos(5t)e^{-\frac{t^2}{2}}.$$

8. https://wikimonde.com/article/Ondelette

Cette ondelette est régulière, complexe et symétrique.

2. L'ondelette mère « chapeau mexicain » donnée par l'équation suivante :

$$\psi(t) = (2 - t^2)e^{-\frac{t^2}{2}}.$$

Dans le cas de la transformée en ondelettes dyadiques, la famille d'ondelettes est constituée de la même façon par dilatation et translation d'une ondelette mère sauf que les deux paramètres a et b doivent suivre une suite géométrique de raison 2 ($a = 2^{-j}$ et $b = k2^{-j}$ avec $j, k \in \mathbb{Z}$). La famille d'ondelettes dyadiques générée par l'ondelette mère prend alors la forme suivante :

$$\left\{\psi_{j,k}(t) = \sqrt{2^j}\psi(2^j t - k) \quad \forall j, k \in \mathbb{Z}\right\}. \tag{7.22}$$

Cette famille d'ondelettes fournit une base orthonormée de l'espace $L^2(\mathbb{R})$. Par suite, toute fonction $f(t) \in L^2(\mathbb{R})$ pourra s'écrire sous la forme :

$$f(t) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}, \tag{7.23}$$

où les produits scalaires $< f, \psi_{j,k} >$ sont appelés les coefficients d'ondelettes dans la base $\{\psi_{j,k}\}$.

En 1989, Stéphane Mallat fit le lien entre les ondelettes et l'analyse multirésolution (Stephane G Mallat, 1989). L'analyse multirésolution d'une fonction $f \in L^2(\mathbb{R})$ permet de la décomposer sur plusieurs niveaux de résolution. Elle consiste à la projection de cette dernière sur des bases de fonctions donnant des approximations de moins en moins fines de la fonction originale. Nous définissons une approximation multirésolution comme une suite de sous-espaces vectoriels fermés $\{V_j\}_{j\in\mathbb{Z}}$ de $L^2(\mathbb{R})$ emboîtés les uns dans les autres tels que le passage de l'un à l'autre résulte d'un changement d'échelle. Ces sous-espaces vectoriels vérifient les propriétés suivantes :

1. $\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots \subset V_{j+1} \subset V_j \subset L^2(\mathbb{R})$.
2. $\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R})$.
3. $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$.
4. $\forall j \in \mathbb{Z}, f(t) \in V_j \Longleftrightarrow f(\frac{t}{2}) \in V_{j+1}$.
5. $\forall k \in \mathbb{Z}, f(t) \in V_0 \Longleftrightarrow f(t-k) \in V_0$.
6. Il existe une fonction $\phi \in V_0$ telle que $\{\phi(t-k), k \in \mathbb{Z}\}$ soit une base orthonormée de V_0.

L'AMR d'une fonction $f(t) \in L^2(\mathbb{R})$ consiste alors à calculer les projections de $f(t)$ appartenant à l'espace V_j sur un sous-espace V_{j+1} et un sous-espace W_{j+1} afin de réduire

la résolution par 2. Le sous-espace V_{j+1} est dit le sous-espace d'approximation et le sous-espace W_{j+1} est dit le sous-espace de détails.
Les relations successives d'inclusion dans la première propriété indiquent que V_{j+1} est un sous-espace de V_j et par suite la projection de f sur V_{j+1} donne une moins bonne approximation par rapport à sa projection sur V_j autrement dit l'approximation sur V_j contient toutes les informations nécessaires pour calculer l'approximation sur V_{j+1}. La deuxième propriété signifie que pour une résolution infinie ($j \to -\infty$), l'approximation converge bien vers la fonction f. Contrairement à la troisième propriété qui indique que pour une résolution nulle ($j \to +\infty$) toute l'information sur la fonction f est perdue.

À partir de ces propriétés, il a été démontré l'existence d'une fonction d'échelle $\phi(t) \in L^2(\mathbb{R})$ et d'une fonction d'ondelettes $\psi(t) \in L^2(\mathbb{R})$ qui engendrent par translation et par dilatation une base orthonormée respectivement de V_{j+1} et W_{j+1}. Ces deux sous-espaces possèdent par construction [9] des propriétés intéressantes, ils sont supplémentaires avec $V_j = V_{j+1} \bigoplus W_{j+1}$. De plus si les bases sont orthogonales alors ils sont orthonormaux $V_{j+1} \perp W_{j+1}$.
Les deux fonctions de bases dilatées et translatées de ϕ et ψ sont définies par $\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k)$ et $\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k)$ avec $k \in \mathbb{Z}$. D'après tout ce qui précède, $\forall j_0 \in \mathbb{Z}$, l'ensemble suivant constitue une base orthonormale de l'espace $L^2(\mathbb{R})$,

$$\{\phi_{j_0,k}, k = 0, 1, \ldots, 2^{j_0} - 1; \psi_{j,k}, j \geq j_0, k = 0, 1, \ldots, 2^j - 1\}.$$

Et par suite, toute fonction f de l'espace $L^2(\mathbb{R})$ admet une décomposition en ondelettes de la forme suivante :

$$f = \sum_{k \in \mathbb{Z}} a_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} d_{j,k}\psi_{j,k}, \tag{7.24}$$

où $a_{j_0,k} =< f, \phi_{j_0,k} >$ les coefficients d'approximation ou d'échelle et $d_{j,k} =< f, \psi_{j,k} >$ les coefficients d'ondelettes ou détails.

Régression à noyau d'un processus multivarié

La régression à noyau dans le cas du processus multivarié est une extension de la méthode de régression à noyau pour les données univariées (voir la partie « La régression à noyau » dans l'annexe 6.2). C'est une méthode flexible et non paramétrique qui permet de modéliser la relation entre une variable réponse multivariée et une ou plusieurs va-

9. La relation d'emboîtement des sous-espaces vectoriels V_j implique que les projections de f sur ces sous-espaces sont de plus en plus grossières avec l'incrémentation de j. La différence alors entre l'approximation sur V_j et celle sur V_{j+1} représente l'information de détail perdue par la transition du niveau de résolution j à $j + 1$.

riables explicatives. Dans cette partie, nous présentons la méthode de régression à noyau dans le cas d'un processus multivarié à la base de la construction du modèle *KWF*. Les informations contenues dans cette partie sont tirées de Cugliari (2011).

L'objectif de cette méthode est de modéliser le comportement d'un processus $(Y_t)_t$ et de prédire la valeur de Y_{n+h} à l'horizon h, en utilisant uniquement les observations passées de $Y_1, \ldots, Y_n$.

Soit $Y = (Y_t, t \in \mathbb{N})$ un processus stochastique discret à valeurs réelles. Supposons que ce processus est un processus de Markov (voir définition dans l'annexe 6.2) alors pour $s \geq 0$ nous avons :

$$\mathbb{E}[Y_n|Y_{n-1}, \ldots, Y_0] = \mathbb{E}[Y_n|Y_{n-1}, \ldots, Y_{n-s}]. \tag{7.25}$$

L'équation de régression du processus $(Y_t)_t$ est donnée par :

$$Y_{n+1} = m(\mathcal{Y}_{n,s}) + \epsilon_n, \tag{7.26}$$

où $\mathcal{Y}_{n,s} = (Y_n, Y_{n-1}, \ldots, Y_{n-s})$, $m(y) = \mathbb{E}(Y_{n+1}|\mathcal{Y}_{n,s})$, y un vecteur de $\mathbb{R}^s$ et ϵ_n un processus vectoriel de bruit blanc de dimension s. L'estimateur à noyau $\hat{m_n}$ de m s'écrit alors de la façon suivante :

$$\hat{m_n}(y) = \sum_{i=s}^{n-h} w_{n,i}(y) Y_{i+h}, \tag{7.27}$$

où $w_{n,i}(y) = \frac{K_{h_n}(\mathcal{Y}_{i,s} - y)}{\sum_{i=s}^{n-h} K_{h_n}(\mathcal{Y}_{i,s} - y)}$ avec $K : \mathbb{R}^s \mapsto \mathbb{R}$ une fonction de densité de probabilité multivariée symétrique et $K_{h_n}(.) = K(./h_n)$, $h_n > 0$. Les poids sont $w_{n,i} \geq 0 \quad \forall i \in \{s, \ldots, n-h\}$ avec $\sum_s^{n-h} w_{n,i} = 1$. Le prédicteur à noyau prend alors la forme suivante :

$$\widehat{Y_{n+h}} = \hat{m_n}(\mathcal{Y}_{n,s}). \tag{7.28}$$

Finalement, la prévision du processus Y à l'horizon h est calculée par :

$$\widehat{Y_{n+h}} = \sum_{i=s}^{n-h} w_{n,i}(\mathcal{Y}_{n,s}) Y_{i+h}. \tag{7.29}$$

Le déploiement massif des compteurs intelligents dans le secteur résidentiel et tertiaire a permis de récolter des données de consommation électrique de haute fréquence à l'échelle des consommateurs (particuliers, professionnels, ...). Ces données constituent une matière première pour les recherches sur la prévision de la consommation de l'électricité à cette échelle. La majorité de ces recherches visent largement à répondre aux besoins du milieu industriel comme les applications dans le contexte des maisons intelligentes ainsi que les programmes de la gestion et de la réduction de la consommation. L'objectif de ce travail est de déployer ou mettre en œuvre des modèles de prévision de la charge électrique à court terme $(J+1)$ à l'échelle des consommateurs qui seront intégrés dans des applications à des fins industrielles. La complexité du sujet réside dans le fait que les données de consommation à cette échelle sont très volatiles. En effet, elles comprennent une grande quantité de bruit et dépendent du mode de vie du consommateur et de ses habitudes de consommation. De plus, le déploiement de ces modèles dans des applications industrielles ajoute une certaine complexité supplémentaire au sujet. Durant cette thèse, plusieurs modèles ont été déployés pour la prévision de charge des ménages en tenant compte de différentes contraintes industrielles. Les modèles ont été testés et évalués sur un grand échantillon de courbes de charges disparates du secteur résidentiel et tertiaire. Une approche a été proposée également pour la prévision des courbes de charge les plus volatiles. Ensuite, nous avons étudié l'impact de l'utilisation des prévisions à l'échelle des ménages sur la prévision de la charge agrégée. Les modèles testés ont été également adaptés pour répondre aux besoins en prévision de l'entreprise pour d'autres portefeuilles.

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
UMR 7501
Université de Strasbourg
CNRS
IRMA, UMR 7501
7 rue René Descartes
F-67000 STRASBOURG

Tél. 03 68 85 01 29

irma.math.unistra.fr
irma@math.unistra.fr

IRMA 2023/004
http://tel.archives-ouvertes.fr/tel-04270352