Fouille de données pour l'extraction de profils d'usage et la prévision dans le domaine de l'énergie

Fateh Melzi

UNIVERSITÉ PARIS-EST

IFSTTAR

École Doctorale Mathématiques et Sciences et Technologies de l'Information et de la Communication

Thèse de doctorat

Discipline : Informatique

présentée par

Fateh Nassim MELZI

Fouille de données pour l'extraction de profils d'usage et la prévision dans le domaine de l'énergie

---

dirigée par **Latifa OUKHELLOU** et **Allou SAME**

Soutenue le 17/10/2018 devant le jury composé de :

M. Moamar Sayed-Mouchaweh	IMT Lille Douai	rapporteur
M. Fahed Abdallah	L'Université libanaise	rapporteur
M. Patrice Aknin	IRT SystemX	examinateur
Mme. Amira Ben hamida	IRT SystemX	examinatrice
M. Frédéric Héliodore	General Electric	examinateur
Mme. Latifa Oukhellou	IFSTTAR	directrice
M. Allou SAMÉ	IFSTTAR	co-directeur

Remerciement

Je tiens tout d'abord à exprimer mes plus profonds remerciements à ma directrice de thèse Mme Latifa OUKHELLOU pour avoir accepté de diriger cette thèse. Je lui exprime toute mon estime et la remercie pour ses précieux conseils et encouragements. J'aimerai également remercier mon co-directeur M. Allou SAME de m'avoir offert un encadrement de qualité, avec une grande disponibilité, merci encore Allou.

Je tiens également à remercier les membres du jury de me faire l'honneur d'évaluer ce travail : M. Patrice AKNIN directeur scientifique de l'IRT SystemX, Mme Amira BEN HAMIDA chef de projet à l'IRT SystemX, Frédéric HELIODORE responsable d'axe de recherche à General Electric ainsi que les deux rapporteurs M. Moamar SAYED-MOUCHAWEH professeur à l'école nationale supérieur des Mines-Télécom Lille-Douai et M. Fahed ABDALLAH professeur à l'université libanaise. Merci pour votre lecture attentive de la thèse et vos remarques constructives.

Je tiens aussi à adresser mes remerciements les plus chaleureux à l'ensemble des membres du projet SCE, en particulier à mon référent Mohamed Haykel ZAYANI pour son aide, son soutien, ses précieux conseils et sa disponibilité à tout moment. J'adresse ma sympathie à l'ensemble des personnes que j'ai côtoyées durant cette thèse : Oussama, Ahmed, Amira, Mustapha, Fallilou, Mouadh, Lilia, Rym, Mallek, Kahina, Alessandro, Aymen, Mouncef, Laura, Anne-sarah, Josquin, Florence, Maxime, Milade, Florian, Alexis, Elise, Ferhat, Abderrahmane, Enoch ...

Un grand merci pour mes amis qui m'ont toujours soutenu, supporté mon stress, mes problèmes et d'avoir tout fait pour m'aider : Moucha, Lilia, Mamia, Nedjmou, Oussama, Haykel...

Je tiens à exprimer mes plus profonds remerciements à mes parents et mes deux sœurs qui m'ont toujours encouragé et soutenu depuis l'autre côté de la méditerranée.

Résumé de la thèse

De nos jours, les pays sont amenés à prendre des mesures visant à une meilleure rationalisation des ressources en électricité dans une optique de développement durable. Des solutions de comptage communicantes (Smart Meters), sont mises en place et autorisent désormais une lecture fine des consommations. Les données spatio-temporelles massives collectées peuvent ainsi aider à mieux connaitre les habitudes de consommation et pouvoir les prévoir de façon précise. Le but est d'être en mesure d'assurer un usage « intelligent » des ressources pour une meilleure consommation : en réduisant par exemple les pointes de consommations ou en ayant recours à des sources d'énergies renouvelables. Les travaux de thèse se situent dans ce contexte et ont pour ambition de développer des outils de fouille de données en vue de mieux comprendre les habitudes de consommation électrique et de prévoir la production d'énergie solaire, permettant ensuite une gestion intelligente de l'énergie.

Le premier volet de la thèse s'intéresse à la classification des comportements types de consommation électrique à l'échelle d'un bâtiment puis d'un territoire. Dans le premier cas, une identification des profils types de consommation électrique journalière a été menée en se basant sur l'algorithme des K-moyennes fonctionnel et sur un modèle de mélange gaussien. A l'échelle d'un territoire et en se plaçant dans un contexte non supervisé, le but est d'identifier des profils de consommation électrique types des usagers résidentiels et de relier ces profils à des variables contextuelles et des métadonnées collectées sur les usagers. Une extension du modèle de mélange gaussien classique a été proposée. Celle-ci permet la prise en compte de variables exogènes telles que le type de jour (samedi, dimanche et jour travaillé,...) dans la classification, conduisant ainsi à un modèle parcimonieux. Le modèle proposé a été comparé à des modèles classiques et appliqué sur une base de données irlandaise incluant à la fois des données de consommations électriques et des enquêtes menées auprès des usagers. Une analyse des résultats sur une période mensuelle a permis d'extraire un ensemble réduit de groupes d'usagers homogènes au sens de leurs habitudes de consommation électrique. Nous nous sommes également attachés à quantifier la régularité des usagers en termes de consommation ainsi que l'évolution temporelle de leurs habitudes de consommation au cours de l'année. Ces deux aspects sont en effet nécessaires à l'évaluation du potentiel de changement de comportement de consommation que requiert une politique d'effacement (décalage des pics de consommations par exemple) mise en place par les fournisseurs d'électricité.

Le deuxième volet de la thèse porte sur la prévision de l'irradiance solaire sur deux horizons temporels : à court et moyen termes. Pour ce faire, plusieurs méthodes ont été utilisées parmi lesquelles des méthodes statistiques classiques et des méthodes d'apprentissage automatique. En vue de tirer profit des différents modèles, une approche hybride combinant les différents modèles a été proposée. Une évaluation exhaustive des différents approches a été menée sur une large base de données incluant des paramètres météorologiques mesurés et des prévisions issues des modèles NWP (*Numerical Weather Predictions*). La grande diversité des jeux de données relatifs à quatre localisations aux climats bien distincts (Carpentras, Brasilia, Pampelune et Ile de la Réunion) a permis de démontrer la pertinence du modèle hybride proposé et ce, pour l'ensemble des localisations.

Abstract

Nowadays, countries are called upon to take measures aimed at a better rationalization of electricity resources with a view to sustainable development. Smart Metering solutions have been implemented and now allow a fine reading of consumption. The massive spatio-temporal data collected can thus help to better understand consumption behaviors, be able to forecast them and manage them precisely. The aim is to be able to ensure "intelligent" use of resources to consume less and consume better, for example by reducing consumption peaks or by using renewable energy sources. The thesis work takes place in this context and aims to develop data mining tools in order to better understand electricity consumption behaviors and to predict solar energy production, then enabling intelligent energy management.

The first part of the thesis focuses on the classification of typical electrical consumption behaviors at the scale of a building and then a territory. In the first case, an identification of typical daily power consumption profiles was conducted based on the functional K-means algorithm and a Gaussian mixture model. On a territorial scale and in an unsupervised context, the aim is to identify typical electricity consumption profiles of residential users and to link these profiles to contextual variables and metadata collected on users. An extension of the classical Gaussian mixture model has been proposed. This allows exogenous variables such as the type of day (Saturday, Sunday and working day,...) to be taken into account in the classification, thus leading to a parsimonious model. The proposed model was compared with classical models and applied to an Irish database including both electricity consumption data and user surveys. An analysis of the results over a monthly period made it possible to extract a reduced set of homogeneous user groups in terms of their electricity consumption behaviors. We have also endeavoured to quantify the regularity of users in terms of consumption as well as the temporal evolution of their consumption behaviors during the year. These two aspects are indeed necessary to evaluate the potential for changing consumption behavior that requires a demand response policy (shift in peak consumption, for example) set up by electricity suppliers.

The second part of the thesis concerns the forecast of solar irradiance over two time horizons : short and medium term. To do this, several approaches have been developed, including autoregressive statistical approaches for modelling time series and machine learning approaches based on neural networks, random forests and support vector machines. In order to take advantage of the different models, a hybrid model combining the different models was proposed. An exhaustive evaluation of the different approaches was conducted on a large database including four locations (Carpentras, Brasilia, Pamplona and Reunion Island), each characterized by a specific climate as well as weather parameters : measured and predicted using NWP models (*Numerical Weather Predictions*). The results obtained showed that the hybrid model improves the results of photovoltaic production forecasts for all locations.

Table des matières

Chapitre 1

Introduction

1.1 Introduction générale

L'avènement des technologies de l'information et de la communication (*Information and Communication Technologies*, TIC) ainsi que la percée d'Internet a changé la façon de stocker, transmettre et traiter l'information. Avec la possibilité de collecter et d'analyser de grandes masses de données, l'enjeu est de valoriser ces données en vue d'adapter automatiquement des produits ou des services à des millions de clients/citoyens en fonction de leurs besoins, leurs situations actuelles et des besoins emergents. Amazone et Netflix sont les principaux moteurs de cette tendance. Ils ont développé des moteurs de recommandation qui suggèrent automatiquement des produits aux clients en fonction de leurs achats et de leurs navigations [readwrite, 2007]. D'autres grandes entreprises dans le secteur des services, comme les banques et les assurances utilisent l'analyse de données pour améliorer la gestion de la relation client [Davenport and Dyché, 2013]. Le concept de personnalisation devient encore plus important avec l'Internet des objets (*Internet of Things*, IoT). Dans la mesure où plusieurs objets sont connectés à Internet [Mattern and Floerkemeier, 2010], et capables de communiquer leurs informations, une multitude de services dans divers domaines deviennent possibles en se basant sur l'analyse des données issues des capteurs en temps réel.

Dans le domaine de l'énergie, l'avènement des TIC, de l'IoT et de l'analyse des données conduit également à la personnalisation des efforts d'économie d'énergie susmentionnés. Ces dernières années, les gouvernements du monde entier œuvrent vers la transition énergétique. Elle est défine comme un changement structurel dans les modes de consommation et de production d'énergie. Le but consiste à réduire les effets négatifs de ce secteur sur l'environnement. De ce fait, les gouvernements ont fixé des objectifs quantitatifs sur l'efficacité énergétique afin de diminuer la quantité du combustibles fossiles brûlés [Salvatore, 2013]. L'Union Européenne (UE), par exemple, a récemment renforcé l'objectif de réduire la consommation d'énergie de 20% jusqu'à 2020 [EEA, 2007, EEA, 2014] et de 27% jusqu'en 2030 [EEA, 2014] par rapport à la situation en 2007. La France a aussi décidé dans sa stratégie de réduire de 30% la consommation d'énergies fossiles en 2030.

Dans les états de l'Union Européenne, le secteur résidentiel représente près d'un tiers de la consommation totale d'énergie [eurostat, 2017]. La réduction de la consommation énergétique résidentielle est donc cruciale pour atteindre les objectifs économiques susmentionnées. Dans le passé, de nombreux efforts ont été fournis pour améliorer l'efficacité énergétique dans les ménages : le chauffage est devenu plus efficace grâce à une meilleure

isolation. Le réglage des appareils électroménagers en mode veille limite maintenant la consommation énergétique, et les appareils électroménagers deviennent de plus en plus éconergétiques.

Malgré ces mesures visant à rendre les appareils et les maisons plus éconergétiques, la réduction de la consommation énergétique résidentielle est toujours inférieure aux attentes. Les raisons qui empêchent les consommateurs d'investir dans ces équipements sont le manque d'information et d'incitation, les coûts d'investissement élevés et le manque d'expertise technique. Pour surmonter ces obstacles, de nombreux pays encouragent de plus en plus l'achat de nouveaux produits économiques en terme d'énergie obligeant ainsi les fournisseurs d'énergie à mettre en œuvre des programmes d'efficacité énergétique pour les ménages. De cette façon, les fournisseurs d'énergie encouragent les ménages à modifier leurs comportements énergétiques en offrant à leurs clients des incitations tels que des rabais pour l'achat des équipements et des technologies éconergétiques.

En plus de promouvoir les équipements économes en énergie, inciter à un changement de comportement est également un moyen prometteur pour réduire la consommation énergétique, et plus particulièrement l'électricité dans le secteur résidentiel. La majeure partie de l'électricité est consommée par le chauffage qui représente environ 55% de la consommation électrique dans un ménage français [monenergie, 2017]. La Figure 1.1 représente la part des différents équipements dans la consommation électrique des foyers français.

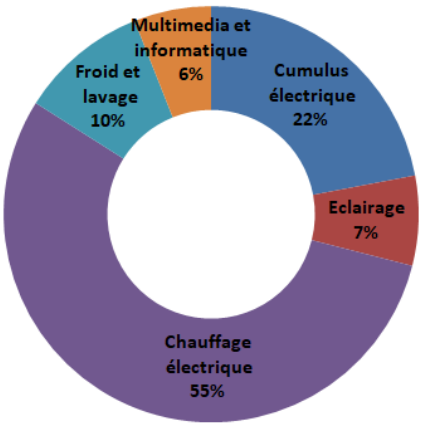


FIGURE 1.1 – La part des différents équipements dans la consommation électrique des ménages français [monenergie, 2017]

Les TIC peuvent jouer un rôle important pour motiver et aider les ménages à économiser leurs énergies [Mattern and Floerkemeier, 2010]. Les compteurs intelligents par exemple, peuvent mesurer la consommation électrique d'un ménage et fournir un retour d'information pour ses occupants, et les aider ainsi à améliorer leurs comportements de consommation. Plusieurs recherches ont examiné l'effet de la rétroaction sur la consommation électrique d'un ménage, celle-ci s'effectuant en temps réel et pouvant être sous la forme d'un affichage à domicile (voir Figure 1.2(a)) ou sur une application smart phone (voir Figure 1.2(b)). Les premières études ont montré que la rétroaction permet de réaliser des économies de l'ordre de 5% à 15% [Darby et al., 2006, Ehrhardt-Martinez et al., 2010].

(a) (b)

FIGURE 1.2 – (a) Smart Energy Kit pour la rétroaction de la consommation électrique [gadgetreview, 2015] ; (b) Retroaction fourni par l'application eMeter [Weiss et al., 2013]

En plus de stimuler le comportement éconergétique, les compteurs intelligents jouent un rôle important pour assurer l'équilibre entre l'offre et la demande, notamment en facilitant l'intégration des énergies renouvelables dans les réseaux électriques [Appelrath et al., 2012]. Ils permettent également de faciliter la procédure de facturation pour les fournisseurs d'énergie. Pour toutes ces raisons, le nombre de compteurs intelligents installés dans les ménages ne cesse d'augmenter. Aux États-Unis, par exemple, plus de 50 millions de compteurs intelligents ont déjà été installés, ce qui représente une couverture de 43% à l'échelle nationale [Cooper, 2014]. Dans l'UE, 45 millions de compteurs intelligents ont été déployés. L'UE vise une couverture de 80% jusqu'à 2020, à condition que l'analyse coûts-avantages réalisée par chaque état membre prouve que l'installation de compteurs intelligents est économiquement raisonnable. Après avoir réalisé ces analyses, 16 pays prévoient d'effectuer un déploiement à grande échelle d'ici 2020 [Covrig et al., 2014, Union, 2009].

Les économies réalisées grâce aux compteurs intelligents jouent un rôle important pour atteindre les objectifs d'économies d'énergie. Pour augmenter ces économies de 3% à 5%, le retour d'information sur l'énergie doit aller au-delà de la simple visualisation. La rétroaction sur la consommation d'énergie devrait contenir des détails utiles en temps réel, avec une manière facile à comprendre et à mettre en œuvre [Ehrhardt-Martinez et al., 2010]. De nombreux chercheurs considèrent que la rétroaction spécifique à un appareil est un élément clé [Armel et al., 2013, Ehrhardt-Martinez et al., 2010, Fischer, 2008]. La décomposition de la consommation globale en fonction de la contribution des appareils individuels, peut aider les ménages à mieux comprendre leurs consommations électriques et à adapter leurs comportements en conséquence. Contrairement aux recommandations génériques en matière d'économie d'énergie, les études indiquent que la rétroaction spécifique à l'appareil permet de réaliser des économies de 9% à 18% [Ehrhardt-Martinez et al., 2010].

L'effet de la rétroaction est également plus important si les conseils et les indices de motivation sont adaptés au destinataire, par exemple en comparant la consommation électrique d'un ménage à celle d'un autre ayant les mêmes caractéristiques [Allcott, 2011, Ayres et al., 2013, Goldstein et al., 2008]. Les caractéristiques d'un ménage jouent un rôle important pour concevoir l'efficacité énergétique. Par exemple, les ménages ayant un revenu élevé sont plus susceptibles d'investir dans les services du changement d'infrastructures, tandis que les personnes âgées de 65 ans et plus ont tendance à être plus critiques en vers ces changements. Les ménages qui sont ouverts aux investissements d'infrastructure pourraient être de bonnes cibles pour une campagne de marketing de pompes à chaleur par exemple [Fei et al., 2013]. De même, les ménages contenant deux personnes qui travaillent pendant la journée ont généralement des horaires régulières. Ces personnes sont des candidats idéaux pour un thermostat intelligent, qui contrôle le système de chauffage d'un ménage en fonction de son état d'occupation [Kleiminger et al., 2014].

La personnalisation des programmes d'efficacité énergétique pour les ménages est nécessaire pour améliorer les économies d'énergie. La rétroaction sur la consommation d'énergie ou les recommandations sur les économies d'énergie exigent une connaissance détaillée sur les caractéristiques des ménages ainsi que sur leurs appareils ménagers utilisés. D'un côté, les technologies telles que les capteurs qui observent les comportements énergétiques des ménages font de plus en plus partie de notre vie [Mattern and Floerkemeier, 2010]. D'un autre coté leur déploiement reste coûteux et les efforts fournis compensent souvent les économies réalisées. De même, les caractéristiques d'un ménage, comme le nombre de personnes qui y habite ou la taille du logement, pourraient être obtenues par le biais d'enquêtes auprès des clients. L'obtention de ces informations prend beaucoup de temps et coûte cher, et souvent, seule une petite fraction des clients y participe [Stoop, 2005].

La situation mondiale en matière de changement climatique s'aggrave du fait de l'exploitation excessive des énergies fossiles. D'après des statistiques faites par IEA (*International Energy Agency*), Plus de 80% de la production mondiale d'énergie a été basée en 2015 sur les combustibles fossiles [international energy agency, 2017]. La Table 1.1 représente la production mondiale d'énergie en 2015. D'après la Table 1.1, nous constatons que 86.8% de la production mondiale d'énergie a été basée essentiellement sur des ressources non renouvelables. Cette production d'énergie a évolué dans le temps depuis 1973. La Figure 1.3 représente l'évolution de la production d'énergie en millions de tep (Mtep) par source entre 1973 et 2015. Si nous prenons comme exemple le charbon, sa part a augmenté de 3.9% par rapport à 1973. L'énergie mondiale produite en 2015 est de 13647 Mtep. Cette production a permis de couvrir une consommation totale de 9384 Mtep [international energy agency, 2017].

Source	**Mtep**	**%**
Pétrole	4326	31.7
Charbon	3835	28.1
Gaz naturel	2984	21.6
Nucléaire	669	4.9
Hydraulique	341	2.5
Renouvelables+déchets	1528	11.2

TABLE 1.1 – Production mondiale d'énergie en 2015 en millions de tep (Mtep)

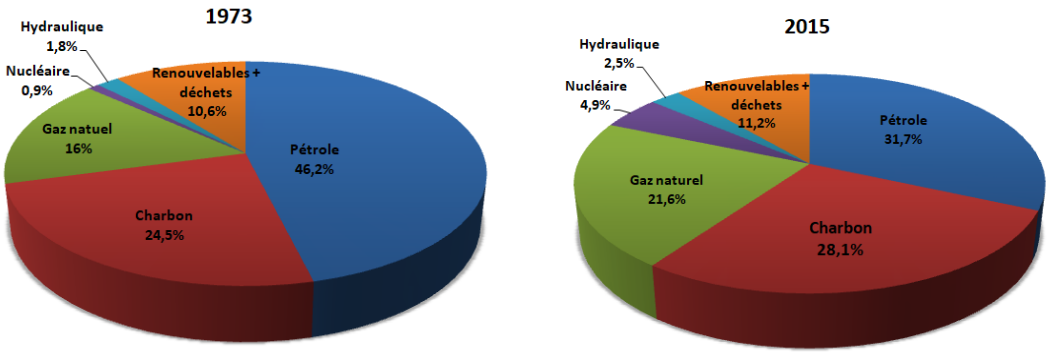


FIGURE 1.3 – Évolution de la production d'énergie (en millions de tep) par source

En plus de l'efficacité énergétique, un autre scénario envisageable pour la transition énergétique, consiste à passer d'un système énergétique reposant sur l'utilisation de ressources non renouvelables vers un mixte énergétique basé principalement sur des énergies vertes. En 2016, 161 GW d'énergies renouvelables ont été installées. Cela a participé à l'augmentation de la capacité totale de production d'électricité d'origine renouvelable de près de 9% par rapport à 2015. Le photovoltaïque a contribué avec 47% de ces capacités additionnelles, suivi de l'éolien (34%) et l'hydroélectricité (15.5%) [REN21, 2017]. En 2017, Les énergies renouvelables ont contribué à la consommation mondiale d'énergie à hauteur de 19.3% contre 78, 4% pour les combustibles fossiles et seulement 2, 3% pour le nucléaire. Les investissements mondiaux dans les technologies renouvelables se sont élevés à plus de 286 milliards de dollars américains en 2015. La Chine et les États-Unis investissent massivement dans l'énergie éolienne, hydroélectrique, solaire et les biocarburants. Environ 7.7% millions d'emplois associés aux industries des énergies renouvelables sont principalement orientés vers le photovoltaïque étant considéré comme la source la plus importante dans ce secteur.

En plus de la transition énergétique, les énergies renouvelables alimentent 90 millions de personnes dans le monde en électricité et s'imposent notamment dans les endroits ruraux où il est difficile de déployer les réseaux électriques classiques. Au Mali par exemple, 10000 villages avaient besoin d'être électrifiés, mais grâce aux énergies renouvelables, ils sont passés de 1% de villages électrifiés en 2004 à 17% aujourd'hui [LATRIBUNE, 2017].

En France, les énergies renouvelables sont en constante progression. Ils fournissent plus de 18.9% du courant électrique français [RTE, 2017]. La Figure 1.4 représente la couverture de la consommation électrique en France par la production renouvelable. Nous remarquons que les régions d'Occitanie et d'Auvergne-Rhône-Alpes utilisent plus de 30% des énergies renouvelables dans leurs consommations, par contre, en Ile de France seulement 1.5% sont exploitées.

Les objectifs fixés pour 2018 visent à atteindre 51.7% GW de la puissance installée. En 2017, cette puissance est déjà à 46.8 GW en cumulant les parcs hydrauliques (25.5 GW), éoliens (12.8 GW), solaires (7.2 GW) et bioénergies (1.9 GW) (voir Figure 1.5).

L'électricité d'origine solaire est souvent utilisée par des particuliers ou pour des habitations éloignées du réseau électrique. Un panneau photovoltaïque (PV) de 1 m^2 produit

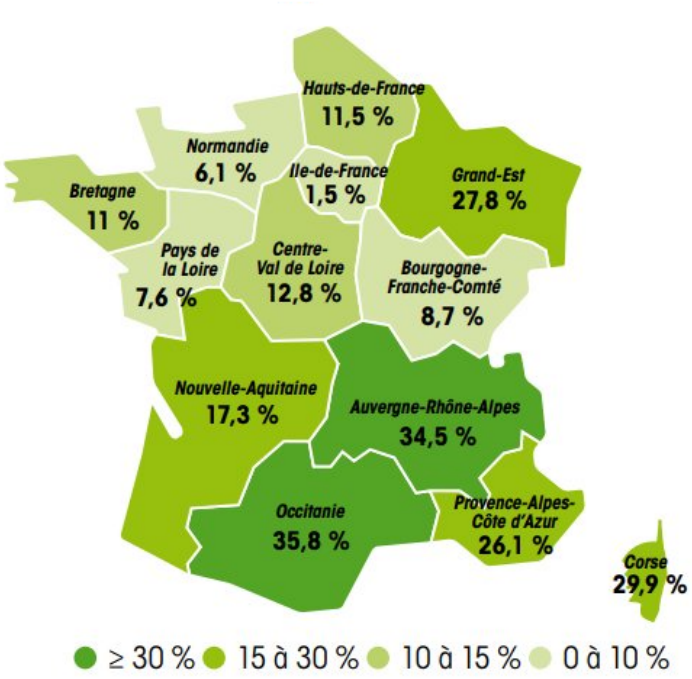


FIGURE 1.4 – Couverture de la consommation électrique en France par la production des énergies renouvelables en 2017. [RTE, 2017]

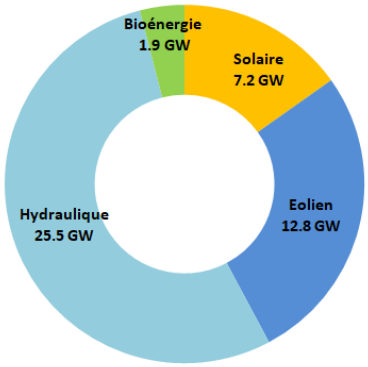
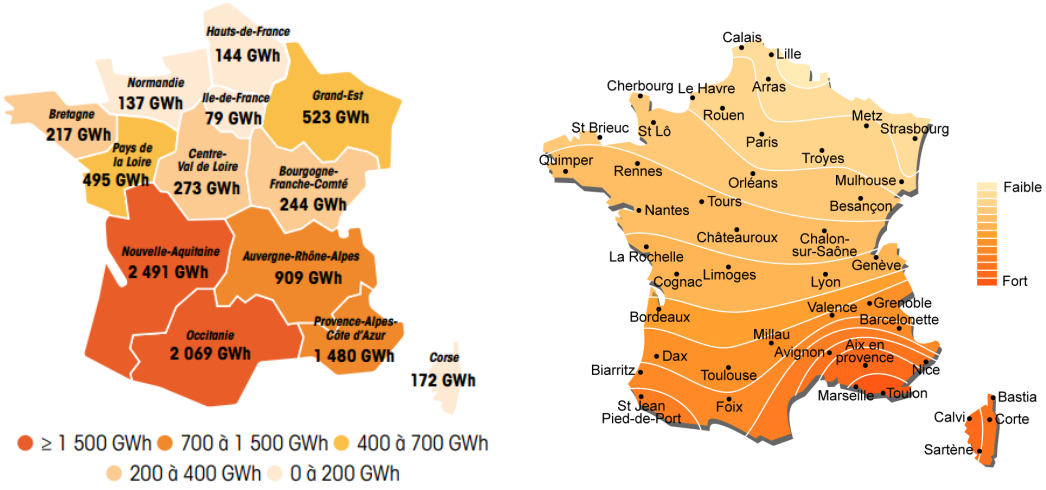


FIGURE 1.5 – Parc renouvelable français en 2017 [RTE, 2017]

en moyenne entre 100 et 200 Watt crête (Wc) de puissance électrique par an, et cela dépend de la disposition des panneaux ainsi que de l'ensoleillement du site. Un panneau installé dans le sud de la France produira en moyenne 40% à 50% d'électricité en plus qu'une installation dans le nord (voir Figure 1.6).

L'intégration d'un nombre important de PVs dans un réseau électrique pose un défi technique en raison de la nature variable de la ressource solaire. A cet effet, une precision de la prévision de l'ensoleillement ou ce qu'on appelle l'irradiance solaire va permettre aux opérateurs des réseaux de mieux planifier l'utilisation des différentes sources d'énergies, tout en favorisant les énergies renouvelables.



(a) (b)

FIGURE 1.6 – (a) Production solaire par région durant une année [RTE, 2017] ; (b) carte d'ensoleillement de la France [EDF, 2017]

1.2 Objectifs et organisation de la thèse

Ces dernières années de nombreux travaux de recherche ont été menés autour des techniques d'analyse de données pour modéliser, analyser et mieux comprendre la consommation et la production électriques. Certains de ces travaux se sont focalisés sur la prévision de la consommation et de la production électrique sur différents horizons temporels. Ces travaux visent à mieux planifier l'utilisation des différentes sources d'énergies. D'autres applications visent à classifier les consommateurs en différents groupes en se basant sur leurs comportements de consommation. Ces applications ont pour but d'améliorer les programmes d'efficacité énergétique et d'optimiser les compagnes de marketing. Pour une analyse plus fine des comportements de consommation dans les ménages, la désagrégation de la courbe de charge totale permet de mieux identifier les appareils ménagers utilisés. La détection d'anomalies à travers des données de consommation électrique est une application très intéressante, car elle permet de révéler les défaillances et les pertes dans le système.

Cette thèse s'inscrit dans le cadre du projet Smart City Energy analytics (SCE) porté par l'institut de recherche technologique SystemX. Le projet SCE regroupe plusieurs partenaires industriels (GE Grid solution, Alstom Transport, Reuniwatt, Engie, Cosmo Tech, G2 Mobility, NovEner, Artelys, Ecogélec, Sherpa Engineering et Solunergie) ainsi que académiques (IFSTTAR, CEA et CentraleSupélec), dans l'objectif de développer des outils d'aide à la décision afin d'optimiser la consommation énergétique.

Cette thèse propose des solutions aux problématiques des deux scénarios de la transition énergétique à savoir : l'efficacité énergétique et l'exploitation des énergies renouvelables plus précisément l'énergie solaire. Les solutions proposées se basent sur l'analyse des données de consommation électrique et de production photovoltaïque.

Le premier volet s'intéresse à la compréhension et l'analyse des comportements de consommation électrique à deux échelles : bâtiment et territoire. Ces travaux s'inscrivent

dans le cadre de l'apprentissage non supervisé visant à extraire un ensemble réduit de profils de consommation électrique à partir des données mesurées par des compteurs intelligents.

A l'échelle du bâtiment Galilée de General Electric, deux méthodes de classification automatique ont été appliquées à savoir : l'algorithme des K-moyennes fonctionnel et le modèle de mélange gaussien. L'objectif de ce travail consiste à l'identification des profils types de consommation électrique journaliers, ainsi qu'à l'analyse de leurs évolutions durant une année. Cette étude représente une étape préliminaire pour la suite des travaux ainsi que pour le développement des nouveaux modèles prédictifs.

A l'échelle d'un territoire, une approche de classification automatique a été développée pour identifier les profils types de consommation électrique. Cette approche est une extension du modèle de mélange gaussien classique dont les paramètres dépendent d'une variable exogène représentant le type de jour (samedi, dimanche et jour ouvré). L'objectif de ce modèle consiste à regrouper les consommateurs ayant des comportements de consommation électrique similaires durant les trois types de jours et non pas séparément. Chaque classe résultante sera caractérisée par trois profils types (samedi, dimanche et jour ouvré). Il est important de mentionner que notre approche diffère de celle qui divise les données selon les trois types de jours, et qui applique par la suite un modèle de mélange gaussien de base sur chaque catégorie de données. Dans le cas de cette méthode alternative, pour chaque type de jour, des profils types de consommation électrique seront déterminés. A cet effet, chaque compteur intelligent va appartenir à trois classes : une pour les samedis, une autre pour les dimanches, et une troisième pour les jours travaillés. Les données utilisées dans cette étude ont été mises à notre disposition par la commission de régulation d'énergie (CER) [1] de l'Irlande. Ces données ont été collectées dans le cadre d'un projet d'installation des compteurs intelligents, lequel a pour but d'évaluer les performances de ces compteurs et leur impact économique. En plus des données sur la consommation électrique de 4232 ménages, des informations sur leurs caractéristiques socio-économiques sont aussi mises à notre disposition. Dans un premier temps, notre approche a été appliquée sur des données de consommation électrique durant le mois de novembre. L'objectif consiste à identifier les comportements de consommation électrique des usagers durant ce mois, et trouver le lien entre ces comportements et les caractéristiques socio-économiques extraites de l'enquête. Dans un deuxième temps, une investigation sur l'évolution des comportements des consommateurs au cours des mois de l'année a été menée à travers notre modèle.

Le deuxième volet est consacré à la prévision de l'irradiance solaire sur deux horizons temporels : court et moyen termes. Dans un premier temps, six approches statistiques ont été appliquées à savoir : Naïve, Calendaire, ARMAX, SVM, RF et NN. Pour construire et valider ces modèles, deux types de données ont été utilisés : des paramètres météorologiques mesurés et des prévisions issues des modèles NWP (*Numerical Weather Predictions*). Ces données sont relatives à quatre localisations (Carpentras, Brasilia, Pampelune et Ile de la Réunion) caractérisées chacune par un type de climat bien spécifique. Les performances des modèles ont été mesurées sur les deux horizons temporels et pour chaque climat. Dans un deuxième temps, pour tirer bénéfice des performances de chaque méthode, un modèle hybride a été proposé. Ce modèle fusionne les résultats de prévision de : ARMAX, SVM, RF et NN, en les pondérant à travers un apprentissage supervisé. Cette approche permet d'améliorer la précision de prévision de l'irradiance solaire pour une meilleure planification de l'utilisation de l'énergie solaire.

Ce manuscrit débute par le chapitre 2 qui présente les volets applicatifs à l'origine de

1. `http://www.ucd.ie/issda/data/commissionforenergyregulationcer/`

ces travaux. Dans un premier temps, nous décrivons les moyens d'acquisition des données, à savoir les compteurs intelligents et les capteurs météo. Dans un deuxième temps, une description, un pré-traitement ainsi qu'une analyse exploratoire sont effectués et détaillés sur deux ensembles de jeux de données. D'une part, les consommations électriques collectées à l'échelle d'un bâtiment (Galilée-GE à la défense), puis d'un territoire (CER-Irlande). D'autre part, il s'agit des irradiances solaires collectées sur 4 différentes localisations géographiques dans le monde (Reuniwatt).

Le chapitre 3 est consacré à la classification des comportements de consommation électrique à deux échelles : bâtiment et territoire. La première partie de ce chapitre est consacrée à l'état de l'art abordant les différents travaux de recherche qui ont appliqué des approches de classification automatique sur des données de consommation électrique des bâtiments. Par la suite, nous évoquons les différents algorithmes de classification classiques utilisés ainsi que le modèle proposé pour mener à bien cette tâche de classification. Nous présentons la mise en œuvre des différentes méthodes pour identifier les profils types de consommation électrique journaliers à l'échelle d'un bâtiment tertiaire, ainsi que les comportements de consommation électrique d'usagers à l'échelle d'un territoire.

La prévision de l'irradiance solaire à court et moyen termes fait l'objet du chapitre 4. Un état de l'art décrivant les différentes méthodes de prévision de l'irradiance solaire est détaillé, en se focalisant sur les approches statistiques. Par la suite, nous évoquons les méthodes de prévision utilisées, et présentons leur mise en œuvre pour prévoir l'irradiance solaire à court terme (horaire) et à moyen terme (journalière), et cela en considérant quatre climats différents. Nous terminons par une description détaillée du modèle hybride proposé ainsi que sa comparaison avec les méthodes utilisées.

Une conclusion de ce manuscrit rappellera les différentes étapes des travaux menés dans la thèse, ainsi que les perspectives de recherche envisagées.

Les travaux de cette thèse ont été valorisés dans les publications listées ci-dessous :

Revues avec comité de lecture

— MELZI, F., SAME, A., ZAYANI, H., OUKHELLOU, L. " A Dedicated Mixture Model for Clustering Smart Meter Data : Identification and Analysis of Electricity Consumption Behaviors", energies. September 2017.

Conférences internationales avec actes

— MELZI, F., TOUATI, T., SAME, A., OUKHELLOU, L. "Hourly Solar Irradiance Forecasting bases on Machine Learning Models", ICMLA 2016 - IEEE 15th International Conference on Machine Learning and Applications, Los Angeles, December 2016.

— MELZI, F., ZAYANI, H., BEN HAMIDA, A., SAME, A., OUKHELLOU, L. " Identifying Daily Electric Consumption Patterns from Smart Meter Data by Means of Clustering Algorithms", ICMLA 2015 - IEEE 14th International Conference on Machine Learning and Applications, Miami, December 2015.

— MELZI, F., ZAYANI, H., BEN HAMIDA, A., STEPHAN, F., SAME, A., OUKHELLOU, L. “ Towards Smart City Energy Analystics : Identification of Consumption Patterns Based on the Clustering of Daily Electric Consumption Curves", CSDM 2015 - Springer 6th International Conference on Complex Systems Design & Management, Paris, November 2015.

Chapitre 2

Données de consommation électrique et d'irradiance solaire

2.1 Introduction

Le but de ce chapitre est de présenter les volets applicatifs à l'origine de ces travaux de thèse. Le premier volet concerne la compréhension et l'analyse des comportements de consommation électrique, tandis que le deuxième s'intéresse à la prévision de la production photovoltaïque. Pour concrétiser ces applications, nous avons besoin non seulement des données concernant la consommation et la production électriques, mais aussi d'informations contextuelles qui ont un impact direct sur ces dernières. Dans ce chapitre, nous allons décrire dans un premier temps les moyens d'acquisition des données, à savoir les compteurs intelligents et les capteurs météo. Dans un deuxième temps, une description, un pré-traitement ainsi qu'une analyse exploratoire sont effectués et détaillés sur deux ensembles de jeux de données. D'une part, nous disposons de deux jeux de données de consommations électriques. Un premier, à l'échelle d'un bâtiment (Galilée de GE), et un deuxième à l'échelle d'un territoire (CER en Irlande). D'autre part, nous nous sommes appuyés sur des données d'irradiance solaire collectées sur 4 différentes localisations géographiques dans le monde (Reuniwatt).

2.2 Sources de données

2.2.1 Compteurs intelligents

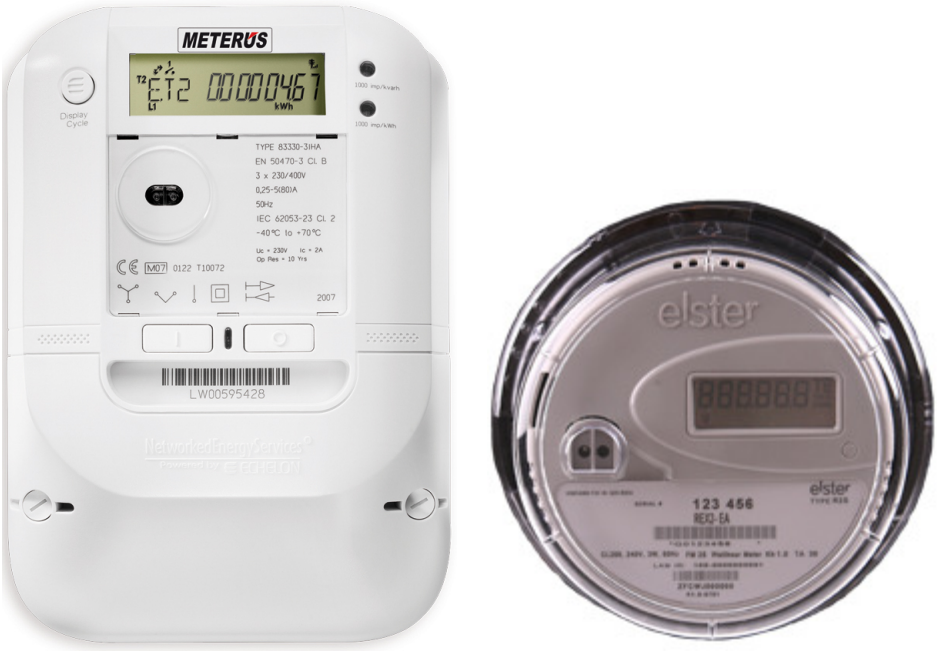
Pour connecter les différents types de bâtiment d'une ville à un réseau électrique intelligent, il est nécessaire de les équiper avec une technologie numérique via des compteurs intelligents. Contrairement aux compteurs analogiques, ces compteurs mesurent la consommation électrique en permanence avec une granularité très fine. Ces mesures sont mises à disposition à travers un réseau de communication. Les Figures 2.1(a) et 2.1(b) représentent deux compteurs intelligents fabriqués par Echelon et Elster destinés aux marchés européen et américain respectivement. L'accès à la consommation électrique en temps réel offre de nombreuses opportunités pour les ménages, les fournisseurs d'énergie et les opérateurs de réseau de distribution [Vasconcelos, 2008].

Les ménages, par exemple, bénéficiant des compteurs intelligents, peuvent recevoir des retours sur leur consommation électrique en temps réel et avec une granularité très fine pouvant aller jusqu'à la seconde. La visualisation de la consommation électrique d'un ménage aide à la sensibilisation de ses occupants à économiser l'électricité. Cela permet

de réduire à la fois la facture d'électricité et l'émission du CO2. En plus des retours sur la consommation, les clients peuvent bénéficier d'une flexibilité tarifaire et d'un changement de fournisseur plus facile. Cela conduit à une libéralisation du marché et à une concurrence entre les fournisseurs en termes de prix et qualité de service.

Les fournisseurs d'énergie exploitent le potentiel des compteurs intelligents de deux façons [Vasconcelos, 2008]. Tout d'abord, les compteurs intelligents rendent la procédure de facturation et le schéma de tarification plus simple. Par exemple, une tarification adaptative peut être mise en œuvre où différents prix de l'électricité sont appliqués à des moments différents de la journée. Deuxièmement, avoir un accès aux données de consommation électrique des consommateurs permet d'assurer l'équilibre entre l'offre et la demande. En se basant sur les comportements de consommation des clients, les fournisseurs d'énergie peuvent fournir des propositions personnalisées aux clients pour changer leurs façons de consommer, notamment en évitant la consommation pendant les heures de pointe et en la décalant durant les périodes creuses.

Les opérateurs de réseaux de distribution peuvent utiliser un déploiement étendu de compteurs intelligents pour optimiser la qualité de service du réseau. Ils peuvent identifier rapidement les emplacements des pannes, détecter les pertes du réseau, améliorer la stabilité de la tension et informer les clients en cas de pannes [Vasconcelos, 2008]. En outre, la disponibilité d'informations sur le réseau de basse tension permet d'améliorer la planification des investissements tout en respectant les nouvelles infrastructures et en renforçant celles qui existent.



(a) (b)

FIGURE 2.1 – (a) Compteur intelligent fabriqué par Echelon [wikimedia, 2008] ; (b) Compteur intelligent fabriqué par Elster [elster solutions, 2017]

2.2.2 Capteurs météo

Les stations météo fournissent des paramètres météorologiques liés à la variation du climat. Ces paramètres englobent l'irradiance solaire, la température, la pression, l'humi-

dité, la vitesse du vent, ..., etc. La météo est utilisée dans divers secteurs, nous pouvons citer l'énergie, l'agriculture et le transport (routier, maritime et aérien). Il existe deux catégories de stations météorologiques : les stations manuelles et les stations automatiques. Concernant les stations manuelles, les mesures sont prélevées par un technicien en météorologie à des horaires réguliers. Pour les stations automatiques (voir la Figure 2.2), les données météorologiques sont rapportées par des capteurs sans intervention humaine. Pour certaines mesures météorologiques comme la couverture nuageuse par exemple, les observations sont plus fiables avec les stations manuelles qu'avec les stations automatiques. L'instrumentation en capteurs d'une station automatique dépend du secteur et de l'objectif visé. Pour une station destinée à l'énergie renouvelable, les capteurs utilisés sont listés ci-dessous :

— Pyranomètre pour mesurer l'irradiance solaire,
— Thermomètre électronique pour mesurer la température de l'air,
— Hygromètre pour l'humidité,
— Baromètre pour mesurer la pression atmosphérique,
— Anémomètre pour mesurer le vent et sa direction,
— Pulviomètre pour mesurer la précipitation,
— Cléomètre pour mesurer la couche nuageuse.

Dans le cadre de cette thèse, les données issues de ces capteurs nous permettront de développer un modèle de prévision dédié à la production photovoltaïque. Pour atteindre cet objectif, nous allons nous focaliser sur la prévision d'un paramètre météorologique très important qui est l'irradiance solaire.

FIGURE 2.2 – Station météo automatique [cimel, 2011]

2.3 Jeux de données

2.3.1 Données Galilée

Description

Dans cette partie, nous nous intéressons aux données de consommation électrique des bureaux du bâtiment Galilée de General Electric (GE). Ce bâtiment est situé en Ile-de-France, plus exactement à la Défense. La consommation électrique, exprimée en Watt (W),

est mesurée par un seul compteur intelligent durant l'année 2013. Ce compteur prélève la consommation électrique avec une fréquence de 10 minutes, qui est traduite par 144 mesures par jour. La Figure 2.3 représente les courbes de consommation électrique journalière durant l'année 2013. A première vue, on note une grande variabilité des courbes, notamment en distinguant des courbes plates et d'autres avec des consommations électriques considérables . Ces observations sont expliquées dans la partie suivante.

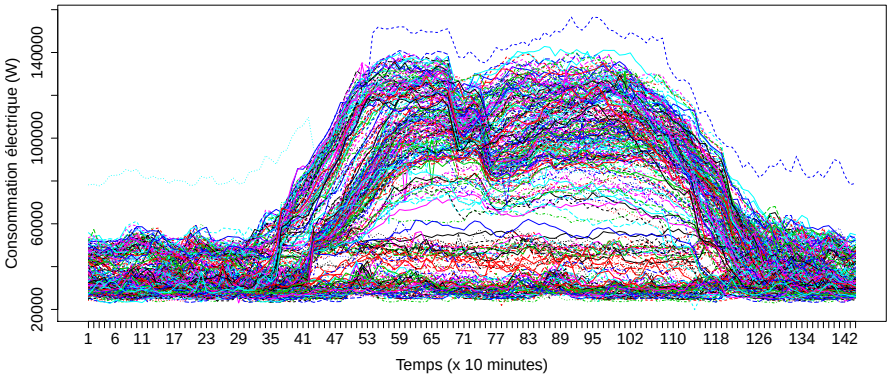


FIGURE 2.3 – Courbes de consommation électrique journalières durant l'année 2013

Pré-traitement et analyse exploratoire

Il est important de noter que 11% des données sont manquantes. La Figure 2.4, représente la consommation électrique du bâtiment durant l'année 2013, où les parties blanches représentent les valeurs manquantes. Pour une meilleure précision d'analyse, nous avons choisi de ne considérer que les jours sans valeurs manquantes (311 jours). Il est remarqué aussi que le compteur du bâtiment ne prend pas en compte la correction du changement du fuseau horaire indiqué par les deux flèches dans la Figure 2.4. Pour résoudre ce problème, un décalage d'une heure a été effectué le 31 mars et le 27 octobre relativement aux dates du passage à l'heure d'été et à l'heure d'hiver. La Figure 2.5 représente les données de consommation électrique du bâtiment après le pré-traitement.

Pour avoir une vue globale sur les données, les Figures 2.6(a) et 2.6(b) représentent respectivement la consommation électrique moyenne de chaque jour de la semaine et les deux profils de consommation électrique moyens durant les jours travaillés et non travaillés (jours de weekend et jours fériés). La consommation électrique durant les jours de semaine est plus élevée que celle des jours de weekend. Concernant les jours travaillés, la consommation électrique augmente le matin entre 9 heures et midi jusqu'à atteindre un pic. Juste après une décroissance pendant le déjeuner est aperçue (entre midi et 14 heures). La consommation électrique croît progressivement jusqu'à 17 heures, puis une diminution est observée. Cela peut être expliqué par la sortie du personnel des bureaux. Pour les jours non travaillés, la consommation électrique est caractérisée par une certaine constance et un faible niveau de consommation. Ces observations reflètent le type de bâtiment qui est tertiaire. Pour une analyse plus fine, le chapitre 3 s'intéressera à l'identification de profils types de consommation journaliers du bâtiment Galilée durant l'année 2013.

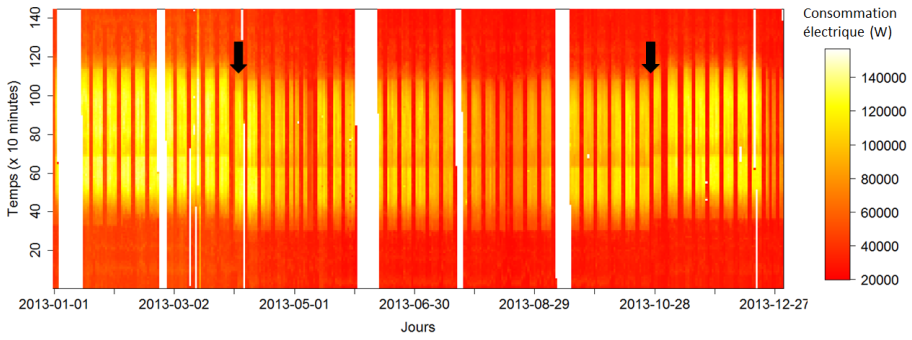


FIGURE 2.4 – Données de consommation électrique du bâtiment Galilée durant l'année 2013 avant le pré-traitement

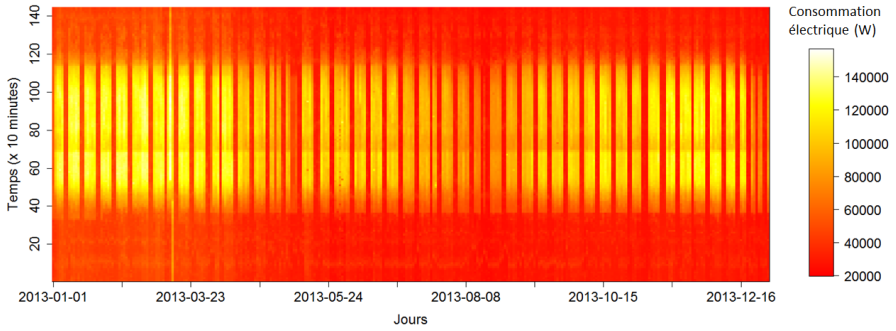


FIGURE 2.5 – Données de consommation électrique du bâtiment Galilée durant l'année 2013 après le pré-traitement

2.3.2 Données CER

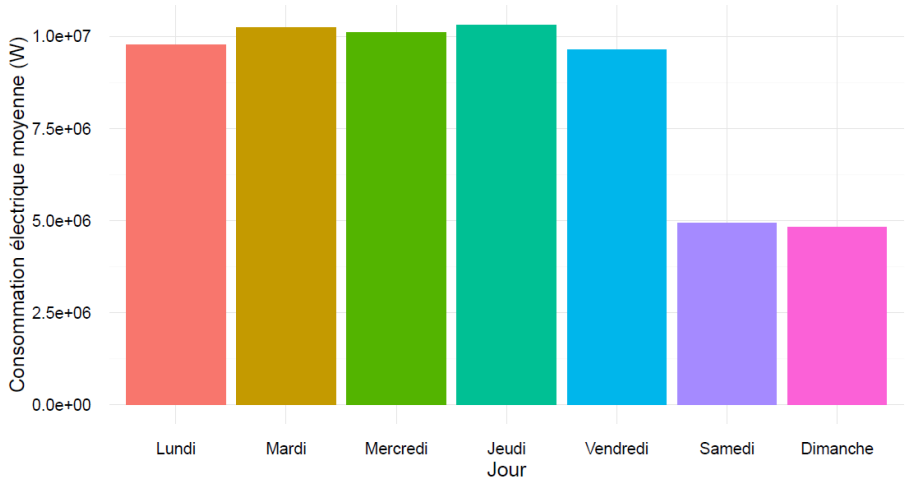
Description

Cette partie s'intéresse aux données qui ont été mises à disposition par la commission de régulation d'énergie (CER) [1] de l'Irlande. Ces données ont été collectées du juillet 2009 jusqu'à décembre 2010 dans le cadre d'un projet d'installation des compteurs intelligents, lequel a pour but d'évaluer les performances de ces compteurs et leur impact économique. Plus de 6000 ménages et entreprises ont participé dans cette enquête. Chaque compteur mesure la consommation électrique en kilo Watt (kW) toutes les 30 minutes. Chaque participant était invité à remplir un questionnaire spécifique à son type de bâtiment (ménage ou entreprise). Dans le cadre de cette thèse, nous nous sommes focalisés sur la catégorie résidentielle (4232 compteurs) durant l'année 2010. Le questionnaire s'y rapportant est basé principalement sur les caractéristiques socio-économiques des ménages. La description des questions, qui nous sera utile, est présentée dans la Table 2.1.

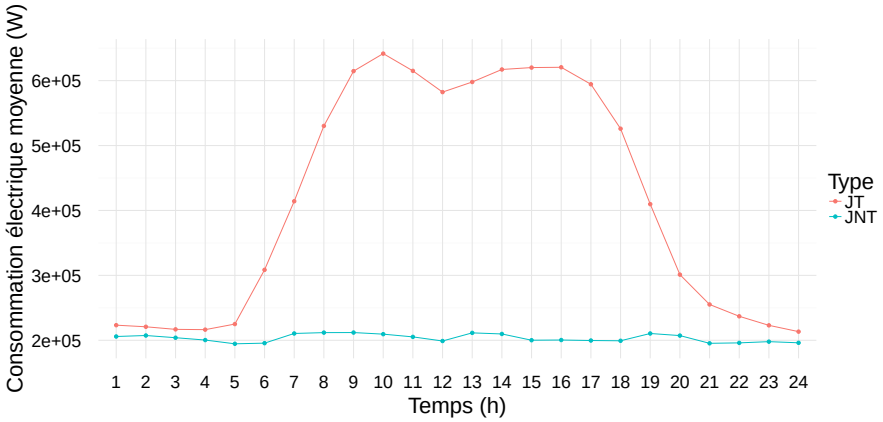
Pré-traitement et analyse exploratoire

Pour une analyse précise, nous avons utilisé uniquement les compteurs qui ne contiennent

1. `http://www.ucd.ie/issda/data/commissionforenergyregulationcer/`



(a)



(b)

FIGURE 2.6 – (a) Consommation électrique moyenne durant les jours de semaine ; (b) Consommation électrique moyenne durant les jours travaillé (JT) et non travaillé (JNT)

pas de valeurs manquantes durant toute l'année 2010. Le nombre de compteurs retenus est de 2995. Pour une vue globale des données pré-traitées, la Figure 2.7 représente les profils de consommation électrique moyens des 2995 compteurs durant les trois types de jours (samedi, dimanche et jour travaillé) du mois de novembre. Pour une analyse plus détaillée, le chapitre 3 s'intéressera à l'identification de profils types de consommation électrique d'usagers en fonction des trois types de jours cités.

2.3.3 Données Reuniwatt

Description

Dans le cadre du projet Smart City Energy (SCE) porté par l'IRT SystemX, la société Reuniwatt [2] a fourni des données d'irradiance solaire mesurées sur 4 localisations dans le monde. Les caractéristiques climatiques ainsi que géographiques de ces régions sont présentées dans la Table 2.2. Cette irradiance qui est exprimée en W/m^2 a été collectée durant 2 ans (du 1 janvier 2012 jusqu'au 31 décembre 2013) avec une fréquence de prélèvement

2. `http://reuniwatt.com/fr/`

Caractéristique	Description	Modalité
Emploi	Activité de la personne qui a la source de revenu principale	Travailleur Chômeur Retraité
Classe-sociale	Classe sociale selon le grade social NRS	A ou B C1 ou C2 D ou E F
Nombre-appareils	Nombre d'appareils électriques	Faible (<6) Moyen(6->8) Élevé (>8)
Nombre-résidents	Nombre de résidents	Seul Faible (2->4) Moyen(5->7) Élevé(>7)
Age-personne	Age de la personne qui a la source de revenu principale	Jeune (<35) Adulte (35->65) Senior (>65)
Usage-Internet	Manière d'utiliser Internet	Régulier Non-régulier
Chauffage	Type de chauffage utilisé	Non-électrique Électrique
nbr-pers-actifs	Nombre de personnes actives	Aucun 1 2 3 >4

TABLE 2.1 – Description des caractéristiques socio-économique des ménages

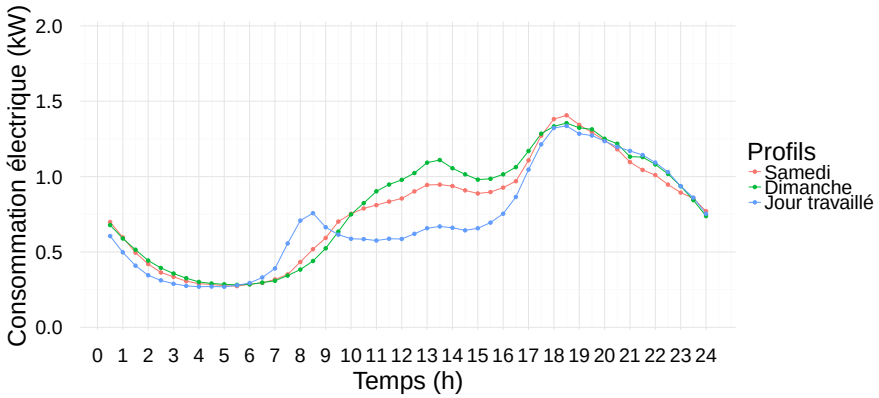


FIGURE 2.7 – Profils de consommation électrique moyens des ménages durant les trois types de jours (samedi, dimanche et jour travaillé) du mois de novembre

d'une heure. D'autres données concernant la météo mesurée et prévue ont été également mises à notre disposition. Ces données englobent la température, la pression, l'humidité et la vitesse du vent. En plus de ces paramètres, la couverture nuageuse est aussi disponible

mais à travers des données prévues uniquement. Les paramètres météorologiques cités sont résumés dans la Table 2.3. En se focalisant principalement sur le jeu de données Carpentras, la Figure 2.8 montre les courbes d'irradiance solaire journalières durant l'année 2013. À première vue, trois parties sont distinguées : la première, représente la nuit avant l'aube, la seconde, est la partie du jour entre l'aube et le couché du soleil, alors que la troisième, correspond à la nuit après le couché du soleil. Comme nous pouvions nous y attendre, les durées de ces parties varient d'une période à une autre dans l'année et elles dépendent aussi de l'endroit dans lequel est situé. Pour une vue globale sur les données, une analyse exploratoire est présentée dans la partie suivante.

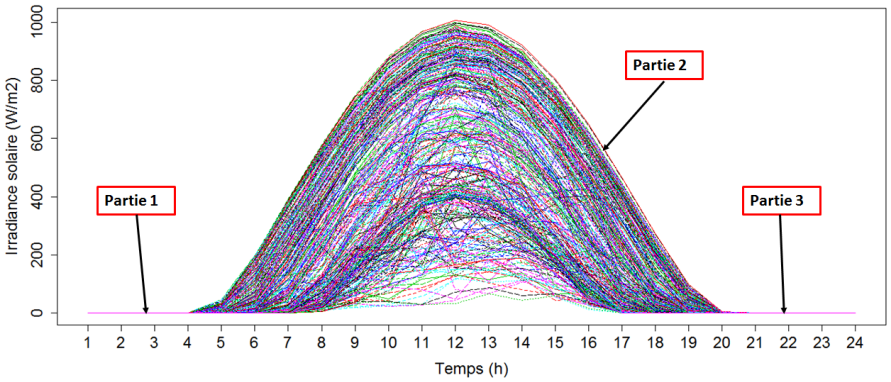


FIGURE 2.8 – Courbes d'irradiance solaire journalières durant l'année 2013

Région	Pays	Climat	latitude	Longitude
Carpentras	France	Méditerranéen	44.0830	5.0590
Pampelune	Espagne	Maritime cote ouest	42.8160	-1.6010
Brasilia	Brésil	Tropical humide et sec	-15.6010	-47.7130
Ile de la réunion	France	Tropical humide	-21.0351	55.7052

TABLE 2.2 – Caractéristiques climatiques et géographiques des 4 localisations

Paramètre météorologique	Unité	Mesuré	Prévu
Irradiance	W/m^2	Oui	Non
Température	K	Oui	Oui
Pression	Pa	Oui	Oui
Humidité	Kg/m^2	Oui	Oui
Vitesse du vent	m/s	Oui	Oui
Couverture nuageuse	-	Non	Oui

TABLE 2.3 – Paramètres météorologiques utilisés dans la prévision de l'irradiance solaire

Pré-traitement et analyse exploratoire

Dans la mesure où l'analyse exploratoire appliquée est similaire pour les 4 jeux de données, nous avons choisi de ne montrer que les résultats obtenus pour la région de Carpentras qui est située en France. La Figure 2.9 représente l'évolution du nombre d'heures

d'ensoleillement au cours des deux ans (2012 et 2013). Par manque d'historique de données, nous avons confirmé sur d'autres données (Alfortville, historique de 12 ans) que la progression du nombre d'heures d'ensoleillement est la même d'une année non bissextile à une autre, et d'une année bissextile à l'autre. La Figure 2.10 montre la répartition du nombre d'heures d'ensoleillement durant les années 2012 et 2013 correspondant aux années bissextiles et non bissextiles respectivement. Il est constaté que pour la région de Carpentras, l'année est divisée en 8 groupes de jours selon leur nombre d'heures d'ensoleillement. Ce nombre varie entre 9 et 16 heures. La Figure 2.11 représente les profils d'irradiation solaire moyens pour chaque groupe de jours au cours de l'année 2013. Comme il est attendu, plus le nombre d'heures d'ensoleillement est élevé, plus le niveau d'irradiance solaire est haut. Ceci est expliqué par la présence des jours ensoleillés durant les périodes chaudes où les jours sont longs, contrairement aux jours nuageux pendant la période froide. Le chapitre 4 s'intéressera à la prévision de l'irradiance solaire à court et à moyen termes.

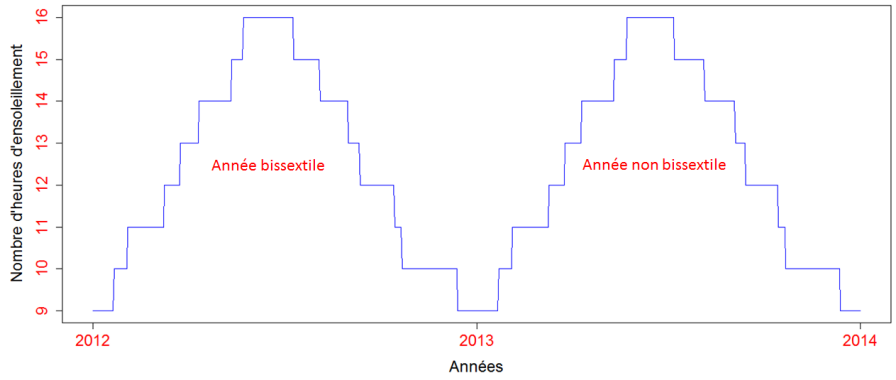
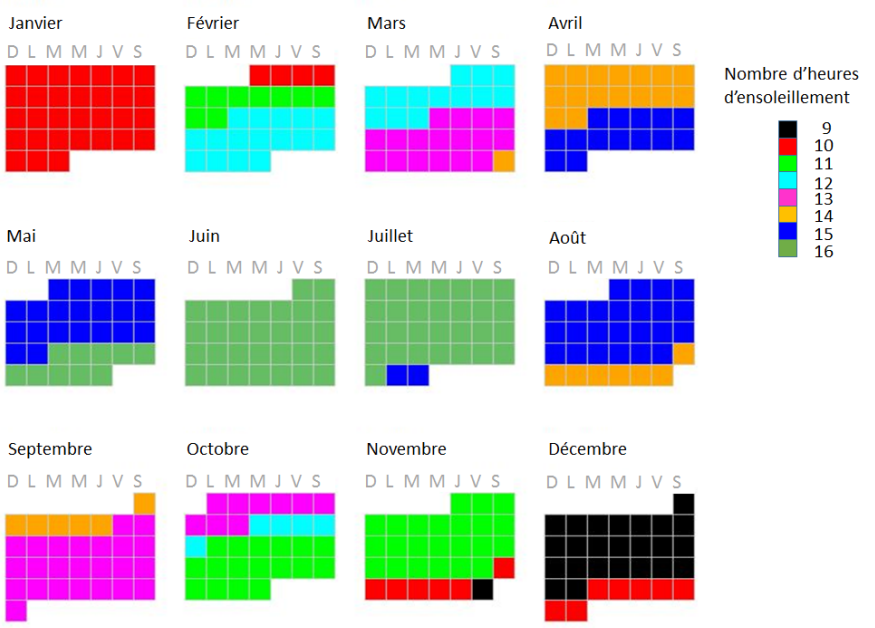
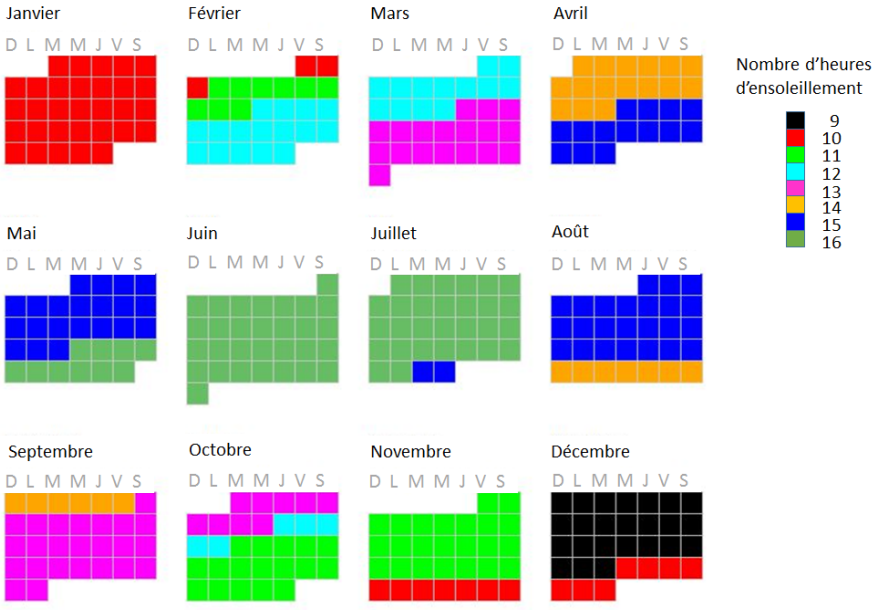


FIGURE 2.9 – Évolution du nombre d'heures d'ensoleillement durant deux ans (2012-2013)



(a)



(b)

FIGURE 2.10 – (a) Distribution du nombre d'heures d'ensoleillement durant une année non bissextile ; (b) une année bissextile

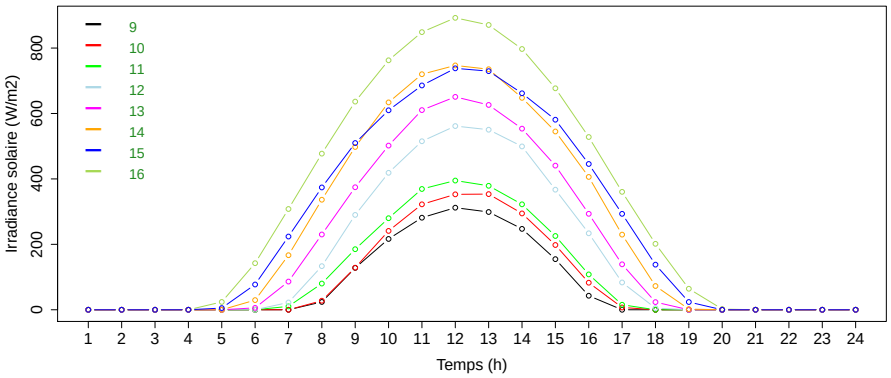


FIGURE 2.11 – profils d'irradiation solaire moyens pour chaque groupe de jours au cours de l'année 2013

2.4 Conclusion

Dans ce chapitre, nous avons décrit les différents outils de recueil de données, à savoir les compteurs intelligents pour la consommation électrique et les capteurs météo pour les paramètres météorologiques. Trois jeux de données recueillis par ces outils sont décrits : Galilée, CER et Reuniwatt. Pour la mise en forme des données, une étape de pré-traitement a été effectuée. Cette partie a permis d'organiser et de filtrer les données. Une analyse exploratoire a également été menée sur les trois jeux de données en vue d'en avoir une vision globale. Dans le chapitre qui suit, nous allons nous intéresser à l'identification des profils types de consommation électrique sur différent échelles (bâtiment et territoire) à travers des approches de classification automatique.

Chapitre 3

Classification des courbes de consommation électrique

3.1 Introduction

DAns ce chapitre nous présentons la classification non supervisé des comportements de consommation électrique à différentes échelles : bâtiment et territoire. Ces travaux de recherche sont utiles à plus d'un titre. En effet, ils participent à une meilleure compréhension des usages, permettant de mieux ajuster les stratégies d'optimisation de consommation d'énergie ou encore ils peuvent servir dans les modèles de simulation de manière à les rendre plus fiables. La première partie de ce chapitre est consacrée à l'état de l'art abordant les différents travaux de recherche qui ont appliqué les méthodes de classification sur des données de consommation électrique des bâtiments. Par la suite, nous évoquons les différents algorithmes de classification classiques utilisés ainsi que le modèle proposé pour mener à bien cette tache de classification. Dans la dernière partie du chapitre, nous présentons la mise en œuvre des différentes méthodes pour identifier les profils types de consommation électrique journaliers à l'échelle d'un bâtiment tertiaire, ainsi que les comportements de consommation électrique d'usagers à l'échelle d'un territoire.

3.2 État de l'art

Plusieurs travaux de recherche en apprentissage automatique ont été développés dans le domaine de l'énergie. Dans cette section nous allons nous focaliser sur certains travaux qui ont utilisé les approches de classification sur les données de consommation électrique des bâtiments (ménage et entreprise).

Dans le cadre d'apprentissage non supervisé, l'algorithme des K-moyennes est l'une des méthodes de classification les plus utilisées dans le domaine de l'énegie [Nizar et al., 2006, Yu et al., 2011, Melzi et al., 2015, Birt et al., 2012, Cao et al., 2013, Kwac et al., 2013].

Les chercheurs dans [Nizar et al., 2006], ont travaillé sur la détermination de la meilleure méthode de classification qui permet d'identifier les différents profils de consommation électrique d'usagers. A cet effet, trois méthodes de classification ont été utilisées, à savoir : l'algorithme des K-moyennes, l'algorithme Espérance Maximisation (EM) et l'algorithme COBWEB. Il a été noté, que COBWEB est la méthode qui obtient les moins bons résultats, alors que les K-moyennes est le meilleur algorithme pour classer cet ensemble de données. Malheureusement, aucune discussion n'a été faite sur les classes et leurs relations

avec les variables contextuelles. Dans [Yu et al., 2011], Yu et al. ont utilisé la classification pour identifier l'impact du comportement des occupants sur la consommation électrique dans les ménages. Pour ce faire, l'algorithme des K-moyennes a été appliqué. Comme résultat, 4 classes ont été obtenues. Par la suite, une étude a été réalisée pour comprendre et interpréter l'effet du comportement des occupants sur la consommation d'énergie. Melzi et al. ont identifié les profils types de consommation électrique journaliers d'un bâtiment tertiaire dans [Melzi et al., 2015]. L'objectif était d'étudier l'évolution de ces profils durant une année. Deux approches de classification ont été utilisées à savoir l'algorithme des K-moyennes et l'algorithme Espérance Maximisation basé sur un modèle de mélange gaussien. Les auteurs ont constaté que les profils types journaliers dépendent fortement de la température durant l'année ainsi que des types de jours (jour travaillé, jour non travaillé, vacances scolaires,..., etc). Dans [Birt et al., 2012], 327 ménages au Canada ont été regroupés en appliquant l'algorithme des K-moyennes et la classification hiérarchique. Les chercheurs ont découvert une forte corrélation entre la consommation électrique résidentielle et les fluctuations de la température en hiver comme en été. Cette corrélation permet de déterminer l'électricité consommée par le chauffage et par le climatiseur. Les mêmes approches de classification ont été utilisées dans [Cao et al., 2013, Kwac et al., 2013]. Le but de ces études est d'identifier des groupes de ménages qui présentent les mêmes pics durant la journée, et aussi de déterminer les profils de consommation qui sont stables pendant certaines périodes.

Restant sur les approches des centroides, plusieurs études ont utilisé des cartes auto-organisatrices (SOMs) pour regrouper les consommateurs
[Figueiredo et al., 2005, Dent et al., 2013, Verdú et al., 2006, Sanchez et al., 2009].

Figueiredo et al. ont utilisé l'algorithme SOMs pour identifier des groupes de consommateurs ayant des comportements de consommation similaires [Figueiredo et al., 2005]. Les auteurs ont construit par la suite un arbre de décision afin d'affecter automatiquement de nouveaux ménages à l'une des classes trouvées au par avant. Dent et al. ont également appliqué l'algorithme SOMs pour classifier 93 ménages au Royaume-Uni [Dent et al., 2013]. Verdu et al. se sont appuyés sur les cartes auto-organisatrices afin d'étudier l'évolution des comportements de consommation au fil du temps [Verdú et al., 2006]. L'objectif de ce travail, est de reconnaitre les comportements de consommation qui s'écartent d'un comportement typique et d'identifier de nouveaux clients. Contrairement aux approches citées précédemment qui utilisent uniquement les données de consommation électrique, Sanchez et al. ajoutent des informations socio-économiques pour alimenter l'algorithme SOMs [Sanchez et al., 2009].

Récemment, quatre publications ont utilisé les mêmes données (CER), que nous avons repris pour notre travail dans ce chapitre [McLoughlin et al., 2015, Haben et al., 2016, Tong et al., 2016, Wang et al., 2016]. Dans le premier papier, les auteurs ont étudié trois méthodes de classification à savoir : K-moyennes, K-médoide et SOMs, et ils ont choisi la méthode la plus performante qui permet de regrouper les ménages en fonction de leur consommation électrique journalière. Cette classification a été effectuée sur une période de six mois. Les résultats obtenus sont représentés par des profils types nommés (PCs). Chaque PC a été lié aux caractéristiques des ménages à travers une régression logistique multinomiale. Le deuxième article présente une analyse détaillée sur les données de consommation électrique. L'objectif de cette analyse, est de mieux comprendre les pics de demande et d'identifier les principales sources qui provoquent la variabilité des comportements de consommation. Le regroupement a été basé sur sept attributs. Ces attributs ont

été calculés pour chaque consommateur sur une période d'un an. En utilisant un modèle de mélange fini basé sur une méthode de classification, un ensemble réduit de classes a été obtenu. Dans [Tong et al., 2016], les auteurs ont mené des recherches en vue d'identifier le lien entre les profils types et les informations sur les ménages en Irlande. Trois groupes de consommateurs ont été identifiés en utilisant l'algorithme des X-moyennes. Ces groupes ont été étiquetés par : groupe du jour, groupe de soirée et groupe de minuit. Les auteurs ont constaté que le comportement de consommation électrique est principalement lié à l'utilisation d'Internet. Une nouvelle approche de classification nommée CFSFDP (*Clustering by Fast Search and Find of Density Peaks*) a été proposée dans [Wang et al., 2016]. Cette technique permet de classifier les comportements de consommation électrique dynamiques, où "les dynamiques" désignent les transitions et les relations entre les comportements de consommation.

Les travaux présentés dans ce chapitre partagent le même objectif que celui des auteurs de [Kwac et al., 2014, Kwac et al., 2017]. Cela revient à ce que les trois travaux visent à synthétiser un ensemble important de données de consommation électrique en un ensemble réduit de profils types journaliers liés aux styles de vie des consommateurs.

Dans [Kwac et al., 2014], la méthodologie utilisée pour créer un dictionnaire de profils de charge fonctionne en deux étapes, où les auteurs exécutent l'algorithme des K-moyennes suivi d'une classification hiérarchique. Le but de cette méthodologie est de fusionner les sous classes qui sont trop proches.

L'approche de classification proposée dans cette thèse est une extension du modèle de mélange gaussien classique dont les paramètres dépendent d'une variable exogène représentant le type de jour (samedi, dimanche et jour travaillé). Le modèle proposé vise à regrouper les consommateurs qui consomment de la même manière durant les trois types de jours et non séparément. La deuxième partie de ce travail porte sur le changement de comportement des consommateurs au fil du temps, qui est quantifié par le biais de l'entropie.

Contrairement aux [Kwac et al., 2014, Kwac et al., 2017], l'ensemble de données utilisé dans cette étude ne couvre qu'une seule année et les emplacements des compteurs ne sont pas connus. La disponibilité de ces informations peut être utile pour une meilleure efficacité du modèle de classification ainsi pour l'interprétation des résultats.

3.3 Modèles de classification automatique

Dans cette section, nous allons nous intéresser aux méthodes de classification automatique. Ces méthodes visent à partitionner un ensemble de données en classes homogènes. Deux catégories de méthodes sont distinguées : approches non probabilistes et approches probabilistes.

Les approches non probabilistes englobent la classification hiérarchique
[Lance and Williams, 1967], la méthode des K-moyennes [MacQueen, 1967], les cartes auto-organisatrices de Kohonen [Kohonen, 1982] et la classification floue [Bezdek, 1974]. Pour les approches probabilistes, elles se basent sur une hypothèse concernant la distribution de probabilité des données à classifier. Les modèles de mélange [McLachlan and Basford, 1988] constituent un cadre adapté à ce type d'approche.

Dans ce chapitre, nous nous intéressons aux méthodes de classification automatique qui visent à optimiser le critère d'inertie intra-classes à savoir : la classification ascendante hiérarchique, les K-moyennes et le modèle de mélange gaussien. Dans un premier temps, nous

rappelons les méthodes citées, et nous décrivons ensuite l'approche que nous proposons qui est basée sur un modèle de mélange gaussien intégrant des variables exogènes.

Données temporelles et notations

Dans cette partie, nous allons définir les notations utilisées tout au long de ce chapitre. Les données à classifier sont représentées par N séries temporelles notées $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. Chaque série temporelle $\mathbf{x}_i$, qui correspond à un compteur intelligent indexée par i, est notée $\mathbf{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iD})$, où D est le nombre de jours, et $\boldsymbol{x}_{id} = (\boldsymbol{x}_{id1}, \boldsymbol{x}_{id2}, \ldots, \boldsymbol{x}_{idT})$ est un vecteur de T mesures appartenant à $\mathbb{R}$.

Les algorithmes classiques de classification décrits dans les parties qui suivent sont utilisés non seulement dans le cas de la classification des consommations électriques des usagers (Compteurs intelligents) $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, mais aussi dans la classification des consommations électriques journalières $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iD})$ d'un compteur i.

3.3.1 Algorithme des K-moyennes

L'algorithme des K-moyennes est l'un des algorithmes d'apprentissage non supervisé les plus simples et les plus communément employé [MacQueen, 1967]. Étant donné un ensemble d'observations et un nombre entier de groupes K fixé à priori, l'objectif est de diviser cet ensemble d'observations en K classes tout en minimisant le critère d'inertie intra-classes. L'idée principale est de commencer par choisir aléatoirement K centres pour chaque classe. Il est recommandé que les centres des classes soient loin les uns des autres pour une meilleure exploration de l'espace de recherche. L'étape suivante, consiste à affecter chaque observation de l'ensemble à la classe dont le centre de gravité est le plus proche. Une fois les classes formées, le centre de gravité de chaque classe est recalculé. Les deux précédentes étapes sont répétées jusqu'à ce que les centres de gravité ne changent pas de placement. La Figure 3.1 représente les étapes de classification de 10 observations en 2 classes ($K = 2$) à travers l'algorithme des K-moyennes.

Pour classifier les observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, l'algorithme des K-moyennes minimise l'inertie intra-classes qui est définie par :

$$C(\mathbf{p}, \mu) = \sum_{i=1}^{N} \sum_{\mathbf{x}_i \in p_k} ||\mathbf{x}_i - \mu_k||^2,$$

où $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K)$ est l'ensemble des classes, et $\mu = (\mu_1, \mu_2, ..., \mu_K)$ est l'ensemble des centres. La norme $||.||$ est associée à la distance euclidienne. Les étapes principales de la méthode sont décrites dans le pseudo-code 1.

Algorithme 1 Pseudo-code de l'algorithme des K-moyennes

1: Initialisation : Choisir aléatoirement K centres $(\mu_1, \mu_2, ..., \mu_K)$
2: **Répéter**
3: Former les classes $\mathbf{p}$ en affectant les observations $\mathbf{x}_i$ aux centres les plus proches
$\mathbf{x}_i \in \mathbf{p}_k \Leftrightarrow k = argmin_l ||\mathbf{x}_i - \mu_l||^2$
4: Mettre à jour les centres des classes μ_k
$\mu_k \leftarrow \frac{\sum_{\mathbf{x}_i \in \mathbf{p}_k} \mathbf{x}_i}{\#\mathbf{p}_k}$
5: **Jusqu'à** Convergence (les centres de gravité ne changent plus)

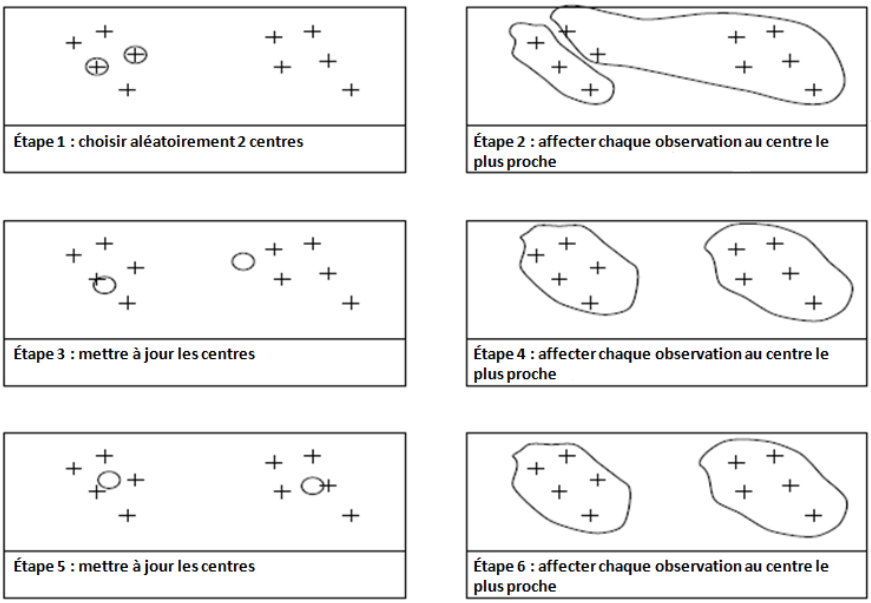


FIGURE 3.1 – Étapes de classification à travers l'algorithme des K-moyennes [Celeux et al., 1989]

3.3.2 Algorithme des K-moyennes fonctionnel

La version fonctionnelle des K-moyennes est une approche qui est capable de prendre en compte l'aspect fonctionnel des données [Ramsay and Silverman, 2005]. Cette méthode consiste à convertir les données de départ en fonctions, et les classifier en groupes juste après. L'implémentation de cette approche est basée sur les étapes suivantes :

- Lissage : Cette étape consiste à lisser l'ensemble des observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ qui sont sous forme de séries temporelles (ou courbes) en utilisant une approche de lissage (lissage polynomial, lissage par spline ou lissage par série de fourrier). Dans cette thèse nous avons utilisé le lissage par spline cubique avec pénalisation [Reinsch, 1967]. A la fin de cette étape un ensemble de courbes lissées $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N)$ est obtenu, où $\mathbf{s}_i = (\mathbf{s}_{i1}, \mathbf{s}_{i2}, \ldots, \mathbf{s}_{if})$ est la $i^{ème}$ courbe lissée sur des instants discrets $(1, 2, \ldots, f)$
- Analyse en Composantes Principales (ACP) : Une ACP [Benzécri and Bellier, 1976] est appliquée sur les données lissées $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N)$ obtenues à l'étape précédente afin de réduire leur dimension. Le nombre d'axes factoriels c retenu est celui qui explique au mieux la variabilité des données. Les scores des composantes principales résultants sont notés par $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$, où $\mathbf{y}_i \in \mathbb{R}^c$, tel que $c << f$.
- Classification : L'algorithme des K-moyennes est ensuite utilisé sur les scores des composantes principales $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$ obtenues à l'étape précédente.

La Figure 3.2 représente les étapes des K-moyennes fonctionnel susmentionnées.

3.3.3 Classification Ascendante Hiérarchique (CAH)

La Classification Ascendante Hiérarchique est une méthode de classification itérative [Lance and Williams, 1967]. Cette méthode regroupe itérativement les classes en partant d'une partition la plus élémentaire (un singleton par classe), puis en fusionnant successivement les classes qui se ressemblent jusqu'à l'obtention d'une seule classe. Le regroupement

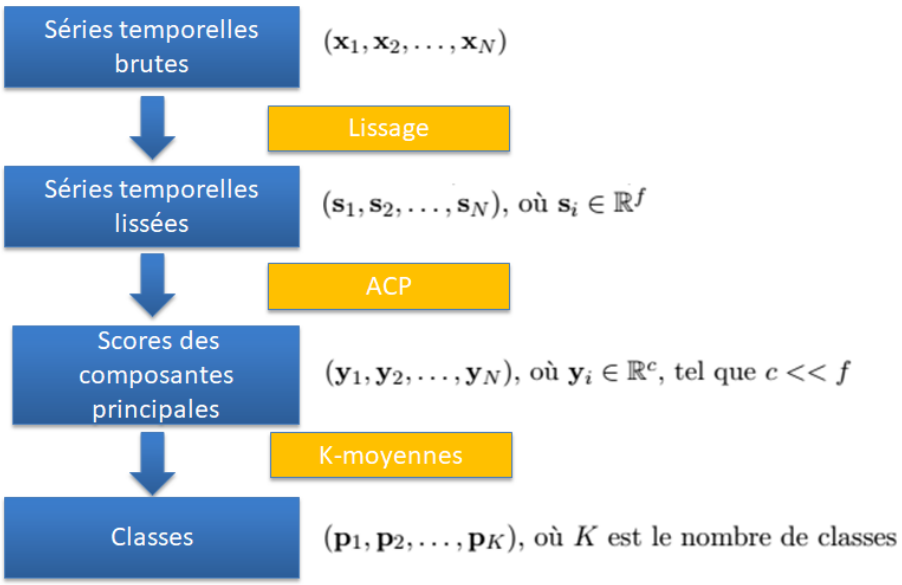


FIGURE 3.2 – Étapes des K-moyennes fonctionnel

des classes se fait de manière à optimiser un critère d'agrégation donné. Les regroupements successifs sont représentés par un dendrogramme où les feuilles correspondent aux singletons. Ce dendrogramme représente une hiérarchie de partitions. Pour déterminer une partition, une coupe au premier saut jugé significatif est recommandée, les sous arbres obtenus forment alors les classes. La Figure 3.3 représente un dendrogramme, où la coupe (trait en pointillé rouge) a permis de diviser les observations en 3 classes.

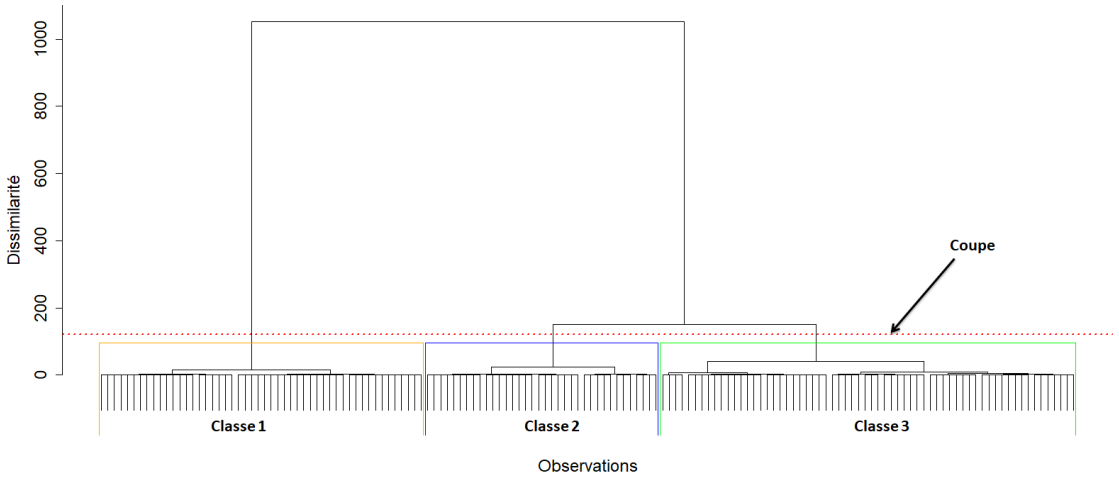


FIGURE 3.3 – Dendrogramme

Pour classifier les données $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, le pseudo-code 2 décrit les étapes principales de la méthode.

Algorithme 2 Pseudo-code de l'algorithme de Classification Ascendante Hiérarchique (CAH)

1: Initialisation : Former les classes initiales : singleton $\{\mathbf{x}_1\}, \{\mathbf{x}_2\}, ..., \{\mathbf{x}_N\}$
2: Calculer la matrice de distance entre singletons
3: **Répéter**
4: Regrouper les deux classes les plus proches
5: Mettre à jour le tableau des distances
6: **Jusqu'à** ce que le nombre de classes soit égal à 1

3.3.4 Modèle de Mélange classique

Un modèle de mélange est un modèle qui suppose que les données proviennent généralement d'un ensemble fini de classes, où chaque classe est modélisée par une loi de probabilité (loi normale, loi de poisson ou loi binomiale). Plusieurs recherches ont été menées sur les modèles de mélange. Parmi ces recherches, nous pouvons citer [Everitt and Hand, 1981, Titterington et al., 1985, McLachlan and Krishnan, 2008, Mengersen et al., 2011].

Définition d'un modèle de mélange

Un modèle de mélange suppose que les observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ sont générées suivant un mélange de K distributions $f_k(.;\theta_k)$, tel que $k = (1, 2, \ldots, K)$, où θ_k représente les paramètres de la $k^{ème}$ distribution. Les K distributions sont mélangées selon les proportions $(\pi_1, \pi_2, \ldots, \pi_K)$, où $\sum_k \pi_k = 1$. La variable latente qui représente l'étiquette de la classe associée à chaque observation i est notée $z_i \in \{1, 2, ..., K\}$, $\forall i = (1, 2, \ldots, N)$. Chaque observation $\mathbf{x}_i$ est ainsi distribuée suivant la densité suivante :

$$f(\mathbf{x}_i; \Theta) = \sum_k \pi_k f_k(\mathbf{x}_i; \theta_k), \tag{3.3.1}$$

où le vecteur $\Theta = (\pi_k, \theta_k)_k$ représente le vecteur des paramètres globaux du modèle à estimer.

Modèle de mélange gaussien

Un modèle de mélange gaussien est un modèle qui suppose que $f_k(.;\theta_k)$ est une densité gaussienne. La distribution de chaque $\mathbf{x}_i$ est ainsi définie par :

$$f(\mathbf{x}_i; \Theta) = \sum_k \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{m}_k, \boldsymbol{C}_k), \tag{3.3.2}$$

où $\mathcal{N}(.; \boldsymbol{m}_k, \boldsymbol{C}_k)$ désigne la densité normale de moyenne $\boldsymbol{m}_k$ et matrice de variance covariance $\boldsymbol{C}_k$. Pour une meilleure illustration, la Figure 3.4 représente un mélange de $K = 2$ distributions gaussiènnes, où chaque courbe colorée représente une densité normale.

Estimation par la méthode du maximum de vraisemblance et l'algorithme EM

Plusieurs approches peuvent être utilisées pour estimer les paramètres d'un modèle de mélange. Nous pouvons citer le maximum de vraisemblance [McLachlan and Peel, 2000] et la méthode bayésienne du Maximum A Posteriori (MAP) [Stephens,]. Dans cette thèse

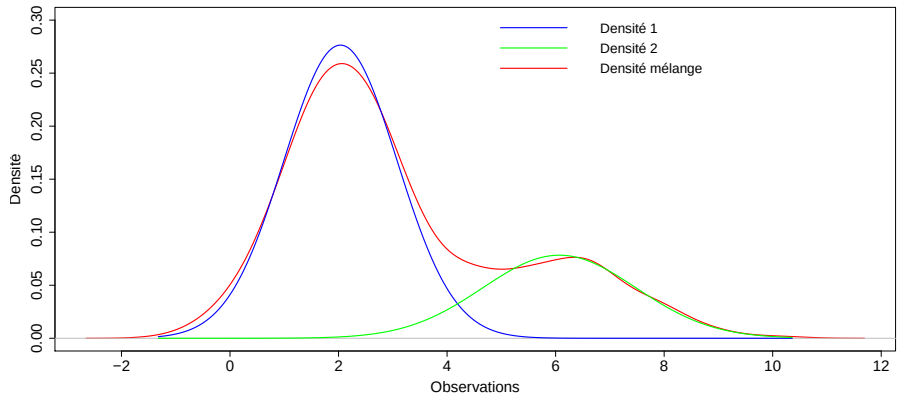


FIGURE 3.4 – Mélange de deux distributions gaussiènnes

nous allons nous intéresser au maximum de vraisemblance. La log-vraisemblance notée $L(\Theta)$ est définie par :

$$\begin{aligned} L(\Theta) &= \log f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; \Theta) \\ &= \sum_i \log f(\mathbf{x}_i; \Theta) \\ &= \sum_i \log \sum_k \pi_k f_k(\mathbf{x}_i; \theta_k). \end{aligned} \tag{3.3.3}$$

La détermination du meilleur paramètre $\widehat{\Theta}$ qui maximise la log-vraisemblance est effectuée via l'algorithme Espérance Maximisation (EM) [McLachlan and Krishnan, 2008]. L'algorithme EM est un algorithme itératif proposé par Dempster [Dempster et al., 1977] dans le but d'estimer les paramètres d'un modèle à variables latentes. Cet algorithme est adapté aux problèmes de classification, et plus particulièrement à l'estimation des paramètres des mélanges de loi. En effet ce dernier prend en compte la structure latente en complétant les données observées par les données non observées d'appartenance aux classes. La log-vraisemblance complétée par les classes non observées s'écrit :

$$L(\Theta, \mathbf{z}) = \sum_i \sum_k z_{ik} \log \Big(\pi_k f_k(\mathbf{x}_i; \theta_k)\Big), \tag{3.3.4}$$

où la variable binaire z_{ik} vaut 1 si $z_i = k$ et 0 sinon.

L'algorithme EM opère itérativement en 2 étapes. Dans un premier temps, les paramètres initiaux $\Theta^{(0)}$ sont fixés. Dans un deuxième temps, les nouveaux paramètres $\Theta^{(q+1)}$ sont calculés à partir des paramètres $\Theta^{(q)}$ trouvés lors de l'itération précédente, de manière à maximiser l'espérance conditionnelle de la log-vraisemblance complétée notée $Q(\theta, \theta^{(q)})$ et définie par :

$$\begin{aligned} Q(\Theta, \Theta^{(q)}) &= E\left[L_c(\Theta, \mathbf{z}) | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \Theta^{(q)}\right] \\ &= \sum_i \sum_k E\left[z_{ik} | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \Theta^{(q)}\right] \log \Big(\pi_k f_k(\mathbf{x}_i; \theta_k^{(q)})\Big) \\ &= \sum_i \sum_k \tau_{ik}^{(q)} \log \Big(\pi_k f_k(\mathbf{x}_i; \theta_k^{(q)})\Big), \end{aligned} \tag{3.3.5}$$

où

$$\tau_{ik}^{(q)} = \frac{\pi_k^{(q)} f_k\left(\mathbf{x}_i; \theta_k^{(q)}\right)}{\sum_k \pi_k^{(q)} f_k\left(\mathbf{x}_i; \theta_k^{(q)}\right)} \tag{3.3.6}$$

est la probabilité *a posteriori* que l'observation $\mathbf{x}_i$ soit dans la classe k avec les paramètres $\Theta^{(q)}$ de l'itération q.

L'algorithme EM consiste à alterner itérativement les deux étapes suivantes jusqu'à la convergence :

— Espérance (étape E) : Calculer la fonction $Q(\Theta, \Theta^{(q)})$

— Maximisation (étape M) : Mettre à jour les paramètres en maximisant la fonction Q

La qualité d'estimation des paramètres dépend fortement de leur initialisation. Dans le cas gaussien, la stratégie la plus utilisée est celle proposée par [Biernacki et al., 2003] où l'algorithme EM est lancé plusieurs fois avec des paramètres initiaux $\Theta^{(0)}$ choisis par l'algorithme des K-moyennes. Les paramètres $\Theta^{(0)}$ retenus sont ceux qui donnent la plus grande log-vraisemblance.

Algorithme EM pour un mélange de lois gaussiennes

Dans le cas d'un modèle de mélange fini où les données de chaque classe suivent une distribution gaussienne, le vecteur de paramètres à estimer s'écrit $\Theta = (\pi_k, \boldsymbol{m}_k, \boldsymbol{C}_k)$. A partir d'un paramètre initial $\Theta^{(0)}$, l'algorithme EM alterne les deux étapes suivantes jusqu'à la convergence :

— Espérance (étape E) : Calcul de l'espérance de la log-vraisemblance complétée conditionnellement aux données observées $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ et au paramètre courant $\Theta^{(q)}$:

$$\begin{aligned} Q(\Theta; \Theta^{(q)}) &= E\left[L_c(\Theta, \mathbf{z}) | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \Theta^{(q)}\right] \\ &= \sum_i \sum_k \tau_{ik}^{(q)} \log\left(\pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{m}_k^{(q)}, \boldsymbol{C}_k^{(q)})\right), \end{aligned} \tag{3.3.7}$$

où

$$\tau_{ik}^{(q)} = \frac{\pi_k^{(q)} \mathcal{N}(\mathbf{x}_i; \boldsymbol{m}_k^{(q)}, \boldsymbol{C}_k^{(q)})}{\sum_k \pi_k^{(q)} \mathcal{N}(\mathbf{x}_i; \boldsymbol{m}_k^{(q)}, \boldsymbol{C}_k^{(q)})} \tag{3.3.8}$$

représente la probabilité *a posteriori* que l'observation $\mathbf{x}_i$ appartient à la classe k.

— Maximisation (étape M) : Maximisation de l'espérance de la log-vraisemblance complétée en tenant compte des données observées $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ et du paramètre courant $\Theta^{(q)}$. La maximisation de la fonction Q par rapport à ($\boldsymbol{m}_k$ et $\boldsymbol{C}_k$) aboutit aux formules de mise à jour suivantes :

$$\pi_k^{(q+1)} = \frac{1}{N} \sum_i \tau_{ik}^{(q)}, \tag{3.3.9}$$

$$\boldsymbol{m}_k^{(q+1)} = \frac{1}{\sum_i \tau_{ik}^{(q)}} \sum_i \tau_{ik}^{(q)} \mathbf{x}_i, \tag{3.3.10}$$

$$\boldsymbol{C}_k^{(q+1)} = \frac{1}{\sum_i \tau_{ik}^{(q)}} \sum_i \tau_{ik}^{(q)} \left(\mathbf{x}_i - \boldsymbol{m}_k^{(q+1)}\right)\left(\mathbf{x}_i - \boldsymbol{m}_k^{(q+1)}\right)'. \tag{3.3.11}$$

Le pseudo-code 3 décrit l'algorithme EM pour un modèle de mélange gaussien. Après la convergence de l'algorithme, chaque observation $\mathbf{x}_i$ sera attribuée à la classe z_i qui maximise la probabilité *a posteriori* τ_{ik}.

Algorithme 3 pseudo-code de l'algorithme EM appliqué à un mélange gaussien

Entrée : Ensemble de données $(\mathbf{x}_i)_{1\leq i\leq N}$
Initialisation : $\Theta^{(0)}$
$q \leftarrow 0$
Répéter
  Étape espérance
  Calcul des probabilités à postériori $\tau_{ik}^{(q)}$ $\forall i,k$ (équation 3.3.8)
  Étape maximisation
  Mise à jour de $\pi_k^{(q+1)}$ $\forall k$ (équation 3.3.9)
  Mise à jour de $\boldsymbol{m}_k^{(q+1)}$ $\forall k$ (équation 3.3.10)
  Mise à jour de $\boldsymbol{C}_k^{(q+1)}$ $\forall k$ (équation 3.3.11)
  $q \leftarrow q+1$
Jusqu'à Convergence
Sortie : Vecteur des paramètres $\widehat{\Theta}$, probabilités *a postériori* $\widehat{\tau}_{ik}$

3.4 Modèle de mélange proposé

3.4.1 Formulation modèle

Le modèle proposé est un mélange à K distributions, qui suppose que les séries temporelles $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ associées aux compteurs intelligents sont distribuées selon la densité mélange décrite précédemment (voir équation 3.3.1). En raison de la grande dimension $D \times T$ des séries temporelles étudiées (Nous rappelons que D et T représentent respectivement le nombre de jour et le nombre de fréquences de prélèvement par jour), il a été supposé que les matrices de variance covariance sont diagonales. Dans ce cas, $\boldsymbol{m}_k$ et $\boldsymbol{C}_k$ sont décomposées comme suit :

$$\begin{array}{rcl} \boldsymbol{m}_k & = & (m_{k1}, \ldots, m_{kD}), \\ \boldsymbol{C}_k & = & \mathrm{diag}(C_{k1}, \ldots, C_{kD}), \end{array} \tag{3.4.1}$$

où $\boldsymbol{m}_{kd}$ $(d = 1, ..., D)$ sont des vecteurs de dimension T et $\mathrm{diag}(C_{k1}, C_{k2}, ..., C_{kD})$ est une matrice diagonale en blocs. Ces blocs représentent les matrices de variance covariance diagonales $C_{k1}, C_{k2}, ..., C_{kD}$ de dimension $T \times T$.

De manière à tenir compte des spécificités applicatives, notamment du fait que le calendrier impacte les consommations électriques, les contraintes suivantes s'imposent à la structure du modèle proposé :

$$\left\{ \begin{array}{rcl} m_{kd} & = & \mu_{kl}, \\ C_{kd} & = & \Sigma_{kl}, \end{array} \right. \quad \text{si } d \text{ correspond au type de jour } l \in \{1, \ldots, L\}, \tag{3.4.2}$$

où $L = 3$ correspond aux types de jours samedi $(l = 1)$, dimanche $(l = 2)$ et jour travaillé $(l = 3)$. μ_{kl} et Σ_{kl} représentent respectivement la moyenne et la matrice de variance covariance relatives au type de jour l. Pour le mois de novembre 2010 qui commence par

un lundi, la moyenne $\boldsymbol{m}_k$ et la matrice de variance covariance $\boldsymbol{C}_k$ sont explicitement écrites par :

$$\begin{aligned}\boldsymbol{m}_k = (&\mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k1}, \quad \mu_{k2},\\ &\mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k1}, \quad \mu_{k2},\\ &\mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k1}, \quad \mu_{k2},\\ &\mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k3}, \quad \mu_{k1}, \quad \mu_{k2},\\ &\mu_{k3}, \quad \mu_{k3}),\end{aligned} \tag{3.4.3}$$

$$\begin{aligned}\boldsymbol{C}_k = \mathrm{diag}(&\Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k1}, \quad \Sigma_{k2},\\ &\Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k1}, \quad \Sigma_{k2},\\ &\Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k1}, \quad \Sigma_{k2},\\ &\Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k3}, \quad \Sigma_{k1}, \quad \Sigma_{k2},\\ &\Sigma_{k3}, \quad \Sigma_{k3}),\end{aligned} \tag{3.4.4}$$

Il est supposé que pour chaque série temporelle $\mathbf{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iD})$ provenant de la $k^{\text{ème}}$ distribution du mélange, la variable $\boldsymbol{x}_{id}$ (La consommation électrique du compteur i durant le jour d) est distribuée selon l'une des L distributions gaussiennes spécifiques à un type jour. L'affectation des variables $\boldsymbol{x}_{id}$ aux différents modèles gaussiens est donnée par des variables exogènes discrètes notées $\delta_d \in \{1, ..., L\}$. La densité de la $k^{\text{ème}}$ distribution s'écrit

$$f_k(\mathbf{x}_i; \theta_k) = \prod_d \mathcal{N}(\boldsymbol{x}_{id}; \boldsymbol{m}_{kd}, \boldsymbol{C}_{kd}). \tag{3.4.5}$$

En utilisant les variables binaires δ_{dl} ($\delta_{dl} = 1$ si $\delta_d = l$ et $\delta_{dl} = 0$ sinon), les moyennes ainsi que les matrices de variance covariance peuvent respectivement s'écrire

$$\boldsymbol{m}_{kd} = \sum_l \delta_{dl} \mu_{kl}, \tag{3.4.6}$$

et

$$\boldsymbol{C}_{kd} = \sum_l \delta_{dl} \Sigma_{kl}, \tag{3.4.7}$$

La densité mélange définit par l'équation 3.3.1 peut se réécrire comme :

$$f(\mathbf{x}_i; \Theta) = \sum_k \pi_k \left(\prod_d \mathcal{N}(\boldsymbol{x}_{id}; \sum_l \delta_{dl} \mu_{kl}, \sum_l \delta_{dl} \Sigma_{kl}) \right). \tag{3.4.8}$$

Le modèle proposé est décrit par le modèle graphique représenté dans la Figure 3.5.

D'autres variables binaires auraient pu être utilisées dans le modèle proposé pour encoder des jours fériés et des vacances scolaires par exemple. Comme les données ne sont disponibles que sur une seule année, l'estimation des paramètres du modèle avec une telle modélisation aura une précision limitée (en raison du nombre limité des jours fériés et des vacances scolaires). Il est à noter que les jours fériés ont été considérés comme des dimanches. Dans le reste du chapitre, on supposera que $\Theta = (\pi_k, (\mu_{kl}, \Sigma_{kl})_{l=1,\ldots,L})_{k=1,\ldots,K}$. L'estimation des paramètres du modèle est décrite en détail dans la partie suivante.

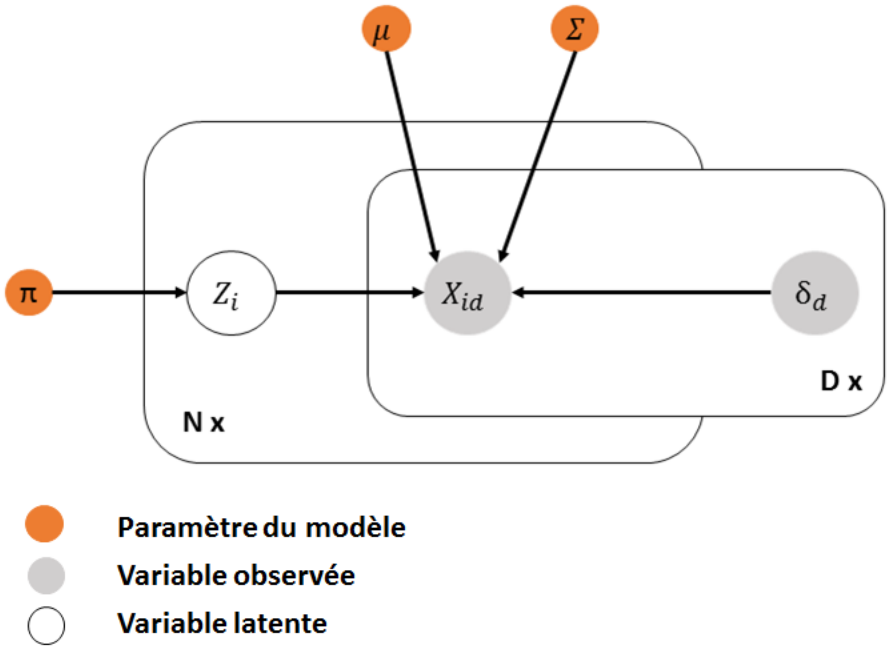


FIGURE 3.5 – Représentation graphique du modèle proposé

3.4.2 Estimation des paramètres du modèle

Pour estimer les paramètres du modèle, l'approche du maximum de vraisemblance via l'algorithme espérance maximisation a été utilisée. Pour le modèle proposé, le critère de log-vraisemblance à maximiser est donné par

$$L(\Theta) = \sum_i \log\left(\sum_k \pi_k \prod_d \mathcal{N}(\boldsymbol{x}_{id}; \sum_l \delta_{dl}\mu_{kl}, \sum_l \delta_{dl}\Sigma_{kl})\right). \tag{3.4.9}$$

La log-vraisemblance complétée est définie par

$$L_c(\Theta) = \sum_i \sum_k z_{ik} \log\left(\pi_k \prod_d \mathcal{N}(\boldsymbol{x}_{id}; \sum_l \delta_{dl}\mu_{kl}, \sum_l \delta_{dl}\Sigma_{kl})\right), \tag{3.4.10}$$

où la variable binaire z_{ik} vaut 1 lorsque $z_i = k$ et 0 sinon.

Comme il est souvent utilisé dans l'initialisation des modèles de mélanges, l'algorithme des K-moyennes a été appliqué pour partitionner les compteurs en K classes. L'estimation préliminaire des paramètres du modèle proposé est donc fournie par les paramètres obtenus de l'algorithme des K-moyennes.

À partir d'un paramètre initial $\Theta^{(0)}$, l'algorithme EM alterne les deux étapes suivantes jusqu'à la convergence :

— Espérance (étape E) : Calcul de la probabilité *a postériori* τ_{ik} qu'un compteur i appartient à une classe k

$$\tau_{ik}^{(q)} = \frac{\pi_k^{(q)} \prod_d \mathcal{N}\left(\boldsymbol{x}_{id}; \sum_l \delta_{dl}\mu_{kl}^{(q)}, \sum_l \delta_{dl}\Sigma_{kl}^{(q)}\right)}{\sum_k \pi_k^{(q)} \prod_d \mathcal{N}\left(\boldsymbol{x}_{id}; \sum_l \delta_{dl}\mu_{kl}^{(q)}, \sum_l \delta_{dl}\Sigma_{kl}^{(q)}\right)}. \tag{3.4.11}$$

— Maximisation (étape M) : Mise à jour des paramètres du modèle

$$\pi_k^{(q+1)} = \frac{1}{N}\sum_i \tau_{ik}^{(q)}, \tag{3.4.12}$$

$$\mu_{kl}^{(q+1)} = \frac{1}{\sum_{i,d}\tau_{ik}^{(q)}\delta_{dl}}\sum_{i,d}\tau_{ik}^{(q)}\delta_{dl}\boldsymbol{x}_{id}, \tag{3.4.13}$$

$$\Sigma_{kl}^{(q+1)} = \frac{1}{\sum_{i,d}\tau_{ik}^{(q)}\delta_{dl}}\sum_{i,d}\tau_{ik}^{(q)}\delta_{dl}\left(\boldsymbol{x}_{id}-\mu_{kl}^{(q+1)}\right)\left(\boldsymbol{x}_{id}-\mu_{kl}^{(q+1)}\right)'. \tag{3.4.14}$$

Le critère d'arrêt utilisé dans l'algorithme EM est basé sur un seuil de log-vraisemblance prédéfini. Le pseudo-code 4 résume la procédure itérative de l'algorithme EM. Nous rappelons que chaque observation $\mathbf{x}_i$ sera affectée à la classe z_i qui maximise la probabilité *a posteriori* τ_{ik}.

Algorithme 4 pseudo-code de l'algorithme EM pour le modèle proposé

1: **Entrée** : données $(\boldsymbol{x}_{id})_{1\leq i\leq N, 1\leq d\leq D}$
2: **Initialisation** : $\Theta^{(0)}$
   $q \leftarrow 0$
3: **Répéter**
4: **Étape espérance**
   Calcul des probabilités *a postériori* $\tau_{ik}^{(q)}$ $\forall i,k$ (équation 3.4.11)
5: **Étape maximisation**
   Mise à jour de $\pi_{kl}^{(q+1)}$ $\forall k,l$ (équation 3.4.12)
   Mise à jour de $\mu_{kl}^{(q+1)}$ $\forall k,l$ (équation 3.4.13)
   Mise à jour de $\Sigma_{kl}^{(q+1)}$ $\forall k,l$ (équation 3.4.14)
   $q \leftarrow q+1$
6: **Jusqu'à** Convergence
7: **Sortie** : Vecteur des paramètres $\widehat{\Theta}$, probabilités *a postériori* $\widehat{\tau}_{ik}$

Il est à noter que l'approche proposée diffère de celle qui consiste à diviser dans un premier temps les données en L sous-ensembles $(P_l)_{l=1,\ldots,L}$ de séries temporelles journalières avec

$$P_l = \{\boldsymbol{x}_{id} \mid 1\leq i\leq N, 1\leq d\leq D,\ d \text{ correspond à un type de jour } l\},$$

et classifier dans un deuxième temps chacun de ces sous-ensembles en utilisant un modèle de mélange gaussien classique. La principale différence, est que cette dernière ne conduit pas directement à une partition de $(\mathbf{x}_i)_{i=1,\ldots,N}$ mais plutôt à une partition de séries temporelles journalières $(\boldsymbol{x}_{id})_{i=1,\ldots,N,d=1,\ldots,D}$. Soit $(\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(L)})$ et $(\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(L)})$ deux ensembles qui représentent respectivement les profils types journaliers et les partitions obtenus par cette méthode, avec $\mathbf{c}^{(l)} = (c_k^{(l)})_{k=1\ldots,K}$ et $\mathbf{z}^{(l)} = (z_j^{(l)})_{i=1,\ldots,M_l}$, où $c_k^{(l)}$ est le profil journalier correspondant à la $k^{ième}$ classe de P_l, $z_j^{(l)} \in \{1,\ldots,K\}$ est l'étiquette de la classe de la $j^{ième}$ série temporelle de P_l, et M_l est le cardinal de P_l. Les partitions $(\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(L)})$ peuvent être réorganisées en une seule partition $\mathbf{z} = (z_{id})_{i=1,\ldots,N,d=1,\ldots,D}$ constitué de $L\times K$ classes, où l'étiquette de la classe $z_{id} \in \{1,\ldots,L\times K\}$ est obtenu à partir de cette réorganisation.

Afin de regrouper les consommateurs qui consomment de la même manière pendant les L types de jours, des étapes supplémentaires sont nécessaires. On peut, par exemple, faire

correspondre l'ensemble des profils $\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(L)}$ pour obtenir la représentation souhaitée des classes, ou partitionner les séries temporelles catégorielles définies par les lignes de la matrice

$$\begin{bmatrix} z_{11} & \cdots & z_{1d} & \cdots & z_{1D} \\ \vdots & & \vdots & & \vdots \\ z_{i1} & \cdots & z_{id} & \cdots & z_{iD} \\ \vdots & & \vdots & & \vdots \\ z_{N1} & \cdots & z_{Nd} & \cdots & z_{ND} \end{bmatrix}$$

en utilisant une méthode appropriée. Malgré la pertinence de ce schéma, nous avons opté pour un modèle global qui prend en compte l'aspect temporel à travers une structure spécifique imposée aux paramètres du modèle.

3.5 Choix du nombre de classes

La détermination du nombre de classes à partir d'un ensemble de données est un problème fondamental en classification automatique. Certains algorithmes comme les K-moyennes [MacQueen, 1967], K-médoides [Kaufman and Rousseeuw, 1987] et
les algorithmes d'espérance maximisation exigent la spécification de ce paramètre au départ. Alors que d'autres algorithmes tels que DBSCAN [Ester et al., 1996] et la classification hiérarchique [Lance and Williams, 1967] ne le nécessitent pas. Le choix correct du nombre de classes est souvent ambigu, à cause des interprétations qui dépendent de la forme des classes et aussi de la résolution de classification souhaitée par l'utilisateur. L'augmentation du nombre de classes réduit l'erreur de classification. Si nous considérons le cas extrême, où chaque classe est composée d'un singleton (c.à.d le nombre de classes sera égale au nombre d'observations), l'erreur de classification sera égale à zéro. Le choix optimal du nombre de classes revient à trouver un équilibre entre la compression des données et la précision. Il existe plusieurs méthodes qui vont nous aider à prendre cette décision. Nous pouvons citer la méthode du coude, la validation croisée [Smyth, 1996], la méthode de la silhouette [Rousseeuw, 1987] et les approches par critère d'information. Dans cette thèse, nous allons nous intéresser à deux méthodes à savoir : la méthode du coude et le critère d'information bayésien (BIC) [Schwarz, 1978].

3.5.1 Méthode du coude

Cette méthode est souvent utilisée conjointement avec les algorithmes de partitionnement tels que l'algorithme des K-moyennes. Elle consiste à exécuter l'algorithme en faisant varier le nombre de classes de 1 jusqu'à K_{max}. Le nombre de classes choisi est celui à partir duquel le critère d'inertie intra-classes ne décroit pas d'une manière significative.

3.5.2 Critère d'information bayésien (BIC)

Ce critère sert à sélectionner un modèle parmi un ensemble fini de modèles. Le meilleur modèle est celui qui minimise ce critère. Cette approche est souvent utilisée pour choisir un nombre de classes dans le cadre des modèles de mélange. Le Critère BIC est défini par la vraisemblance pénalisée par un terme qui dépend du nombre de paramètres. Le critère BIC est défini par :

$$BIC\,(K) = -2L\,(\Theta) + \nu_K \log(N), \tag{3.5.1}$$

où Θ est le vecteur des paramètres globaux à estimer par l'algorithme EM, N le nombre d'observations et ν_K le nombre de paramètres des K composantes du modèle de mélange.

3.6 Résultats de classification à l'échelle d'un bâtiment

Dans cette section, nous allons nous intéresser aux résultats de classification automatique des consommations électriques journalières au niveau du bâtiment Galilée de Genral Electric (GE). L'objectif de cette classification est d'identifier les différents profils types de consommation électrique journaliers. Pour atteindre cet objectif, deux méthodes de classification automatique ont été utilisées à savoir : l'algorithme des K-moyennes fonctionnel et le modèle de mélange gaussien. Nous rappelons que les consommations électriques ont été prélevées par un seul compteur durant l'année 2013. Le travail décrit dans cette section peut être appliqué sur n'importe quel type de bâtiment (ménage, entreprise,... etc). Dans les parties qui suivent, nous déterminons le nombre de classes et interprétons les résultats de classification obtenus.

3.6.1 Choix du nombre de classes

K-moyennes fonctionnel

Pour le choix du nombre de classes, la méthode du coude a été utilisée (voir Section 3.5.1). La Figure 3.6 représente l'évolution de l'inertie intra-classes en fonction du nombre de classes utilisé par l'algorithme des K-moyennes fonctionnel. Comme on peut s'y attendre, l'inertie intra-classes décroit avec l'augmentation du nombre de classes, mais à partir d'un nombre de classes égal à 7, on observe que la variation du critère n'est pas significative. Le nombre de classes a ainsi été fixé à 7.

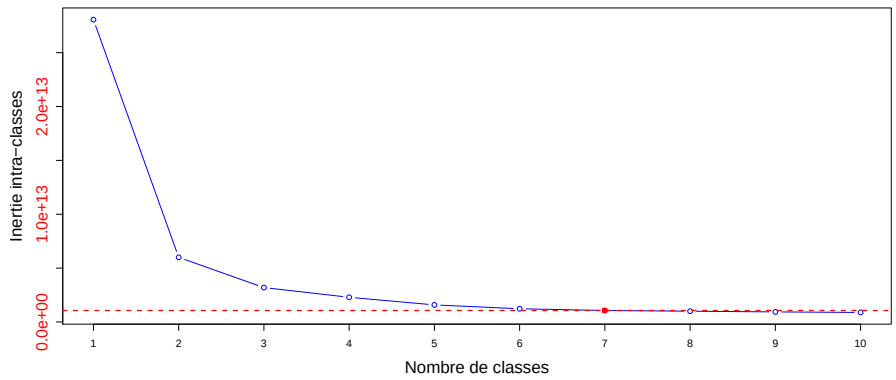


FIGURE 3.6 – Évolution de l'inertie intra-classes en fonction du nombre de classes (K-moyennes fonctionnel)

Cas du modèle de mélange gaussien

La sélection du nombre de classes associé au modèle de mélange gaussien a été menée en minimisant le critère d'information bayésien (BIC) (voir Section 3.5.2). La Figure 3.7 montre l'évolution de la log-vraisemblance et du critère BIC en fonction du nombre de classes du modèle de mélange gaussien. Il est observé que la log-vraisemblance croît avec le nombre de classes, alors que le critère BIC montre une décroissance jusqu'à un nombre de classes égale à 14, qui est suivi par une légère augmentation. Vu la variation peu significative des deux critères à partir d'un nombre de classes égal à 7 et aussi par souci de cohérence d'interprétation des classes, le nombre de classes a été fixé à 7.

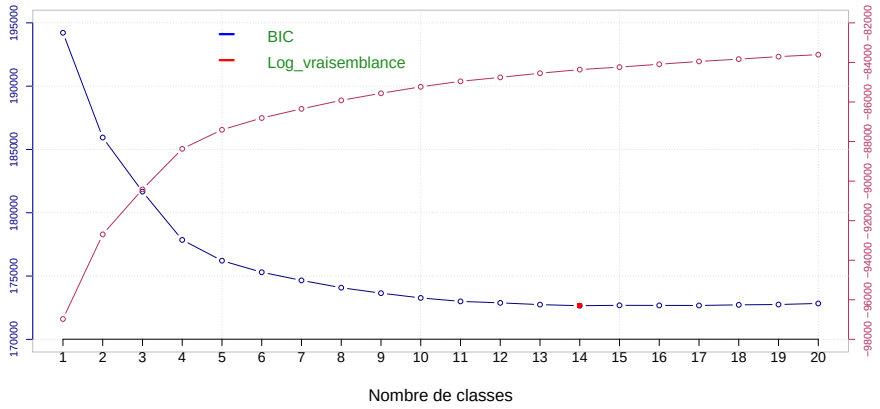


FIGURE 3.7 – Évolution du critère d'information bayésien (BIC) et de la log-vraisemblance en fonction du nombre de classes (modèle de mélange gaussien)

3.6.2 Interprétation des classes

Les résultats issus de l'application de l'algorithme des K-moyennes fonctionnel (voir Section 3.3.2) sont représentés dans la Figure 3.8 (a). Deux formes de profils de consommation électrique sont identifiés : le premier est lié à la consommation durant les jours de weekend (classes 1 et 2), tandis que le deuxième reflète celle durant les jours de semaine (classes 3, 4, 5, 6 et 7). Pour confirmer cette interprétation, les résultats de classification ont été croisés avec une variable catégorielle à deux modalités : jours de semaine (JS) et jours de weekend (JW) (voir la Figure 3.9 (a)). Les classes 3, 4, 5, 6 et 7 sont entièrement composées des jours de semaine, alors qu'une petite proportion des jours de semaine est retrouvée dans les classes 1 et 2 qui sont majoritairement constituées des jours de weekend. Cela est expliqué par la présence de jours fériés. Pour prouver cette explication, d'autres modalités liées aux vacances (jours fériés et vacances scolaires) ont été rajoutées. La Figure 3.9 (b) montre la répartition des jours travaillés, jours non travaillés (Jours de weekend et jours fériés) et vacances scolaires en fonction des classes. Les classes 1 et 2 sont composées de jours non travaillés avec une petite proportion des vacances scolaires. La classe 3 est constituée entièrement de vacances scolaires, alors que les observations de la classe 6 correspondent aux jours travaillés. Au contraire, les classes 4, 5 et 7 mélangent des jours travaillés et des vacances scolaires avec des proportions spécifiques à chacune. Un autre paramètre important à considérer dans cette interprétation est l'aspect saison. Pour cela, l'année a été divisée en trois saisons : saison froide (SF, entre 21 décembre et 20 mars), saison intermédiaire (SI, entre 21 mars et 20 juin et entre 21 septembre et 20 décembre) et saison chaude (SC, entre 21 juin et 20 septembre). La Figure 3.9 (c) représente ces saisons en fonction des classes obtenues. Il est observé la présence de jours chauds dans les classes

3 et 4, de jours tempérés dans les classes 1, 5 et 6 et des jours froids dans les classes 2 et 7.

Pour résumer, la Figure 3.8 (b) représente la distribution des classes durant l'année 2013 à travers un calendrier. Chaque couleur est associée à un jour appartenant à une classe. Les jours avec des valeurs manquantes sont représentés par la couleur blanche. Plusieurs prototypes de semaines sont observés. Ces prototypes dépendent bien sûr, de la température durant l'année 2013 et aussi des types de jours (jours travaillés, jours non travaillés et vacances scolaires). Il est remarqué la présence des classes 2 et 7 entre janvier et mars qui correspondent à la combinaison de jours de weekend et de jours de semaine durant la saison froide. La classe 2 représente aussi les jours fériés durant cette période (e.g., 1^{er} janvier). Pendant le mois d'avril, on observe l'apparition de deux nouvelles classes (1 et 5). Ce mélange de classes (1, 2, 5 et 7) est expliqué par la transition d'une période froide vers une autre plus chaude. Il est constaté également que les jours de la classe 4 sont souvent adjacents aux jours fériés de mai et à ceux des vacances scolaires d'été (juillet), où les bureaux sont moins occupés. Une autre classe qui peut également être attribuée à l'occupation du bâtiment est la classe 3. Les jours de cette classe sont observés en août et pendant les vacances de Noël. La différence entre les classes 3 et 4 est le taux d'occupation dans les bureaux, qui est plus faible pendant les jours de la classe 3. La classe 6 est observée en novembre et décembre, où la température commence à diminuer. Cette première classification calendaire peut servir de base pour une prévision long-terme.

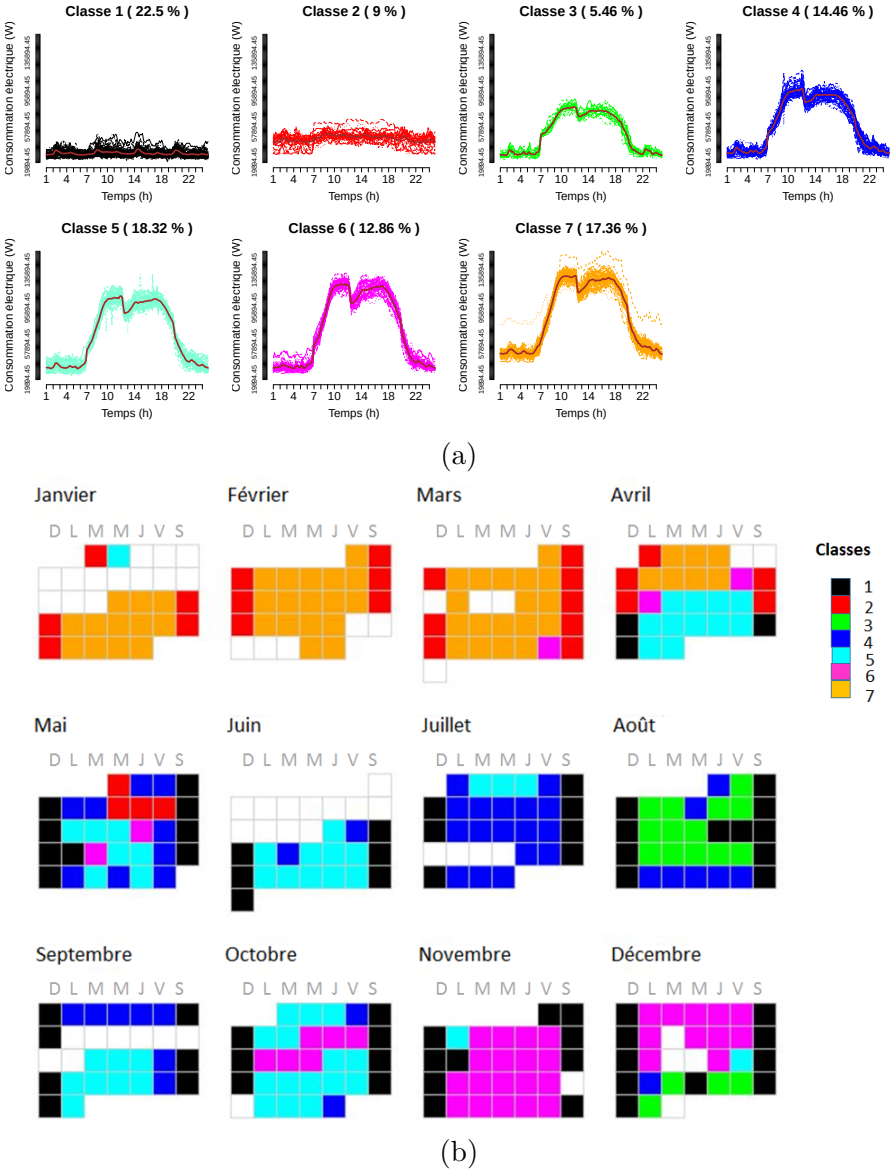
Les résultats obtenus avec le modèle de mélange gaussien (MMG) (voir Section 3.3.4) sont affichés dans la Figure 3.10 (a). Pour interpréter ces résultats de classification, nous les avons croisés avec les mêmes variables catégorielles utilisées avec l'algorithme des K-moyennes fonctionnel. Nous trouvons que l'interprétation des classes est identique à l'analyse précédente. La distribution des classes en fonction des jours de l'année 2013 est affichée dans la Figure 3.10 (b). Il convient de noter que la répartition est la même que celle obtenue avec l'algorithme des K-moyennes fonctionnel, à l'exception de quelques jours (par exemple, 21 avril , 5 mai , 16 août, ...) qui correspondent à des jours non travaillés.

3.7 Résultats de classification à l'échelle d'un territoire

Cette section présente les résultats de classification automatique qui vont nous permettre de mieux comprendre les comportements de consommation électrique d'usagers résidentiels. Nous considérons ici les données collectées à l'échelle d'un pays, Irlande. Rappelons que nous avons développé une approche qui est basée sur un modèle de mélange gaussien, dont les paramètres dépendent d'une variable exogène représentant le type de jour (samedi, dimanche et jour travaillé). Nous avons envisagé deux variantes. Dans un premier temps, l'approche a été appliquée sur des données de consommation électrique du mois de novembre, afin d'extraire des groupes de consommateurs ayant le même comportement de consommation. Dans un deuxième temps, la même approche a été adaptée à 12 mois de données de consommation, dans le but d'étudier l'évolution des comportements des consommateurs au fil des mois.

3.7.1 Classification appliquée aux données non-normalisées du mois de novembre

Dans cette partie, nous allons présenter les résultats obtenus après l'application du modèle proposé (voir Section 3.4) sur des données de consommation électrique des usagers durant le mois de novembre. Le choix de ce mois est lié à l'absence de jours spéciaux (jours



(a)

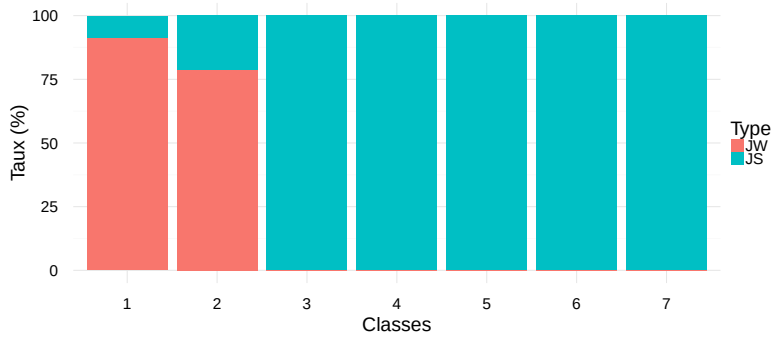
(b)

FIGURE 3.8 – (a) Courbes de consommation électrique pour chaque classe en utilisant l'algorithme des K-moyennes fonctionnel ; (b) Distribution des jours de chaque classe durant l'année 2013 en utilisant l'algorithme des K-moyennes fonctionnel

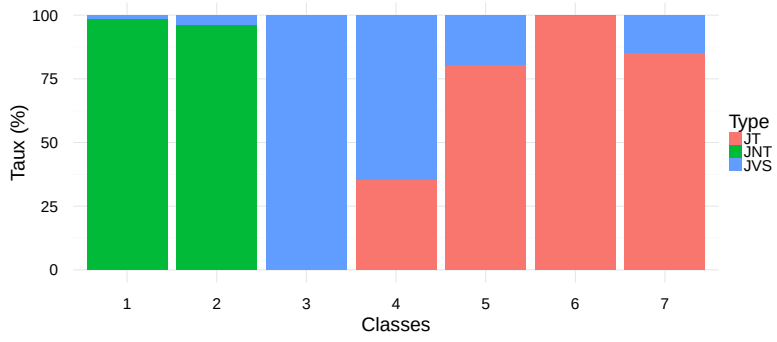
fériés, jours de vacance). Le choix du nombre de classes, l'évaluation du modèle proposé ainsi que l'interprétation des classes sont décrits dans les parties qui suivent.

Choix du nombre de classes

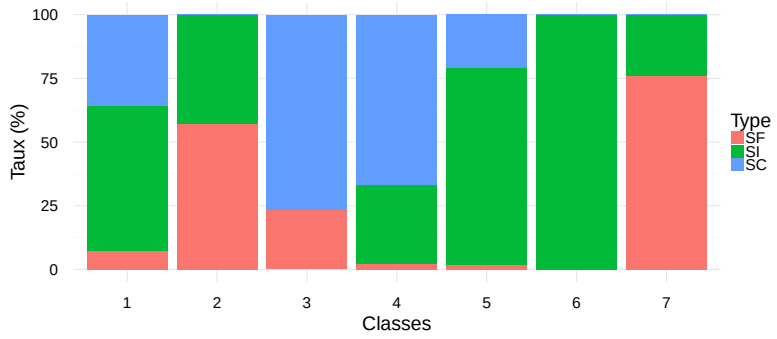
Pour sélectionner le nombre de classes, le critère d'information Bayesian (BIC) a été utilisé (voir Section 3.5.2). Le nombre de paramètres à estimer pour le modèle proposé est donné par $\nu_K = K(1 + 2LT) - 1$ où K est le nombre de classes, L est le nombre de types de jours et T est le nombre de mesures par jour. La Figure 3.11 représente l'évolution du critère BIC en fonction du nombre de classes. On peut constater que le critère diminue jusqu'à un nombre de classes $K = 6$, puis augmente progressivement. Par conséquent, le



(a)



(b)



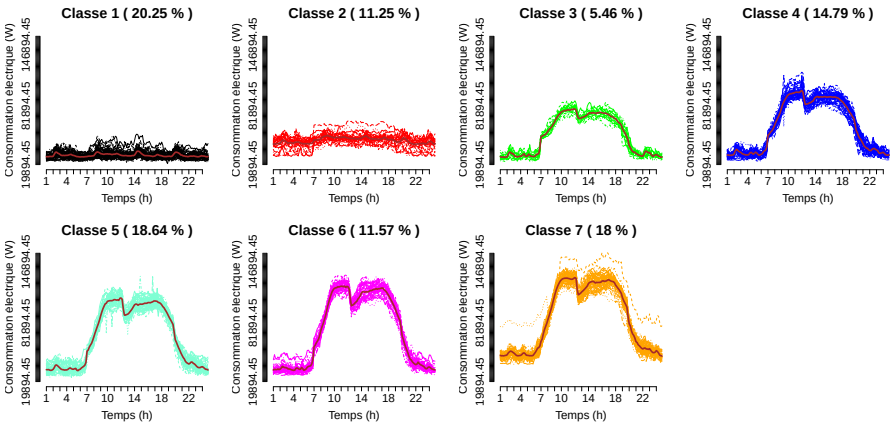
(c)

FIGURE 3.9 – (a) Représentation des classes en fonction des jours de weekend (JW) et jours de semaine (JS) ; (b) Jours travaillés (JT), jours non travaillés (JNT) et jours de vacances scolaires (JVS) ; (c) Jours d'une saison froide (SF), saison intermédiaire (SI) et saison chaude (SC)

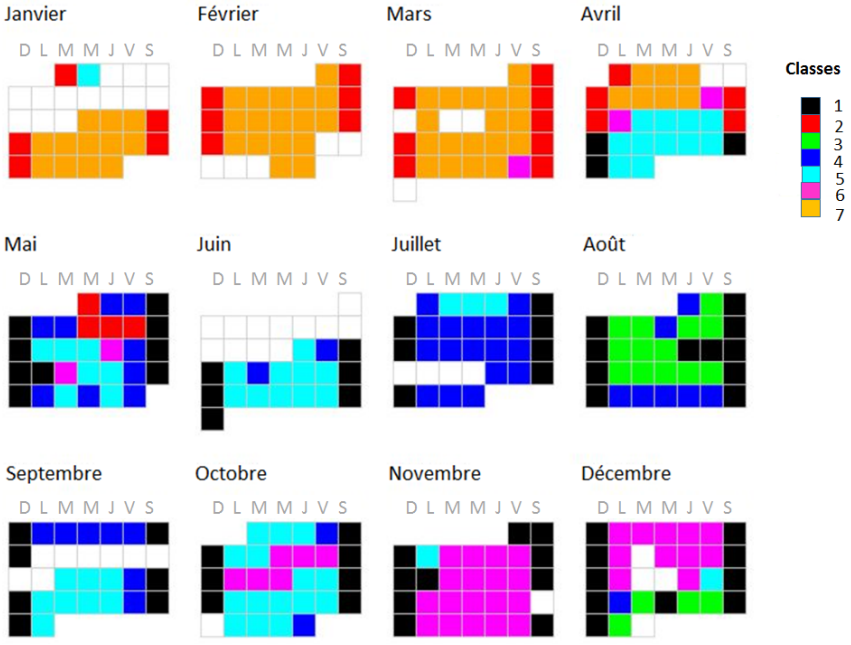
nombre de classes retenu est $K = 6$.

Évaluation des performances du modèle

Pour évaluer les performances du modèle proposé, deux indicateurs ont été utilisés pour mieux caractériser les classes : la divergence de Kullback-Leibler entre les densités des classes et la proportion de chaque classe (voir Table 4.1). La version symétrique de



(a)



(b)

FIGURE 3.10 – (a) Courbes de consommation électrique pour chaque classe en utilisant le modèle de mélange gaussien ; (b) Distribution des jours de chaque classe durant l'année 2013 en utilisant le modèle de mélange gaussien

cette divergence est définie par :

$$KL(\mathcal{N}(\cdot;\boldsymbol{m}_k,\boldsymbol{C}_k),\mathcal{N}(\cdot;\boldsymbol{m}_j,\boldsymbol{C}_j)) = \int_{\mathbb{R}} \mathcal{N}(\mathbf{x}_i;\boldsymbol{m}_k,\boldsymbol{C}_k)\log\frac{\mathcal{N}(\mathbf{x}_i;\boldsymbol{m}_k,\boldsymbol{C}_k)}{\mathcal{N}(\mathbf{x}_i;\boldsymbol{m}_j,\boldsymbol{C}_j)}d\mathbf{x}_i + \int_{\mathbb{R}} \mathcal{N}(\mathbf{x}_i;\boldsymbol{m}_j,\boldsymbol{C}_j)\log\frac{\mathcal{N}(\mathbf{x}_i;\boldsymbol{m}_j,\boldsymbol{C}_j)}{\mathcal{N}(\mathbf{x}_i;\boldsymbol{m}_k,\boldsymbol{C}_k)}d\mathbf{x}_i. \quad (3.7.1)$$

Nous avons choisi d'attribuer une étiquette à chaque classe en fonction de son niveau de consommation moyen : plus l'étiquette de la classe est faible, moins la consommation électrique moyenne est importante. Pour mesurer la taille des classes par rapport à l'ensemble des données, nous avons calculé les proportions de chacune. Nous pouvons noter que les

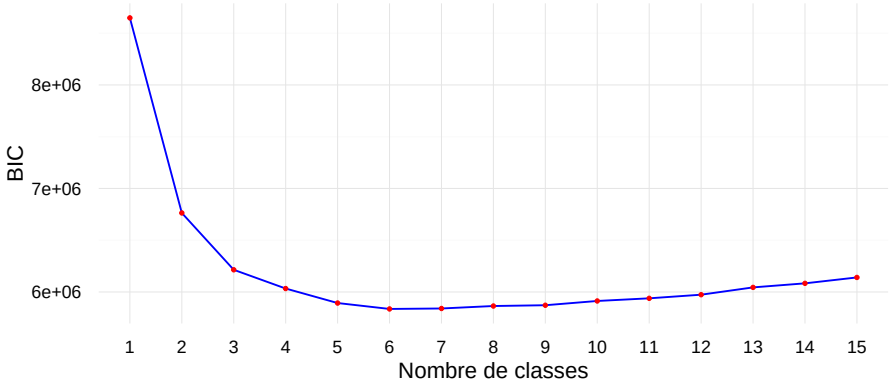


FIGURE 3.11 – Évolution du BIC en fonction du nombre de classes

Classes	1	2	3	4	5	6	Proportions(%)
1	0	-	-	-	-	-	11.36
2	433	0	-	-	-	-	19.07
3	743	**243**	0	-	-	-	20.48
4	1871	449	349	0	-	-	19.47
5	2634	1095	445	**245**	0	-	20.19
6	**6151**	3357	1853	1226	441	0	9.44

TABLE 3.1 – La divergence de Kullback-Leibler entre les densités des classes et les proportions de chaque classes

proportions des classes 2, 3, 4 et 5 sont supérieures à 19 %, ce qui représente la moitié pour les classes 1 et 6. Comme la divergence de Kullback-Leibler mesure la dissimilarité entre les densités des classes, plus la divergence est grande, plus les densités des classes sont éloignées. Les valeurs des densités des classes les plus proches et les plus éloignées sont indiquées en caractères gras dans la Table 4.1. Il est à noter que certaines classes ont des distributions proches comme les classes (2 et 3) et (4 et 5). En revanche, les classes 1 et 6 ont des distributions éloignées.

Une comparaison du modèle proposé avec les algorithmes de classification classiques, à savoir : K-moyennes, Classification Ascendante Hiérarchique (CAH) et Modèle de Mélange Gaussien classique (MMG-classique) a été menée. Celle-ci est basée sur l'inertie intra-classes, le temps de calcul et le nombre de paramètres à estimer pour chaque méthode. L'inertie intra-classe est définie par $I_w = \sum_k I_k$ avec $I_k = \sum_{\mathbf{x}_i \in P_k} ||\mathbf{x}_i - \boldsymbol{m}_k||^2$, où $\boldsymbol{m}_k = (\boldsymbol{m}_{kd})_{d=1,\ldots,D}$ est le centre de la $k^{ième}$ classe. Ce critère permet d'estimer la concentration des séries $\mathbf{x}_i$ autour de leur centre de classe. Plus le critère I_w est petit, moins les points sont dispersés autour de leur centre de classe et plus la classification est bonne. Afin de rendre les résultats comparables pour toutes les méthodes, les centres des classes $\boldsymbol{m}_k$ obtenus avec K-moyennes, CAH et MMG-classique ont été agrégés sous la forme de l'équation 3.4.3, avec $\mu_{kl} = \sum_d \delta_{dl} m_{kd} / \sum_d \delta_{dl}$. La Table 3.2 montre que la méthode proposée améliore de 24% l'inertie intra-classes totale comparée aux méthodes K-moyennes et CAH, et de 10% par rapport au MMG-classique. Les expérimentations ont été menées en utilisant le langage R (version 3.3.2) sur un ordinateur standard avec un processeur Intel (R) Core (TM) i5 CPU @ 1,80 GHz et 8 Go de RAM. Le temps de calcul de la méthode proposée est plus long qu'avec l'algorithme des K-moyennes, mais plus court que ceux de la CAH

et de MMG-classique. Le nombre de paramètres à estimer pour le modèle est inférieur à ceux des K-moyennes et de MMG-classique. le nombre de paramètres du modèle proposé est donné par $\nu_K = K(1 + 2LT) - 1$ alors que pour les K-moyennes $\nu_K = KDT$, et pour MMG-classique $\nu_K = K(1 + 2DT) - 1$, où K est le nombre de classes, L est le nombre de types de jours et T est le nombre de mesures par jour. Concernant la CAH, la notion de paramètres n'est pas définie.

Classes	**Inertie**			
	Modèle proposé	**K-moyennes**	**CAH**	**MMG-classique**
Classe 1	**27224**	437851	52990	35570
Classe 2	**173957**	270131	235645	260631
Classe 3	189603	**173658**	556865	183004
Classe 4	459959	582254	**448206**	547664
Classe 5	456532	**430745**	769036	465702
Classe 6	492079	483594	**309410**	512367
Inertie totale (I_w)	**1809356**	2378233	2372152	2004938
Temps de calcul (sec)	138 ± 34	**7 ± 2**	219 ± 4	154 ± 7
Nombre de paramètres	**1733**	8640	-	17285

TABLE 3.2 – Comparaison entre le modèle proposé, K-moyennes, CAH et MMG-classique par rapport à l'inertie intra-classes, le temps de calcul et le nombre de paramètres

Interprétation des classes

La Figure 3.12 représente les 6 classes obtenues par le modèle proposé. Chaque classe est caractérisé par 3 profils types de consommation électrique relativement aux samedi, dimanche et jour travaillé. Pour une meilleure comparaison entre ces patterns, la Figure 3.13 représente les 6 profils types durant chaque type de jour. Comme mentionné précédemment, nous avons choisi d'attribuer une étiquette à chaque classe en fonction de son niveau de consommation moyen. A première vue, 4 types de profils sont distingués :

- La classe 1 est principalement caractérisée par un profil de consommation électrique très faible et lissé. Aucun pic de consommation n'est observé, et les profils semblent similaires pendant les jours de semaine et les weekend.
- Le premier groupe qui est constitué des classes 2 et 3 est caractérisé par un niveau de consommation relativement faible avec un pic le matin à 8h30 pendant les jours de semaine. Ces pics ne sont pas prononcés, et ils sont suivis par une légère baisse. Cela atteste que les résidents de ces ménages ne se réveillent pas aussi tôt que ceux appartenant au deuxième groupe (décrits ci-dessous), et une minorité d'entre eux quittent la maison pendant la journée. L'heure du déjeuner est également observable par une légère augmentation entre 11h30 et 13h. En soirée, la consommation électrique augmente progressivement de 14h30 à 18h jusqu'à atteindre un pic. Après le pic du soir, la consommation électrique décroit graduellement. Cette baisse est expliquée par une période d'inactivité liée au sommeil. Il est noté que les deux classes diffèrent principalement par leur comportement du soir pour la période de temps entre 18h et minuit (voir Figure 3.14).
- Le deuxième groupe qui englobe les classes 4 et 5 présente un pic de consommation électrique prononcé vers 8h pendant les jours de semaine. Juste après le pic du matin, la consommation électrique diminue de manière significative. Ces observations reflètent que les résidents des ménages concernés se réveillent tôt le matin, et la

majorité d'entre eux partent de la maison après l'heure du pic. L'écart significatif entre la valeur du pic et le niveau de consommation après la baisse est également lié au nombre d'occupants dans le ménage. Pour ces classes, une légère augmentation de la consommation électrique pendant l'heure du déjeuner est observée. Dans la soirée, la consommation électrique augmente jusqu'à atteindre un pic entre 14h et 18h. Cela est expliqué par le retour progressif des résidents. Après le pic du soir, une baisse de consommation électrique est constatée, et qui correspond à la période où les occupants sont susceptibles de dormir. De manière similaire au premier groupe, les comportements du soir sont différents pour les deux classes, pour la période de temps entre 18h et minuit comme le montre la Figure 3.14.

— Le comportement de la classe 6 est assez semblable à celui des classes (4 et 5), mais ce qu'il la caractérise est son niveau de consommation plus élevé.

Une analyse similaire a été menée pour les jours de weekend (voir les Figures 3.13 (Samedi, Dimanche)). Comme attendu, les profils de consommation électrique du weekend diffèrent de ceux de la semaine sauf pour la classe 1. En outre, le niveau de consommation électrique du dimanche est plus élevé que celui du samedi pour la période de temps comprise entre 10h30 et 16h30. En se focalisant sur les classes 5 et 6, la consommation électrique du samedi est moins importante que celle du dimanche entre 19h30 et 23h30. Cela est expliqué par le fait que les occupants appartenant à ces classes sont plus susceptibles d'être absents de la maison le samedi soir que le dimanche soir. Les observations vues précédemment montrent le lien fort entre les profils de consommation électrique et le style de vie quotidien des résidents.

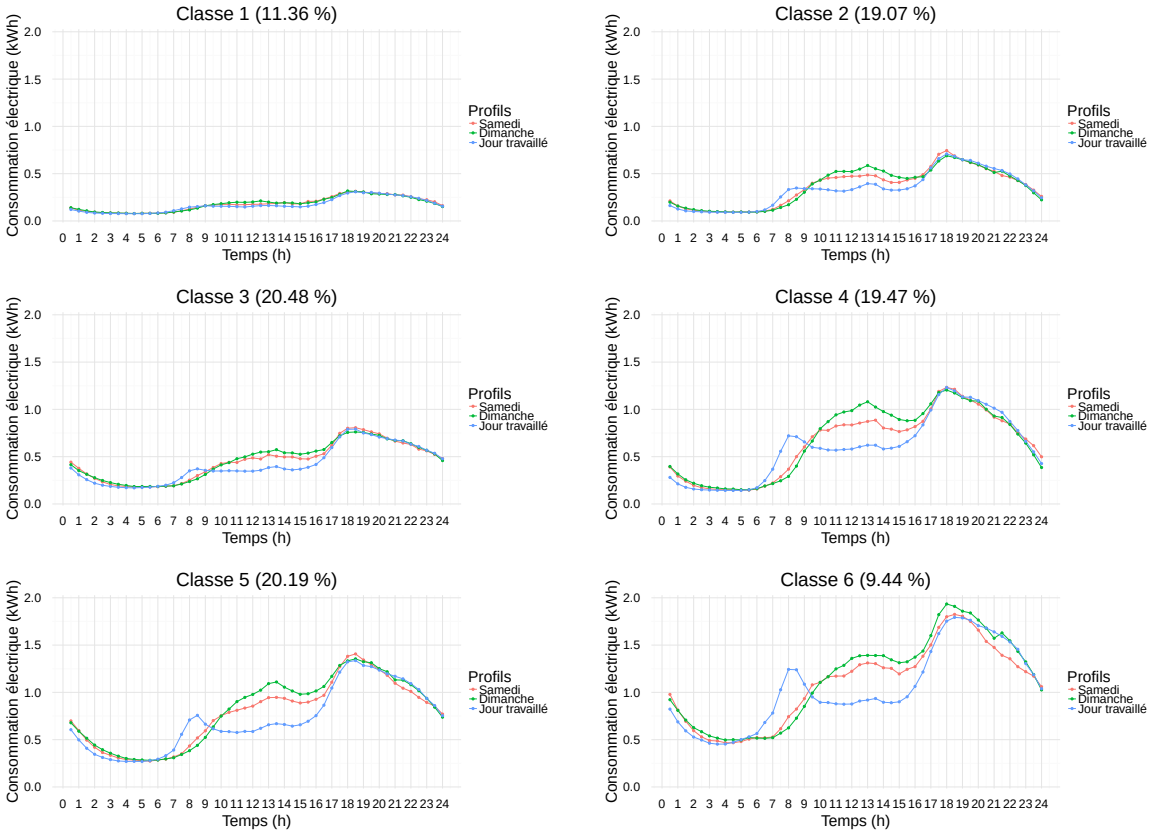


FIGURE 3.12 – Profils types de consommation électrique des 6 classes obtenues par le modèle proposé

Une évaluation quantitative de ces observations est effectuée en croisant les résultats de la classification avec des variables catégorielles extraites du questionnaire résidentiel. Le choix des questions pertinentes a été basé sur des travaux antérieurs, en particulier ceux présentés dans [Tong et al., 2016, Beckel et al., 2014]. Les auteurs de [Beckel et al., 2014] ont sélectionné les caractéristiques qui sont intéressantes pour les services publics en menant des entretiens avec quatre consultants en énergie [Beckel et al., 2012]. La Figure 3.15 représente le croisement des résultats de classification avec huit caractéristiques socio-économiques.

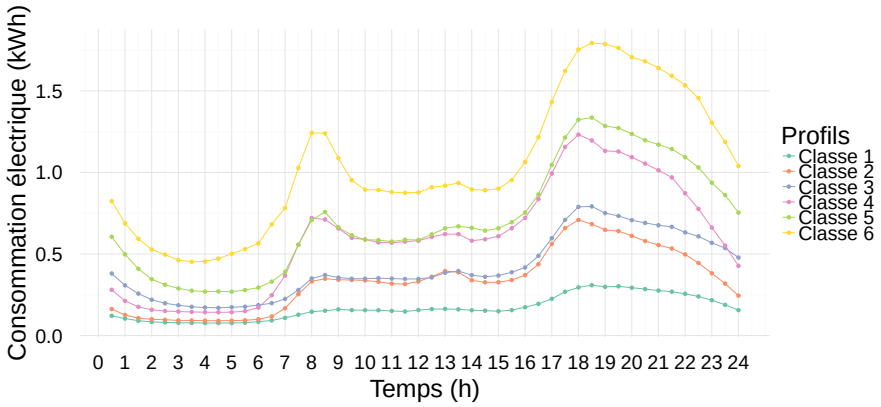
Les résidents sans emploi (Figure 3.15 (a)), ainsi que les résidents seuls (Figure 3.15 (d)), sont moins importants dans les classes qui présentent des profils de consommation électrique avec 3 pics : le matin, à l'heure de déjeuner et le soir (classes 4, 5 et 6). Ces classes sont aussi caractérisées par un pourcentage élevé de résidents appartenant à une classe sociale élevée par rapport aux classes 1, 2 et 3 comme le montre la figure 3.15 (b). Quatre modalités sont utilisées pour cette variable, à savoir cadres supérieurs et professions libérales (AB), employé et profession intermédiaire (C1C2), ouvriers (DE) et les agriculteurs (F). En outre, il est noté que le nombre de retraités est légèrement plus important dans la classe 1, où les pics de consommation électrique sont diffus, et que la différence d'utilisation de l'électricité entre les jours de semaine et les jours de weekend n'est pas significative. On peut noter que la majorité des consommateurs (93%) de toutes les classes utilisent des moyens de chauffage non électriques comme le montre la Figure 3.15 (g). La différence dans les comportements de consommation électrique dépend également de l'âge des résidents, comme illustré dans la Figure 3.15 (e). En particulier, le pourcentage des personnes âgées est plus élevé dans les classes 1 et 2. L'âge a également un impact sur l'utilisation de certaines nouvelles technologies telles que Internet (voir Figure 3.15 (f)). En effet, la majorité des consommateurs dans les classes 4, 5 et 6 utilisent Internet régulièrement. Il est observé que les classes 4, 5 et 6 avec un pic prononcé le matin ont une proportion plus élevée d'employés par rapport aux autres classes, comme le montre la Figure 3.15 (h). Enfin, ces profils de consommation dépendent bien évidemment du nombre de résidents dans chaque ménage, qui a un impact direct sur le nombre d'appareils électriques dans les ménages. Toutes ces observations confirment que les profils de consommation électrique reflètent les modes de vie des citoyens et qui peuvent être associés directement aux caractéristiques socio-économiques des ménages. Les profils types ainsi obtenus peuvent être implémentés dans les modèles de simulation utilisés pour les villes intelligentes.

3.7.2 Classification des données normalisées du mois de novembre

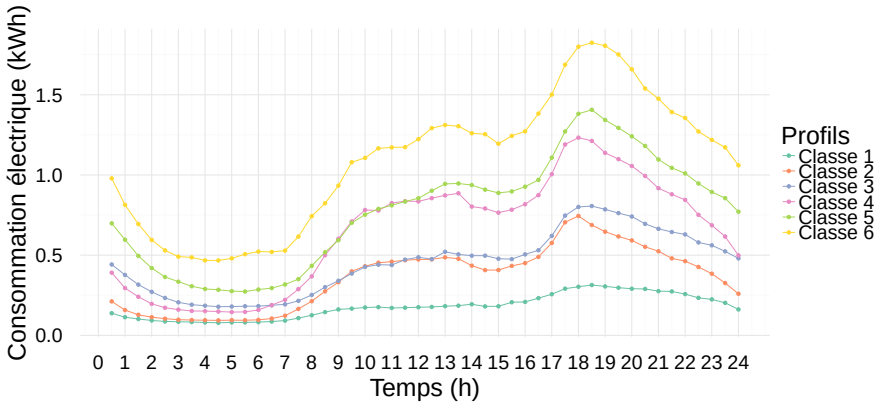
Cette partie s'intéresse à l'impact de la normalisation des données sur la classification des consommateurs. Deux types de normalisations ont été utilisés, à savoir la normalisation centrée réduite, et la normalisation sur une échelle de 0 et 1. Dans cette section, nous considérons qu'un comportement de consommation électrique est défini par la forme de la courbe de consommation et non pas par son niveau. Les résultats de la classification et l'interprétation des classes sont présentés dans les parties qui suivent.

Données normalisées

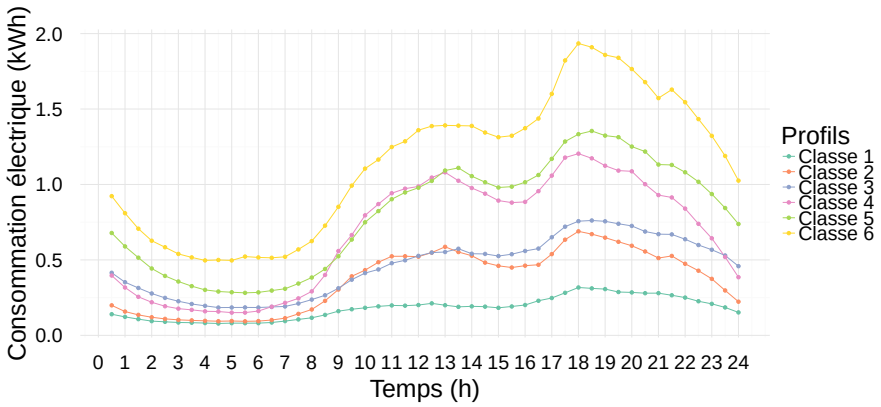
Deux types de normalisations ont été appliqués à la consommation électrique journalière $\boldsymbol{x}_{id} = (x_{id1}, \ldots, x_{idT})$ de chaque consommateur $\boldsymbol{x}_i$. Le premier type est la normalisation centrée réduite. Cette normalisation transforme les valeurs du vecteur $\boldsymbol{x}_{id}$ afin



(a)



(b)



(c)

FIGURE 3.13 – (a) Profils types de consommation électrique durant un jour travaillé ; (b) Samedi et (c) Dimanche

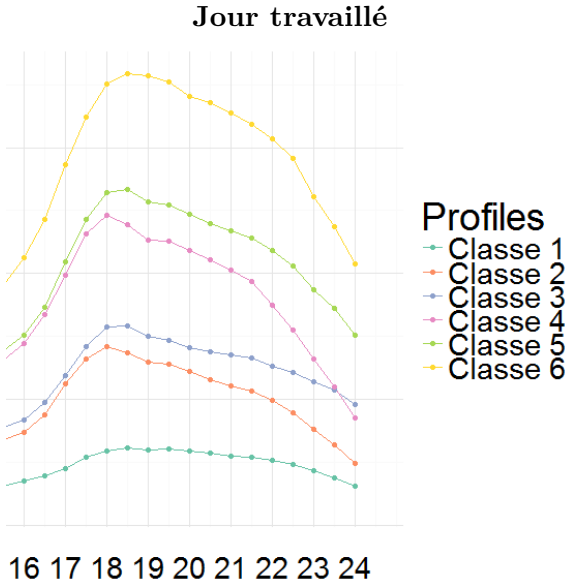


FIGURE 3.14 – Gros plan sur les profils de consommation électrique sans normalisation durant le soir des jours travaillés.

d'obtenir une moyenne et un écart type de la consommation électrique journalière égaux à 0 et 1 respectivement comme indiqué ci-dessous :

$$x^*_{idt} = \frac{x_{idt} - \overline{x}_{id}}{std(x_{id})}, \tag{3.7.2}$$

où x_{idt} et x^*_{idt} représentent respectivement la consommation électrique journalière réelle et normalisée durant un jour d à l'instant t. $\overline{x}_{id}$ et $std(x_{id})$ désignent la moyenne et l'écart type de la consommation électrique journalière $\boldsymbol{x}_{id}$ respectivement.

Le deuxième type de normalisation a été suggéré dans [Wang et al., 2016], où les auteurs ont utilisé le même jeu de données CER. Il consiste à transformer les valeurs du vecteur $\boldsymbol{x}_{id}$ sur une plage de $[0, 1]$ comme suit :

$$x^{'}_{idt} = \frac{x_{idt} - x_{min}}{x_{max} - x_{min}}, \tag{3.7.3}$$

où x_{min} et x_{max} représentent la consommation électrique minimale et maximale pendant un jour d respectivement.

Interprétation des classes

Après avoir appliqué notre approche à des données normalisées avec un nombre de classes fixé à 6, les profils types résultants sont assez similaires à ceux obtenus sans normalisation en terme d'allure de courbe. Dans cette étude, nous ne considérerons que les résultats obtenus à partir de la normalisation centrée réduite. La Figure 3.16 représente les profils de consommation électrique à partir des données normalisées et non normalisées. La quantification de la différence entre les résultats de classification obtenus avec et sans normalisation est affiché dans la Table 3.3. Chaque ligne représente la répartition des consommateurs des classes non normalisées par rapport aux six classes normalisées (A, B, C, D, E et F). Plusieurs tendances remarquables sont observées. Par exemple, plus de 38% de la population dans la classe 1 est présente dans la classe C. Des tendances similaires sont

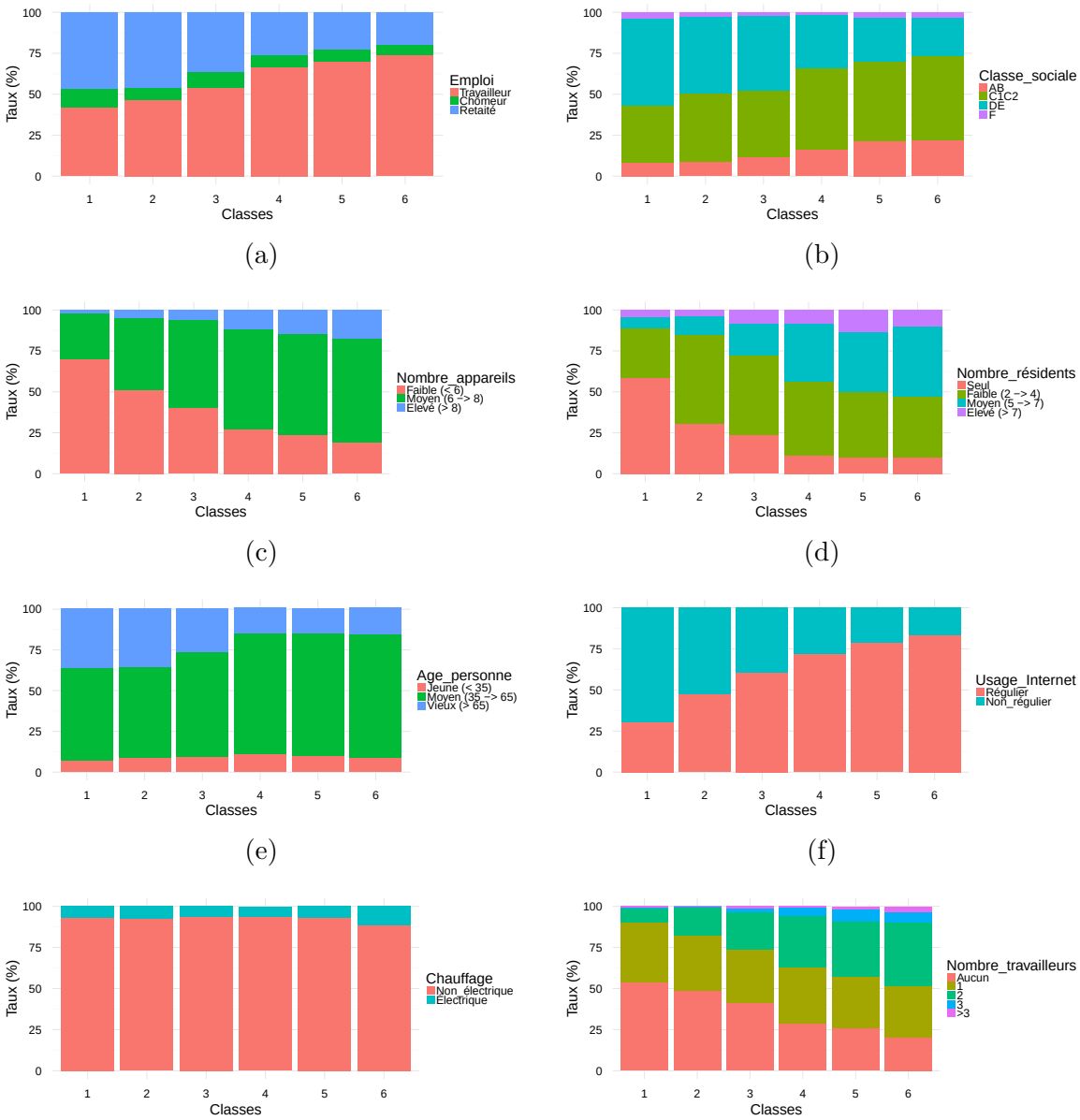


FIGURE 3.15 – (a) Représentation des classes en fonction de l'emploi, (b) classe sociale (AB : cadres supérieurs ; C1C2 : profession intermédiaire ; DE : ouvriers ; F : agriculteurs), (c) nombre d'appareils, (d) Nombre de résidents, (e) age, (f) usage d'Internet, (g) chauffage et (h) nombre de travailleurs

mises en évidence par des pourcentages de transition en caractères gras. L'usage ou non de la normalisation doit dépendre de l'objectif ciblé. La combinaison des résultats avec et sans normalisation va aider à cibler les consommateurs de manière plus précise en se basant sur certaines caractéristiques en terme de formes de pics et de niveaux de consommation (dans le cas de données non normalisées). Cette étude peut fournir des informations pertinentes pour l'application d'une demande d'effacement [Balijepalli et al., 2011].

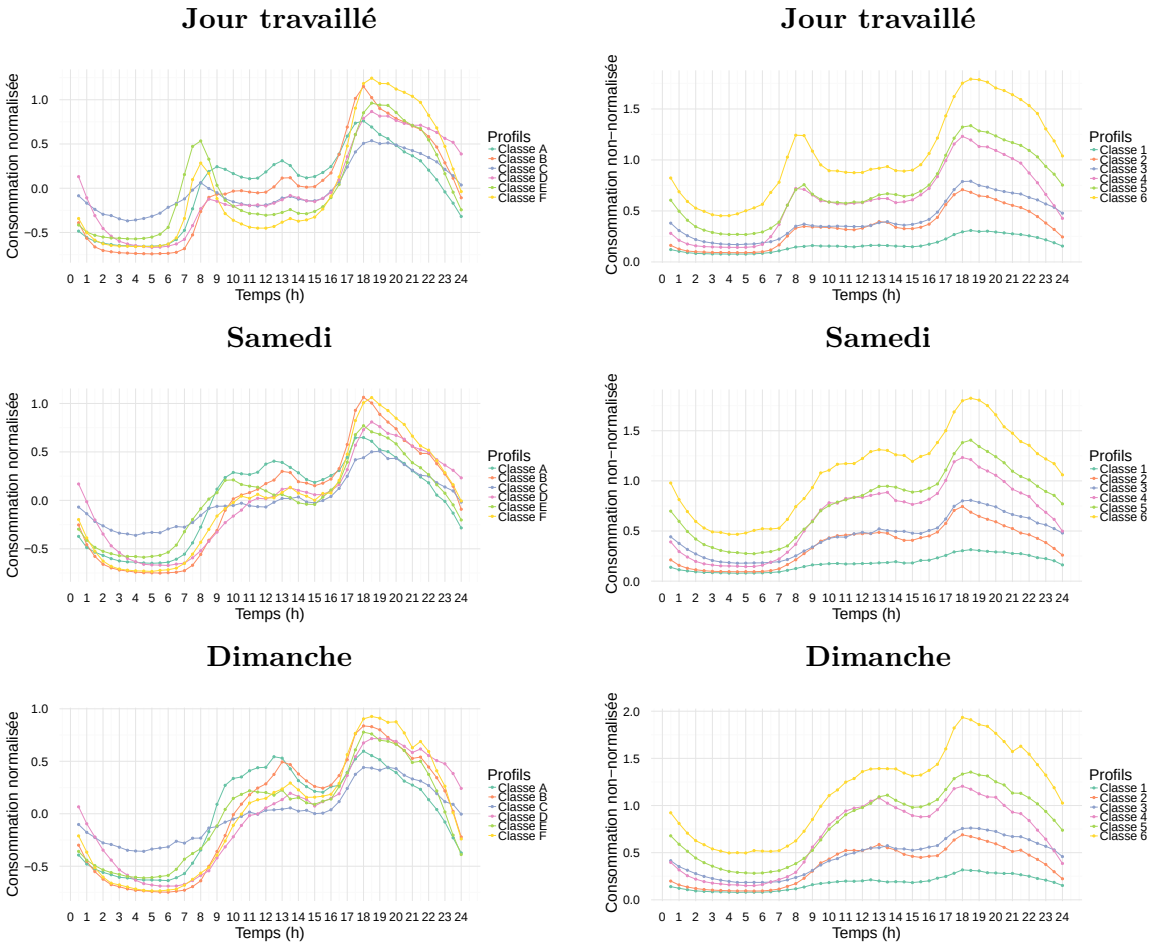


FIGURE 3.16 – Profils de consommation électrique avec et sans normalisation durant les trois types de jour : Samedi, Dimanche et jour travaillé

Classes	A	B	C	D	E	F
1	15.84	9.87	**38.70**	18.70	12.98	3.89
2	**33.63**	**24.28**	7.69	7.54	18.40	8.44
3	9.97	12.64	**21.76**	**39.60**	8.70	7.30
4	**27.17**	**29.54**	3.84	4.28	**20.97**	14.18
5	12.07	15.05	17.18	**35.51**	9.09	11.07
6	10.63	11.55	**37.68**	**20.06**	12.76	7.29

TABLE 3.3 – Table de contingence entre des classes non-normalisées et normalisées

3.7.3 Changement de comportement des habitations résidentielles

Dans cette partie, nous allons nous intéresser à l'évolution des comportements de consommation électrique des consommateurs au fil des mois de l'année 2010. Cette étude vise à mieux comprendre la variabilité de la consommation électrique des usagers au cours de l'année. Ce travail peut être utile pour les fournisseurs d'électricité afin de cibler les consommateurs qui vont participer à une demande d'effacement.

Méthodologie

Pour étudier les changements de comportement résidentiel au fil des mois de l'année 2010, notre approche de classification a été appliquée sur une nouvelle structure de données composée de $N \times M$ séries temporelles $(\mathbf{x}_{im})_{1\leq i\leq N, 1\leq m\leq M}$, où N est le nombre de compteurs intelligents (2870 compteurs intelligents) et M est le nombre de mois (12 mois). Cette formulation permet à un compteur intelligent de changer de classe d'un mois vers un autre, et aussi de différencier les compteurs intelligents en fonction du mois. Dans notre cas, chaque série temporelle $\mathbf{x}_{im}$ représente la consommation électrique d'un consommateur i durant un mois m. La longueur de chaque $\mathbf{x}_{im}$ est de $T \times D_m$, où T est le nombre de mesures par jour (48 mesures) et $D_m \in \{28, 30, 31\}$ est le nombre de jours durant un mois m. Pour adapter notre modèle à ces données, des variables discrètes $\delta_{dm} \in \{1, \ldots, L\}$ qui représentent le type de jour ont été utilisées. Ces variables sont représentées par les variables binaires δ_{dml}, où $\delta_{dml} = 1$ si $\delta_{dm} = l$ et 0 sinon. Les équations décrites dans ce qui suit, représentent les étapes principales pour la construction d'un modèle dédié à la classification des consommations électriques mensuelles des usagers durant une année.

La log-vraisemblance peut s'écrire comme :

$$L(\Theta) = \sum_{m,i} \log\Big(\sum_k \pi_k \prod_d \mathcal{N}(\boldsymbol{x}_{imd}; \sum_l \delta_{dml}\mu_{kl}, \sum_l \delta_{dml}\Sigma_{kl})\Big). \tag{3.7.4}$$

L'étape E de l'algorithme EM consiste à calculer la probabilité à posteriori τ_{imk} qu'un consommateur i durant un mois m appartient à la classe k :

$$\tau_{imk}^{(q)} = \frac{\pi_k^{(q)} \prod_d \mathcal{N}\Big(\boldsymbol{x}_{imd}; \sum_l \delta_{dml}\mu_{kl}^{(q)}, \sum_l \delta_{dml}\Sigma_{kl}^{(q)}\Big)}{\sum_k \pi_k^{(q)} \prod_d \mathcal{N}\Big(\boldsymbol{x}_{imd}; \sum_l \delta_{dml}\mu_{kl}^{(q)}, \sum_l \delta_{dml}\Sigma_{kl}^{(q)}\Big)}. \tag{3.7.5}$$

Pour l'étape maximisation, les paramètres sont calculés comme suit :

$$\pi_k^{(q+1)} = \frac{1}{NM}\sum_{i,m} \tau_{imk}^{(q)}, \tag{3.7.6}$$

$$\mu_{kl}^{(q+1)} = \frac{1}{\sum_{m,i,d} \tau_{imk}^{(q)}\delta_{dml}} \sum_{i,m,d} \tau_{imk}^{(q)}\delta_{dml}\boldsymbol{x}_{imd}, \tag{3.7.7}$$

$$\Sigma_{kl}^{(q+1)} = \frac{1}{\sum_{m,i,d} \tau_{imk}^{(q)}\delta_{dml}} \sum_{m,i,d} \tau_{imk}^{(q)}\delta_{dml}\Big(\boldsymbol{x}_{imd} - \mu_{kl}^{(q+1)}\Big)\Big(\boldsymbol{x}_{imd} - \mu_{kl}^{(q+1)}\Big)^T \tag{3.7.8}$$

Comme la classification est réalisée sur une année, les profils types journaliers obtenus reflètent le comportement moyen des consommateurs au cours de l'année. Pour obtenir une description plus précise, il faudrait ajouter des informations supplémentaires sur les

saisons, les jours fériés et les vacances scolaires. Par conséquent, cela nécessiterait des données couvrant une période plus longue afin d'estimer les paramètres du modèle.

Le nombre de classes a été fixé à $K = 6$ conformément au critère d'information Bayésien (BIC). Comme mentionné précédemment, nous avons choisi d'attribuer une étiquette à chaque classe en fonction de son niveau de consommation moyen : plus l'étiquette de la classe est faible, moins la consommation électrique moyenne est importante. La Figure 3.17 représente les profils de consommation électrique obtenus à partir d'une année de données.

La variabilité du comportement de chaque consommateur au cours des mois de 2010 a été évaluée à l'aide de l'entropie :

$$H(\boldsymbol{\tau}_i) = -\sum_{k=1}^{K}\sum_{m=1}^{M} \tau_{imk}\log(\tau_{imk}), \tag{3.7.9}$$

avec $\boldsymbol{\tau}_i = (\tau_{imk})_{1\leq m\leq M, 1\leq k\leq K}$, où τ_{imk} est la probabilité *a posteriori* qu'un consommateur $i = 1, \ldots, N$ appartienne à la classes $k = 1, \ldots, K$ durant le mois $m = 1, \ldots, M$. Le résultat de cette analyse est utile pour avoir une idée sur l'évolution de la consommation électrique durant une année.

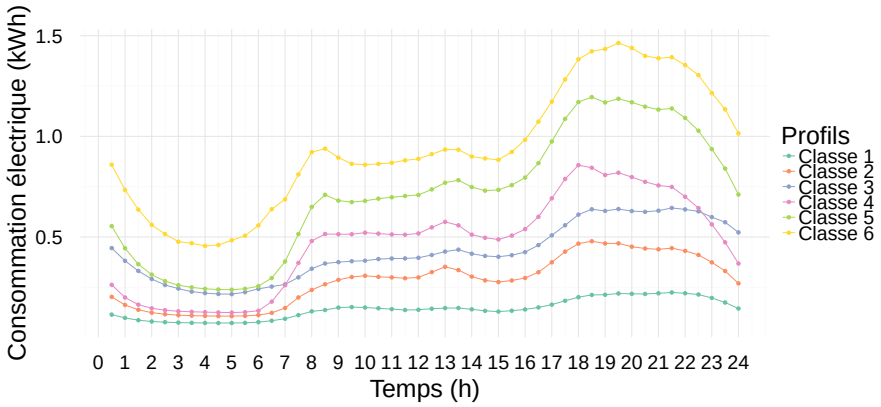
Interprétation

Dans cette partie de l'étude, notre objectif, n'est pas simplement d'identifier les comportements des consommateurs, mais plutôt l'évolution de leurs comportements au cours des mois de l'année 2010.

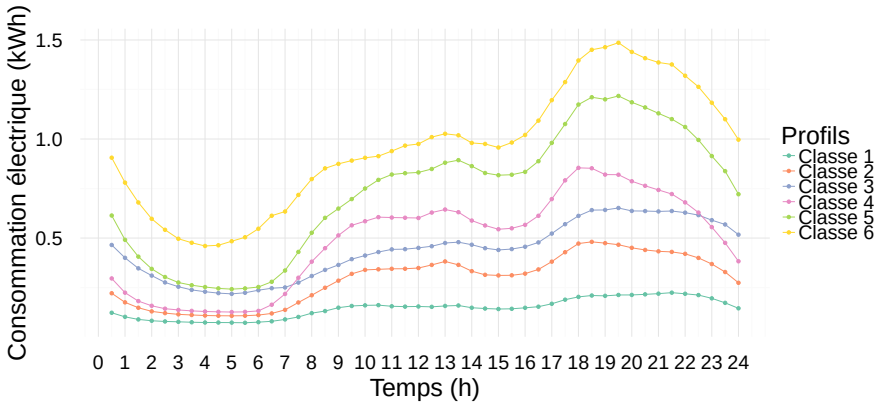
La Figure 3.18 montre la distribution des classes par mois. On peut remarquer que les proportions des classes 1, 2, 3 et 4 augmentent de janvier jusqu'à juin, alors que celles des classes 5 et 6 diminuent. De plus, les proportions de toutes les classes sont stables entre juin et août. Inversement, à partir de septembre jusqu'à décembre, les proportions des classes 1, 2, 3 et 4 diminuent, tandis que les proportions des classes 5 et 6 montrent une augmentation.

Afin de mieux comprendre le changement des proportions des classes au cours de l'année, la Figure 3.19 représente l'évolution des comportements des consommateurs d'une classe vers une autre au fil des mois. Pour rendre la Figure 3.19 plus facile à interpréter, seuls 4 mois de chaque saison sont retenus : janvier (Jan) pour l'hiver, avril (Avr) pour le printemps, juillet (Juil) pour l'été et octobre (Oct) pour l'automne. La migration des consommateurs d'une classe vers une autre est expliquée par le changement de comportement qui dépend fortement de la température et des événements calendaires. Par exemple, si nous nous focalisons uniquement sur la migration des consommateurs de la classe 5 entre janvier et avril (voir Figure 3.20), nous pouvons remarquer que 44,15% des consommateurs se déplacent vers des classes avec des niveaux de consommation électrique plus faibles (classes 1, 2, 3 et 4), plus spécialement les classes 3 et 4 qui ont des profils de consommation très proches à celui de la classe 5. Il est à noter aussi que la majorité des consommateurs de la classe 5 (50.87%) y restent, et seulement 4.97% migrent vers une classe avec un niveau de consommation électrique plus élevé (classe 6).

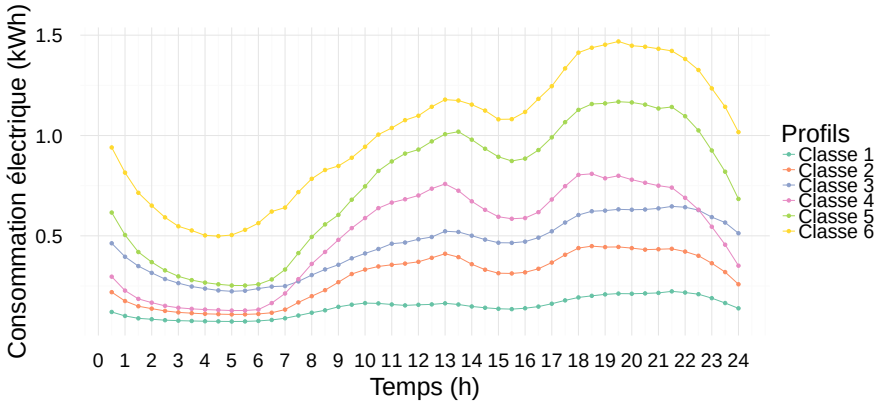
Pour quantifier le déplacement des consommateurs d'une classe vers une autre, les taux de transition globale entre les classes ont été calculés à partir des résultats de classification (voir la Table 3.4). Comme expliqué précédemment, il est constaté que pour la majorité des consommateurs, la probabilité de rester dans la même classe est élevée (voir la diagonale de la Table 3.4). Le déplacement des consommateurs est généralement entre les classes adjacentes. Cela dépend de plusieurs raisons telles que le changement de climat



(a)



(b)



(c)

FIGURE 3.17 – (a) Profils types de consommation électrique des six classes durant un jour travaillé ; (b) Samedi et (c) Dimanche à partir d'une année de données

("mois froid vers mois chaud" ou "mois chaud vers mois chaud"). L'histogramme qui décrit empiriquement la distribution des entropies $H(\boldsymbol{\tau}_1), \ldots, H(\boldsymbol{\tau}_N)$ est affiché dans la Figure 3.21. Trois exemples d'entropies : faible, moyenne et élevée sont également représentées dans la Figure 3.22. Une entropie proche de 0 représente des consommateurs ayant une consommation régulière. Tandis que l'entropie d'une grande proportion de consommateurs se situe autour de 0,9. Ce type d'indicateur peut également aider les fournisseurs d'énergie à identifier les ménages qui peuvent être impliqués dans une procédure de demande d'effacement.

Classes	**1**	**2**	**3**	**4**	**5**	**6**
1	**77.21**	**18.03**	3.16	0.72	0.24	0.63
2	8.58	**69.97**	9.51	**10.71**	0.73	0.47
3	1.71	**11.27**	**65.13**	8.72	**10.26**	2.89
4	0.32	**10.85**	7.56	**67.94**	**11.91**	1.39
5	0.11	0.43	8.45	**10.72**	**69.83**	**10.44**
6	0.46	0.65	4.98	2.17	**20.54**	**71.17**

TABLE 3.4 – Table de probabilités de transition entre les classes, d'un mois vers un autre

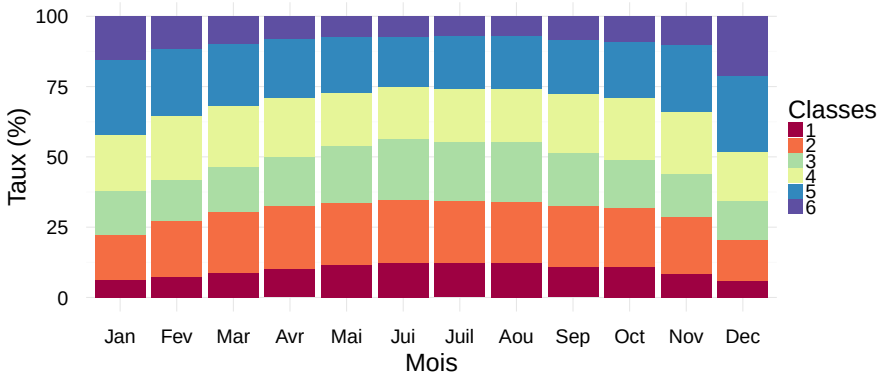


FIGURE 3.18 – Distribution des classes par mois

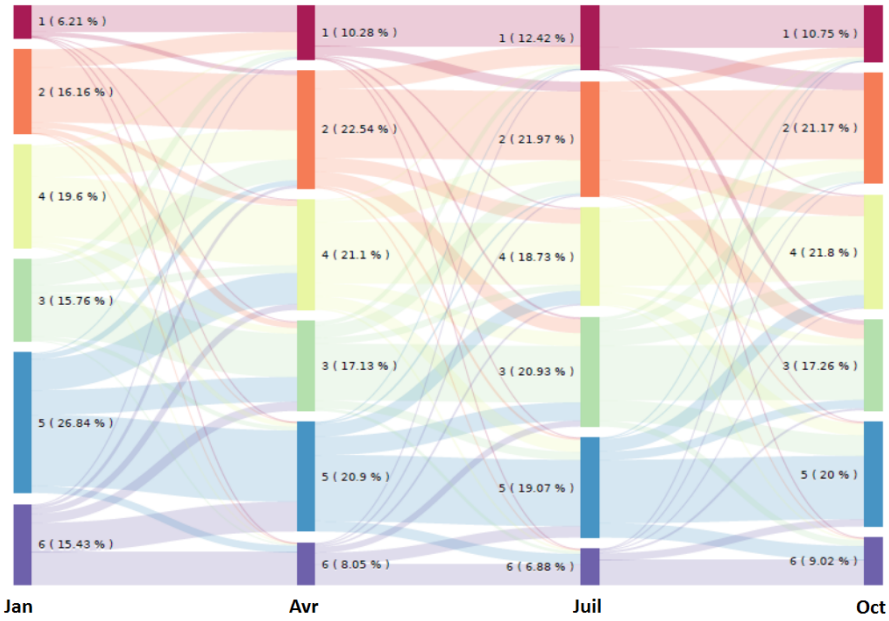


FIGURE 3.19 – Évolution des comportements des consommateurs au cours des mois de l'année

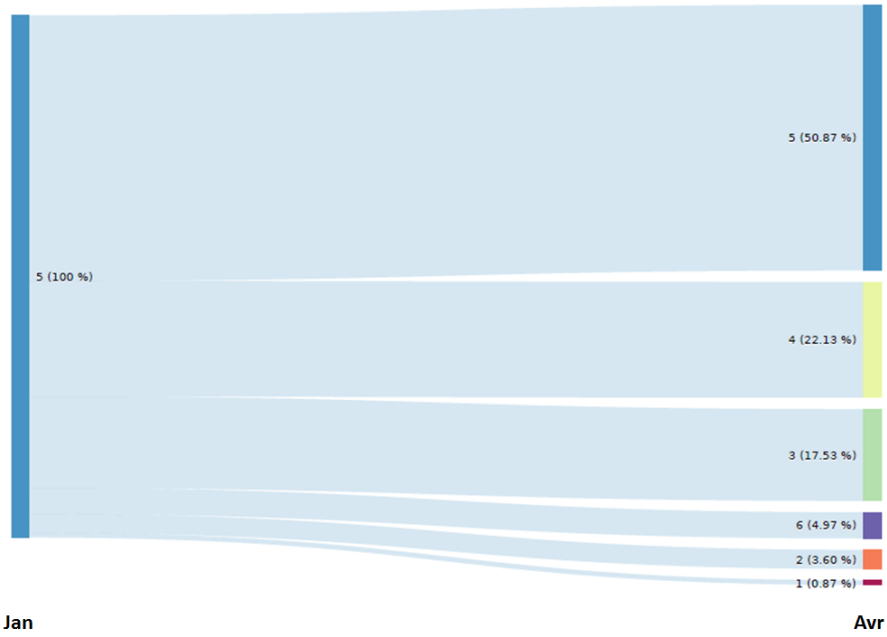


FIGURE 3.20 – Évolution des comportements des consommateurs de la classe 5 entre janvier et avril

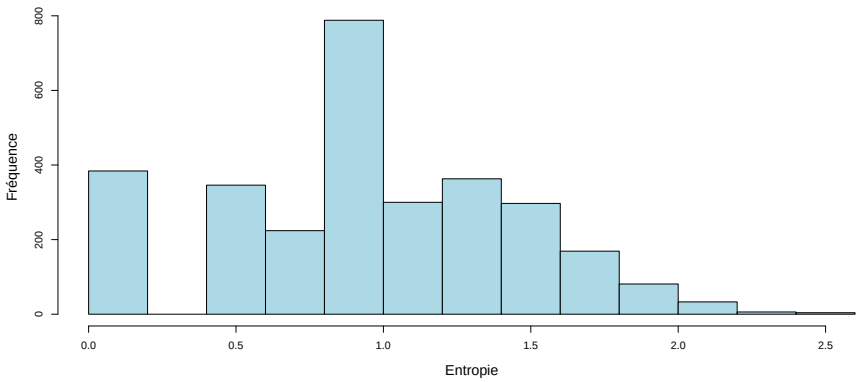


FIGURE 3.21 – Histogramme des entropies

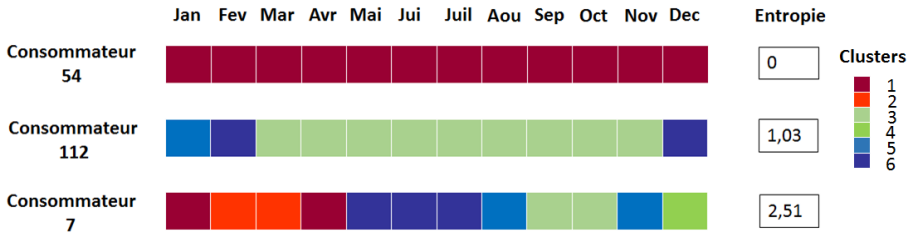


FIGURE 3.22 – Évolution des comportements de consommation électrique mensuel de trois consommateurs durant l'année 2010

3.8 Conclusion

Dans ce chapitre, nous nous somme intéressés à la classification automatique des comportements de consommation électrique à deux échelles : bâtiment et territoire. Concernant le bâtiment, une identification des profils types de consommation électrique journaliers a été menée en utilisant deux méthodes de classification, à savoir l'algorithme des K-moyennes fonctionnel et le modèle de mélange gaussien. Sept classes ont été identifiées pour les deux méthodes. Chaque classe est composée d'un sous groupe de courbes de consommation électrique journalières, représentées par un profil type. Il a été constaté que ces profils types dépendent fortement de la température et les types de jour (jour travaillé, jour non travaillé et vacances scolaires). Des résultats similaires ont été obtenus pour les deux méthodes, sauf pour quelques jours non travaillés.

A l'échelle d'un territoire, une approche de classification automatique a été développée pour identifier les profils types de consommation électrique. L'objectif était d'analyser les comportements de consommation électrique des usagers résidentiels irlandais. Un modèle génératif basé sur un modèle de mélange gaussien spécifique a été développé. Ce modèle intègre le type de jour (samedi, dimanche et jour travaillé) comme variable exogène. Chaque classe obtenue est caractérisée par trois profils types relativement aux types de jours. Six classes ont ainsi été identifiées à partir des données non normalisées durant le mois de novembre, et une analyse approfondie a été menée sur les profils types obtenus. Les comportements résidentiels dépendent principalement des caractéristiques socio-économiques des ménages ainsi que leurs temps d'occupation pendant la journée. Une analyse de l'évolution des consommations durant une année a permis de voir que les comportements des

consommateurs évoluent au fil du temps en fonction de la température et des événements calendaires. Les travaux présentés dans cette partie représentent une première étape vers un cadre plus général basé sur des modèles génératifs. Les travaux futurs devraient tenir compte de certaines considérations supplémentaires pour prolonger cette étude. Il serait intéressant d'incorporer d'autres variables spécifiques à des jours particuliers. Une façon d'y parvenir serait d'utiliser des informations plus précises sur le calendrier. Comme on peut s'y attendre, la consommation électrique dépend de plusieurs facteurs comme les jours de la semaine (lundi,..., dimanche), les jours fériés, les ponts et les vacances scolaires. Ces informations calendaires pourraient être fixées manuellement ou extraites à partir d'une classification préliminaire sur les jours de l'année. En terme de modélisation, cela pourrait facilement être incorporé dans le modèle en augmentant le nombre de types de jours encodés par la variable observée δ_d. Cependant, pour une meilleure précision, ce type d'approche nécessite des données collectées sur une longue période (plus d'un an). De la même manière, il serait également motivant d'étudier l'intégration des saisons dans le modèle. Ceci permet de considérer le comportement périodique des séries temporelles.

Les travaux de fouille de données détaillés dans ce chapitre peuvent servir les fournisseurs d'énergie afin de développer de nouvelles politiques de tarification, améliorer les performances des modèles de prévision et identifier les cibles potentielles pour une procédure de demande d'effacement. Une meilleure compréhension des comportements de consommation d'usagers peut être également bénéfique pour la modélisation urbaine, en fournissant aux modèles de simulation des profils de consommation précis. Cette compréhension détaillée est essentielle pour la construction des futures villes intelligentes.

Chapitre 4

Prévision de l'irradiance solaire à court et moyen termes

4.1 Introduction

Dans ce chapitre, nous nous intéressons à l'énergie renouvelable et plus précisément à la prévision de la production électrique générée par les panneaux photovoltaïques (PVs). L'intégration d'un nombre important de PVs dans un réseau électrique pose un défi technique en raison de la nature variable de la ressource solaire. Une meilleure prévision de l'irradiance solaire permet aux opérateurs du réseau de mieux planifier l'utilisation de l'énergie solaire. La première partie de ce chapitre est consacrée à l'état de l'art des différentes méthodes utilisées dans la prévision de l'irradiance solaire en se focalisant sur les approches statistiques. Par la suite, nous évoquons les méthodes de prévision utilisées. Après, nous présentons la mise en œuvre des différentes approches de prévision de l'irradiance solaire sur deux horizons temporels : à court terme (horaire) et à moyen terme (journalière). Dans la dernière partie du chapitre, nous proposons un modèle hybride qui tire bénéfices des performances de chaque méthode afin d'améliorer la précision de prévision. Les expérimentations seront menées sur quatre climats différents permettant ainsi une large évaluation des approches proposées.

4.2 État de l'art

La prévision de l'irradiance solaire est un élément clé dans la majorité des systèmes de prévision de production photovoltaïque. De nombreux travaux de recherche sont dédiés à cette problématique où l'on peut catégoriser les modèles de prévision de l'irradiance en deux groupes : les modèles physiques et les modèles statistiques.

Les modèles physiques sont basés sur des équations mathématiques qui décrivent l'état physique et le mouvement dynamique de l'atmosphère. Ces modèles sont basés sur des équations non-linéaires complexes dont la résolution nécessite une forte puissance de calcul. Des méthodes numériques ont été utilisées pour obtenir des solutions approximatives à ces équations. Ces méthodes sont nommées modèles numériques de prévision météorologique *"Numerical Weather Predictions (NWP)"*. Les erreurs de prévision de l'irradiance solaire basées sur les modèles (NWP) varient en fonction des climats et du mouvement dynamique de l'atmosphère pour une localisation donnée [Wittmann et al., 2008, Ciabattoni et al., 2013].

Les modèles statistiques se basent sur des données historiques qui peuvent être is-

sues de stations météorologiques, de satellites [Hammer et al., 1999] ou des images du ciel [Nova et al., 2005]. Les modèles statistiques basés sur des données satellites et des images du ciel détectent les mouvement des nuages à l'aide du champ de vecteurs de mouvement *"motion vector fields"*. Les erreurs de prévision issues de ce type de modèles augmentent dans le cas d'un faible ensoleillement et pour de fortes variations spatiales [Hammer et al., 1999].

Les modèles statistiques sont moins complexes comparés aux modèles physiques. Cela revient au fait qu'ils nécessitent une quantité de données moins importante ainsi qu'un temps de calcul plus court. Concernant la qualité de prévision, les modèles physiques fournissent une bonne estimation de l'irradiance solaire sur un horizon de plus d'un jour, alors que pour les horizons à court et moyen termes (heure et jour) où l'influence du déplacement des nuages est importante, les modèles statistiques sont plus performants [Kleissl, 2010].

Plusieurs travaux de recherche se sont basés sur les approches statistiques pour prévoir l'irradiance solaire. Elles vont des méthodes d'apprentissage automatique comme les réseaux de neurones (NN) et les machines à vecteurs de support (SVM) jusqu'aux méthodes de séries temporelles comme les modèles auto-régressifs (AR). Dans la suite, nous décrivons les principaux travaux de recherche qui ont utilisé les méthodes citées.

Les réseaux de neurones (NN) sont considérés comme des méthodes de référence pour résoudre des problèmes de régression non linéaire rencontrés dans divers domaines. C'est la raison pour laquelle leur utilisation dans le domaine de la prévision solaire est fréquente [Mellit and Pavan, 2010, Deng et al., 2010, Rao et al., 2012, Ghanbarzadeh et al., 2009].

Mellit et *al* ont étudié le perceptron multicouche (MLP) pour prévoir l'irradiance solaire sur un horizon de 24 heures en Italie [Mellit and Pavan, 2010]. Les données d'entrée étaient l'irradiance et la température journalières moyennes. Les résultats ont montré que le meilleur modèle a été obtenu avec une couche d'entrée et deux couches cachées. Deng et al ont utilisé un réseau de neurones avec rétro-propagation dans [Deng et al., 2010] pour prévoir l'irradiance solaire journalière en Chine. Le modèle proposé comporte trois couches cachées et l'algorithme d'optimisation utilisé est celui de Levenberg-Marquardt. Les entrées du réseau de neurones sont constituées des paramètres météorologiques et géographiques. Les auteurs ont constaté que la durée d'ensoleillement, les paramètres géographiques et le jour de l'année sont les données d'entrée les plus pertinentes. Dans [Rao et al., 2012], les auteurs ont trouvé que le modèle à base de (NN) était performant quand la température, l'humidité, le mois et le jour étaient utilisés en entrée du réseau de neurones, et moins performant quand la date était rajoutée. Dans [Ghanbarzadeh et al., 2009] une combinaison de paramètres météorologiques comme entrées du réseau de neurones a été exploitée pour prévoir l'irradiance solaire. Il a été constaté que les meilleurs performances ont été obtenues lorsque la durée d'ensoleillement, la température moyenne journalière et l'humidité sont intégrées.

L'utilisation des machines à vecteurs de support (SVM) est relativement récente dans le domaine de la prévision de l'irradiance solaire [Zeng and Qiao, 2013], [Wolff et al., 2016], [da Silva Fonseca Jr. et al., 2013].

Les machines à vecteurs de support ont été utilisées dans [Zeng and Qiao, 2013] pour une prévision à court terme. L'entrée du modèle comporte des données historiques sur la transmissivité atmosphérique ainsi que des variables météorologiques comprenant la couverture du ciel, l'humidité et la vitesse du vent. La sortie du modèle est la transmission atmosphérique prévue, qui est convertie par la suite en irradiance solaire selon la latitude du site et l'heure du jour. Les résultats montrent que le modèle proposé surpasse le modèle auto-régressif (AR) ainsi que les réseaux de neurones. Dans [Wolff et al., 2016], les

machines à vecteurs de support et les K plus proches voisins ont été mis en oeuvre pour prévoir la production photovoltaïque. Les chercheurs se sont basés sur deux types de données à savoir : des données météorologiques mesurées et des prévisions issues des modèles (NWP). Après avoir optimisé les paramètres et comparé les deux modèles, Wolff et al. ont construit un modèle hybride qui utilise les prévisions des deux approches employées. Joao et al. ont appliqué les SVM avec une fonction gaussienne comme noyau, pour prévoir l'irradiance solaire au niveau de la ville de Tsukuba (Japon) [da Silva Fonseca Jr. et al., 2013]. L'ajustement des paramètres du modèle a été effectué en utilisant les K validations croisées et la recherche par grille. L'analyse a nécessité l'usage de données météorologiques.

Concernant les méthodes de séries temporelles, les modèles auto-régressifs (AR) et les modèles auto-régressifs et moyenne mobile (ARMA) sont efficaces pour prévoir les futures valeurs des séries temporelles auto-corrélées. Ces modèles nécessitent que la condition de stationnarité soit satisfaite. Dans [Ji and Chee, 2011], le test augmenté de Dickey-Fuller (ADF) [Said and Dickey, 1984] a été utilisé pour mesurer la stationnarité des séries temporelles. Comme les séries temporelles d'irradiance solaire ne sont pas stationnaires, un pré-traitement est nécessaire. Lauret et al. [Lauret et al., 2015] ont proposé une analyse comparative de plusieurs méthodes d'apprentissage automatique avec un modèle (AR). Pour rendre les séries de l'irradiance solaire stationnaires, ils ont utilisé le modèle Bird [Bird and Hulstrom, 1981] pour les normaliser. Récemment, David et al. [David et al., 2016] ont utilisé le même pré-traitement pour prévoir l'irradiance solaire avec des modèles récursifs ARMA-GARCH.

Lauret et al. [Lauret et al., 2015] ont effectué une étude comparative de plusieurs approches statistiques à savoir : une méthode naïve, un modèle auto-régressif, un modèle gaussien, les machines à vecteurs de support et un réseau de neurones. Ces modèles sont calibrés et validés en utilisant uniquement des données d'irradiance solaire issues de trois îles françaises : la Corse, la Guadeloupe et la Réunion. Les auteurs ont trouvé que les méthodes d'apprentissage automatique améliorent légèrement les performances des modèles (AR) et celles des modèles naïfs dans le cas d'une prévision horaire. Par contre, cette amélioration est plus significative dans le cas des conditions instables du ciel.

En s'inscrivant dans le même contexte que les travaux cités précédemment, notre travail consiste à prévoir l'irradiance solaire sur deux horizons temporels (court et moyen termes). Six approches statistiques sont utilisées à savoir : la méthode naïve, la méthode calendaire, le modèle ARMA intégrant des variables exogènes (ARMAX), les machines à vecteurs de support (SVM), les forêts aléatoires (RF) et les réseaux de neurones (NN). Pour construire et valider les modèles, deux types de données sont disponibles : des paramètres météorologiques mesurés et des prévisions issues des modèles NWP. Ces données sont relatives à quatre localisations caractérisées chacune par un type de climat bien spécifique. La particularité de ce travail est l'étude des performances d'un ensemble élargi de méthodes incluant des méthodes références, une méthode de séries temporelles et des méthodes d'apprentissage automatique. Ces performances sont mesurées sur deux horizons temporels pour quatre climats différents. Ce travail donnera également des indications utiles pour la mise en place d'un modèle hybride de prévision plus performant.

4.3 Modèles de prévision

Dans cette section, nous allons nous intéresser aux modèles de prévision. Ces modèles visent à prévoir une situation future en se basant sur des données historiques. Comme cela a été déjà mentionné, les approches de prévision de l'irradiance solaire peuvent être scindées en deux groupes : les approches basées sur un modèle physique et les approches

statistiques.

Dans ce chapitre, nous nous intéressons à six méthodes de prévision à savoir :

— la méthode naïve,
— la méthode calendaire,
— le modèle ARMA intégrant des variables exogènes (ARMAX),
— les machines à vecteurs de support (SVM),
— les forêts aléatoires (RF),
— les réseaux de neurones (NN).

Les données historiques de l'irradiance solaire sont représentées par une série temporelle $(I_1, I_2, ..., I_N)$, où I_t correspond à l'irradiance solaire à une heure t donnée. Le signe$\widehat{}$sera utilisé pour désigner la prévision de I_t et h indiquera l'horizon de prévision.

4.3.1 Modèles références

Méthode naïve

La méthode naïve est une méthode de prévision très simple qui est souvent utilisée comme référence pour évaluer les performances d'autres méthodes. Elle suppose que les valeurs à prévoir sont égales à celles observées (c.-à-d. les conditions météorologiques restent inchangées entre les heures observées et les heures futures). La prévision de l'irradiance solaire sur un horizon h est donnée par :

$$\widehat{I}_{t+h} = \begin{cases} I_t & \text{Si } h = 1, \\ I_{t+h-24} & \text{Si } h > 1 \text{ et } t \geq 24. \end{cases} \tag{4.3.1}$$

Remarque : Pour une prévision de l'irradiance solaire sur un horizon $h > 1$, il est nécessaire d'avoir au minimum une série d'un jour (24 heures).

Méthode calendaire

La méthode calendaire est une méthode basée sur des informations issues du calendrier. Comme vu précédemment dans l'analyse exploratoire (voir chapitre 2), l'année est divisée en G groupes de jours $(P_1, P_2, ..., P_G)$ selon leurs nombres d'heures d'ensoleillement. Connaissant le groupe P_g du jour de l'instant $t + h$ à prévoir, l'irradiance solaire à cet instant est égale à la moyenne de toutes les irradiances I_l des jours appartenant au groupe P_g. La formule est comme suit :

$$\widehat{I}_{t+h} = \frac{\sum_{I \in P_g} I_l}{\#P_g}, \tag{4.3.2}$$

telle que $l = (t + h) \mod 24$ et $\#$ représente le cardinal du groupe P_g.

4.3.2 Modèles de séries temporelles et d'apprentissage automatique

Taux d'irradiance par rapport à un ciel clair

Pour respecter la condition de stationnarité de certaines méthodes comme ARMAX, et afin d'améliorer les performances d'autres telles que SVM, RF et NN, nous avons choisi d'utiliser un paramètre clé qui est le taux d'irradiance par rapport à un ciel clair. Ce paramètre est déterminé en divisant l'irradiance mesurée par l'irradiance sous un ciel clair. La formule est donnée par :

$$\tau_t = \begin{cases} I_t/I_t(clair) & \text{Si } I_t(clair) \neq 0 \\ 0 & \text{Sinon,} \end{cases} \tag{4.3.3}$$

où τ_t et $I_t(clair)$ représentent respectivement le taux d'irradiance et l'irradiance par rapport à un ciel clair. I_t correspond à l'irradiance solaire mesurée. $I_t(clair)$ est déterminée à partir d'un modèle de ciel clair appelé modèle Bird [Bird and Hulstrom, 1981]. Ce modèle génère l'irradiance sous un ciel dégagé avec un précision acceptable et avec peu de données d'entrées [Badescu et al., 2013].

Méthodologie

Pour prévoir l'irradiance solaire sur un horizon h, nous nous sommes basés sur les quatre étapes suivantes :

— Calcul du taux d'irradiance par rapport à un ciel clair τ_t en utilisant l'équation (4.3.3) ;
— Construction du modèle de prévision sur le taux d'irradiance ;
— Prévision du taux d'irradiance $\widehat{\tau}_{t+h}$;
— Prévision de l'irradiance $\widehat{I}_{t+h}$, en multipliant le taux d'irradiance prévu $\widehat{\tau}_{t+h}$ par l'irradiance sous un ciel clair $I_{t+h}(clair)$.

La formule de prévision adoptée est la suivante :

$$\widehat{I}_{t+h} = \widehat{\tau}_{t+h} \times I_{t+h}(clair). \tag{4.3.4}$$

Désignons l'ensemble d'apprentissage par $A = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^{N_A}$, où N_A est la taille de l'ensemble et $\boldsymbol{x}_t \in \mathbb{R}^v$ est une donnée d'entrée contenant v attributs et $y_t \in \mathbb{R}$ est sa donnée de sortie. Les données d'entrée peuvent être représentées par une matrice $\mathbf{X}$ de dimension $N_A \times v$ qui correspond à un vecteur de sortie $\boldsymbol{y}$. L'ensemble d'apprentissage peut s'écrire $A = \{\mathbf{X}, \boldsymbol{y}\}$.

Dans cette thèse, nous avons adapté le type d'attributs ainsi que la taille de la donnée d'entrée $\boldsymbol{x}_t$ selon l'horizon de prévision (court et moyen termes) :

— Pour un horizon de prévision court (horaire) où l'objectif est d'estimer y_{t+1}, chaque donnée d'entrée $\boldsymbol{x}_t$ comporte des paramètres à l'heure actuelle t (taux d'irradiance et météo observée) et des paramètres à l'heure prochaine $t+1$ (météo prévue). En plus des paramètres météo, une variable binaire qui encode l'ensoleillement à l'heure $t+1$ (1 durant le jour et 0 sinon) a été intégrée. Dans ce cas, la taille de la donnée d'entrée $\boldsymbol{x}_t$ est $v = 11$ attributs (voir Table 4.1). Concernant les données de sortie, chaque y_t représente le taux d'irradiance à l'heure $t+1$.
— Pour un horizon de prévision moyen (journalier) où l'objectif est d'estimer $y_{t+1}, y_{t+2}, ..., y_{t+24}$, la donnée d'entrée est constituée de paramètres météo prévus et d'ensoleillement correspondant à l'heure t du prochain jour. Pour le moyen terme, la donnée d'entrée $\boldsymbol{x}_t$ est représentée par un vecteur de taille $v = 6$ attributs (voir Table 4.1). Pour les données de sortie, chaque y_t représente le taux d'irradiance à l'heure t du jour à prévoir.

Pour évaluer les modèles de prévision construits, nous considérons $V = \{\mathbf{X}_*, \boldsymbol{y}_*\}$ un ensemble de N_V données de validation.

Type d'attribut	Court terme	Moyen terme
Paramètres météo observés	Température (t)	-
	Pression (t)	-
	Humidité (t)	-
	Vitesse du vent (t)	-
Paramètres météo prévus	Température (t+1)	Température (t)
	Pression (t+1)	Pression (t)
	Humidité (t+1)	Humidité (t)
	Vitesse du vent (t+1)	Vitesse du vent (t)
	Couverture nuageuse (t+1)	Couverture nuageuse (t)
Paramètres déterminés	Ensoleillement (t+1)	Ensoleillement (t)
	Taux d'irradiance (t)	-

TABLE 4.1 – Attributs de la donnée d'entrée $\boldsymbol{x}_t$ utilisés dans les prévisions court et moyen termes

Modèle auto-régressif et moyenne mobile intégrant des variables exogènes (ARMAX)

Les modèles auto-régressifs et moyenne mobile (ARMA) sont des modèles conçus principalement pour comprendre et prévoir des valeurs futures d'une série temporelle [Box and Jenkins, 1994]. Le modèle ARMA est divisé en une partie "auto-régressive" (AR) et une partie " moyenne mobile". Le modèle ARMA est généralement noté ARMA(p,q) où p et q représentent les ordres des parties (AR) et (MA) respectivement. Dans le cas du taux d'irradiance solaire, le modèle ARMA(p,q) s'écrit comme suit :

$$\tau_t = \sum_{i=1}^{p} \phi_i \tau_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{4.3.5}$$

Pour une meilleure précision du modèle ARMA, nous avons rajouté v variables exogènes qui dépendent de l'horizon de prévision (voir Table 4.1). L'extension du modèle ARMA en incluant ces variables exogènes est nommée ARMAX(p,q). Elle est définie par :

$$\tau_t = \sum_{i=1}^{p} \phi_i \tau_{t-i} + \sum_{r=1}^{v} \beta_r \boldsymbol{x}_{tr} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \tag{4.3.6}$$

où β_r est le coefficient de la variable exogène r et $(\boldsymbol{x}_{tr})_{r=1,\dots,v}$ est le vecteur qui englobe les v variables exogènes.

Le défi majeur dans la construction des modèles ARMA est de déterminer les ordres p et q appropriés. Dans le cas de notre travail, nous avons opté pour la minimisation du critère d'Information d'Akaike (AIC) par rapport aux ordres p et q [Akaike, 1973].

La prévision du taux d'irradiance sur un horizon h est ainsi donnée par :

$$\widehat{\tau}_{t+h} = \sum_{i=1}^{p} \widehat{\phi}_i \tau_{t-i} + \sum_{r=1}^{v} \widehat{\beta}_r \boldsymbol{x}_{tr} + \epsilon_t + \sum_{j=1}^{q} \widehat{\theta}_j \epsilon_{t-j}, \tag{4.3.7}$$

où les $\widehat{\phi}_i$, $\widehat{\beta}_r$ et $\widehat{\theta}_j$ sont les paramètres estimés.

Machines à vecteurs de support (SVM)

Les machines à vecteurs de support constituent une technique d'apprentissage à base de noyau souvent utilisée dans les problèmes de classification et de régression [Vapnik, 1995]. La résolution des deux problèmes consiste à construire une fonction f qui fait correspondre une entrée $\boldsymbol{x}$ à une sortie y. Dans le cas d'une régression linéaire on a par exemple $f(\boldsymbol{x}) = \boldsymbol{x}w+b$. L'écart entre les données et le modèle est défini par une fonction de perte $L_\epsilon(y, f(\boldsymbol{x}))$ suivante (voir Figure 4.1) :

$$L_\epsilon(y, f(\boldsymbol{x})) = \begin{cases} 0 & \text{Si } |y - f(\boldsymbol{x})| \leq \epsilon \\ |y - f(\boldsymbol{x})| - \epsilon & \text{Sinon.} \end{cases} \tag{4.3.8}$$

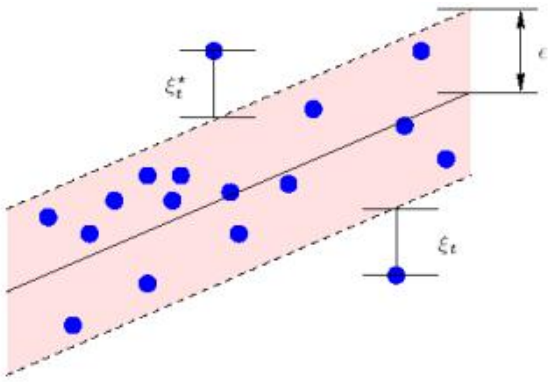


FIGURE 4.1 – Support à vaste marge dans le cas d'une régression linéaire

Pour mesurer l'erreur des observations qui sont en dehors de la zone de tolérance bornée par ϵ, deux variables de relaxation "ressort" ξ_t et ξ_t^* sont rajoutées pour assouplir et contrôler au travers de la variable C (voir Figure 4.1). Le support à vaste marge de régression est formulée comme une minimisation de la fonction suivante par rapport à :

$$\frac{1}{2}||w||^2 + C\sum_{t=1}^{N_A}(\xi_t + \xi_t^*), \tag{4.3.9}$$

tout en respectant les contraintes suivantes :

$$\begin{cases} y_t - f(\boldsymbol{x}_t) \leq \epsilon + \xi_t^* \\ f(\boldsymbol{x}_t) - y_t \leq \epsilon + \xi_t \\ \xi_t, \xi_t^* \geq 0, t = 1, ..., N_A. \end{cases}$$

Ce problème d'optimisation est transformé en un problème dual dont la solution est donnée par :

$$f(\boldsymbol{x}) = \sum_{t=1}^{N_A} \alpha_t \langle \boldsymbol{x}_t, \boldsymbol{x} \rangle + b, \tag{4.3.10}$$

tel que les coefficients α_t correspondent à la différence entre deux multiplicateurs de Lagrange.

Dans le cas d'une régression non linéaire, Vapnik propose de projeter les observations dans un espace de dimension supérieure à travers une fonction de transformation non

linéaire φ. Cette transformation est implicite à travers le calcul d'une fonction noyau qui vérifie [Vapnik, 1995] :

$$K(\boldsymbol{x}_t, \boldsymbol{x}) = \langle \varphi(\boldsymbol{x}_t).\varphi(\boldsymbol{x}) \rangle. \tag{4.3.11}$$

Plusieurs noyaux peuvent être utilisés. Nous pouvons citer : le noyau linéaire, le noyau polynomial et le noyaux RBF (Radial Basis Fonction).

La solution du problème d'optimisation dans le cas d'une régression non linéaire est exprimée par :

$$f(\boldsymbol{x}) = \sum_{t=1}^{N_A} \alpha_t K(\boldsymbol{x}_t, \boldsymbol{x}) + b. \tag{4.3.12}$$

Dans le cadre de notre travail, la prévision du taux d'irradiance sur un horizon h est positionnée dans un cadre de régression non linéaire qui est décrit par :

$$\tau_{t+h} = \sum_{t=1}^{N_A} \alpha_t K_{RBF}(\boldsymbol{x}_t, \boldsymbol{x}_*) + b, \tag{4.3.13}$$

tel que K_{RBF} représente le noyau RBF qui est calculé par [Vapnik, 1995] :

$$\begin{aligned} K_{RBF} &= exp(-\frac{||\boldsymbol{x}_t - \boldsymbol{x}_*||^2}{2\sigma^2}) \\ &= exp(-\gamma||\boldsymbol{x}_t - \boldsymbol{x}_*||^2), \end{aligned} \tag{4.3.14}$$

avec $\gamma = \frac{1}{2\sigma^2}$ et $||.||$ est la norme euclidienne.

Forets aléatoires (RF)

Avant d'évoquer les forêts aléatoires, il est nécessaire d'introduire les arbres de décision. Les arbres de décision sont des algorithmes de classification et de régression souvent utilisés en apprentissage automatique [Breiman et al., 1984].

Un arbre est construit avec un seul nœud racine qui n'a pas de branches entrantes. Tous les autres nœuds ont une unique branche entrante. Les nœuds avec des branches sortantes sont appelés nœuds internes. Les autres nœuds sont nommés feuilles.

L'arbre de décision le plus utilisé est l'algorithme CART (Classification And Regression Trees) qui inclut la classification et la régression [Breiman et al., 1984]. L'arbre de décision généré par CART est binaire, où chaque nœuds interne a deux branches sortantes. Le critère de segmentation utilisé par cet algorithme est l'indice de Gini [Breiman et al., 1984]. Après la construction de l'arbre, un élagage est effectué pour supprimer les branches les moins informatives en terme d'erreur quadratique de prévision.

Le principe des modèles forêts aléatoires (Random Forest) consiste à construire un ensemble d'arbres de décision afin d'améliorer les performances en termes de precision [Breiman, 2001].

Soit un ensemble de données d'apprentissage $A = \{\boldsymbol{x}_t, y_t\}_{t=1}^{N_A}$. L'algorithme des forêts aléatoires combine la méthode du bagging [Breiman, 1996] et la sélection aléatoire des attributs de partitionnement des nœuds qui participent à la construction de l'arbre de décision [Amit and Wilder, 1997]. Le bagging est une technique d'ensemble basée sur le reéchantillonnage pour générer un ensemble de T régresseurs. Cet ensemble est construit à partir de plusieurs échantillons boostrap $(S_j)_{j=1,...,M}$ composés de N_A-uplets $(\boldsymbol{x}_t, y_t)$ tirés d'une manière aléatoire avec remise dans A. L'ensemble de M régresseurs sera utilisé par

la suite pour prévoir le taux d'irradiance solaire via un calcul de moyenne. Les étapes principales sont décrites dans le pseudo-code 5.

Algorithme 5 Pseudo-code de construction d'une forêt aléatoire

1: **Entrées** : $A = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^{N_A}$ un ensemble d'apprentissage
M nombre d'arbres
v nombre d'attributs
2: **Pour** $j = 1$ à M **Faire**
3: Générer un échantillon Boostrap S_j
4: Construire un arbre de décision de type CART sur l'échantillon Boostrap S_j
5: Choisir pour chaque nœud une variable de partitionnement en tirant aléatoirement parmi les v attributs
6: **Fin Pour**
7: **Sorties** : $(f_j)_{j=1,\ldots,M}$. (f_j est le résultat de prévision de l'arbre de décision j)

La prévision du taux d'irradiance sur un horizon h est donnée par :

$$\widehat{\tau}_{t+h} = \frac{1}{M} \sum_{j=1}^{M} f_j, \tag{4.3.15}$$

où f_j est le résultat de prévision du taux d'irradiance d'un arbre de décision j sur un horizon h.

Réseaux de neurones (NN)

Un réseau de neurones (NN) est une méthode d'apprentissage automatique qui a été conçue pour imiter le fonctionnement du cerveau humain [McCulloch and Pitts, 1943]. Le cerveau se compose de nombreux neurones qui sont reliés par des axones, des synapses et des dendrites. Un réseau de neurones est composé de neurones qui sont liés par des poids et des biais. L'objectif d'un réseau de neurones est de créer une relation entre les entrées $\boldsymbol{x}_t$ et les sorties y_t. Chaque valeur de l'attribut r est multipliée par un poids w, qui sert de connexion entre une entrée et un neurone, ainsi qu'entre les différentes couches de neurones. Dans l'étape suivante, les produits des binômes (valeurs des attributs et poids) sont additionnés. Un poids supplémentaire qui représente le coefficient de biais b est rajouté à la somme. Une fonction d'activation f est appliquée au résultat trouvé pour obtenir la sortie a. Pour une meilleure compréhension, la Figure 4.2 représente un réseau de neurone de base.

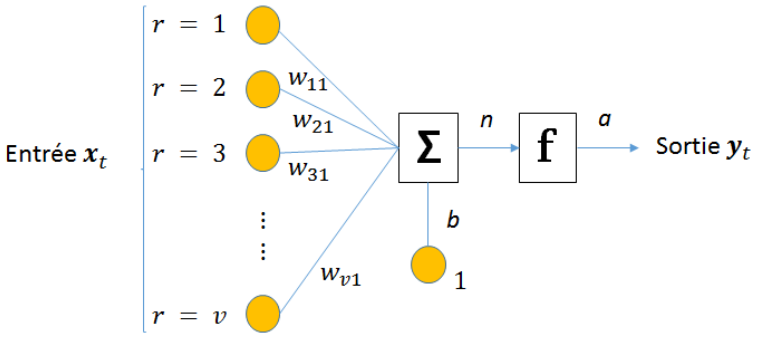


FIGURE 4.2 – Réseau de neurones de base

La relation entre l'entrée $\boldsymbol{x}_t$ et la sortie a d'un réseau de neurone de base est exprimée par :

$$a = f(n), \tag{4.3.16}$$

où

$$n = \boldsymbol{x}_{t1}w_{11} + \boldsymbol{x}_{t2}w_{21} + ... + \boldsymbol{x}_{tv}w_{v1} + b, \tag{4.3.17}$$

tel que v est le nombre d'attributs d'une donnée d'entrée $\boldsymbol{x}_t$.

Durant la phase d'apprentissage, les poids et les biais sont mises à jour à travers les équations suivantes :

$$w_{r1}^{(c+1)} = w_{r1}^{(c)} + 2\alpha e^{(c)}\boldsymbol{x}_{tr}^{(c)}, \tag{4.3.18}$$
$$b^{(c+1)} = b^{(c)} + 2\alpha e^{(c)}, \tag{4.3.19}$$

où w est le poids, b est le biais, e est l'erreur et α est le taux d'apprentissage. Il existe plusieurs fonctions d'activations, mais dans la pratique la fonction qui est principalement utilisée est la fonction sigmoïde (voir Figure 4.3).

La fonction sigmoïde est définit par : $\forall n \in \mathbb{R}$,

$$f(n) = \frac{1}{1 + e^{-n}} \tag{4.3.20}$$

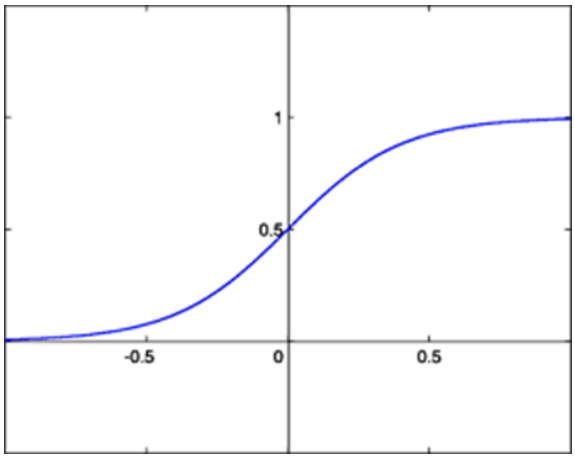


FIGURE 4.3 – Graphe de la fonction sigmoïde

Les réseaux de neurones Feed-forward sont des réseaux de neurones multi-couches. Dans cette étude, un réseau de neurone Feed-forward à trois couches a été utilisé comme le montre la Figure 4.4. Ce réseau est constitué d'une couche d'entrée composée de v éléments correspondants aux attributs de la donnée d'entrée $\boldsymbol{x}_t$. La deuxième couche est une couche cachée à m neurones. La dernière couche correspond à la couche de sortie. La fonction d'activation utilisée est la fonction sigmoïde.

La rétro-propagation de l'erreur est considérée comme l'approche d'apprentissage la plus populaire pour apprendre un réseau de neurones multi-couches [Rumelhart et al., 1988]. Cet algorithme permet de réduire l'erreur du réseau en ajustant les poids et les biais. Cette approche utilise la méthode de descente du gradient pour minimiser l'écart quadratique moyen entre la sortie du réseau a et la sortie souhaitée y_t.

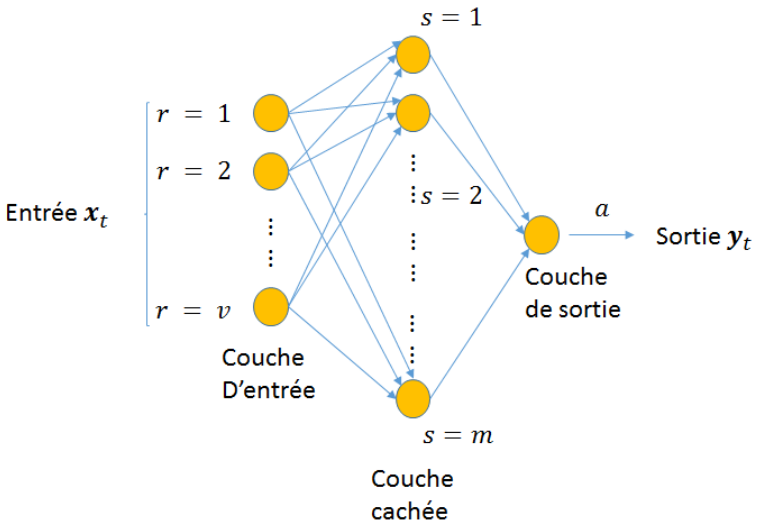


FIGURE 4.4 – Structure du réseau de neurones Feed forward utilisé

La prévision du taux d'irradiance sur un horizon h est donnée par :

$$\widehat{\tau}_{t+h} = \sum_{s=1}^{m} w_s f(\sum_{r=1}^{v} \boldsymbol{x}_{tr} w_{rs} + b_1) + b_2, \tag{4.3.21}$$

où w_s, w_{rs} représentent les poids et b_1, b_2 sont les biais.

4.4 Évaluation des modèles

4.4.1 Mesure de la performance de généralisation

La validation croisée est une méthode d'évaluation des modèles qui est basé sur des techniques de rééchantillonnage. Elle consiste dans un premier temps à diviser l'ensemble de données en trois parties :

— Un sous-ensemble d'apprentissage *(A)* dont les données seront utilisées pour la construction du modèle ;

— Un sous-ensemble de validation *(V)* où les données permettront d'estimer les paramètres du modèle ;

— Un sous-ensemble test *(T)* dont les données serviront uniquement pour évaluer la performance du modèle construit.

A partir d'un ensemble de données de taille N, il existe plusieurs méthodes (appelées techniques de rééchantillonnage) pour estimer la qualité de l'apprentissage. Parmi ces méthodes on peut citer :

— **Validation simple** : cette technique consiste à diviser l'ensemble de données en deux sous ensembles (base d'apprentissage et base de test). Souvent on garde 70% de données pour l'apprentissage et 30% pour le test [Ljung, 1986]. Le score de performance du modèle est calculé sur l'échantillon de test (voir Figure 4.5).

— **K-validations croisées** : Cette technique consiste à partitionner l'ensemble de données initiale en K sous ensembles de tailles approximativement identiques N/K [Stone, 1974]. Un des K échantillons est sélectionné comme jeu de test et les $(K-1)$ restants comme base d'apprentissage. Cette opération est répétée K fois et pour

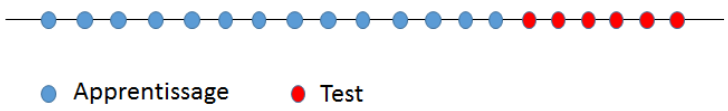


FIGURE 4.5 – Validation simple

chaque répétition des nouveaux jeux de test et d'apprentissage sont proposés. Le score de performance global du modèle est la moyenne des K scores obtenus sur chaque échantillon de test (voir Figure 4.6).

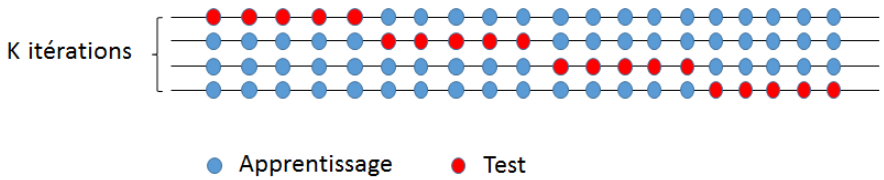


FIGURE 4.6 – K-validations croisées

— **Leave-one-out** : Cette méthode est un cas particulier de la K-validation croisée où $K = N$ [Kohavi, 1995] ; l'apprentissage est effectué sur $(N-1)$ observations et le test sur la $N^{ième}$ restante. Cette opération est répétée N fois pour que chacune des observations figure en tant que observation test. Cette technique est souvent appliquée lorsque le nombre d'observations n'est pas élevé (voir Figure 4.7).

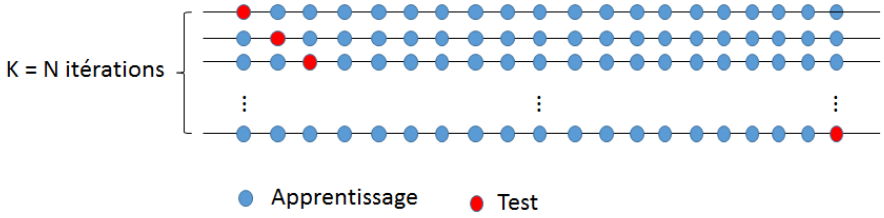


FIGURE 4.7 – Leave-one-out

— **Validation croisée pour les séries temporelles** : cette méthode prend en compte le sequencement chronologique des observations [Hyndman, 2018]. Elle consiste à sélectionner les observations aux instants $(1, 2, ..., k+i-1)$ en tant qu'ensemble d'apprentissage (estimation du modèle) et à tester l'observation à l'instant $k+i$. Le paramètre k désigne le nombre minimum d'observations dans l'ensemble d'apprentissage. Cette étape est réitérée en faisant évoluer l'indice $i = 1, 2, ..., N-k$ et calculant pour chaque itération le score de performance du modèle (voir Figure 4.8).

Dans le cadre de la thèse, vu que l'irradiance solaire est représentée par une série chronologique, il est nécessaire d'appliquer la validation croisée pour les séries temporelles. Pour trouver les meilleurs paramètres des modèles d'apprentissage automatique (SVM, RF et NN), la K-validation croisée est utilisée durant chaque phase d'apprentissage. Nous rappelons aussi que les ordres p et q du modèle ARMAX sont déterminés en utilisant le Critère d'Information d'Akaike (AIC).

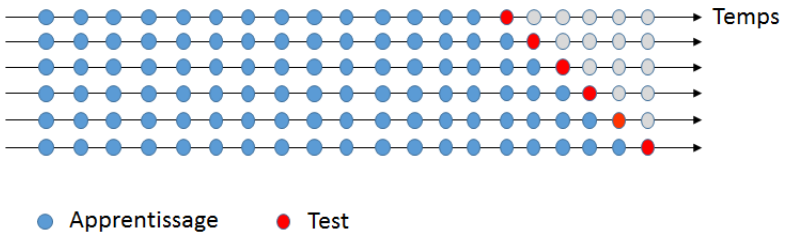


FIGURE 4.8 – Validation croisée pour les séries temporelles

4.4.2 Mesure de la performance d'un modèle de prévision

Pour évaluer les performances des modèles de prévision, il existe plusieurs métriques permettant de quantifier la différence entre les valeurs prédites $\widehat{I}_t$ par le modèle et les valeurs réelles observées I_t, tel que $t = (1, 2, ..., N_T)$. Dans le cas de la prévision journalière,N_T est multiple de 24. Parmi ces métriques nous pouvons citer :

— **Erreur quadratique moyenne (*Mean Square Error, MSE*)** : cette métrique consiste à calculer la moyenne arithmétique des carrés des écarts entre les observations prédites et les observations réelles. La formule s'écrit comme suit :

$$MSE = \frac{1}{N_T}\sum_{t=1}^{N_T}(\widehat{I}_t - I_t)^2. \tag{4.4.1}$$

— **Racine carrée de l'erreur quadratique moyenne (*Root Mean Square Error, RMSE*)** : comme son nom l'indique, cette mesure correspond à appliquer une racine carrée au MSE. La formule est donnée par :

$$\begin{aligned} RMSE &= \sqrt{MSE} \\ &= \sqrt{\frac{1}{N_T}\sum_{t=1}^{N_T}(\widehat{I}_t - I_t)^2}. \end{aligned} \tag{4.4.2}$$

— **Erreur absolue moyenne (*Mean Absolute Error, MAE*)** : est la moyenne arithmétique des valeurs absolues des écarts entre les observations prédites et les observations réelles. La MAE est formulée par :

$$MAE = \frac{1}{N_T}\sum_{t=1}^{N_T} \mid \widehat{I}_t - I_t \mid . \tag{4.4.3}$$

— **Erreur absolue moyenne en pourcentage (*Mean Absolute Percentage Error, MAPE*)** : elle consiste à calculer une moyenne des écarts en valeur absolue par rapport aux valeurs observées. Pour exprimer cette mesure en pourcentage, le résultat de calcul sera multiplié par 100 comme indiqué dans la formule suivante :

$$MAPE = \frac{100}{N_T}\sum_{t=1}^{N_T}\frac{\mid \widehat{I}_t - I_t \mid}{I_t}, \tag{4.4.4}$$

avec $I_t > 0$.

Le MAPE ne peut pas être utilisé dans notre cas, à cause des valeurs nulles que peut prendre l'irradiance solaire pendant certaines périodes de la journée.

Dans le cadre de cette thèse, deux métriques d'évaluation de performance sont utilisées à savoir : RMSE et MAE.

4.5 Résultats de prévision à court et moyen termes

Dans cette section, nous présentons les résultats de prévision à court et moyen termes (horaire et journalière) après l'application des six méthodes (Naive, calendaire, ARMAX, SVM, RF et NN) sur les quatre localisations (Carpentras, Pampelune, Brasilia et Ile de la Réunion) décrites dans l'analyse exploratoire.

Deux ans (2012-2013) de données sont considérés pour apprendre et valider les modèles de prévision. Comme nous l'avons déjà mentionné dans la Section 4.4.1, nous avons utilisé la validation croisée de série temporelle pour prendre en compte l'aspect temporel des données.

L'apprentissage des modèles s'effectue sur une fenêtre d'un an. Une fenêtre glissante est ensuite utilisée sur toute l'année 2012 lors de la première itération. Le décalage de cette fenêtre dépend de l'horizon de prévision (horaire ou journalier). La validation sera faite sur la prochaine observation dans le cas de prévision horaire et sur les 24 prochaines observations pour une prévision journalière.

Nous rappelons que pour évaluer les performances des modèles de prévision, deux métriques sont utilisées à savoir : la racine carrée de l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). Ces mesures sont calculées sur toute l'année de validation (2013). Pour l'interprétation des résultats, nous nous sommes basés uniquement sur le RMSE.

4.5.1 Résultats de prévision à moyen terme

Dans cette partie, nous nous intéressons aux résultats de prévision moyen terme (journalière). La Table 4.2 montre les performances des méthodes dans le cas de prévision journalière pour quatre localisations différentes (Carpentras, Pampelune, Brasilia et Ile de la Réunion). Nous remarquons que quelque soit le point d'observation, la méthode naïve est moins efficace que les autres méthodes. Celle-ci peut être écartée dans les futures évaluations. Les résultats des modèles obtenant les meilleures performances sont mis en caractères gras. C'est le cas des NN pour les quatre localisations. Ces performances sont aussi atteintes par SVM pour Pampelune et ARMAX pour l'Ile de la Réunion. L'efficacité des modèles varient d'un site à un autre. Par exemple, une différence de RMSE d'environ 40 est observée entre Carpentras et l'Ile de la Réunion. Les différences de performances entre les différentes localisations sont directement liées au type de climat. Le climat tropical humide de l'Ile de la Réunion se caractérise par son irrégularité et des pluies durant toute l'année. A l'inverse, le climat méditerranéen qui règne sur la région de Carpentras connait un été sec et chaud et un hiver froid et humide. C'est pour cette raison que les RMSE obtenus pour les climats assez stables comme ceux de Carpentras et de Pampelune sont plus faibles que ceux obtenus pour les climats tropicaux (Brasilia et Ile de la Réunion).

4.5.2 Résultats de prévision à court terme

Dans cette partie, nous présentons les résultats obtenus par les modèles de prévision sur un horizon court terme (horaire). La Table 4.3 fournit les performances des modèles dans le cas d'une prévision horaire appliquée sur les quatre sites. En comparant les Tables 4.2 et 4.3, nous remarquons que les modèles horaires améliorent les performances des

		Méthodes de prévision					
Localisation	**Métrique**	**Naïve**	**Calendaire**	**ARMAX**	**SVM**	**RF**	**NN**
Carpentras	**RMSE**	110,87	92,94	63,38	57,60	57,78	**54,34**
	MAE	44,85	43,53	28,37	23,36	24,53	**22,18**
Brasilia	**RMSE**	134,65	116,08	91,38	90,11	89,76	**88,94**
	MAE	56,61	54,85	41,69	40,63	40,56	**38,44**
Pampelune	**RMSE**	121,50	108,71	73,22	**66,19**	69,14	**66,69**
	MAE	55,58	54,30	36,80	**30,22**	33,29	31,46
Ile de la Réunion	**RMSE**	140,42	116,50	**96,68**	97,34	98,86	**96,38**
	MAE	62,36	60,27	**47,87**	48,55	49,44	**47,68**

TABLE 4.2 – Performances des modèles dans le cas d'une prévision journalière appliquée sur quatre localisations différentes

modèles journaliers et cela est valable pour toutes les localisations. Cette amélioration s'explique par l'intégration des données météo mesurées et le taux d'irradiance à l'heure courante (t). Ces variables permettent en effet de décrire et de comprendre une situation actuelle et aident aussi à mieux prévoir l'irradiance solaire à l'heure prochaine (t+1). Nous remarquons également que les méthodes naïve et calendaire ont les moins bons résultats par rapport aux autres méthodes (ARMAX, SVM, RF et NN). Il est important de noter que la méthode naïve fournit de meilleurs résultats par rapport à la méthode calendaire dans le cas des climats méditerranéen et maritime cote-ouest (Carpentras et pampelune), et inversement dans le cas des climats tropicaux (Brasilia et Ile de la réunion). Cela est expliqué par l'irrégularité du climat et ce même à une échelle temporelle courte (horaire) dans le cas des régions tropicales. Cette irrégularité pénalise la méthode naïve et favorise la méthode calendaire, et inversement dans le cas des climats assez stables. Concernant les autres méthodes, nous remarquons que les modèles RF obtiennent les meilleurs résultats pour Carpentras, Pampelune et l'Ile de la Réunion. Ces performances sont aussi atteintes par NN pour Carpentras et SVM pour l'Ile de la Réunion.

		Méthodes de prévision					
Localisation	**Métrique**	**Naïve**	**Calendaire**	**ARMAX**	**SVM**	**RF**	**NN**
Carpentras	**RMSE**	88,47	92,94	37,62	36,89	**35,14**	**35,75**
	MAE	43,62	54,53	15,83	14,96	**13,67**	**13,78**
Brasilia	**RMSE**	123,62	116,08	59,65	58,52	**57,64**	58,50
	MAE	75,66	56,85	26,86	24,65	**23,76**	24,35
Pampelune	**RMSE**	88,52	108,71	46,81	45,52	45,22	**44,93**
	MAE	52,81	55,30	22,03	20,00	19,82	**19,41**
Ile de la Réunion	**RMSE**	122,79	116,50	65,91	**64,49**	**64,98**	65,39
	MAE	74,63	62,27	30,39	**29,30**	**29,59**	29,98

TABLE 4.3 – Performances des modèles dans le cas d'une prévision horaire appliquée sur quatre localisations différentes

4.5.3 Discussion

En observant les résultats affichés dans la Table 4.2, nous avons remarqué que les NN ont les meilleurs performances en termes de RMSE annuel pour les quatre localisations. Mais cette déduction n'est pas toujours valable si nous nous focalisons sur les performances

des modèles par jour. La Table 4.4 représente la fréquence du meilleur modèle par jour durant l'année 2013. Ces résultats sont obtenus en calculant des RMSE journaliers pour les quatre méthodes durant l'année. Par la suite, nous déterminons le nombre d'occurrences où ces modèles sont meilleurs. Le RMSE journalier est calculé comme suit :

$$RMSE = \sqrt{\frac{1}{24}\sum_{t=1}^{24}(\widehat{I}_t - I_t)^2}, \tag{4.5.1}$$

où 24 est le nombre d'heures durant la journée.

D'après la Table 4.4, si nous prenons comme exemple la localisation Brasilia, nous remarquons que ARMAX est le modèle le plus fréquent en terme de meilleure performance journalière suivie par celles de SVM, RF et NN.

	Méthodes de prévision			
Localisation	**ARMAX**	**SVM**	**RF**	**NN**
Carpentras	63	103	72	**127**
Brasilia	**112**	101	75	77
Pampelune	84	**128**	67	86
Ile de la Réunion	100	**146**	65	54

TABLE 4.4 – Fréquence du meilleur modèle par jour durant l'année 2013

Pour une analyse plus fine, la Figure 4.9 représente la répartition des meilleurs modèles sur l'année 2013 pour les quatre localisations. Nous remarquons que l'efficacité de certaines méthodes est dépendante de la période que nous considérons. C'est le cas des NN pour Carpentras dont une concentration est observée durant les mois de juillet et octobre. Il est noté aussi que certaines méthodes ont de meilleurs performances pour une série de jours consécutifs (exemple : SVM pour l'Ile de la Réunion durant le mois de juin), alors que pour certaines périodes, une alternance de méthodes est observée (exemple : le mois de novembre de Brasilia).

Suivant le même principe que le cas journalier, la Table 4.5 représente la fréquence du meilleur modèle par heure durant l'année 2013. Cette fréquence est calculée uniquement sur les parties non nulles de l'irradiance. La Figure 4.10 représente l'évolution des RMSE horaires des quatre modèles par rapport aux quatre localisations durant la journée du 1^{er} janvier 2013. D'après la Figure 4.10, nous remarquons que le modèle avec les meilleures performances change d'une heure à une autre pour certains cas (exemple : Carpentras entre 9h et 10h) et reste le même pour une séquence d'heures dans d'autres (exemple : Ile de la Réunion entre 5h et 13h).

	Méthodes de prévision			
Localisation	**ARMAX**	**SVM**	**RF**	**NN**
Carpentras	1010	1104	**1526**	1090
Brasilia	900	1440	**1282**	1102
Pampelune	974	**1332**	1268	1184
Ile de la Réunion	1175	**1583**	1077	897

TABLE 4.5 – Fréquence du meilleur modèle par heure durant l'année 2013

D'après les observations citées précédemment soit pour le cas journalier ou le cas horaire, nous avons constaté que chaque modèle de prévision a montré son efficacité pour

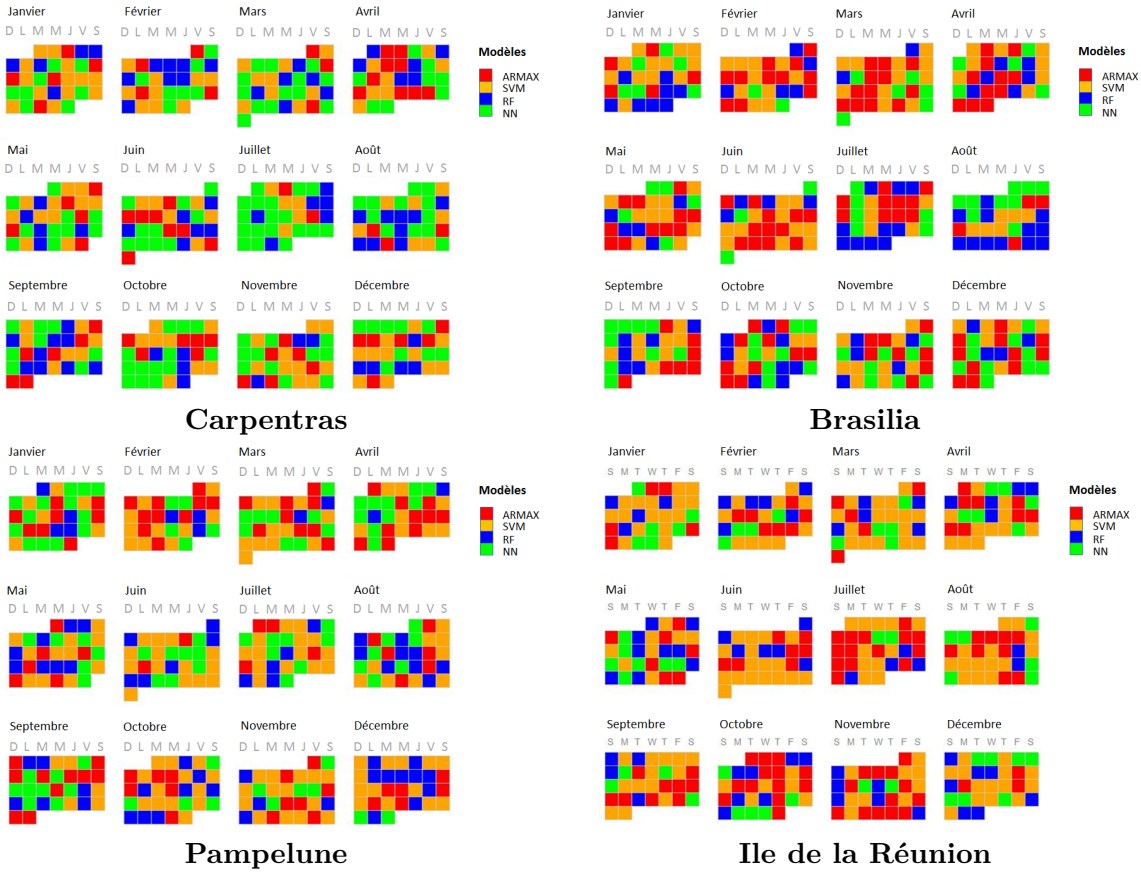


FIGURE 4.9 – Répartition des meilleurs modèles sur le calendrier de l'année 2013

certaines situations météorologiques. En tirant bénéfice des performances de chaque méthode, un modèle hybride qui prend en compte les résultats de chacune peut améliorer les résultats de prévision.

4.6 Modèle hybride

Comme cité précédemment, pour tirer bénéfices des performances de chaque modèle de prévision, nous avons décidé de fusionner les différentes décisions issues des quatre modèles : ARMAX, SVM, RF et NN. La contribution de chaque modèle dans cette fusion est mesurée par un poids qui représente son efficacité pour une situation météorologique donnée. Ces poids sont estimés par un apprentissage supervisé (classe connues) en supposant que chaque classe suit une densité gaussienne. Ce modèle est construit à partir d'un ensemble d'apprentissage qui est formé au fur et à mesure, en se basant sur les résultats de prévisions obtenus des quatre modèles. Cet ensemble d'apprentissage est noté $A' = \{(\boldsymbol{x}_t, \mathbf{z}_t)\}_{t=1}^{N_{A'}}$, où $\boldsymbol{x}_t \in \mathbb{R}^v$ est une donnée d'entrée contenant v attributs (voir Table 4.1) et $\mathbf{z}_t \in \mathbb{N}$ représente la classe correspondant au meilleur modèle de prévision à l'instant t, tel que $\mathbf{z} \in \{ARMAX, SVM, RF, NN\}$. Le meilleur modèle à un instant t donnée est celui qui minimise le RMSE. Une fois que les densités gaussienne ont été estimés : la proportion π_k (voir équation 4.6.1), la moyenne μ_k (voir équation 4.6.2) et la matrice variance covariance Σ_k (voir équation 4.6.3), les poids attribués aux $K = 4$ modèles pour une nouvelle situation météorologique ne sont que les ρ_{tk} (voir équation 4.6.4) qui représentent

la probabilité *a posteriori* qu'un modèle $k = \{1, 2, ..., K\}$ soit le meilleur à l'instant t.

$$\pi_k = \frac{\#\mathbf{p}_k}{N_{A'}} \tag{4.6.1}$$

$$\mu_k = \frac{1}{\#\mathbf{p}_k} \sum_t \mathbf{x}_t, \tag{4.6.2}$$

$$\Sigma_k = \frac{1}{\#\mathbf{p}_k} \sum_t (\mathbf{x}_t - \mu_k)(\mathbf{x}_t - \mu_k)', \tag{4.6.3}$$

$$\rho_{tk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_t; \mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k}, \tag{4.6.4}$$

où $\#\mathbf{p}_k$ représente le cardinal de la classe k et $\mathcal{N}(.; \mu_k, \Sigma_k)$ désigne la densité normale de moyenne μ_k et de matrice variance covariance Σ_k.

La prévision de l'irradiance solaire sur un horizon h est donnée par :

$$\widehat{I}_{t+h} = \sum_{k=1}^{K} \rho_{tk} \widehat{I}_{t+h}(k), \tag{4.6.5}$$

où $\widehat{I}_{t+h}(k)$ est la prévision de l'irradiance solaire obtenue par le modèle k sur un horizon h.

4.6.1 Discussion

Dans cette partie, nous présentons les résultats de prévision après application du modèle hybride sur les quatre localisations (Carpentras, Pampelune, Brasilia et Ile de la Réunion). Par manque de données, nous avons choisi d'appliquer le modèle hybride uniquement dans le cas de la prévision à court terme.

Une année (2013) de données a été considérées pour apprendre et valider le modèle. L'apprentissage s'effectue sur les six premiers mois et la validation sur les six mois restants.

		Méthodes de prévision				
Localisation	**Métrique**	**ARMAX**	**SVM**	**RF**	**NN**	**Hybride**
Carpentras	**RMSE**	34,08	33,45	32,16	32,74	**31,53**
	MAE	13,66	13,09	11,75	11,74	**11,39**
Brasilia	**RMSE**	60,39	60,55	59,60	60,09	**58,44**
	MAE	26,83	24,34	24,68	24,71	**23,65**
Pampelune	**RMSE**	39,79	40,00	39,02	39,18	**38,60**
	MAE	18,05	16,58	16,18	16,13	**15,90**
Ile de la Réunion	**RMSE**	66,47	64,32	64,94	64,74	**63,58**
	MAE	30,46	27,79	29,20	28,88	**27,58**

TABLE 4.6 – Performances des modèles dans le cas d'une prévision horaire appliquée sur quatre localisations différentes

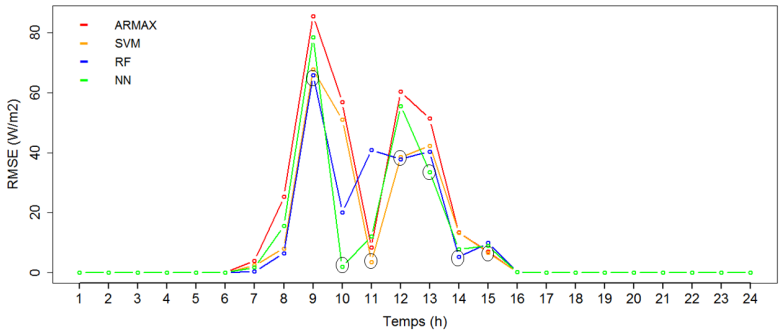
D'après la Table 4.6, nous remarquons que le modèle hybride améliore les résultats de prévision, et cela est valable pour les quatre localisations.

4.7 Conclusion

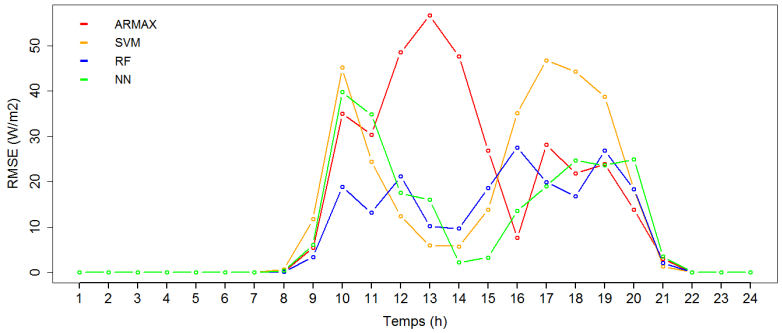
Dans ce chapitre, nous nous sommes intéressés à la prévision de l'irradiance solaire sur deux horizons temporels (court et moyen termes). Six approches statistiques ont été appliquées à savoir : Naïve, Calendaire, ARMAX, SVM, RF et NN.

Pour construire et valider les modèles, deux types de données ont été utilisés : des paramètres météorologiques mesurés et des prévisions issues des modèles NWP (*Numerical Weather Predictions*). Ces données sont relatives à quatre localisations (Carpentras, Brasilia, Pampelune et Ile de la Réunion) caractérisées chacune par un type de climat bien spécifique. Les performances des modèles ont été mesurées sur les deux horizons temporels et pour chaque climat. En faisant une comparaison entre les modèles, il a été remarqué que ARMAX, SVM, RF et NN obtiennent des résultats meilleurs que les méthodes de références (Naive et Calendaire). Cette remarque est valable pour les deux horizons et pour les quatre localisations. Nous avons observé également que les modèles de prévision horaire améliorent les performances des modèles journaliers. Ceci est expliqué par l'intégration des paramètres météorologiques et du taux d'irradiance à l'heure actuelle (t).

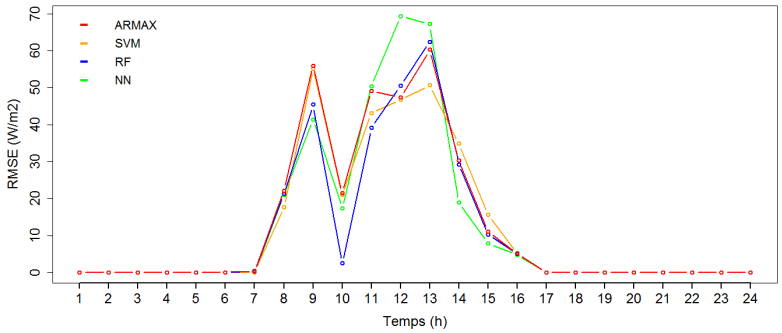
En se focalisant uniquement sur les quatre modèles : ARMAX, SVM, RF et NN, nous avons remarqué que chacun d'eux était efficace dans certaines situations météorologiques. En exploitant les performances de chaque méthode ainsi que les indications obtenues à travers cette étude, il est légitime de penser qu'un modèle hybride puisse améliorer les performances de prévision. Pour tirer bénéfice des performances de chaque méthode, nous avons proposé un modèle hybride qui combine les résultats de chacune, en leurs affectant des poids. Ceux-ci sont définis comme les probabilités *a posteriori* associées à une méthode de discrimination bayésienne supposant des classes gaussienne. Comme cela était prévu, le modèle hybride a fourni des résultats meilleurs que les modèles seuls et cela est valable pour les quatre localisations. Cependant, pour une meilleure précision des modèles seuls ainsi que pour le modèle hybride, des données collectées sur une longue période (plus de 2 ans) sera nécessaire. Une perspective de ce travail consiste à utiliser l'apprentissage profond et plus précisément les réseaux de neurones récurrents qui permettront d'exploiter l'aspect temporelles des données de façon plus explicite.



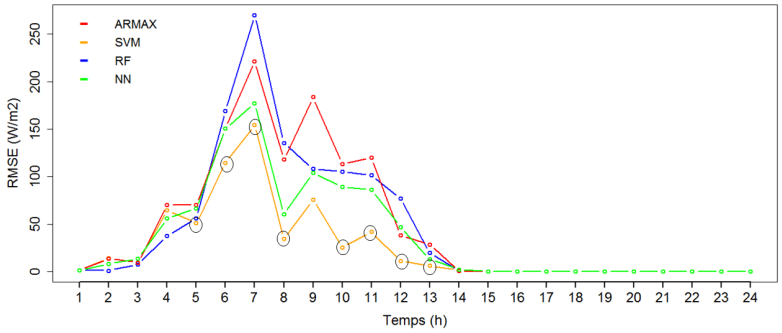
Carpentras



Brasilia



Pampelune



Ile de la Réunion

FIGURE 4.10 – Évolution des RMSE horaires des 4 modèles (ARMAX, SVM, RF et NN) pour la journée du 1^{er} janvier 2013 pour les quatre localisations (Carpentras, Brasilia, Pampelune et Ile de la Réunion)

Chapitre 5

Conclusion

Ces travaux de thèse avaient pour objectif de proposer des méthodes pour la fouille de données de consommation électrique et la prévision de la production photovoltaïque, problématiques clés dans la perspective d'une gestion intelligente de l'énergie. Les capteurs intelligents (compteurs intelligents et capteurs météo) permettent en effet de remonter des données sur la consommation et la production, qui sont suffisamment fines pour mener des analyses avancées et proposer des outils d'aide à la décision en vue d'une gestion intelligente de l'énergie.

Après une présentation des données réelles qui ont été exploitées dans la thèse, les tendances globales de consommation et de production ont d'abord été mises en évidence à l'aide des statistiques exploratoires. Suite à cette analyse préliminaire, les deux volets principaux de la thèse ont été développés.

Le premier volet traité porte sur la classification automatique des comportements de consommation électrique à l'échelle d'un bâtiment puis à l'échelle d'un territoire. Un algorithme de classification automatique à base de modèle de mélange Gaussien a été proposé. L'originalité de ce modèle réside dans sa capacité à prendre en compte des variables calendaires lors de la classification, conduisant ainsi à identifier de façon automatique des profils types de consommation en nombre réduit. Une évaluation du modèle sur une base de données réelles collectées en Irlande a été investiguée. Le croisement des résultats de classification avec des métadonnées riches sur les usagers a permis d'établir des liens entre profils de consommation et caractéristiques socio-économiques. Le modèle proposé a également servi à mener une analyse longitudinale sur l'année afin d'identifier les variations des habitudes de consommations. Les résultats de ce type d'analyse peuvent aider les fournisseurs d'énergie à développer de nouvelles politiques de tarification, améliorer les performances des modèles de prévision et identifier les cibles potentielles pour une procédure de demande d'effacement. Une compréhension fine des comportements de consommation électrique peut également être bénéfique pour la modélisation urbaine, en fournissant aux modèles de simulation des profils de consommation réalistes et dynamiques dans le temps. Cette compréhension est essentielle pour la construction des futures villes intelligentes et durables.

Le second volet de ces travaux de thèse porte sur la prévision de l'irradiance solaire sur deux horizons temporels : court terme $(t+1)$ et moyen terme $(t+1, t+2, ..., t+24)$. En effet, les modèles prédictifs peuvent aider les opérateurs des réseaux à mieux planifier leurs stratégies d'utilisation des différentes sources d'énergies tout en favorisant les énergies vertes.

Pour ce faire, des modèles de prévision basés sur des approches statistiques classiques, des approches d'apprentissage (RF, SVM et NN) ainsi qu'un modèle hybride combinant différentes approches ont été développés. En plus des données historiques sur l'irradiance, la prévision utilise des paramètres météorologiques mesurés et des prévisions issues des modèles NWP (*Numerical Weather Predictions*). La grande diversité des jeux de données relatifs à quatre localisations aux climats bien distincts (Carpentras, Brasilia, Pampelune et Ile de la Réunion) a permis de démontrer la pertinence du modèle hybride proposé et de mettre en évidence l'apport significatif des données météorologiques notamment pour les climats tropicaux.

Ce travail de thèse ouvre plusieurs perspectives. Le modèle génératif proposé pour l'identification des comportements types de consommations peut être étendu de plusieurs manières. Il peut en effet être intéressant d'y incorporer d'autres variables contextuelles correspondant à des jours particuliers comme des jours fériés, jours de pont vacances scolaires, des jours avec des événements spécifiques. De la même manière, il serait pertinent d'intégrer des saisons dans le modèle. Ceci permettra de considérer le comportement périodique des séries temporelles. L'estimation des paramètres d'un tel modèle nécessiterait toute fois de disposer d'une base de données collectées sur une longue période (plus d'un an).

Une autre perspective de ce travail consiste à utiliser l'apprentissage profond et plus précisément les réseaux de neurones récurrents qui permettront d'exploiter l'aspect temporelles des données de façon plus explicite.

Bibliographie

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281.

[Allcott, 2011] Allcott, H. (2011). Social norms and energy conservation. *Journal of public Economics*, 95(9) :1082–1095.

[Amit and Wilder, 1997] Amit, Y. and Wilder, K. (1997). Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11) :1300–1305.

[Appelrath et al., 2012] Appelrath, H.-J., Kagermann, H., and Mayer, C. (2012). Future energy grid. *Migrationspfade ins Internet der Energie. acatech Studie. Deutsche Akademie der Technikwissenschaften.*

[Armel et al., 2013] Armel, K. C., Gupta, A., Shrimali, G., and Albert, A. (2013). Is disaggregation the holy grail of energy efficiency ? the case of electricity. *Energy Policy*, 52 :213–234.

[Ayres et al., 2013] Ayres, I., Raseman, S., and Shih, A. (2013). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *The Journal of Law, Economics, and Organization*, 29(5) :992–1022.

[Badescu et al., 2013] Badescu, V., Gueymard, C. A., Cheval, S., Oprea, C., Baciu, M., Dumitrescu, A., Iacobescu, F., Milos, I., and Rada, C. (2013). Accuracy analysis for fifty-four clear-sky solar radiation models using routine hourly global irradiance measurements in romania. *RenewableEnergy*, 55 :85–103.

[Balijepalli et al., 2011] Balijepalli, V. M., Pradhan, V., Khaparde, S., and Shereef, R. (2011). Review of demand response under smart grid paradigm. In *Innovative Smart Grid Technologies-India (ISGT India), 2011 IEEE PES*, pages 236–243. IEEE.

[Beckel et al., 2012] Beckel, C., Sadamori, L., and Santini, S. (2012). Towards automatic classification of private households using electricity consumption data. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '12). Toronto, Canada*, pages 169–176. ACM.

[Beckel et al., 2014] Beckel, C., Sadamori, L., Staake, T., and Santini, S. (2014). Revealing household characteristics from smart meter data. *Energy : the international journal : technologies, resources , reserves, demand, impact, conservation, management, policy*, 78 :397–410.

[Benzécri and Bellier, 1976] Benzécri, J.-P. and Bellier, L. (1976). *L'analyse des données*, volume 1. Dunod Paris.

[Bezdek, 1974] Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 1(1) :57–71.

[Biernacki et al., 2003] Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41(3–4) :561 – 575.

[Bird and Hulstrom, 1981] Bird, R. E. and Hulstrom, R. L. (1981). Simplified clear sky model for direct and diffuse insolation on horizontal surfaces. Technical report, Solar Energy Research Inst., Golden, CO (USA).

[Birt et al., 2012] Birt, B. J., Newsham, G. R., Beausoleil-Morrison, I., Armstrong, M. M., Saldanha, N., and Rowlands, I. H. (2012). Disaggregating categories of electrical energy end-use from whole-house hourly data. *Energy and Buildings*, 50 :93–102.

[Box and Jenkins, 1994] Box, G. E. P. and Jenkins, G. M. (1994). *Time Series Analysis : Forecasting and Control*. Prentice Hall.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. In *Machine Learning*, pages 123–140.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

[Cao et al., 2013] Cao, H.-Â., Beckel, C., and Staake, T. (2013). Are domestic load profiles stable over time ? an attempt to identify target households for demand side management campaigns. In *Proceedings of the 39th IECON (IEEE Industrial Electronics Society)*. IEEE.

[Celeux et al., 1989] Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., and Ralambondrainy, H. (1989). *Classification automatique des données :[environnement statistique et informatique]*. Bordas Paris.

[Ciabattoni et al., 2013] Ciabattoni, L., Ippoliti, G., Longhi, S., Pirro, M., and Cavalletti, M. (2013). *Solar Irradiation Forecasting for PV Systems by Fully Tuned Minimal RBF Neural Networks*, pages 289–300.

[cimel, 2011] cimel (2011). Stations météorologiques. `https://www.cimel.fr/?family=stations-meteo`.

[Cooper, 2014] Cooper, A. (2014). Utility-scale smart meter deployments : Building block of the evolving power grid. *Inst. for Electron. Innovation, Washington, DC. IEI Rep.*

[Covrig et al., 2014] Covrig, C. F., Ardelean, M., Vasiljevska, J., Mengolini, A., Fulli, G., Amoiralis, E., Jiménez, M., and Filiou, C. (2014). Smart grid projects outlook 2014. *JRC science and policy reports*.

[da Silva Fonseca Jr. et al., 2013] da Silva Fonseca Jr., J. G., Takashi, O., Hideaki, O., Ken-ichi, S., Takumi, T., and Kazuhiko, O. (2013). Analysis of different techniques to set support vector regression to forecast insolation in tsukuba, japan. *Journal of International Council on Electrical Engineering*, pages 121–128.

[Darby et al., 2006] Darby, S. et al. (2006). The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 486(2006).

[Davenport and Dyché, 2013] Davenport, T. H. and Dyché, J. (2013). Big data in big companies. *International Institute for Analytics*, 3.

[David et al., 2016] David, M., Ramahatana, F., Trombe, P.-J., and Lauret, P. (2016). Probabilistic forecasting of the solar irradiance with recursive arma and garch models. *Solar Energy*, pages 55–72.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B*, 39 :1–38.

[Deng et al., 2010] Deng, F., Su, G., Liu, C., and Wang, Z. (2010). Global solar radiation modeling using the artificial neural network technique. In *2010 Asia-Pacific Power and Energy Engineering Conference*, pages 1–5.

[Dent et al., 2013] Dent, I., Aickelin, U., and Rodden, T. (2013). The application of a data mining framework to energy usage profiling in domestic residences using uk data. *arXiv preprint arXiv :1307.1380*.

[EDF, 2017] EDF (2017). Carte d'ensoleillement de la france. `https://www.edf.fr/edf/carte-d-ensoleillement-de-la-france`.

[EEA, 2007] EEA (2007). Presidency conclusions of the brussels european council of 8/9 march 2007. `https://www.eea.europa.eu/policy-documents/presidency-conclusions-of-the-brussels`.

[EEA, 2014] EEA (2014). European council 23-24/10/2014 - conclusions on 2030 climate and energy policy framework. `https://www.eea.europa.eu/policy-documents/european-council-23-24-10`.

[Ehrhardt-Martinez et al., 2010] Ehrhardt-Martinez, K., Donnelly, K. A., Laitner, S., et al. (2010). Advanced metering initiatives and residential feedback programs : a meta-review for household electricity-saving opportunities. American Council for an Energy-Efficient Economy Washington, DC.

[elster solutions, 2017] elster solutions (2017). Rexuniversal meter. `https://www.elstersolutions.com/en/product-details-na/1154/en/REX_Universal_Meter#sbox0=;`.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pages 226–231.

[eurostat, 2017] eurostat (2017). Final energy consumption in households. `http://ec.europa.eu/eurostat/web/products-datasets/-/t2020_rk200`.

[Everitt and Hand, 1981] Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions*. Monographs on applied probability and statistics. Chapman and Hall.

[Fei et al., 2013] Fei, H., Kim, Y., Sahu, S., Naphade, M., Mamidipalli, S. K., and Hutchinson, J. (2013). Heat pump detection from coarse grained smart meter data with positive and unlabeled learning. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1330–1338. ACM.

[Figueiredo et al., 2005] Figueiredo, V., Rodrigues, F., Vale, Z., and Gouveia, J. B. (2005). An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems,20(2) :596–602,2005*.

[Fischer, 2008] Fischer, C. (2008). Feedback on household electricity consumption : a tool for saving energy ? *Energy efficiency*, 1(1) :79–104.

[gadgetreview, 2015] gadgetreview (2015). Onzo smart energy kit helps you track your home's electricity usage. `http://www.gadgetreview.com/onzo-smart-energy-kit-helps-you-track-your-homes-electricity-usage`.

[Ghanbarzadeh et al., 2009] Ghanbarzadeh, A., Noghrehabadi, A. R., Assareh, E., and Behrang, M. A. (2009). Solar radiation forecasting based on meteorological data using artificial neural networks. In *2009 7th IEEE International Conference on Industrial Informatics*, pages 227–231.

[Goldstein et al., 2008] Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint : Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3) :472–482.

[Haben et al., 2016] Haben, S., Singleton, C., and Grindrod, P. (2016). Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, 7(1) :136–144.

[Hammer et al., 1999] Hammer, A., Heinemann, D., Lorenz, E., and Lückehe, B. (1999). Short-term forecasting of solar radiation : a statistical approach using satellite data. *Solar Energy*, pages 139–150.

[Hyndman, 2018] Hyndman, R. (2018). *Forecasting : Principles and Practice*. OTexts, Australia, 2nd edition.

[international energy agency, 2017] international energy agency (2017). Key world energy statistics. `https://www.iea.org/publications/freepublications/publication/KeyWorld2017.pdf`.

[Ji and Chee, 2011] Ji, W. and Chee, K. C. (2011). Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn. *Solar Energy*, pages 808–817.

[Kaufman and Rousseeuw, 1987] Kaufman, L. and Rousseeuw, P. (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*.

[Kleiminger et al., 2014] Kleiminger, W., Mattern, F., and Santini, S. (2014). Predicting household occupancy for smart heating control : A comparative performance analysis of state-of-the-art approaches. *Energy and Buildings*, 85 :493–505.

[Kleissl, 2010] Kleissl, J. (2010). Current state of the art in solar forecasting. *California Renewable Energy Forecasting, Resource Data and Mapping. University of California.*

[Kohavi, 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143.

[Kohonen, 1982] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1) :59–69.

[Kwac et al., 2014] Kwac, J., Flora, J., and Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1) :420–430.

[Kwac et al., 2017] Kwac, J., Flora, J., and Rajagopal, R. (2017). Lifestyle segmentation based on energy consumption data. *IEEE Transactions on Smart Grid*.

[Kwac et al., 2013] Kwac, J., Tan, C.-W., Sintov, N., Flora, J. A., and Rajagopal, R. (2013). Utility customer segmentation based on smart meter data : Empirical study. In *SmartGridComm*, pages 720–725. IEEE.

[Lance and Williams, 1967] Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies 1. hierarchical systems. *The Computer Journal*, 9(4) :373–380.

[LATRIBUNE, 2017] LATRIBUNE (2017). Les investissements dans les énergies renouvelables restent insuffisants. `https://www.latribune.fr/economie/international/les-investissements-dans-les-energies-renouvelables-restent-insuffisants.html`.

[Lauret et al., 2015] Lauret, P., Voyant, C., Soubdhan, T., David, M., and Poggi, P. (2015). A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Solar Energy*, pages 446–457.

[Ljung, 1986] Ljung, L. (1986). *System Identification : Theory for the User*. Prentice-Hall.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, pages 281–297, Berkeley, Calif. University of California Press.

[Mattern and Floerkemeier, 2010] Mattern, F. and Floerkemeier, C. (2010). From the internet of computers to the internet of things. *From active data management to event-based systems and more*, pages 242–259.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5 :115–133.

[McLachlan and Basford, 1988] McLachlan, G. and Basford, K. (1988). *Mixture Models : Inference and Applications to Clustering*. Marcel Dekker, New York.

[McLachlan and Krishnan, 2008] McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*, volume 382 of *Wiley series in probability and statistics*. Wiley.

[McLachlan and Peel, 2000] McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics, New York.

[McLoughlin et al., 2015] McLoughlin, F., Duffy, A., and Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied energy*, 141 :190–199.

[Mellit and Pavan, 2010] Mellit, A. and Pavan, A. M. (2010). A 24-h forecast of solar irradiance using artificial neural network : Application for performance prediction of a grid-connected pv plant at trieste, italy. *Solar Energy*.

[Melzi et al., 2015] Melzi, F. N., Zayani, M.-H., Hamida, A. B., Samé, A., and Oukhellou, L. (2015). Identifying daily electric consumption patterns from smart meter data by means of clustering algorithms. In *ICMLA*, pages 1136–1141. IEEE.

[Mengersen et al., 2011] Mengersen, K., Robert, C., and Titterington, M. (2011). *Mixtures : Estimation and Applications*. John Wiley and Sons.

[monenergie, 2017] monenergie (2017). Consommation moyenne d'électricité. `https://www.monenergie.net/consommation-moyenne-electricite.php`.

[Nizar et al., 2006] Nizar, A., Dong, Z. Y., and Zhao, J. (2006). Load profiling and data mining techniques in electricity deregulated market. In *2006 IEEE Power Engineering Society General Meeting*, page 7. IEEE.

[Nova et al., 2005] Nova, J. C., Cunha, J. B., and de Moura Oliveira, P. (2005). Solar irradiation forecast model using time series analysis and sky images. In *Proceedings of the 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment*.

[Ramsay and Silverman, 2005] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition.

[Rao et al., 2012] Rao, K. D. V. S. K., Rani, B. I., and Ilango, G. S. (2012). Estimation of daily global solar radiation using temperature, relative humidity and seasons with ann for indian stations. In *2012 International Conference on Power, Signals, Controls and Computation*, pages 1–6.

[readwrite, 2007] readwrite (2007). The art, science and business of recommendation engines. `http://readwrite.com/2007/01/16/recommendation_engines`.

[Reinsch, 1967] Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, 10(3) :177–183.

[REN21, 2017] REN21 (2017). renewables 2017 global status report. `https://www.iea.org/publications/freepublications/publication/KeyWorld2017.pdf`.

[Rousseeuw, 1987] Rousseeuw, P. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1) :53–65.

[RTE, 2017] RTE (2017). Panorama de l'électricité renouvelable au 30 septembre 2017. http://www.rte-france.com/sites/default/files/panorama_09-17-web.pdf.

[Rumelhart et al., 1988] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3) :1.

[Said and Dickey, 1984] Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71 :599–607.

[Salvatore, 2013] Salvatore, J. (2013). World energy perspective : Cost of energy technologies. *World Energy Council*.

[Sanchez et al., 2009] Sanchez, I. B., Espinós, I. D., Sarrión, L. M., López, A. Q., and Burgos, I. N. (2009). Clients segmentation according to their domestic energy consumption by the use of self-organizing maps. In *Energy Market, 2009. EEM 2009. 6th International Conference on the European*, pages 1–6. IEEE.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

[Smyth, 1996] Smyth, P. (1996). Clustering using monte carlo cross-validation. In *KDD*, pages 126–133.

[Stephens,] Stephens, M. Bayesian methods for mixtures of normal distributions.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(2) :111–147.

[Stoop, 2005] Stoop, I. A. (2005). *The hunt for the last respondent : Nonresponse in sample surveys*. Sociaal en Cultureel Planbu.

[Titterington et al., 1985] Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*, volume 7. Wiley, New York.

[Tong et al., 2016] Tong, X., Li, R., Li, F., and Kang, C. (2016). Cross-domain feature selection and coding for household energy behavior. *Energy*, 107 :9 – 16.

[Union, 2009] Union, E. (2009). Directive 2009/72/ec of the european parliament and of the council of 13 july 2009 concerning common rules for the internal market in electricity and repealing directive 2003/54/ec. *Off. J. Eur. Union L*, 211 :55–93.

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

[Vasconcelos, 2008] Vasconcelos, J. (2008). Survey of regulatory and technological developments concerning smart metering in the european union electricity market.

[Verdú et al., 2006] Verdú, S. V., Garcia, M. O., Senabre, C., Marin, A. G., and Franco, F. J. G. (2006). Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems, 21(4) :1672-1682, November 2006*, 21(4) :1672–1682.

[Wang et al., 2016] Wang, Y., Chen, Q., Kang, C., and Xia, Q. (2016). Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE Transactions on Smart Grid*, 7(5) :2437–2447.

[Weiss et al., 2013] Weiss, M., Mattern, F., and Beckel, C. (2013). Smart energy consumption feedback-connecting smartphones to smart meters. *ERCIM news*, 14.

[wikimedia, 2008] wikimedia (2008). Intelligenter zaehler- smart meter. https://commons.wikimedia.org/wiki/File:Intelligenter_zaehler_Smart_meter.jpg.

[Wittmann et al., 2008] Wittmann, M., Breitkreuz, H., Schroedter-Homscheidt, M., and Eck, M. (2008). Case studies on the use of solar irradiance forecast for optimized operation strategies of solar thermal power plants. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 18–27.

[Wold, 1938] Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Almqvist & Wiksells.

[Wolff et al., 2016] Wolff, B., Lorenz, E., and Kramer, O. (2016). *Statistical Learning for Short-Term Photovoltaic Power Predictions*, pages 31–45. Springer International Publishing.

[Yu et al., 2011] Yu, Z., Fung, B., Haghighat, F., Yoshino, H., and Morofsky, E. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, 43(6) :1409–1417.

[Zeng and Qiao, 2013] Zeng, J. and Qiao, W. (2013). Short-term solar power prediction using a support vector machine. *Renewable Energy*, pages 118 – 127.