Estimation et prédiction des productions d'énergies renouvelables et des consommations d'un réseau de distribution d'électricité

Mohamed Tribak

THÈSE

pour l'obtention du Grade de

DOCTEUR DE L'ÉCOLE NATIONALE SUPÉRIEURE DE MÉCANIQUE ET D'AÉROTECHNIQUE

(Diplôme National - Arrêté du 25 mai 2016)

Ecole Doctorale : Sciences et Ingéniérie pour l'Information et Mathématiques

Secteur de Recherche : Informatique et Applications

Présentée par :

Mohamed TRIBAK

****

Estimation et prédiction des productions d'énergies renouvelables et des consommations d'un réseau de distribution d'électricité

****

Directeurs de thèse :

Emmanuel GROLLEAU

Thierry POINOT

Co-encadrant de thèse :

Brice CHARDIN

****

Soutenue le 26 novembre 2021

devant la Commission d'Examen

****

JURY

Président :

Thomas DEVOGELE	Professeur, Université François Rabelais Tours, Tours

Rapporteurs :

Sadok BEN YAHIA	Professeur, Tallinn University of Technology, Estonia
Marie-Jeanne LESOT	Maître de conférences HDR, Sorbonne Université, Paris

Membres du jury :

Emmanuel GROLLEAU	Professeur, ISAE-ENSMA, Poitiers
Thierry POINOT	Professeur, Université de Poitiers, Poitiers
Brice CHARDIN	Maître de Conférences, ISAE-ENSMA, Poitiers

"À mes grands-parents Mohamed et Amina TRIBAK"

Remerciements

J'adresse mes vifs remerciements à mon directeur de thèse Emmanuel GROLLEAU, professeur à l'ISAE-ENSMA, et à mon co-directeur de thèse Thierry POINOT, professeur à l'université de Poitiers, pour leurs directives, leurs disponibilités et leur soutien constant au cours de la thèse. Je tiens également à remercier mon encadrant Brice CHARDIN, maître de conférence à l'ISAE-ENSMA, pour son encadrement pertinent, sa disponibilité et son exigence scientifique. Ils m'ont transmis leur méthodologie scientifique en m'apprenant comment le travail de recherche devrait être mené.

J'adresse tous mes remerciements aux membres du jury : les rapporteurs à Monsieur Sadok BEN YAHIA, professeur à Tallinn University of Technology, ainsi qu'à Madame Marie-Jeanne LESOT, maître de conférences, HDR à Sorbonne Université, de l'honneur qu'ils m'ont fait en rapportant ce travail de thèse. A Monsieur Thomas DEVOGELE, professeur à l'université François Rabelais Tours, d'avoir accepté de présider le jury de thèse.

Ma reconnaissance s'adresse également à Fabien PETIT, directeur des infrastructures et de la stratégie de développement à SRD et à Yassine EL BAZ, ingénieur smart grid à SRD, pour leurs conseils du génie électrique qu'ils m'ont prodigués et pour leur judicieux encadrement métier. Ils m'ont accordé une grande autonomie, et j'ai été fier de la confiance qu'ils m'ont témoignée. Je remercie également la direction de SRD pour la chance qu'elle m'a accordée de travailler dans ce projet d'innovation pointu.

Je remercie vivement tout le personnel de SRD pour son écoute bienveillante, sa cordialité et toutes les connaissances techniques qu'ils m'ont appris. En particulier, aux équipes de la conduite, de la métrologie et de la cartographie avec qui j'ai eu de nombreuses collaborations et discussions enrichissantes sur le métier du réseau électrique.

Je remercie chaleureusement toutes les équipes du laboratoire LIAS pour leur accueil et pour tous les bons moments que nous avons passés ensemble. Merci aussi à Bénédicte BOINOT pour sa gentillesse et son aide dans les différentes démarches administratives. Je remercie également Mickael BARON pour toutes les aides techniques et les moyens matériels qu'il m'a accordés durant la thèse.

Mes plus profonds remerciements s'adressent à ma famille, en particulier à mes parents qui m'ont aidé et accompagné pendant toutes mes études et qui m'ont encouragé à me lancer dans cette aventure de thèse.

Un grand merci à mon épouse Hadia qui est toujours à mes cotés et qui a largement contribué à la réussite de cette thèse. Merci à elle pour sa confiance et m'avoir poussé à résister jusqu'à la fin de ma thèse. Je remercie également mes beaux-parents pour leur encouragement et leur soutien inconditionnel.

Que tous ceux et celles qui ont contribué de près ou de loin à l'accomplissement de ce travail trouvent l'expression de mes remerciements les plus chaleureux.

Résumé

Les gestionnaires de réseaux de distribution (GRD) d'électricité ont connu ces dernières années une intégration importante des moyens de production d'énergie renouvelable (EnR). De plus, avec l'apparition de nouveaux usages de l'énergie, notamment la mobilité électrique, les microgrids et les technologies de stockage, la gestion du réseau est devenue de plus en plus complexe, complexité qui ira croissante dans les années à venir. Dans ce contexte, SRD, GRD dans le département de la Vienne, a financé des travaux de recherche dans le domaine des smartgrids, notamment sur l'optimisation dynamique du schéma d'exploitation de son réseau de distribution d'électricité haute et moyenne tension, et cette thèse sur la prévision, la prédiction et l'estimation des valeurs des flux énergétiques circulant sur ce réseau.

Dans une première phase, la thèse propose une approche de sélection des données de la consommation d'énergie les plus pertinentes et étudie leur influence sur l'efficacité de l'optimisation. Pour cela une méthodologie de réduction de dimensionnalité de données est proposée. Elle utilise des techniques issues du domaine de l'apprentissage automatique non supervisé et d'analyse de données temporelles. Cette méthodologie permet de détecter des similitudes dans les données afin de les regrouper dans des groupes homogènes.

La seconde phase élabore une méthodologie d'estimation de la production d'énergie des installations photovoltaïques (PV) distribuées dans le réseau de distribution de SRD en utilisant les méthodes d'interpolation spatiale. En effet, la plupart des producteurs avec des capacités de production basse tension ne sont pas instrumentés pour une mesure en temps réel. Au contraire, les producteurs moyenne et haute tension sont instrumentés et permettent des mesures à grain fin sous forme de séries temporelles. Le but de cette étude est d'estimer les productions de milliers de petits producteurs en exploitant les données des moyens et gros producteurs de référence équipés de compteurs communicants.

Finalement, le problème de la prévision de la production photovoltaïque est abordé. Le but de cette étude est d'élaborer une prévision ponctuelle court terme d'un horizon d'une heure pour gérer l'intermittence de la production solaire et une prévision probabiliste long terme pour planifier et optimiser le réseau sur un horizon d'un à trois mois. Nous montrons que les algorithmes d'apprentissage automatique avec une approche globale améliorent les prévisions fournies par des méthodes naïves. La pertinence des prévisions obtenues par rapport au cadre applicatif a été validée à l'aide d'un estimateur d'état du réseau pour quantifier les différences de pertes, de chutes de tension et d'élévation de tension entre un état prévu et un état réel.

Abstract

In recent years, electricity distribution system operators (DSOs) have seen a significant integration of renewable energy production. In addition, with the development of new energy uses, such as electric mobility, microgrids and storage technologies, grid management has become more and more complex, a complexity that will increase in the years to come. In this context, SRD, DSO in the department of Vienne in France, has funded research in the field of smartgrids, notably on the dynamic optimization of the operating scheme of its high and medium voltage electricity distribution network, and this thesis on the forecasting, prediction and estimation of the values of the energy flows circulating on this network.

In the first phase, the thesis proposes an approach to select the most relevant energy consumption data and study their influence on the efficiency of the optimization. For this purpose, a data dimensionality reduction approach is proposed. It uses techniques from the unsupervised machine learning and temporal data analysis fields. This methodology allows to detect similarities in the data in order to group them in homogeneous groups.

The second phase develops a methodology for estimating the energy production of photovoltaic installations distributed over the SRD distribution network using spatial interpolation methods. In fact, most of the generators with low voltage production capacities are not instrumented for real time measurement. On the other hand, medium and high voltage generators are instrumented and allow fine-grained time series measurements. The goal of this study is to estimate the production of thousands of small generators by exploiting the data of medium and large generators of reference equipped with communicating meters.

Finally, the problem of forecasting photovoltaic production is considered. The goal of this study is to develop a short-term point forecast with a horizon of one hour to manage the intermittency of the solar production and a long-term probabilistic forecast to schedule and optimize the network over a horizon of one to three months. We show that machine learning algorithms with a global approach improve the forecasts provided by naive methods. The pertinence of the obtained forecasts to the application framework has been validated using a loadflow estimator to quantify the differences in losses, voltage drops and voltage rises between a forecasted state and an actual state.

Table des matières

Introduction générale

> Un grand réseau de distribution peut être comparé à un système composé de deux réseaux, l'un transmettant l'intelligence, l'autre transmettant la puissance [. . .] On ne pourra pas exploiter d'une façon satisfaisante le réseau de puissance, si l'on ne peut d'abord exploiter le réseau de l'intelligence
>
> *C.E. Bennett, rapport CIGRE, 1923* [1]

Avec le changement climatique, nous vivons aujourd'hui une transformation remarquable dans notre manière d'utiliser l'énergie. Les préoccupations environnementales ont incité les états à adopter des politiques ambitieuses en matière d'énergie. L'Union européenne, comme exemple, a fixé de nombreux objectifs à l'horizon 2030. Elle prévoit une réduction d'au moins 55% des émissions de gaz à effet de serre par rapport à 1990. Elle s'engage à améliorer l'efficacité énergétique des différents secteurs économiques d'au moins 32.5 % et à augmenter la part des moyens de production de l'énergie verte à au moins 32 % {UE, 2021a,b}.

Dans le département de la Vienne, le parc d'énergie renouvelable (photovoltaïque, éolienne, biogaz, etc.) a connu une croissance remarquable ces dernières années. Il est passé de 119 MW en 2014 à 340 MW fin 2020, soit une évolution de 185%. Ce parc, raccordé au réseau de distribution de l'électricité, a injecté une énergie annuelle de 725 GWh (en 2020), soit 55 % de la consommation annuelle d'énergie du territoire. Cette multiplication des installations des productions EnR distribuées sur tout le réseau entraîne une complexité de gestion du système électrique. En effet, le réseau électrique reliant plusieurs acteurs énergétiques (producteurs et consommateurs) a été conçu dans une perspective d'acheminement unidirectionnel de l'énergie, depuis les producteurs en amont vers les consommateurs finaux en aval. Avec la pénétration de ces moyens de production EnR décentralisés, le transit de l'énergie est devenu bidirectionnel et multi-sens. SRD, le gestionnaire de réseau de distribution (GRD) d'électricité du département de la Vienne, est confronté à ces nouvelles problématiques de gestion des flux d'énergie.

Dans ce contexte, SRD a financé de multiples travaux de recherche dans le domaine des réseaux électriques intelligents pour pouvoir intégrer efficacement ces moyens de production EnR. D'une part, l'optimisation dynamique du schéma d'exploitation du réseau de distribution d'électricité haute et moyenne tension a été développée dans une première thèse {Ali Zazou, 2017}. D'autre part, la prévision et l'estimation des valeurs des flux énergétiques circulant sur le réseau ont été élaborées dans la présente thèse. Ces travaux visent conjointement à déterminer un schéma d'exploitation optimal du réseau électrique,

1. Voir page 3 du livre : « Le système nerveux du réseau français de transport d'électricité : 1946-2006 : 60 années de contrôle électrique » {Bouneau *et al.*, 2012}

en intégrant les moyens de production EnR décentralisés et en respectant les différentes contraintes techniques et réglementaires de gestion du système électrique. L'objectif est de passer d'un fonctionnement classique du réseau électrique vers un fonctionnement plus intelligent.

Cette transition d'un système électrique classique vers un réseau électrique intelligent se décline dans une première phase par une instrumentation du réseau avec des systèmes de communication et de collecte de données. Dans une deuxième phase, elle passe par l'élaboration des outils d'aide à la décision en exploitant et en surveillant ces différentes informations d'état instantané du réseau.

Les intérêts de ces outils d'aide à la décision pour les GRD sont multiples. Premièrement, la reconfiguration urgente du réseau en cas d'incident permettra une meilleure qualité de fourniture de l'électricité au bénéfice des consommateurs finaux. La solution proposée assure un schéma d'exploitation respectant les différentes contraintes techniques du fonctionnement du système électrique. Deuxièmement, l'absence de certaines données de l'état instantané du réseau risque de générer des surtensions et de surcharger les lignes électriques, principalement les données des moyens de production intermittente dépendant des aléas de la météo. Leur estimation permet une meilleure connaissance des états actuels de tous les points du réseau afin de proposer d'autres schémas d'exploitation valorisant une consommation locale des énergies décentralisées.

Ces données exploitées par les différents outils sont non seulement des données statiques des caractéristiques physiques et topologiques du réseau, mais aussi des données dynamiques comme la tension, le courant et la puissance, pour lesquelles un nombre limité de points de mesure instantanée sont disponibles. Les points de mesure se trouvant principalement au niveau des transformateurs haute tension, l'information sous forme de séries temporelles des flux agrégés entrants et sortants sur le réseau est connue, mais seule un portion des productions ayant lieu localement sur le réseau est mesurée. De plus, les flux sont mesurés en temps réel sous forme agrégée, mais ne sont pas mesurés individuellement au niveau des postes de distribution-transformateurs reliés au réseau basse tension.

L'objectif de cette thèse est donc d'estimer et prévoir (1) la consommation au niveau des postes de distribution et (2) la production EnR dont la majeure partie de l'installation diffuse n'est pas mesurée en temps réel. Ces estimations et prédictions de flux étant destinées à être utilisées pour l'optimisation du schéma d'exploitation du réseau de distribution, leur précision doit être suffisante pour que le schéma d'exploitation optimal calculé soit le plus proche possible de l'optimal réel. Cette combinaison de l'optimisation du réseau avec ces techniques de prévision et d'estimation des données permet de mieux gérer la variabilité spatio-temporelle des producteurs et des consommateurs. Cela améliore l'efficacité énergétique du système électrique et de tous ses usages et donc réduit leur empreinte carbone.

Afin de répondre à ces problématiques, il faut développer un processus de sélection de données d'entrée pertinentes parmi toutes les données disponibles et étudier leur influence, leur volume et leur niveau de détail sur l'efficience de l'optimisation. Par la suite, il faut élaborer des approches d'estimation et de prévision de la production EnR et de la consommation d'électricité en exploitant les différentes données internes.

Chapitre 1

Contexte général de la thèse

Le réseau électrique français et ses différents niveaux et composants sont présentés dans ce chapitre. Nous exposons ainsi le contexte général de la thèse et ses différents objectifs.
Le réseau électrique français est divisé en deux catégories de réseau, un réseau de transport et un réseau de distribution. Ils sont composés de plusieurs ouvrages électrotechniques dont le but est d'acheminer l'énergie depuis les sources de production vers les consommateurs finaux.
La thèse s'intègre dans un projet de recherche et développement dans le domaine du smart grid pour le gestionnaire de réseau de distribution du département de la Vienne, SRD, en collaboration avec le laboratoire LIAS. L'objectif est de développer un outil de prévision et d'estimation de la production des énergies renouvelables (EnR) et de la consommation d'électricité sur le réseau de SRD.

Sommaire

1.1 Introduction du chapitre

L'électricité est l'une des grandes découvertes dans l'histoire de l'humanité. Elle joue aujourd'hui un rôle essentiel dans notre vie quotidienne. Sans l'électricité par exemple, nous ne pouvons pas communiquer par téléphone ou utiliser internet.

Pour se servir de cette électricité, il faut l'acheminer des producteurs vers les consommateurs. Pour ce besoin, l'humanité a inventé le réseau électrique comme moyen de transport et de distribution de l'électricité. Ce réseau est un système complexe composé de plusieurs éléments et unités et qui a comme but le transit de l'électricité vers les consommateurs finaux.

Dans ce chapitre, nous exposerons dans la première partie le réseau électrique français, son histoire, son organisation et ses différents composants. Ensuite, nous présenterons les différentes missions d'un gestionnaire de réseau de distribution de l'électricité. Finalement, le contexte général de cette thèse sera présenté dans la dernière partie.

1.2 Réseau électrique français

1.2.1 Histoire

L'histoire du réseau électrique français remonte au début de *XX* siècle avec la croissance de la demande sur l'électricité, notamment avec l'électrification des chemins de fer. A cette époque, il n'existait pas un réseau électrique national, seulement plusieurs entreprises privées, chacune gèrant son propre réseau avec ses propres normes comme la fréquence et le niveau de tension. En 1901, par exemple, la consommation électrique annuelle était de 350 GWh, elle est passée à 690 GWh en 1906. La région parisienne était la plus consommatrice en électricité avec plus de 600 km de lignes électriques. La fréquence a été normalisée à 50 Hz en 1918 avec des niveaux de tension allant de 40 kV à 70 kV {Bouneau *et al.*, 2012}.

Après la deuxième guerre mondiale, le 8 avril 1946, une loi est votée sur la nationalisation des entreprises de l'électricité [1]. Cette loi a créé l'électricité de France (EDF) comme une nouvelle organisation publique gérant la production, la distribution et le transport de l'électricité en intégrant toutes les entreprises privées. Toutefois, elle a exclu de la nationalisation les installations réalisées par les collectivités locales, des établissements publics ou de leurs groupements. Les compagnies fonctionnant en régie ne sont pas donc nationalisées. Elles forment aujourd'hui les entreprises locales de distribution (ELD) et elles gèrent 5 % du réseau électrique français {Enedis, 2020}. SRD fait partie de ces ELD en gérant le réseau de distribution du département de la Vienne {SRD, 2020}.

Le 10 février 2000 une loi introduit l'ouverture du marché de l'électricité à la concurrence [2]. Cette loi prévoit la création d'un gestionnaire de réseau de transport d'électricité indépendant, elle crée la commission de régulation de l'énergie (CRE), elle prévoit un accès non discriminatoire au réseau et elle assure pour tout consommateur final un libre choix de fournisseur d'électricité.

En décembre 2006 une nouvelle loi relative au secteur de l'énergie prévoit la séparation juridique des gestionnaires de réseau de distribution ayant plus de 100 000 clients [3]. Cette loi impose aux entreprises desservant plus de 100 000 clients de séparer juridiquement leurs activités d'acheminement d'électricité

1. Loi n° 46-628 du 8 avril 1946 sur la nationalisation de l'électricité et du gaz
2. Loi n° 2000-108 du 10 février 2000 relative à la modernisation et au développement du service public de l'électricité
3. Loi n°2006-1537 du 7 décembre 2006 relative au secteur de l'énergie

de celles de fourniture. SOREGIES RESEAUX DE DISTRIBUTION est donc crée et qui devient, le 1 janvier 2010, SRD Réseaux de Distribution.

1.2.2 Généralités

Un réseau électrique est un système complexe composé d'un ensemble d'équipements électrotechniques sur plusieurs niveaux de tension. Ce système permet d'acheminer l'électricité depuis les producteurs vers les consommateurs finaux en formant un réseau maillé ou arborescent.

Nous distinguons dans ce réseau plusieurs niveaux de gestion de transit et d'acheminement de l'électricité, la figure 1.1 illustre ces trois niveaux.

Le premier réseau, dit réseau de transport et de répartition, a pour rôle d'acheminer les grandes quantités d'énergie depuis les grands producteurs vers les gros consommateurs. Ces producteurs ont une capacité de production supérieure à 12 MW, comme les centrales nucléaires, les centrales hydroélectriques, les éoliennes offshore, etc. Les grands consommateurs sont généralement des grandes usines, des lignes TGV et des gestionnaires de réseau de distribution (GRD). Nous distinguons dans ce réseau trois niveaux de tensions différents (voir tableau 1.1). Le premier est appelé niveau HTB3 avec une tension égale à 400 kV, le deuxième est appelé niveau HTB2 est une tension égale à 225 kV ou 150 kV, le troisième est le niveau HTB1 avec une tension égale à 90 kV ou 63 kV. Ces niveaux de tension très élevés permettent de transporter sur de longues distances l'énergie en réduisant les pertes électriques générées par effet Joule. En France, le réseau de transport est géré par RTE (Réseau de transport d'électricité). Avec une longueur totale de 100 000 km de lignes électriques, RTE achemine l'électricité de 170 producteurs vers 15 entreprises ferroviaires, 380 grosses industries et 130 distributeurs, avec 2 200 postes de transformation au total {RTE, 2020}.

Le deuxième type de réseau est le réseau de distribution de l'électricité. Géré par les GRD, il a comme rôle l'alimentation des moyens et petits consommateurs comme les hôpitaux, les petites industries et les maisons individuelles. Il assure aussi l'acheminement des quantités d'énergies produites par les moyens et les petits producteurs décentralisés. Dans ce réseau, nous distinguons deux niveaux de tensions : niveau haute tension A (HTA) et niveau base tension (BT). Le niveau HTA (anciennement appelé moyenne tension) (voir tableau 1.1) regroupe les ouvrages dont le niveau de tension est compris entre 1 kV et 50 kV ; sur le réseau SRD, les ouvrages HTA sont exploités à une tension de 20 kV. En France, les ouvrages du niveau BT sont exploités à une tension égale à 400 V (en tension composée, 230 V en tension simple).

Le réseau de distribution est organisé en général selon une architecture arborescente. Il est composé de plus de 1.3 million de kilomètres de lignes électriques et plus de 700 000 postes de distribution publics {CRE, 2020}. Avec les nouveaux usages de l'énergie, comme le stockage, l'autoproduction, l'électromobilité et le développement des EnR, les GRD ont connu plusieurs défis et changements dans leur manière de gérer le réseau. L'enjeu pour les GRD aujourd'hui est de faire transiter l'énergie produite par les producteurs EnR décentralisés et de la faire consommer localement.

Depuis 1946, les réseaux de distribution en France sont gérés sur 95 % du territoire par ENEDIS (ex ERDF) et 5% gérés par les entreprises locales de distribution (ELD) tel que SRD pour le département de la Vienne {CRE, 2020}.

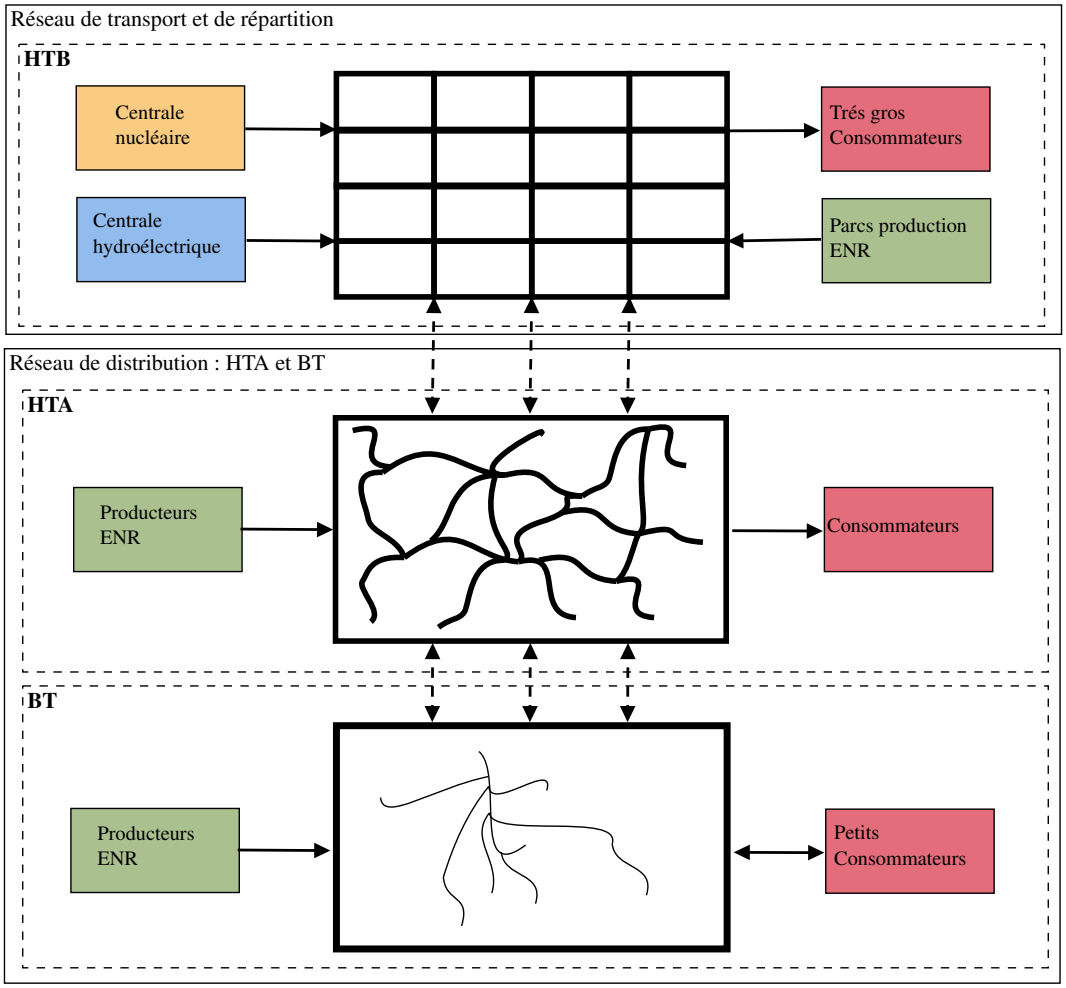


FIGURE 1.1 – Réseau électrique français

Niveau tension	Appellation	Catégorie	Tension
Très haute tension	HTB	HTB3	400 kV
		HTB2	225 kV
		HTB1	63 et 90 kV
Haute tension	HTA	-	1 kV à 50 kV
Basse tension	BT	-	50 V à 1000 V

TABLEAU 1.1 – Les différents niveaux de tension et leur appellation

1.2.3 Structure et composition du réseau électrique

Le réseau électrique est un système contenant plusieurs composants électrotechniques. Il peut être présenté comme un graphe constitué de sommets et d'arcs. La structure du réseau électrique est simplement la forme du graphe formé par ce réseau. Nous avons généralement deux catégories de structures, une structure maillée et une structure arborescente. Le réseau maillé est formé par un graphe contenant plusieurs chemins entre les sommets. Par contre, deux sommets dans le graphe du réseau arborescent sont liés par un seul chemin (arc) ou un enchaînement d'arcs (sauts), c'est-à-dire qu'il n'y a pas de boucles entre ces deux sommets.

Cette structure de graphe ne permet pas de définir complètement le réseau. Il faut alors ajouter à un réseau son schéma d'exploitation. En effet, un réseau peut être exploité de plusieurs façons indépendamment de sa structure. Par exemple, un réseau maillé peut être exploité d'une façon arborescente. Ce schéma d'exploitation arborescente se définit par la manière d'utiliser le réseau pour acheminer l'énergie depuis les producteurs vers les consommateurs.

Pour le réseau HTA, ce schéma est appelé schéma radial. Il permet en cas d'incident d'acheminer l'énergie par d'autres sources en utilisant d'autres chemins possibles dans le réseau. Par contre dans le cas du réseau BT, étant presque toujours exploité sous forme arborescente, en cas d'incident, il n'est pas possible d'alimenter les consommateurs par d'autres sources d'énergie. Le réseau rural est généralement exploité par une structure arborescente composée majoritairement de lignes aériennes, alors que le réseau urbain est exploité par une structure maillée avec des lignes souterraines. En outre, selon l'état du réseau, nous distinguons deux régimes d'exploitation différents : un régime normal d'exploitation qui est le fonctionnement normal du réseau, ce schéma est appelé « schéma normal » ; un régime dégradé d'exploitation qui est un fonctionnement contraint du réseau, lors des travaux ou d'incidents par exemple, ce schéma est appelé « schéma de secours ». Le schéma d'exploitation dépend aussi de la nature du réseau électrique (rural ou urbain) et de la période (été ou hiver). Il est modifiable par les agents chargés de la conduite de réseau en manipulant plusieurs interrupteurs en fonction du transit de l'énergie et de l'état de réseau {Ali Zazou, 2017}.

Le réseau électrique est donc composé de plusieurs matériels électrotechniques permettant l'acheminement de l'électricité. Il est constitué de postes transformateurs, de lignes électriques, de charges, de productions, d'interrupteurs et d'autres composants permettant la sécurité du réseau et le raccordement des lignes électriques entre elles. Par exemple, le réseau de SRD est formé de 16 postes HTB/HTA (postes sources), de plus de 8 777 postes HTA/BT, de 4 791 km de lignes BT, de 7 400 km de lignes HTA et de plus de 8000 interrupteurs permettant la modification du schéma d'exploitation. La figure 1.2 présente la structure du réseau électrique et ces composants électrotechniques, selon les différents niveaux de tension.

Les lignes électriques

Les lignes électriques forment un élément prépondérant pour le fonctionnement du réseau. Elles constituent le conduit sur lequel transite l'énergie. Elles ont des caractéristiques différentes selon le niveau de tension (HTA ou BT) et selon leur nature (ligne aérienne ou ligne souterraine). Généralement, nous distinguons ces câbles par leur section et leur matière (Aluminium, Almélec, Cuivre...). Ces informations permettent de connaître la résistance et la réactance de la ligne électrique et d'en déduire ses pertes engendrées par effet Joule. Elles sont dimensionnées pour passer une quantité d'énergie maximale, et si cette limite est dépassée, cela peut engendrer des dysfonctionnements dans le réseau.

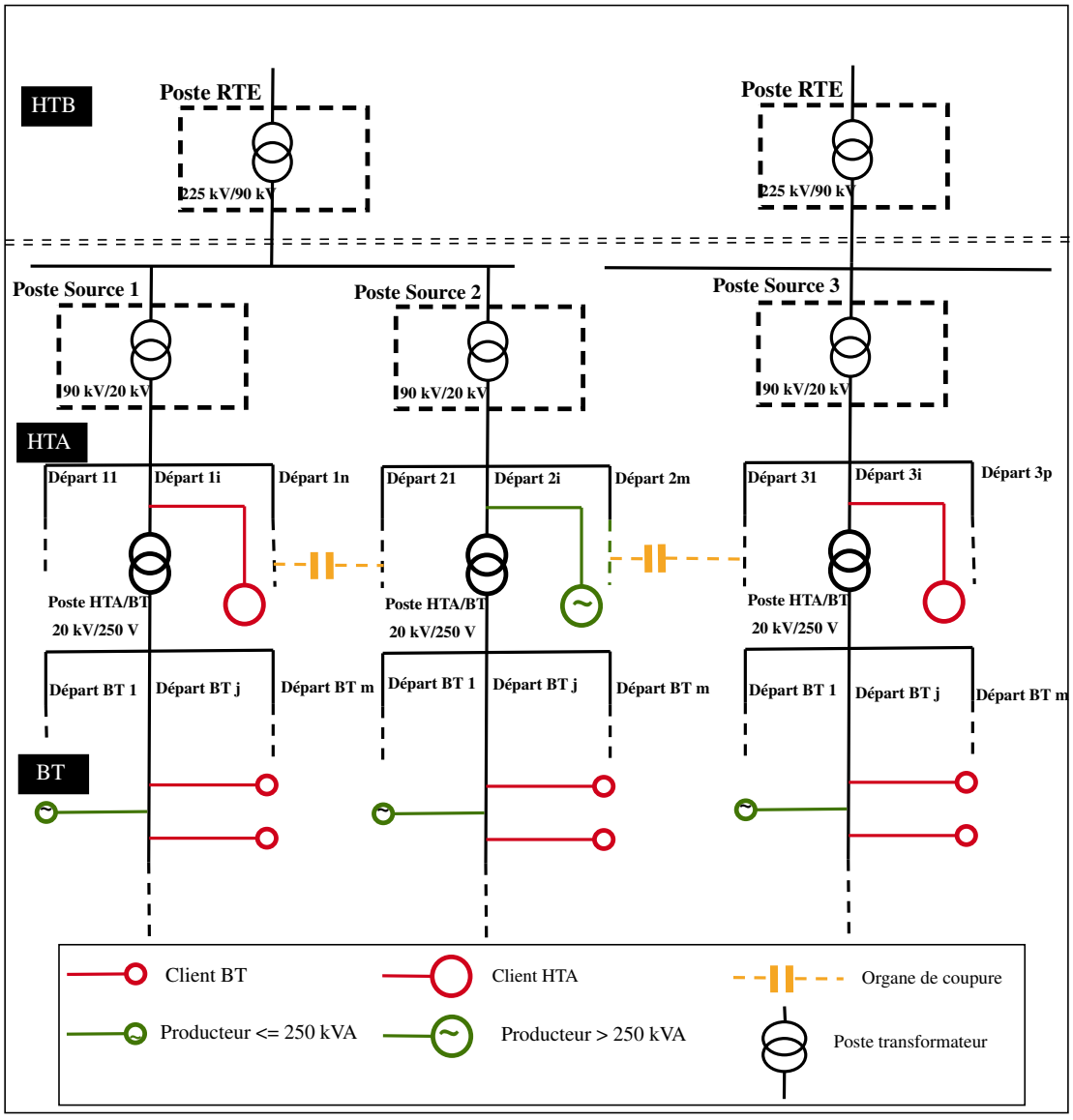


FIGURE 1.2 – Composition du réseau électrique

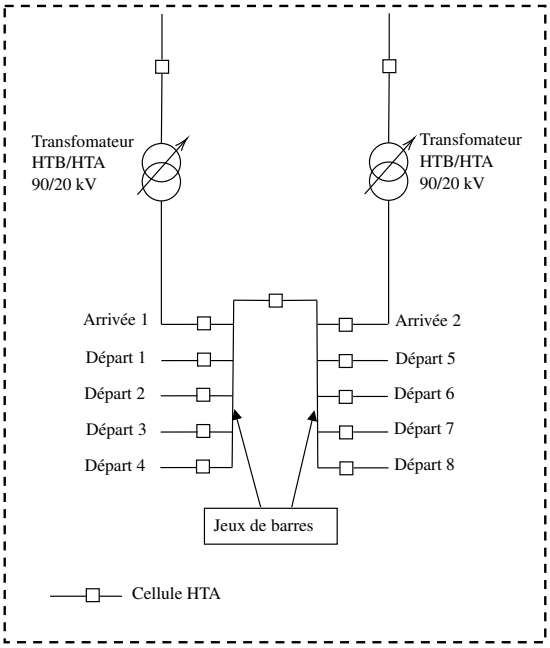


FIGURE 1.3 – Exemple de schéma d'un poste source

Poste transformateur HTB/HTA

Le poste transformateur HTB/HTA est appelé aussi poste source, car pour le GRD il s'agit d'une source d'énergie acheminée par le réseau de transport. Le poste source est un ouvrage électrique en amont du réseau de HTA (figure 1.2) qui joue un rôle important dans le fonctionnement du réseau. Il est l'interface avec le réseau de transport en permettant la réduction de la tension HTB en tension HTA (20 kV) afin de la répartir sur le réseau de distribution selon les besoins des consommateurs.

Le poste source assure ainsi la sécurité et la protection des ouvrages et des tiers. Il contient généralement deux transformateurs 90 kV/20 kV. En cas de défaillance d'un transformateur, la charge est reprise par l'autre transformateur du poste source, et éventuellement par d'autres postes sources voisins. Nous appelons ce principe le principe $N-1$, c'est-à-dire que le réseau doit être opérationnel en desservant les clients en cas de panne d'un transformateur.

Dans un poste source (voir figure 1.3), nous trouvons plusieurs composants. Des jeux de barres qui permettent de connecter plusieurs lignes électriques que nous appelons des départs HTA. Ces départs HTA sont précédés par des disjoncteurs (appelés aussi des cellules HTA) permettant d'ouvrir ou de fermer le départ. C'est-à-dire qu'ils permettent de passer ou de couper le courant selon les besoins de la conduite du réseau. Dans les zones éloignées des postes sources, les postes d'étoilement sont mis en place afin de réduire les longueurs des lignes HTA. Ces postes d'étoilement ont les mêmes caractéristiques que les postes sources (équipements de surveillance, de protection, de télécommande, etc). Ils contiennent aussi des départs HTA vers une continuité du réseau HTA.

Poste transformateur HTA/BT

Le poste HTA/BT joue le rôle d'interface entre le réseau HTA et le réseau BT en réduisant la tension HTA en tension BT (400 V/250 V) utilisable par les petits consommateurs comme les maisons individuelles et les petites industries. Dans certains cas, il lie des grands clients ou des grands producteurs au réseau HTA, il est appelé dans ce cas « poste client » ou « poste producteur ». Le poste HTA/BT englobe plusieurs petits consommateurs et parfois des petits et moyens producteurs raccordés au réseau BT, il est

appelé aussi « poste de distribution public » (poste DP). Du point de vue du GRD, les consommateurs sont appelés des charges et les producteurs des productions.

Les charges

Les charges désignent tous les consommateurs finaux d'énergie dans le réseau. Elles sont caractérisées par les puissances souscrites qui déterminent les besoins maximaux en termes de consommation dans le réseau. Cette puissance souscrite permet de déterminer la catégorie tarifaire des clients. Généralement, nous distinguons trois catégories de tarifs :

— Clients BT < 36 *kVA* (entre 3 et 36 kVA), clients consommateurs du segment C5 ou producteurs du segment P4, anciennement appelés tarifs bleus ;

— Clients BT ≥ 36 *kVA* (entre 36 et 250 kVA), clients consommateurs du segment C4 ou producteurs du segment P3, anciennement appelés tarifs jaunes ;

— Clients HTA, anciennement appelés tarifs verts.

Les clients BT < 36 *kVA* sont les particuliers ou les petites industries. Les clients BT ≥ 36 *kVA* et HTA sont principalement les grands consommateurs comme les usines ou les installations agricoles.

Le GRD doit connaître la puissance souscrite de toutes ces charges et sa somme au niveau des postes HTA/BT ou poste source. C'est-à-dire, au niveau d'un poste HTA/BT par exemple, le GRD doit connaître le nombre des charges et la somme totale de leurs puissances souscrites, ainsi que leur type tarifaire.

SRD gère à ce jour 140 000 clients au total, avec 90% des clients BT < 36 *kVA* sur un réseau majoritairement rural.

Les productions

Les productions désignent tous les producteurs décentralisés raccordés au réseau HTA ou BT. Ce sont principalement des producteurs EnR comme les parcs photovoltaïques (PV), les éoliennes, les centrales hydroélectriques et les usines biogaz. Ils sont caractérisés par leur capacité maximale de production, l'équivalent de la puissance souscrite pour les charges. Ces producteurs injectent leur production directement dans le réseau HTA ou BT. Cette production peut être consommée directement par les consommateurs voisins ou peut être renvoyée dans le réseau supérieur dans le cas de surproduction, c'est-à-dire quand la production est supérieure à la consommation. Cette opération est appelée le refoulement. Ces producteurs sont principalement raccordés au réseau HTA, sauf pour les producteurs PV dont un certain nombre est raccordé directement au réseau BT.

Dans le réseau de SRD, à ce jour il y a plus de 4 272 producteurs PV avec une capacité totale de 156 MW, 16 parcs éoliens avec une capacité de 201 MW, 4 producteurs biogaz avec une capacité de 4 MW et 9 centrales hydroélectriques avec une capacité de 1.5 MW. Ces producteurs injectent annuellement une énergie totale de 725 GWh dans le réseau de SRD (données 2020).

1.2.4 Missions d'un gestionnaire de réseau de distribution

Le gestionnaire d'un réseau distribution a plusieurs missions pour maintenir l'acheminement et le transit de l'énergie dans ce réseau.

Il assure le développement de son réseau en raccordant des nouveaux consommateurs ou producteurs. Il doit renforcer et entretenir le réseau pour bien les accueillir. Il assure la desserte de ses clients avec

une qualité de fourniture et de distribution, en respectant les normes sur les chutes ou les élévations de tensions et en limitant les coupures d'électricité. Par exemple, la tension du réseau BT doit être dans un intervalle de $\pm 10\%$ de la tension nominale, soit entre 207 et 253 V.

Pour développer son réseau, le GRD peut créer de nouveaux réseaux en ajoutant de nouveaux ouvrages comme des postes sources, par exemple. Il doit renouveler les vieux ouvrages et renforcer les réseaux existants. Finalement, il doit automatiser au maximum le réseau en installant des appareils commandés à distance comme des interrupteurs télécommandés afin de réduire les temps de coupure en cas de panne {Doulet, 2010}.

Le GRD assure le contrôle et la supervision de son réseau. Nous avons vu dans la section précédente que le réseau est composé de plusieurs éléments permettant son fonctionnement. L'exploitation de ces composants consiste à les utiliser en assurant une meilleure qualité de distribution. Le réseau peut être exploité de plusieurs façons selon le cas de fonctionnement (normal ou incident). Il doit donc être bien géré et contrôlé en temps réel, afin de connaître l'état de ses composants et sa structure instantanée {Doulet, 2010}. Cette mission de conduite est assurée par le bureau d'information et de conduite (BIC) de SRD. Ce bureau contrôle et conduit le réseau en 24h/24 et 7j/7.

Conduite et exploitation du réseau

Le GRD assure, via le service de conduite, le fonctionnement du réseau et l'écoulement de l'énergie. Son rôle principal est la supervision du réseau afin d'établir le transit de l'énergie dans les cas d'incidents. Il gère aussi la protection des personnes et les différents ouvrages et composants du réseau.

Il exploite en temps réel les données de capteurs permettant le pilotage du réseau. Ces capteurs sont généralement situés au niveau des postes sources et mesurent en temps réel plusieurs valeurs comme le courant, la tension et la puissance dans chaque départ HTA du poste source.

Il assure la conduite du réseau en fonction des travaux ou des pannes. Cela grâce à des interrupteurs appelés « organes de coupure » (voir figure 1.2) pilotés à distance ou manuellement. Ces organes de coupure se trouvent généralement dans les postes HTA/BT ou dans des armoires de coupure. A SRD, il y a à ce jour plus de 8000 organes de coupure répartis sur tout le réseau.

Les interrupteurs qui se trouvent au niveau des postes sources sont les cellules HTA (voir figure 1.3). Ils sont composés d'un interrupteur manipulable à distance et intégrant des dispositifs de protection et de mesure. La figure 1.4 illustre le rôle du service de la conduite et de la supervision du réseau en actionnant plusieurs interrupteurs et en surveillant plusieurs indicateurs comme la puissance, la tension et le courant. Il modifie et configure le schéma d'exploitation selon les contraintes de transit de l'énergie. Il joue un rôle majeur dans la performance et la qualité de gestion de réseau.

Qualité de distribution et de fourniture

Le GRD doit assurer pour ses clients une fourniture de qualité selon leurs besoins énergétiques.

La qualité de fourniture se définit par une distribution continue de l'électricité en respectant les normes sur la distribution. Le GRD doit réduire le temps des coupures de l'électricité. Ces coupures se divisent en trois catégories, les coupures longues qui dépassent les trois minutes, les coupures brèves qui sont comprises entre une seconde et trois minutes et les coupures très courtes qui sont inférieures à une seconde. Cette continuité de fourniture dépend de plusieurs facteurs, comme la structure de réseau et la qualité de

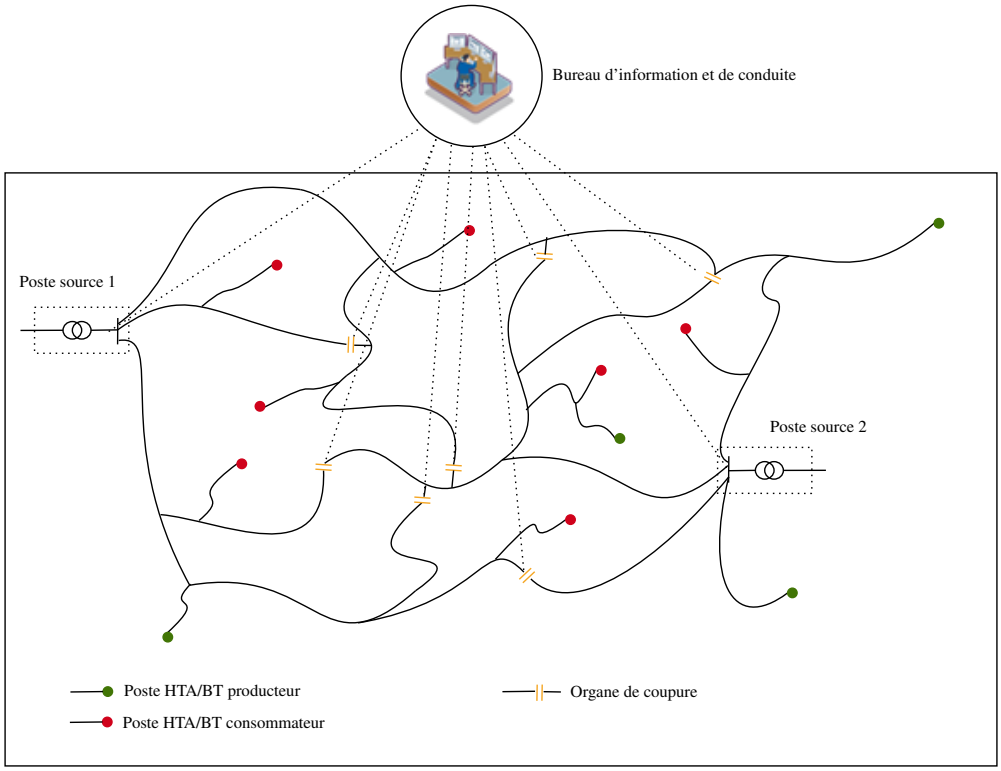


FIGURE 1.4 – Conduite de réseau de distribution

ses ouvrages. Les normes de la distribution de l'électricité imposent le contrôle des phénomènes électriques, comme les variations et les déséquilibres de tension. Ces normes constituent un indicateur de qualité majeur pour le GRD qui doit assurer en permanence la gestion de ces variations même pour les clients qui se trouvent au bout des lignes électriques et qui sont loin des sources d'énergies {Doulet, 2010}.

Gestions des systèmes d'informations

Le GRD, pour une gestion performante de son réseau, doit gérer plusieurs systèmes d'information selon ces différents services.

Le système d'information géographiques du réseau, essentiel pour le GRD, constitue la cartographie complète du réseau. Il contient toutes les informations sur la topologie et la description des différents ouvrages. Il doit être toujours mis à jour en fonction du développement et de la maintenance du réseau.

Le système d'information de téléconduite et de fonctionnement du réseau gère toutes les informations sur l'état du réseau en temps réel. Il s'agit d'un système primordial pour l'exploitation et la conduite de réseau, car il contient toutes les données des charges et des productions ainsi que l'état des interrupteurs (ouvert ou fermé) du réseau en temps réel.

Le GRD gère aussi le système d'information des flux commerciaux. Il s'agit d'un système de comptage des énergies produites et consommées par les clients finaux. Il s'appuie sur des systèmes de type AMM (Automatic Meter Management) pour mesurer les données de comptage {Doulet, 2010}. Aujourd'hui, les clients HTA et BT ≥ 36 *kVA* sont tous équipés de compteurs AMM qui mesurent les énergies consommées sur des pas de temps courts (10 minutes). Ces données sont télérelevées périodiquement par le système d'information (mais pas en temps réel). Les clients BT < 36 *kVA* sont généralement équipés de compteurs électroniques, enregistrant des index de consommation relevés annuellement. Ces

Caractéristiques des réseaux électriques classiques	Caractéristiques des réseaux électriques intelligents
Contrôle commande analogique	Contrôle commande numérique
Acheminement d'énergie unidirectionnel	Acheminement d'énergie bidirectionnel
Production centralisée	Production décentralisée
Communication sur une partie du réseau	Communication sur l'ensemble du réseau
Le consommateur est peu impliqué dans la gestion du système électrique	Le consommateur devient un consomm'acteur avec son implication dans la gestion du réseau

TABLEAU 1.2 – Développement du réseau électrique vers le smart grid.
Source : {CRE, 2021} et {DOULET et HORSON, 2019}

compteurs sont en cours de remplacement par des compteurs de type AMM.

En plus de ces systèmes d'information, nous trouvons d'autres systèmes comme le système d'information des études réseau, le système de la maintenance et des travaux et le système du suivi des processus qui gère tous les indicateurs de qualité des différents processus de GRD.

1.2.5 Réseau électrique intelligent : Smart Grid

Le smart grid ou le réseau électrique intelligent est un nouveau concept qui émerge d'une évolution du réseau électrique classique. D'après la définition du ministère de l'énergie des États-Unis, le smart grid présente une occasion de faire évoluer l'industrie énergétique dans une nouvelle ère de performance, d'efficacité et de fiabilité qui contribuera à l'économie et l'environnement {SmartGrid.gov, 2019}.

Ce réseau intelligent permet un acheminement plus efficace de l'énergie et un rétablissement plus rapide de l'électricité dans les cas d'incidents en améliorant ainsi la sécurité. Il réduit les coûts d'exploitation et gestion du réseau ce qui implique une réduction des prix d'énergie pour les consommateurs finaux. Finalement le smart grid assure une meilleure intégration des moyens de productions renouvelables et décentralisées à grande échelle {SmartGrid.gov, 2019}.

En France, la Commission de régulation de l'énergie (CRE) définit le smart grid comme « *un réseau d'énergie qui intègre des technologies de l'information et de la communication, ce qui concourt à une amélioration de son exploitation et au développement de nouveaux usages tels que l'autoconsommation, le véhicule électrique ou le stockage. Désormais, à la couche physique pour le transit d'énergie des réseaux vient se superposer une couche numérique qui joue un rôle de plus en plus important pour son pilotage* » {CRE, 2021}. Ce mariage entre le réseau physique de transit de l'électricité et le réseau numérique de communication rend le réseau plus performant et plus disponible (voir tableau 1.2). Le consommateur final devient donc un acteur dans le fonctionnement de réseau qui, grâce à des outils numériques, peut adapter ses besoins énergétiques en fonction de la production, ce qui change le fonctionnement traditionnel de pilotage de l'offre de la production en fonction de la demande (la consommation) {CRE, 2021}.

Le smart grid est donc un ensemble de systèmes physiques et numériques dont le but est d'améliorer l'efficacité énergétique et l'exploitation de réseau (voir figure 1.5). L'objectif du système numérique n'est pas uniquement de rendre le réseau communicant, mais aussi de mettre à profit toutes les possibilités qu'offrent les données collectées pour les acteurs de réseau. Ainsi, les données de réseau permettent d'intégrer de nouvelles méthodes et approches d'optimisation et d'exploitation de réseau. Par exemple,

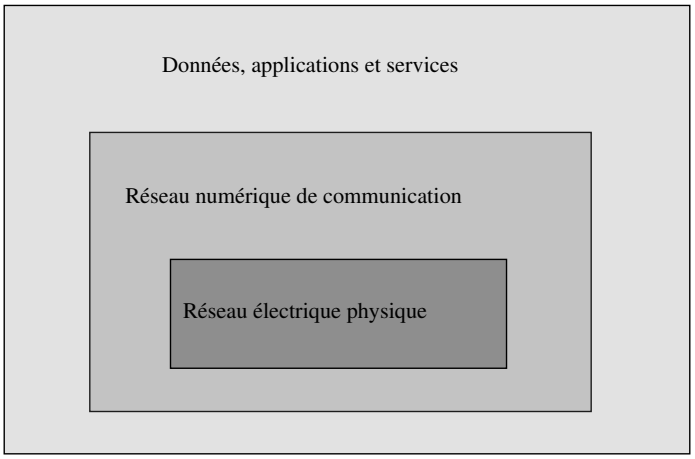


FIGURE 1.5 – Différents niveaux du Smart Grid

grâce à des applications de l'intelligence artificielle, le smart grid peut réduire les coûts d'acheminement et gérer les productions EnR intermittentes et décentralisées. La présente thèse s'inscrit dans le contexte des expérimentations de SRD dans le domaine des Smart Grids. Elle porte plus précisément sur le développement de méthodes de prévision et d'estimation des productions EnR intermittentes et de la consommation d'électricité.

1.3 Contexte et enjeux de la thèse

Le réseau français de distribution de l'électricité a été construit dans une perspective de gestion centralisée et unidirectionnelle. Aujourd'hui, avec l'intégration des EnR dans le réseau, cette gestion est devenue décentralisée et bidirectionnelle. Conséquemment, l'état instantané du réseau dépend à la fois de la consommation et de la production locale.

Ce changement implique une modification majeure dans la gestion du réseau et l'utilisation de l'énergie. D'une part, l'usage de l'électricité connaît des évolutions importantes, comme le stockage et la mobilité électrique. D'autre part, l'intégration des producteurs EnR en France connaît un développement important ces dernières années. Ces producteurs ont évolué de +84 % entre 1990 et 2019. Ils ont présenté en 2019 une part de 17.2 % dans la consommation finale brute d'énergie {SDES, 2021a}. Dans le cas de SRD, le parc EnR a injecté en 2020 une énergie totale de 725 GWh, soit 55% de l'énergie totale consommée cette année-là. Tous ces nouveaux modes de consommation et de production ont rendu la gestion de réseau de distribution de plus en plus complexe. Le GRD se trouve aujourd'hui dans l'obligation de gestion de ces nouveaux usages et mutations dans le système énergétique.

Pour faire face à ces nouveaux défis de gestion, l'une des solutions est de renforcer le réseau ou de construire des nouveaux ouvrages. Par exemple, avec le doublement des lignes électriques ou la construction des nouveaux postes sources, le GRD peut intégrer des nouveaux producteurs ou augmenter la part d'énergie consommée localement. Cependant, ces solutions sont coûteuses financièrement, non optimales et difficilement réalisables. L'une des alternatives pour le GRD est d'adopter des nouvelles technologies de gestion de réseau, notamment les nouvelles techniques du domaine du smart grid.

Le smart grid qui est à la fois communicant et intelligent permet une interaction avec les différents composants du réseau. Cette automatisation du réseau permet de savoir ou d'estimer en temps réel l'état et l'activité de tous les acteurs du système électrique. Il permet de développer de nouveaux services et nouveaux outils d'aide à la décision, afin de satisfaire les besoins et les contraintes de gestion.

C'est dans ce contexte que SRD veut moderniser son réseau électrique en privilégiant le développement des technologies innovantes de pilotage du réseau. Ali Zazou {2017} présente les résultats d'une thèse Cifre (2014-2017) en collaboration avec le laboratoire LIAS, où un outil d'optimisation dynamique du schéma d'exploitation du réseau de distribution d'électricité haute tension (HTA) a été développé. Cet outil a permis à l'outil de conduite de SRD de passer du statut « outil de distribution multi-cibles » au statut « outil de distribution multi-cibles et de collecte multi-sources ». Cette thèse s'inscrit dans le cadre du projet IMAGE (système Intelligent de Management et Gestion de l'Energie) de SRD.

1.3.1 Projet IMAGE

Dans le cadre de ses projets de recherche et développement dans le domaine des réseaux électriques intelligents, SRD a développé le projet IMAGE. Il s'agit d'un système d'optimisation de réseau composé de plusieurs éléments. Ce système doit permettre au GRD de bien gérer et exploiter son réseau en intégrant une énergie renouvelable décentralisée. La consommation de l'énergie produite localement est bénéfique pour le GRD. En effet, elle sollicite moins le réseau amont (réseau de transport) tant en consommation qu'en refoulement ; elle génère donc moins de coûts pour la collectivité, notamment de pertes. En outre, augmenter la part d'énergie renouvelable consommée localement permet d'accroître les capacités de raccordement de nouvelles installations.

Ce système innovant vise à déterminer le schéma d'exploitation optimal du réseau électrique, en intégrant les EnR et en respectant les différentes contraintes techniques et réglementaires de gestion du système électrique. Il doit déterminer une topologie optimale pour un instant de fonctionnement donné en exploitant plusieurs données d'entrée, notamment, les données statiques comme les caractéristiques physiques et topologiques du réseau et les données dynamiques comme les données de la tension, du courant ou de la puissance (voir figure 1.6). Son efficacité d'optimisation est à la fois la différence entre le schéma optimal résultant et le schéma normal réel, et le temps nécessaire au calcul. Cette efficacité dépend de la complexité des données d'entrée et des différentes hypothèses de modélisation de ces données.

1.3.2 Problématique et objectifs de la thèse

Cette thèse s'est déroulée dans le cadre d'une convention CIFRE entre SRD, le gestionnaire de réseau de distribution de l'électricité du département de la Vienne, et le laboratoire LIAS à Poitiers. SRD s'est inscrit ces dernières années dans une démarche de recherche et développement dans le domaine du smart grid, notamment avec le projet IMAGE. Il a modernisé ainsi ses outils de conduite et d'exploitation de réseau avec un nouvel outil SCADA (Supervisory control and data acquisition). Les enjeux de ces projets pour SRD sont importants, puisqu'ils permettent à la fois sa transition énergétique et sa transition numérique. Le laboratoire LIAS apporte son expertise dans le domaine de l'ingénierie de données, de l'optimisation et du génie électrique. L'exploitation combinée de ces domaines est une particularité pour le laboratoire car elle réunit les compétences de ces trois équipes.

SRD cherche, avec ce projet de thèse, à s'équiper d'un outil d'estimation et de prévision des productions des EnR et de la consommation de l'électricité de son réseau (voir figure 1.6). L'objectif de cette thèse est de résoudre certaines problématiques de modélisation de données en lien avec l'outil de l'optimisation, à savoir :

1. la conception d'un processus de sélection de données pertinentes parmi toutes les données archivées,

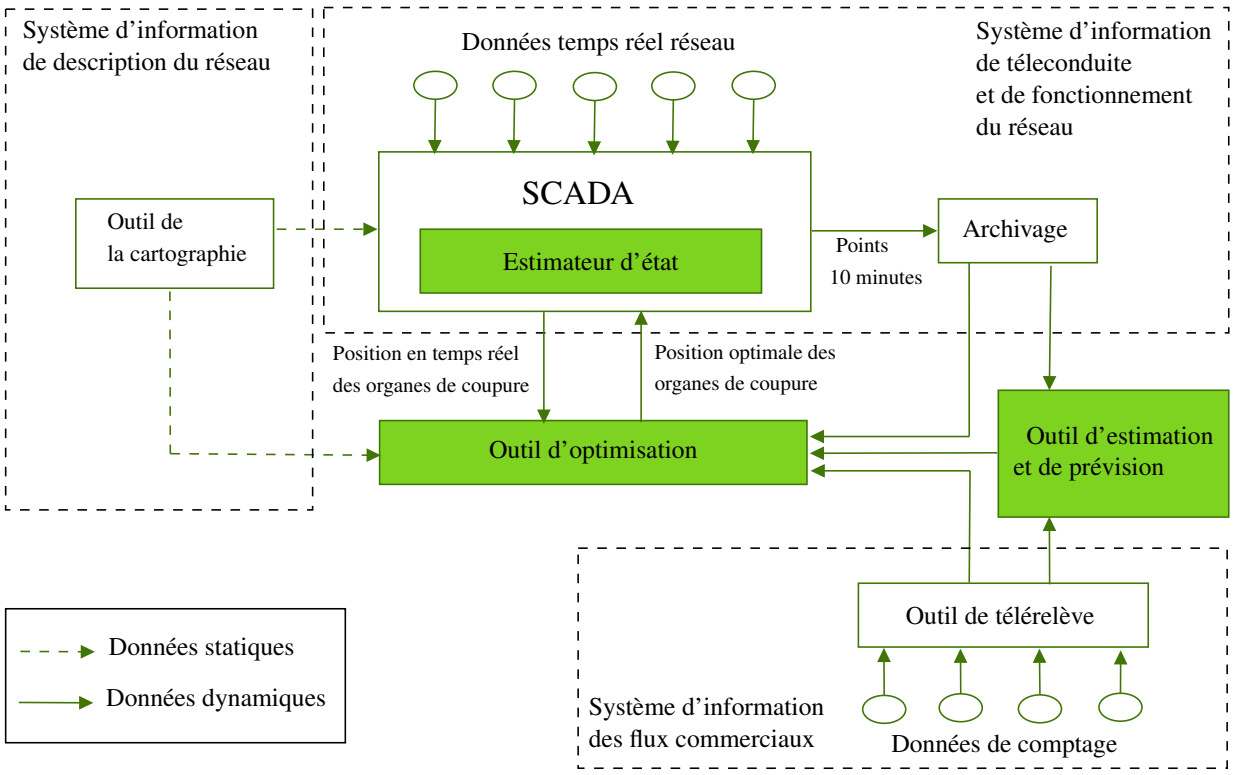


FIGURE 1.6 – Les différents composants du système IMAGE.
L'outil IMAGE se compose de 3 sous-parties : un estimateur d'état (ou calculateur d'état), un optimiseur, et un outil d'estimation et de prédiction. Ces trois outils sont interfacés avec les différents systèmes d'information de SRD.

2. l'étude de l'influence des différentes structures possibles de données d'entrée, ainsi que leur volume et leur niveau de détail sur l'efficience de l'optimisation,
3. la proposition d'une méthodologie d'estimation des données non disponibles en exploitant les différentes données internes et externes,
4. la définition d'un algorithme de prévision de la production des EnR et de la consommation, en exploitant les données archivées et des facteurs externes.

1.3.3 Contributions et organisation du manuscrit

Le présent manuscrit présente nos travaux réalisés durant la thèse. Il est organisé en cinq chapitres :

Dans le deuxième chapitre, des éléments de vocabulaire métier sont exposés. Ensuite, les différentes données utilisées dans la thèse sont présentées, ainsi que leur méthodologie d'acquisition et de préparation. Finalement, un exposé sur la qualité de ces données, avec des exemples d'anomalies, est réalisé. Cette qualité de données est essentielle dans un tel projet fondé sur la donnée.

Dans le troisième chapitre, nous proposons une approche de sélection des données de la consommation d'énergie les plus pertinentes en étudiant leur influence sur l'efficacité de l'optimisation. Cette méthodologie permet de détecter des similitudes (ou des dissimilitudes) dans les données afin de les regrouper dans des groupes homogènes. Un algorithme original est ainsi développé permettant un regroupement de données en se fondant sur des contraintes métiers des erreurs maximales entre les différentes données à regrouper.

Le quatrième chapitre présente une étude d'estimation de données de la production d'énergie des producteurs photovoltaïques distribuées dans un réseau de distribution d'électricité. Cette estimation

est fondée sur les méthodes d'interpolation spatiale en utilisant les différentes données disponibles des autres producteurs comme références. L'évaluation des méthodes est réalisée avec les données réelles des producteurs de SRD.

Dans le cinquième chapitre, les approches de prévision de la production solaire sont élaborées et évaluées avec des données réelles. Le but de cette étude est d'élaborer une prévision ponctuelle court terme d'un horizon d'une heure pour gérer l'intermittence de la production solaire et une prévision probabiliste long terme pour planifier et optimiser le réseau sur un horizon d'un mois à trois mois. Les prévisions ont été validées ainsi par un estimateur d'état du réseau.

Chapitre 2

Données et généralités

Les données utilisées dans la thèse sont exposées dans ce chapitre. Des éléments de vocabulaire du génie électrique ainsi que des fondements théoriques des séries temporelles sont présentés. Les données sont divisées en deux catégories : des données internes de SRD incluant toutes les données constituées à SRD par les différents systèmes d'informations des différents services et des données externes issues des bases de données ouvertes (open data). Une synthèse de la qualité de ces données est ainsi détaillée.

Sommaire

2.1 Introduction du chapitre

Dans le chapitre précédent, le réseau électrique français et ses différents composants ont été présentés, ainsi que les différentes missions des gestionnaires du réseau de distribution.

Dans ce chapitre, une présentation des différentes données utilisées dans la thèse est réalisée, ainsi que leur méthodologie de prétraitement. Les données sont divisées en deux catégories : des données internes de SRD incluant toutes les données constituées à SRD par les différents systèmes d'informations des différents services et des données externes issues des bases de données ouvertes (open data). Une synthèse de la qualité de ces données est ainsi exposée.

2.2 Éléments de vocabulaire

2.2.1 Électrotechnique

Selon le dictionnaire en ligne de l'académie française l'électrotechnique est : « *l'Étude des applications techniques et industrielles de l'électricité.*» {Académie, 2021}

L'électrotechnique, qui s'appelle également le génie électrique, est une science étudiant l'application technique et industrielle de l'électricité dans toute la chaîne de valeur de l'énergie, de la production jusqu'à la consommation finale. Elle utilise de multiples techniques issues de domaines différents comme la mécanique, l'automatique, l'informatique industrielle, etc. pour permettre la production, le transit et la consommation de l'électricité {Marty *et al.*, 2005}.

Dans l'étude de la distribution de l'électricité, nous nous intéressons au calcul de plusieurs grandeurs des phénomènes électriques circulant dans le réseau. Ces grandeurs sont essentiellement le courant et la tension. En France, le réseau de distribution de l'électricité fonctionne sous un régime alternatif, sinusoïdal et triphasé avec une fréquence f de 50 Hz. La tension sinusoïdale $u(t)$ en volt (V) et le courant sinusoïdal $i(t)$ en Ampère (A) sont donnés par

$$u(t) = u_m \, sin(2\pi f t)$$

$$i(t) = i_m \, sin(2\pi f t + \varphi)$$

Où u_m et i_m sont les valeurs crêtes de la tension et du courant. φ est l'angle de déphasage en degrés, elle représente le déphasage entre le courant et la tension. f est la fréquence en hertz et t est le temps en seconde.

Dans l'étude des réseaux électriques, nous nous intéressons souvent à d'autres grandeurs calculées à partir du courant et de la tension. Ces grandeurs sont la puissance active, la puissance réactive et la puissance apparente.

Dans un circuit à courant alternatif, la valeur efficace de la puissance apparente S en Volt-ampère (VA) est égale au produit de la tension efficace U et du courant efficace circulant I, soit

$$S = UI$$

avec $U = \frac{u_m}{\sqrt{2}}$ et $I = \frac{i_m}{\sqrt{2}}$

La puissance active P est égale à

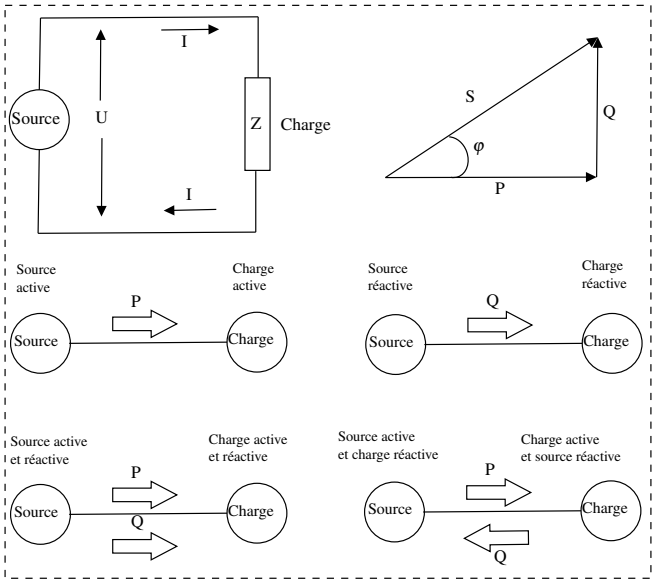


FIGURE 2.1 – Puissance active, puissance réactive et puissance apparente (adapté de {Wildi et Sybille, 2000}).

$$P = UIcos(\varphi)$$

La puissance réactive est égale à

$$Q = UIsin(\varphi)$$

Les trois puissances sont liées par la formule suivante

$$S = \sqrt{P^2 + Q^2}$$

Le terme $cos(\varphi)$ est appelé facteur de puissance. Le GRD cherche à le maintenir autour de 0,9 car plus il est petit, plus la puissance apparente S transitant sur le réseau est grande pour une même puissance active P ; les pertes par effet Joule sont alors plus importantes et des problèmes de baisse ou d'élévation de tension peuvent apparaître.

La figure 2.1 illustre ces trois puissances dans un circuit électrique constitué d'une source et d'une charge. Une charge active est une charge absorbant une puissance active, lorsque le courant i est en phase avec la tension u et que le courant circule de la source vers la charge. Les sources actives sont généralement des générateurs à courant alternatif. Les charges actives sont généralement des éléments résistifs comme le chauffage, le four, etc. Une charge réactive (appelée charge inductive) est une charge absorbant une puissance réactive Q, lorsque le courant i est en retard de 90 degrés avec la tension u. Les charges réactives sont principalement des éléments nécessitant un champ magnétique alternatif comme les bobines, les moteurs, etc. Une source réactive (appelée charge capacitive) est une source produisant une puissance réactive Q, lorsque le courant i est en avance de 90 degrés avec la tension u. Les sources réactives sont principalement les condensateurs. Pour plus de détails sur les calculs électriques, se référer au livre de Wildi et Sybille {2000}.

Pour une gestion efficace du réseau électrique, le GRD doit connaître les valeurs de ces trois puis-

sances acheminées dans tous les points de réseau. Cela permet de respecter la qualité de fourniture de l'électricité et d'éviter la surcharge des lignes électriques. Comme le réseau est instrumenté partiellement de capteurs mesurant ces grandeurs électriques, dans un niveau haut de réseau, une méthode d'estimation de ces puissances dans les autres points du réseau (dans un niveau bas du réseau) est utilisée. Cette méthode se fonde sur le calcul du rapport entre la puissance active P consommée dans un point du réseau et la puissance souscrite *PST* totale cumulée en ce point. Au niveau d'un départ HTA *i* par exemple, ce rapport *k* est égale à

$$k = \frac{P_i}{PST_i}$$

Ce ratio étant homogène au coefficient de foisonnement utilisé pour le dimensionnement du réseau, il sera appelé "foisonnement", "coefficient de foisonnement" ou "foisonnement instantané" dans la suite du document. Il sert à estimer donc les charges inconnues à partir de la charge totale en supposant qu'elles sont proportionnelles à la puissance souscrite. P_i est la consommation réelle au niveau du départ et PST_i est la somme des puissances souscrites de tous les clients raccordés à ce départ $PST_i = \sum_{j=1}^{N} PS_j$ (N est le nombre total des clients et PS_j la puissance souscrite de chaque client). La charge réelle P_i d'un départ HTA égale la charge mesurée P_{TM} au niveau de ce départ moins la somme de toutes les productions HTA et BT raccordées au départ. La figure 2.2 illustre les données de charge et de production d'un départ HTA *i*.

$$P_i = P_{TM} - \sum(Prod_{HTA} + Prod_{BT})$$

Le foisonnement représente un pourcentage de la consommation par rapport à la puissance souscrite maximale des clients. Le foisonnement instantané $k(t)$ est le foisonnement calculé pour chaque instant t

$$k(t) = \frac{P_i(t)}{PST_i}$$

Ce foisonnement est utilisé pour estimer la puissance consommée dans les niveaux inférieurs du réseau non instrumentés par des compteurs permettant la mesure des données de puissance avec une granularité temporelle fine. Ces niveaux inférieurs sont les postes HTA/BT dont une partie desservant les grands clients (postes clients). Dans cette estimation, nous supposons que le foisonnement dans un niveau supérieur $k_{sup}(t)$ de réseau est le même dans un niveau inférieur de réseau $k_{inf}(t)$.

$$k_{sup}(t) = k_{inf}(t)$$

Cela implique que

$$\frac{P_{sup}(t)}{PST_{sup}} = \frac{P_{inf}(t)}{PST_{inf}}$$

Donc, la puissance dans le niveau inférieur est égale à

$$P_{inf}(t) = k_{sup}(t) PST_{inf}$$

.

$P_{inf}(t)$ est représentée par $Conso_{HTA}$ et $Conso_{BT}$ dans la figure 2.2. Le niveau supérieur peut être aussi une agence ou un poste source, les niveaux inférieurs sont les postes HTA/BT raccordés à ce niveau

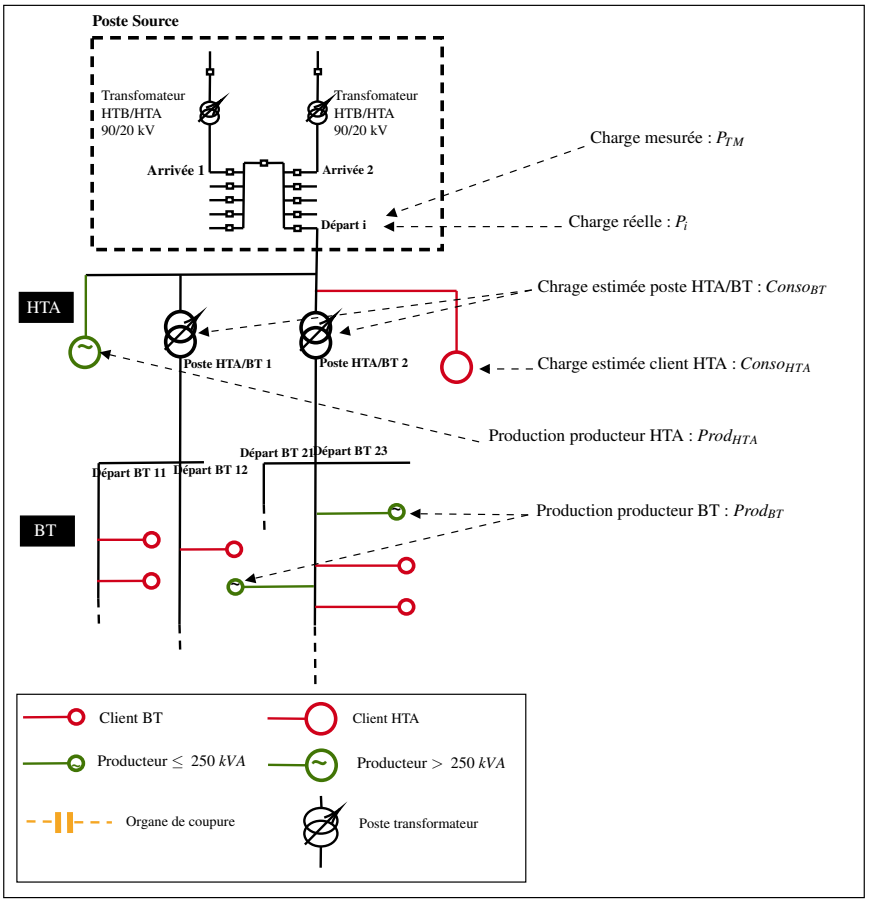


FIGURE 2.2 – Schéma de données d'un départ HTA

supérieur.

La puissance réactive des niveaux inférieurs est obtenue ensuite à partir de la puissance active estimée P_{inf} en supposant que le facteur de puissance $cos(\varphi)$ au niveau des postes HTA/BT est égale à celui du départ HTA de rattachement.

$$Q_{inf} = UIsin(\varphi)$$

Or $sin(\varphi) = \pm\sqrt{1 - cos^2(\varphi)}$, on obtient donc

$$Q_{inf} = \pm UI\sqrt{1 - cos^2(\varphi)}$$

Les valeurs P_{inf} et Q_{inf} estimées au niveau de chaque poste HTA/BT sont nécessaires pour optimiser ou estimer l'état du réseau pour un point de fonctionnement à un instant t.

L'optimiseur utilise ce foisonnement pour estimer la charge des niveaux inférieurs. Aujourd'hui, la solution développée utilise le foisonnement au niveau d'une agence pour pouvoir optimiser le réseau pour un point de fonctionnement. Pour une optimisation utilisant une courbe de charge, il utilise le foisonnement au niveau des postes sources.

2.2.2 Estimation d'état du réseau

L'estimation d'état de tous les points du réseau électrique est réalisée par des méthodes dites de répartition de charge : "Loadflow" ou "Powerflow". Ces méthodes varient selon la nature de la structure du réseau électrique. Ce sont principalement des méthodes issues du domaine de l'analyse numérique. Dans la littérature, il existe de nombreux outils informatique permettant le calcul de loadflow, ils sont appelés estimateurs d'état de réseau. Vogt *et al.* {2018} ont présenté une synthèse approfondie des différents estimateurs d'état utilisés dans le domaine de smartgrid. Certains de ces calculateurs d'états sont payants comme l'outil PowerFactory {Gonzalez-Longatt et Rueda, 2014}, et d'autres sont open source comme Matpower {Zimmerman *et al.*, 2010} développé en Matlab et Pandapower {Thurner *et al.*, 2018} développé en Python. Pour pouvoir calculer le loadflow, les différentes méthodes de calcul nécessitent la topologie complète du réseau avec toutes les informations sur les caractéristiques des lignes électriques, des productions et des charges. Au niveau du réseau HTA par exemple avec une structure radiale, les postes HTA/BT présentent les nœuds de réseau. Ils constituent les charges et dans le cas de présence de producteurs, ils constituent aussi des productions. Les postes sources sont considérés comme des sources de production infinies. Tous ces nœuds sont liés par des lignes électriques. L'estimateur d'état utilise comme entrée les puissances actives et réactives estimées ou mesurées dans chaque charge et production et la résistance de chaque ligne électrique, afin de calculer en sortie les pertes, les variations de tensions et d'autres paramètres électriques de réseau.

Dans notre thèse, nous avons utilisé l'outil Pandapower {Thurner *et al.*, 2018} permettant une automatisation des simulations utilisant des courbes temporelles. Dans les différentes simulations réalisées dans les chapitres 3 et 5, nous avons utilisé le réseau électrique de l'agence 6 de SRD. Cette agence est desservie par 23 départs HTA de 4 postes sources différents.

2.2.3 Séries temporelles

Dans le domaine de l'électrotechnique, les données temporelles de la consommation et de la production sont appelées courbes de charge et courbes de production. Elles sont constituées d'une suite de valeurs de charge ou de production mesurées à un instant t décrivant l'évolution de cette charge ou production dans le temps. Dans la théorie de la statistique, elles sont appelées séries temporelles ou séries chronologiques. Pour une étude détaillée des séries temporelles, nous pouvons se référer au livre de Brockwell et Davis {2016}.

Une série temporelle y est un ensemble de valeurs $\{y_t\}$ mesurées sur une période T, $t \in T$. Une série temporelle peut être discrète ou continue. Dans le cas d'une série discrète les observations sont collectées dans des intervalles de temps fixes, tandis que pour les séries continues les informations sont collectées en continu sur une période temporelle avec une granularité continue. Dans la plupart des cas, en particulier notre cas d'étude, les séries sont discrètes. Les différentes valeurs enregistrées sont collectées avec un intervalle fixe. Cet intervalle temporel est déterminé selon les moyens informatiques de traitement et de stockage utilisés, ainsi que de la variation naturelle de la grandeur mesurée. Le volume de données stockées dépend du nombre de séries enregistrées et de leur granularité de mesure.

Dans l'analyse d'une série temporelle, nous cherchons à déterminer un modèle probabiliste approprié pour les données. Nous supposons que les observations y_t sont des réalisations d'une certaine variable aléatoire Y_t. Un modèle d'une série temporelle pour la valeur observée y_t est une spécification de la distribution d'une variable aléatoire Y_t dont y_t est supposée une réalisation {Brockwell et Davis, 2016}.

Un modèle probabiliste complet d'une série temporelle pour une séquence de variables aléatoires

$Y_1, Y_2, ...$ doit préciser toutes les distributions du vecteur aléatoire $(Y_1, Y_2, ..., Y_n)$, pour tout $n = 1, 2, ...$ c'est-à-dire qu'il doit de façon équivalente indiquer toutes les probabilités

$$P[Y_1 \leq y_1, ..., Y_n \leq y_n]$$

avec $-\infty < y_1, ..., y_n < \infty$ et $n = 1, 2, ...$

Une telle modélisation d'une série temporelle est rarement utilisée dans la réalité puisqu'elle nécessite l'estimation de plusieurs paramètres à partir de données disponibles. Dans l'analyse de séries temporelles, nous ne présentons en général que les moments du premier et du second ordre des distributions conjointes. C'est-à-dire, nous déterminons les valeurs de $E[Y_t]$ et les produits $E[Y_{t+h}Y_t]$ pour tous $t = 1, 2, ...$ et $h = 0, 1, 2, ...$

Un exemple de modèle est le bruit aléatoire. Ce modèle est un modèle basique de modélisation d'une série temporelle dans laquelle les observations sont simplement des variables aléatoires indépendantes et identiquement distribuées (*i.i.d*) avec une moyenne nulle. Nous pouvons écrire, pour tout entier n et pour toutes observations $y_1, y_2, ..., y_n$:

$$P[Y_1 \leq y_1, ..., Y_n \leq y_n] = P[Y_1 \leq y_1]...P[Y_n \leq y_n] = F(y_1)...F(y_n)$$

où F est la fonction de répartition (cumulative distribution function CDF) de chaque variable aléatoire Y_t. Ce modèle basique est un modèle sans mémoire, c'est-à-dire qu'il n'y a aucune dépendance entre les différentes observations.

$$P[Y_{n+h} \leq y | Y_1 = y_1, ..., Y_n = y_n] = P[Y_{n+h} \leq y]$$

Lorsque les variables aléatoires $Y_1, Y_2, ...$ sont non corrélées entre elles, nous appelons ce modèle un bruit blanc. Il est clair que chaque modèle bruit *i.i.d* est un bruit blanc mais pas inversement, car l'indépendance entre les variables aléatoires implique la non corrélation, mais pas forcement l'inverse.

Cela signifie que les informations sur les $Y_1, ..., Y_n$ ne sont pas utiles pour prévoir le comportement de Y_{n+h}. Ce modèle de bruit *i.i.d* est inintéressant pour réaliser une prévision, cependant il joue un rôle essentiel comme élément de base de construction d'autres modèles plus complexes.

Nous disons qu'une série temporelle Y_t est stationnaire, lorsque la série décalée dans le temps Y_{t+h} a les mêmes caractéristiques statistiques que Y_t pour tout h dans $\mathbb{N}$. Nous pouvons exprimer la stationnarité d'une série temporelle en utilisant les moments du premier et du second ordre {Brockwell et Davis, 2016}. Soit Y_t une série temporelle avec $E[Y_t^2] < \infty$. La fonction moyenne de Y_t se définit par

$$\mu_Y(t) = E[Y_t]$$

La fonction de covariance de Y_t se définit par

$$\gamma_Y(r, s) = Cov(Y_r, Y_s) = E[(Y_r - \mu_Y(r))(Y_s - \mu_Y(s))]$$

pour tout r et s dans $\mathbb{N}$. Nous disons que la série Y_t est une série faiblement stationnaire lorsque

(i) $\mu_Y(t)$ est indépendant du temps t

(ii) $\gamma_Y(t+h, t)$ est indépendant de t pour chaque h

La stationnarité signifie que les propriétés statistiques d'une série temporelle ne changent pas dans le temps.

Il existe de multiples familles de modèles dans la modélisation statistique des séries temporelles. Un bon modèle est un modèle qui se rapproche au mieux de la réalité de la série.

En général, une série temporelle peut être décomposée en plusieurs composantes. Une première composante m_t décrit la tendance de la série, c'est-à-dire son évolution globale dans le temps. Cette tendance peut être par exemple linéaire ou quadratique. Une deuxième composante s_t décrit la saisonnalité de la série, lorsque le phénomène étudié est périodique dans le temps d'une période connue. Finalement, un terme de bruit ε_t définit la partie aléatoire de la série. La série temporelle Y_t peut donc se décomposer, dans le cas additif par :

$$Y_t = m_t + s_t + \varepsilon_t$$

et dans le cas multiplicatif par :

$$Y_t = m_t \times s_t \times \varepsilon_t$$

Nous pouvons estimer la tendance par exemple en appliquant la méthode de moyenne mobile qui consiste à diviser la série sous forme de multiples intervalles égaux pour ensuite calculer la moyenne de la série sur chaque intervalle. Les moyennes mobiles nous permettent d'estimer la tendance globale de la série en supprimant les fluctuations et les variations dans la série.

$$MA_t = \frac{1}{m} \sum_{i=-k}^{k} Y_{t+i}$$

où m est l'ordre de la moyenne mobile $m = 2k+1$ (m représente la taille de l'intervalle).

Il existe plusieurs autres méthodes pour estimer à la fois la tendance et la saisonnalité comme les méthodes STL {Cleveland *et al.*, 1990}, X11 {Shiskin, 1967}, SEAT {Dagum et Bianconcini, 2016}, etc. Pour plus d'informations sur les méthodes de décomposition des séries temporelles, nous pouvons nous référer au livre de Dagum et Bianconcini {2016}.

2.2.4 Métriques de comparaison

Pour pouvoir comparer deux séries temporelles entre elles, des métriques statistiques sont utilisées pour mesurer leur degré de ressemblance. Ces métriques sont utilisées dans les différents projets élaborés dans le cadre de cette thèse pour quantifier la qualité d'estimation ou de prévision des modèles proposés.

Nous notons $e_{s,t}$ l'erreur de prévision ou d'estimation d'un modèle pour une série temporelle $y_{s,t}$

$$e_{s,t} = (y_{s,t} - \hat{y}_{s,t}) \tag{2.1}$$

avec $y_{s,t}$ les valeurs réelles observées, $\hat{y}_{s,t}$ les valeurs prévues ou estimées par le modèle, t dans $\{1,...,T\}$ et s dans $\{1,...,S\}$, S est le nombre total des séries. La série $y_{s,t}$ peut être une courbe de production ou une courbe de charge.

L'erreur moyenne absolue (MAE) est définie par

$$MAE = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} |e_{s,t}|$$

Cette mesure statistique représente la moyenne absolue de l'erreur entre la courbe observée et la courbe estimée. Elle ne donne pas l'indication sur le signe de l'erreur. Par exemple, dans le cas de la production PV, elle ne permet pas de savoir si l'énergie produite est sur-estimée ou sous-estimée.

L'erreur de biais moyen (MBE) est définie par

$$MBE = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} e_{s,t}$$

Les mesures de *MBE* peuvent être utilisées pour déterminer si la production est sous-estimée ou sur-estimée. Cependant les erreurs de signes opposés se compensent.

Nous avons également utilisé l'erreur quadratique moyenne (Root-mean-square deviation RMSE), qui est la métrique la plus utilisée dans le cas de comparaison entre séries temporelles. Sa particularité est la pénalisation des grandes erreurs individuelles d'estimation. Comme la *MAE*, elle est cependant sans signe.

$$RMSE = \sqrt{\frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} e_{s,t}^2}$$

Nous avons aussi la métrique MAPE (Mean absolute percentage error) qui est donnée par

$$MAPE = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} \left| \frac{e_{s,t}}{y_{s,t}} \right|$$

L'avantage de MAPE est de donner un indicateur en pourcentage de l'erreur par rapport à la quantité réelle observée. Cependant, dans le cas de la production solaire par exemple, une production nulle ou infime est mesurée dans certaines périodes d'année. Donc, il est impossible dans ce cas de calculer la MAPE. Nous n'avons utilisé MAPE que dans les résultats de l'estimateur d'état.

Nous avons calculé deux autres métriques demandées par les agents de SRD pour quantifier la sur-prévision ou la sous-prévision d'un modèle. Pour cela, nous avons calculé les métriques suivantes

$$MBE^{+} = \frac{1}{N} \sum_{s=1}^{S} \sum_{t=1}^{T} e_{s,t} \mathbb{1}(\{e_{s,t} > 0\}) \tag{2.2}$$

avec $N = \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbb{1}(\{e_{s,t} > 0\}$, et

$$MBE^{-} = \frac{1}{N} \sum_{s=1}^{S} \sum_{t=1}^{T} e_{s,t} \mathbb{1}(\{e_{s,t} < 0\}) \tag{2.3}$$

avec $N = \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbb{1}(\{e_{s,t} < 0\}$

La MBE^{+} indique une moyenne de la sous-prévision ou de la sous-estimation d'un modèle lorsque y_t est supérieur à $\hat{y}_t$. Inversement MBE^{-} calcule une moyenne de la sur-prévision lorsque y_t est inférieur à $\hat{y}_t$.

Dans le cas de la production, nous avons normalisé ces métriques par la capacité maximale de production des producteurs. $e_{s,t}$ est divisé dans les équations précédentes par la capacité C_s du producteur s.

$$ne_{s,t} = \frac{e_{s,t}}{C_s}$$

L'erreur $ne_{s,t}$ indique le pourcentage d'erreur d'un modèle par rapport à la capacité installée des producteurs.

Au final, cette erreur est multipliée par 100 pour représenter les métriques par un pourcentage. Par exemple, dans le cas de la RMSE, nous obtenons la nRMSE par

$$nRMSE = 100 \times \sqrt{\frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} ne_{s,t}^2}$$

Nous obtenons de la même manière les autres les métriques $nMBE^{+}$+ , $nMBE^{-}$, $nMBE$ et $nMAE$

Dans le cas d'un producteur s, l'erreur $nRMSE_s$ d'un modèle se définit par

$$nRMSE_s = \sqrt{\frac{1}{T} \sum_{t=1}^{T} ne_{s,t}^2}$$

De même nous définissons $nMBE_s^{+}$, $nMBE_s^{-}$, $nMBE_s$ et $nMAE_s$.

2.3 Données

2.3.1 Données internes

Les données internes sont gérées par plusieurs systèmes d'informations (SI) des différents services de SRD. Dans le cadre de cette thèse, nous avons utilisé essentiellement les données de SI de la cartographie, le SI des flux commerciaux et le SI de la téléconduite et de fonctionnement du réseau. Ces données sont de nature différente, statique ou dynamique. Ce sont généralement des données tabulaires. C'est-à-dire, qu'elles sont représentées sous forme de tableaux, notamment des fichiers CSV ou Excel. Elles sont stockées soit directement sous forme fichier, soit dans des systèmes de gestion de bases de données.

Données de description du réseau

Les données de description du réseau de SRD sont gérées par le service de la cartographie réseau. Elles sont utilisées par les différents services pour la modélisation et les calculs sur le réseau.

Les données de la cartographie sont gérées par l'outil Editop. Il permet de transformer ces données géographiques sous forme d'un graphe dont les arcs sont les lignes électriques et les sommets sont les différents appareils électrotechniques (poste source, poste HTA/BT, organe de coupure, ...etc.). Ce graphe contient des informations sur les différents ouvrages selon leur type.

Les lignes électriques sont représentées par leur longueur en mètre, leur nature (aérienne ou souterraine), leur section en mm^2 et leur matière, ainsi que les postes qu'elles relient. Ces postes peuvent être des poste sources, des postes HTA/BT ou d'autres types d'objets ponctuels de communication ou de protection du réseau.

Concernant les postes HTA/BT, nous trouvons toutes les informations sur le nombre de clients raccordés au poste et leur puissance souscrite cumulée selon les catégories tarifaires (BT < 36 kVA, BT ≥ 36 kVA ou HTA). Nous trouvons aussi les informations sur le nombre de producteurs rattachés à ce poste et leur capacité de production cumulée. De même, pour les postes HTA/BT client nous avons la puissance souscrite du client selon la saison, été ou hiver. En effet, certains clients choisissent leur puissance souscrite en fonction de la saison, par exemple certains agriculteurs utilisant l'irrigation en été, choisissent leur abonnement en fonction de ce besoin.

De même que pour les postes HTA/BT, nous avons toutes les caractéristiques des organes de coupure (les interrupteurs) dans le réseau. Pour chaque organe de coupure, nous trouvons ses coordonnées GPS, son type (télécommandé ou manuel) et son état (ouvert ou fermé). L'état de ces organes de coupure détermine la configuration du réseau.

De la même façon, nous avons les informations sur toutes les productions raccordées au réseau HTA ou réseau BT, telles que les cordonnées GPS, les capacités maximales, les dates de mise en service et les numéros de PDL (point de livraison) qui constituent des identifiants uniques des producteurs. En outre, les caractéristiques de tous les autres objets de réseau, comme les boîtes de jonction, les éléments de communication, les jeux de barres, etc, sont également disponibles.

Données de comptage

Les données de comptage de SRD font partie du système d'information des flux commerciaux, elles sont appelées aussi index de consommation ou de production. Elles sont divisées en deux catégories.

Premièrement, nous avons les données des comptages des petits consommateurs et les petits producteurs. Pour les consommateurs, les données sont collectées chaque année via des campagnes de relève d'index en présentiel chez les clients. Les producteurs communiquent leurs index selon une période donnée.

Deuxièmement, nous avons les données des moyens et grands consommateurs et producteurs. Les données sont collectées avec une granularité temporelle de 10 minutes via un outil de télérelève géré par le service métrologie de SRD. Cet outil de télérelève appelé Saturne (figure 2.3), est un outil multi-protocoles pour collecter les données depuis les compteurs communicants. Il permet la télérelève des courbes de charge et de production afin de les afficher, de les exporter ou de les communiquer directement par messagerie ou par internet. Il permet aussi d'autres types d'opérations comme la correction des courbes dans le cas d'anomalies ou l'agrégation de ces courbes selon certains critères. En outre, l'outil réalise aussi le profilage qui s'appuie sur une technique de modélisation et d'estimation des consommations ou des productions pour un groupe de clients. Ces clients sont généralement des clients qui n'ont pas de compteurs communicants, le profilage permet donc d'estimer la consommation ou la production avec une granularité fine de 10 min.

Dans le cadre de la thèse, nous avons utilisé les courbes de production de tous les moyens et grands producteurs avec une granularité de 10 minutes et un historique de la date de mise en service du producteur jusqu'à décembre 2020.

Données de télécoduite et de fonctionnement réseau

Nous avons vu dans le chapitre 1 que le bureau d'information de téléconduite est chargé de la conduite et de l'exploitation du réseau. Ce service surveille et conduit l'ensemble de réseau et ses différents équi-

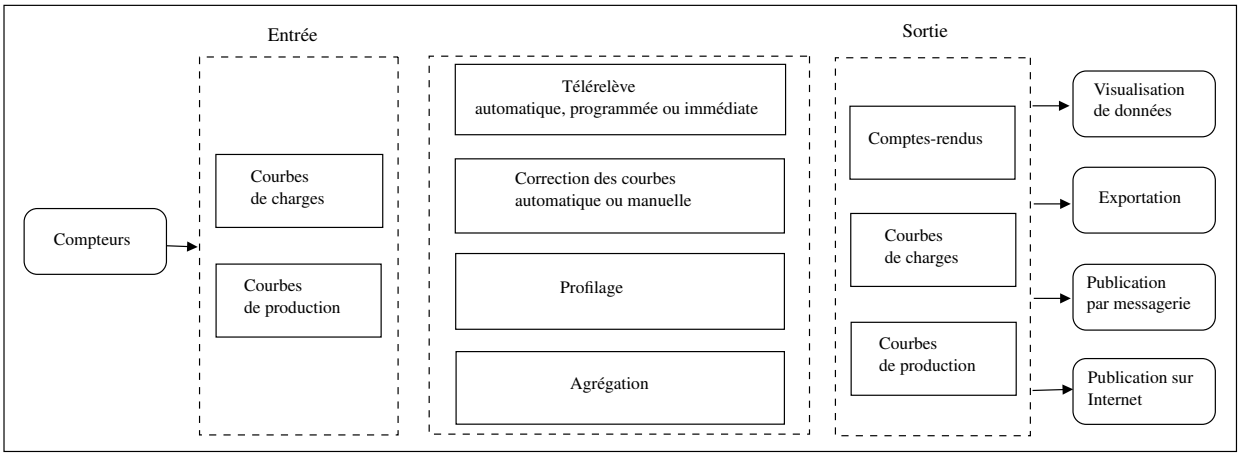


FIGURE 2.3 – Outil de télérelève de SRD

pements. Il gère l'ensemble des données nécessaires au fonctionnement du réseau telles que la position des organes de coupure, les alarmes et les différentes mesures des flux d'énergie (le courant, la tension et les puissances).

Ce service est alimenté par plusieurs données issues de sources différentes. D'une part, par les données de la cartographie décrivant toutes les ouvrages du réseau, d'autre part, par les données de télémesure reçues en temps réel.

Les données de télémesure sont des données mesurées en temps réel dans certains points du réseau, principalement au niveau du poste source et au niveau de certains grands producteurs raccordés au réseau HTA. Ces données sont transmises en temps réel puis stockées dans une base d'archivage.

Dans ces données, nous trouvons les mesures du courant au niveau de tous les départs HTA. Nous avons aussi les mesures de la tension, de la puissance et de la réactance au niveau de toutes les arrivées HTA et pour certains départs. En effet, la disponibilité de ces données dépend de la technologie de mesure utilisée. Certains postes sources sont équipés par des vieilles technologies comme des convertisseurs analogiques numériques à huit bits. Ces appareils permettent la quantification du signal électrique en plusieurs intervalles discrets, avec une fréquence d'échantillonnage donnée. La conversion de ce signal en valeurs discrètes est donc limitée en fonction du nombre de bits disponibles. Dans d'autres postes sources, nous utilisons une autre technologie dite PCCN (Palier Contrôle Commande numérique). Ce projet a été développé par EDF en 1996 {Boisnault *et al.*} et {Jaray *et al.*}, il assure la protection, la télémesure et la télécommande des postes sources. La conversion du signal électrique dans PCCN est plus sophistiquée que celle avec les anciens convertisseurs.

Les différentes valeurs mesurées sont communiquées au SCADA (Supervisory control and data acquisition) du service de téléconduite en temps réel via des réseaux IP en utilisant des protocoles dédiés tels que le HNZ (selon des spécifications EDF) pour les postes les plus anciens ou l'IEC 60870-5-104 pour les plus récents.

Données EnR

Les données de certains producteurs EnR sont stockées dans la base de données Epices. Il s'agit de données concernant les producteurs EnR qui ont confié, par convention, l'exploitation de leurs moyens de production à SRD. L'outil Epices est utilisé pour superviser et contrôler les productions des installations PV et éoliens.

Dans cet outil de monitoring, les caractéristiques techniques des producteurs sont données comme l'orientation et l'inclinaison des panneaux solaires. Nous trouvons ainsi les courbes de production et des données météorologiques comme l'irradiance, la vitesse et la direction du vent.

Données ouvertes SRD

En juin 2020, SRD a mis une partie de ses données en accessibilité via son portail open data. Ces données constituées de dix-huit jeux de données sont réparties en trois catégories : la production des EnR, la consommation d'énergie et la cartographie du réseau {SRD, 2021}.

2.3.2 Données externes

Données SDES

Le service des données et études statistiques (SDES) assure les missions de service statistique des ministères chargés de l'environnement, de l'énergie, de la construction, du logement et des transports. Il est rattaché au Commissariat général au développement durable (CGDD), du ministère de la Transition écologique.

SDES met à disposition les données locales d'énergie suivant l'article 179 de la loi de transition énergitique pour une croissance verte (LTECV). [1]

Le SDES propose, via un site internet, plusieurs catégories de données ouvertes selon plusieurs thèmes : climat, énergie, environnement, logement et transport. Dans le domaine de l'énergie, il existe des jeux de données sur le bilan énergétique, le prix de l'énergie et plusieurs données régionales et locales sur la consommation et la production d'énergie.

Concernant les données locales de consommation d'énergie, SDES met à disposition les données annuelles de consommation énergétique à la maille géographique IRIS. [2]

Ces données sont disponibles pour plusieurs années (2008-2019) suivant plusieurs vecteurs énergétiques comme l'électricité, le gaz, la chaleur, le froid et les produits pétroliers. Dans les données sur la consommation d'électricité, nous avons les informations sur le secteur d'activité des clients (agriculture, industrie, tertiaire, résidentiel et non affecté) et sur l'opérateur chargé du transport ou de la distribution (RTE ou GRD). Dans ces jeux de données, nous avons plusieurs informations sur l'année, l'IRIS, le secteur, la consommation en MWh, le nombre de points PDL, la thermosensibilité dans le résidentiel, la part de la consommation thermosensible dans la consommation résidentielle et des indices sur la qualité de données. Les données sont disponibles sous format CSV sur le site de SDES, ainsi que leur qualité et la méthodologie suivie dans la collecte de données {SDES, 2021b}. Les données de SDES sont disponibles gratuitement sous la licence ouverte Etalab 2.0 {Etalab, 2021} telle que décrite dans le décret n°2017-638 [3].

1. LOI n° 2015-992 du 17 août 2015 relative à la transition énergétique pour la croissance verte
2. Définition Insee : **Iris** « constitue la brique de base en matière de diffusion de données infra-communales. Il doit respecter des critères géographiques et démographiques et avoir des contours identifiables sans ambiguïté et stables dans le temps.Les communes d'au moins 10 000 habitants et une forte proportion des communes de 5 000 à 10 000 habitants sont découpées en IRIS »
3. Décret n° 2017-638 du 27 avril 2017 relatif aux licences de réutilisation à titre gratuit des informations publiques et aux modalités de leur homologation

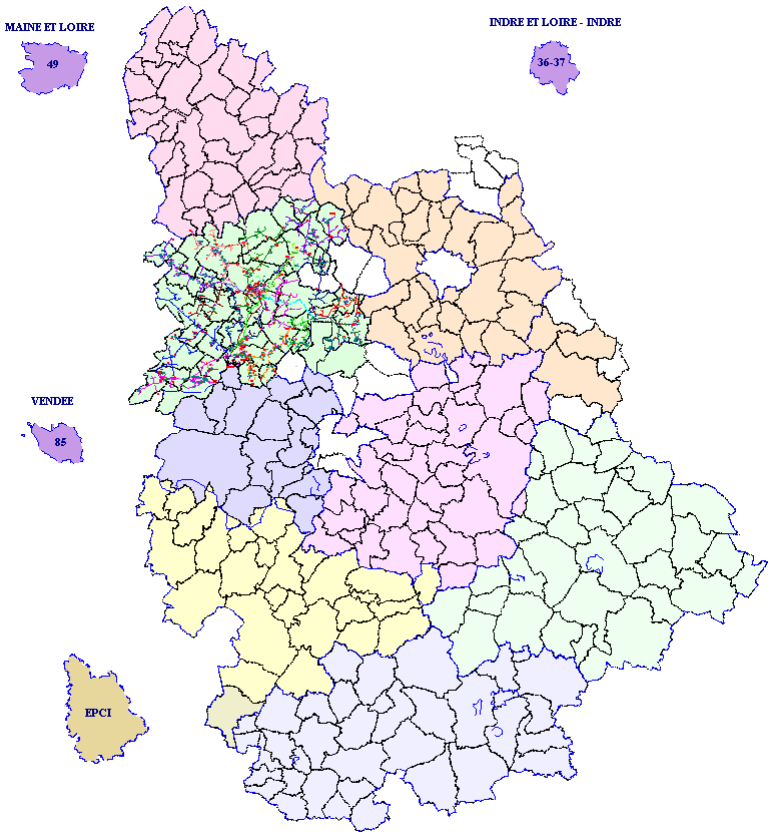


FIGURE 2.4 – Interface graphique de l'outil de la cartographie

Fonds de cartes IRIS

Les données sur les contours IRIS sont présentées sur des fonds de cartes édités par l'Insee et l'Institut national de l'information géographique et forestière (IGN). Elles sont diffusées par l'Ign sous la licence ouverte Etalab 2.0. Elles constituent les fonds numérisés des îlots IRIS définis par l'Insee. En plus des données sur les contours, nous trouvons les données des résultats du recensement de la population par IRIS. Ces données sont disponibles sous format Shapefile. Pour plus d'information sur ces données, on peut se référer à {IGN, 2021}.

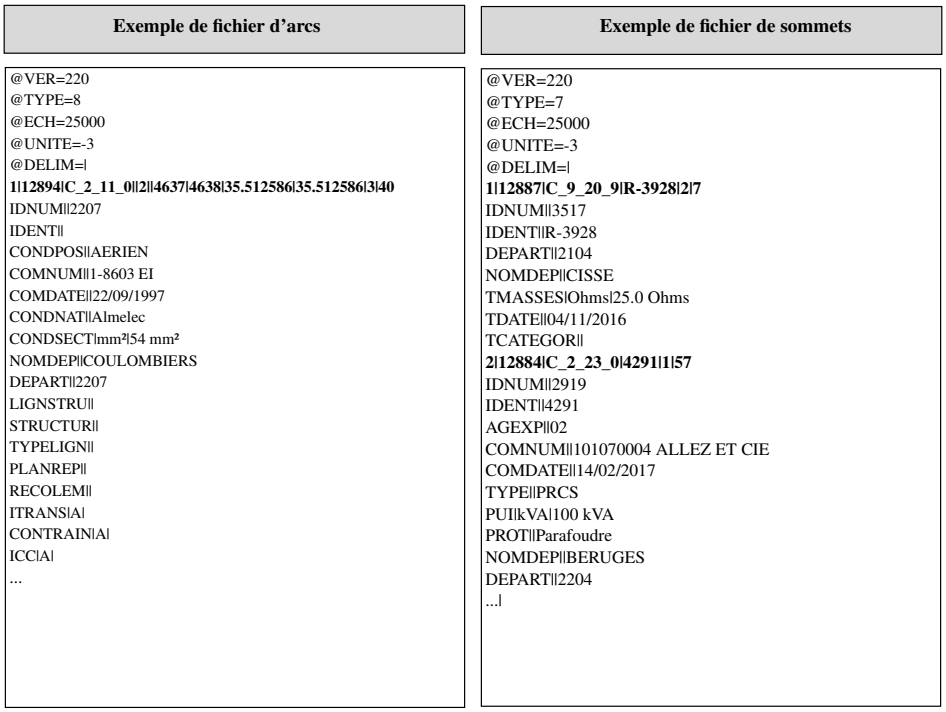
2.4 Prétraitement de données

Dans cette section, nous présentons les différents prétraitements de données utilisées pendant la thèse. L'objectif de ces traitements est de rendre les données exploitables et analysables via des étapes de transformation et d'enrichissement.

2.4.1 Prétraitement de données de la cartographie

Comme nous l'avons présenté dans la section précédente 2.3, les données de description de réseau sont disponibles dans l'outil de la cartographie interne de SRD. Cette cartographie est divisée en huit zones différentes. Ces zones sont appelées des agences, elles ont été choisies pour des raisons historiques de construction et de conduite de réseau.

Depuis l'interface graphique de la cartographie 2.4, nous devons choisir l'agence d'exploitation et ensuite nous exportons les données manuellement via l'outil. Il faut extraire les données de deux schémas différents. Un premier schéma dit schéma radial ou schéma normal. Il s'agit du schéma d'exploitation



Exemple de fichier d'arcs

```
@VER=220
@TYPE=8
@ECH=25000
@UNITE=-3
@DELIM=|
1|12894|C_2_11_0||2||4637|4638|35.512586|35.512586|3|40
IDNUM||2207
IDENT||
CONDPOS||AERIEN
COMNUM||1-8603 EI
COMDATE||22/09/1997
CONDNAT||Almelec
CONDSECT|mm²|54 mm²
NOMDEP||COULOMBIERS
DEPART||2207
LIGNSTRU||
STRUCTUR||
TYPELIGN||
PLANREP||
RECOLEM||
ITRANS|A|
CONTRAIN|A|
ICC|A|
...
```

Exemple de fichier de sommets

```
@VER=220
@TYPE=7
@ECH=25000
@UNITE=-3
@DELIM=|
1|12887|C_9_20_9|R-3928|2|7
IDNUM||3517
IDENT||R-3928
DEPART||2104
NOMDEP||CISSE
TMASSES|Ohms|25.0 Ohms
TDATE||04/11/2016
TCATEGOR||
2|12884|C_2_23_0|4291|1|57
IDNUM||2919
IDENT||4291
AGEXP||02
COMNUM||101070004 ALLEZ ET CIE
COMDATE||14/02/2017
TYPE||PRCS
PUI|kVA|100 kVA
PROT||Parafoudre
NOMDEP||BERUGES
DEPART||2204
...|
```

FIGURE 2.5 – Exemple des fichiers d'arcs et de sommets

normal du réseau. C'est-à-dire, qu'il forme la structure arborescente normale du réseau en considérant les informations sur l'état des organes de coupures. Le deuxième schéma, dit schéma bouclé, considère que le réseau est totalement bouclé et présente la structure maillée du réseau.

Après l'exportation des données, nous obtenons deux fichiers différents décrivant le graphe calculé depuis l'outil Editop. Premièrement, un fichier contient toutes les informations sur les objets linéaires du réseau, c'est-à-dire, les arcs dans le graphe formé par le réseau. Deuxièmement, un fichier contient toutes les données sur les objets ponctuels du réseau qui forment les sommets dans le graphe. Les données exportées dans les deux fichiers sont présentées ligne par ligne (voir figure 2.5). Chaque ligne du fichier présente une information sur un type d'objet donné (ligne souterraine, ligne aérienne, poste HTA/BT, Organe de coupure, etc.). Un travail de transformation est donc nécessaire pour rendre ces données exploitables. C'est-à-dire, nous transformons les données de tous les types d'objets sous forme de données tabulaires pour rendre leur utilisation facile par les outils d'analyse de données. La figure 2.6 illustre les différents blocs de traitements de ces données.

2.4.2 Prétraitement de données de comptage

Dans cette section, nous présentons les différents blocs de traitement et de la préparation des données de comptages. Ces données sont utilisées dans le projet estimation et prévision de la production PV.

Nous avons vu dans la section 2.3.1 que les données de comptage sont gérées par l'équipe métrologie via l'outil Saturne. Nous avons demandé une extraction de toutes les données des producteurs. Cependant, l'outil de télérelève permet, pour une plage donnée, la récupération des courbes de production seulement pour les producteurs mis en service avant cette période. Par exemple, si nous cherchons à exporter les données entre 2012 et 2020, nous ne récupérons que les courbes des producteurs mis en service avant janvier 2012. Nous avons proposé une alternative d'exportation en récupérant les données pour chaque année. C'est-à-dire, nous découpons la période 2012-2020 en neuf années de 2012 à 2020, et ensuite nous exportons les données pour chaque année. Cependant, nous perdons une partie de don-

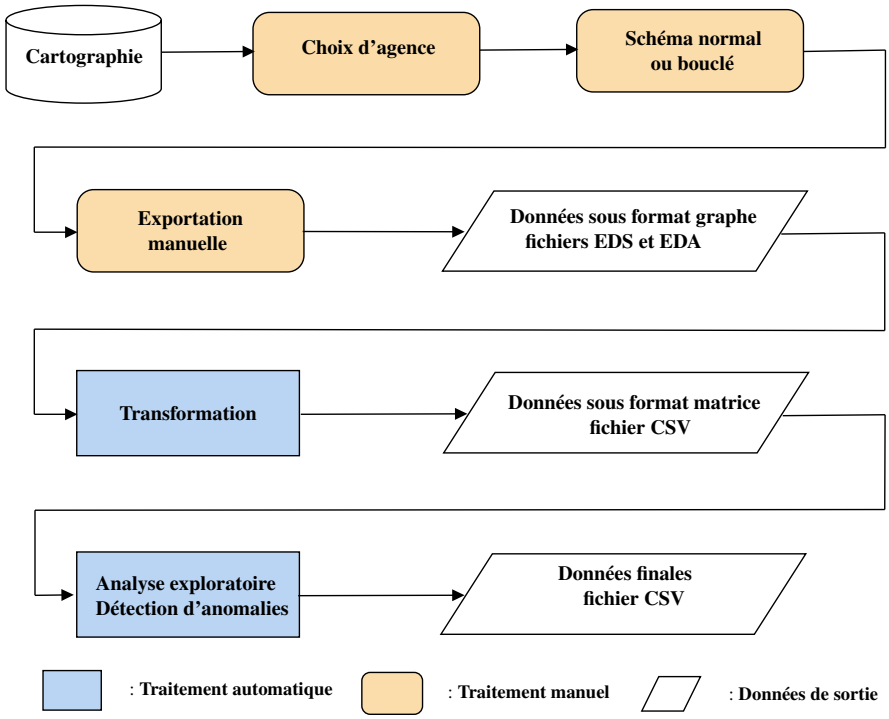


FIGURE 2.6 – Prétraitement des données de la cartographie

nées pour certains producteurs. Par exemple, pour un producteur mis en service en février 2018, nous ne récupérons que la courbe de production de janvier 2019 à décembre 2020. La partie février 2018 à décembre 2018 n'est pas exportée parmi les données. Nous pouvons récupérer cette partie directement via l'interface de l'outil. Comme nous avons plus de 700 producteurs au total, une telle exportation pour toutes les courbes est compliquée. Dans le futur, avec le développement d'un entrepôt de données des comptages, nous n'aurons pas cette problématique.

Après l'exportation des courbes de productions de chaque producteur selon les années, nous avons besoin de les rassembler dans une seule matrice de données (voir figure 2.7). Ensuite, nous devons fusionner les matrices annuelles en une seule grande matrice de données. Dans cette matrice, les colonnes représentent les courbes de production des producteurs et les lignes représentent les mesures correspondantes avec une granularité de 10 min et un historique de 2012 à 2020. Dans ce cas de présentation, le format de cette matrice est appelé format large (Wide format).

Finalement, selon les besoins de modélisation, nous avons ajouté les informations statiques sur les producteurs. Ces informations sont issues de la cartographie, comme les coordonnées GPS et la capacité installée des producteurs. Pour migrer ces données statiques avec les données de comptage, il faut transformer la matrice de données de format large en format long (long format).

La figure 2.7 illustre un exemple des formats de ces matrices de données. D'un côté nous avons les données statiques issues de la cartographie et de l'autre côté nous avons les données des courbes de productions en format large. Nous transformons cette matrice en format long et nous migrons les données statiques en utilisant le numéro PDL des producteurs. La figure 2.8 illustre la méthodologie de prétraitement des données de comptage.

2.4.3 Prétraitement de données de télémesures

Les données télémesures présentées dans la partie 2.3.1 sont stockées dans une base de données d'archivage. Aujourd'hui les données de télémesures ne sont disponibles qu'au niveau des départs HTA

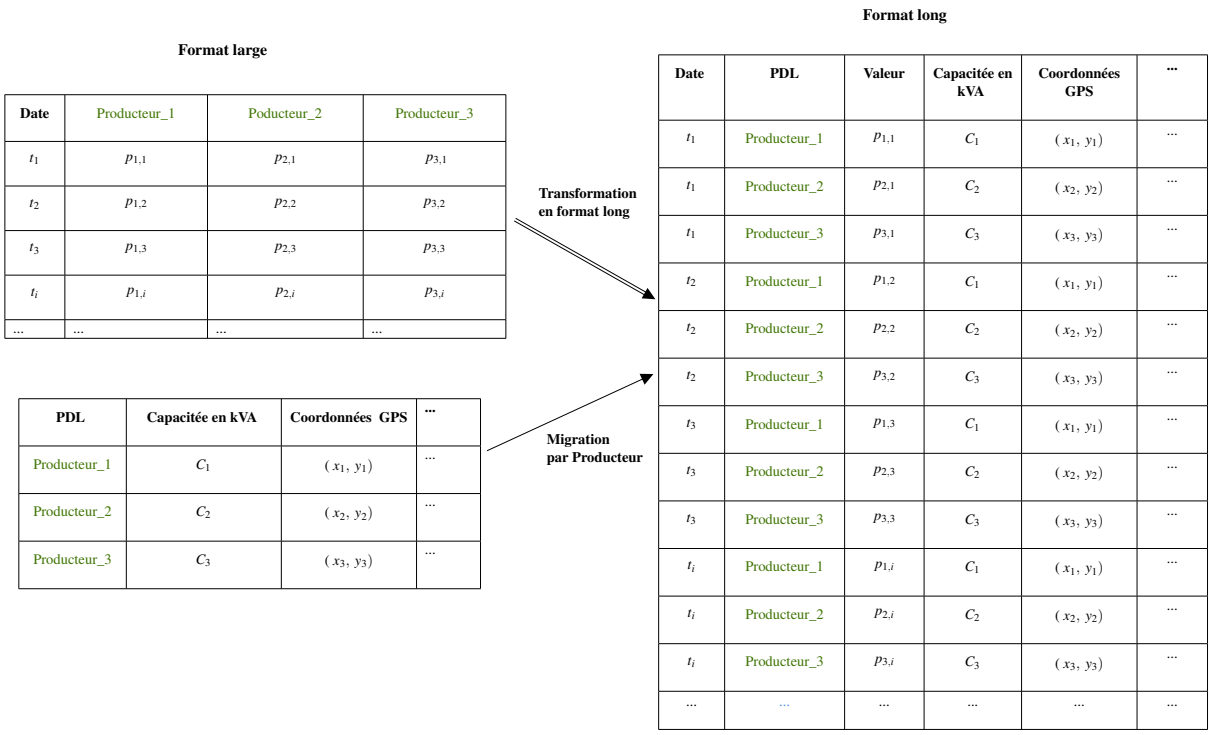


FIGURE 2.7 – Exemple de format de données

et au niveau de certains grands producteurs. Nous avons utilisé ces données dans divers projets de thèse car elles constituent les seules données temporelles sur la consommation du réseau au niveau des départs HTA.

Afin de calculer le foisonnement au niveau des départs HTA, nous avons besoin des données des puissances souscrites disponibles dans la cartographie. Par exemple, pour un départ HTA donné, la puissance souscrite cumulée des clients raccordés à un départ HTA est calculée en agrégeant les données des puissance cumulées des postes HTA/BT raccordés à ce départ HTA. Par la suite, une migration de ces données de la cartographie avec celles de télémesures est nécessaire pour calculer les foisonnements. Cependant, les deux bases d'archivage et de la cartographie ne permettent pas aisément d'associer ces deux types de données, du fait qu'il n'y a pas un identifiant unique commun dans les deux bases. Un rapprochement entre les deux bases de données est donc réalisé. Ce rapprochement semi-automatique utilise les informations disponibles sur les noms des départs et leur poste source. Pourtant, ces noms ne s'écrivent pas de la même façon dans les deux bases car dans l'outil de télémesures les champs sur les noms des départs sont limités en nombre de caractères. Le tableau 2.1 illustre une comparaison entre les noms des départs dans les deux bases de données. La figure 2.9 récapitule tous les blocs de traitement et de transformation de ces données.

Les différentes données vues précédemment sont analysées et visualisées afin de dégager des tendances globales et de détecter des anomalies qui peuvent fausser nos études.

2.5 Qualité de données

La qualité de données fait référence à la validité et à l'intégrité des données en représentant les phénomènes réels étudiés. Elle consiste à identifier les incohérences et les problèmes dans les données {Fan et Geerts, 2012}. Elle est constituée de plusieurs étapes permettant la détection de ces incohérences :

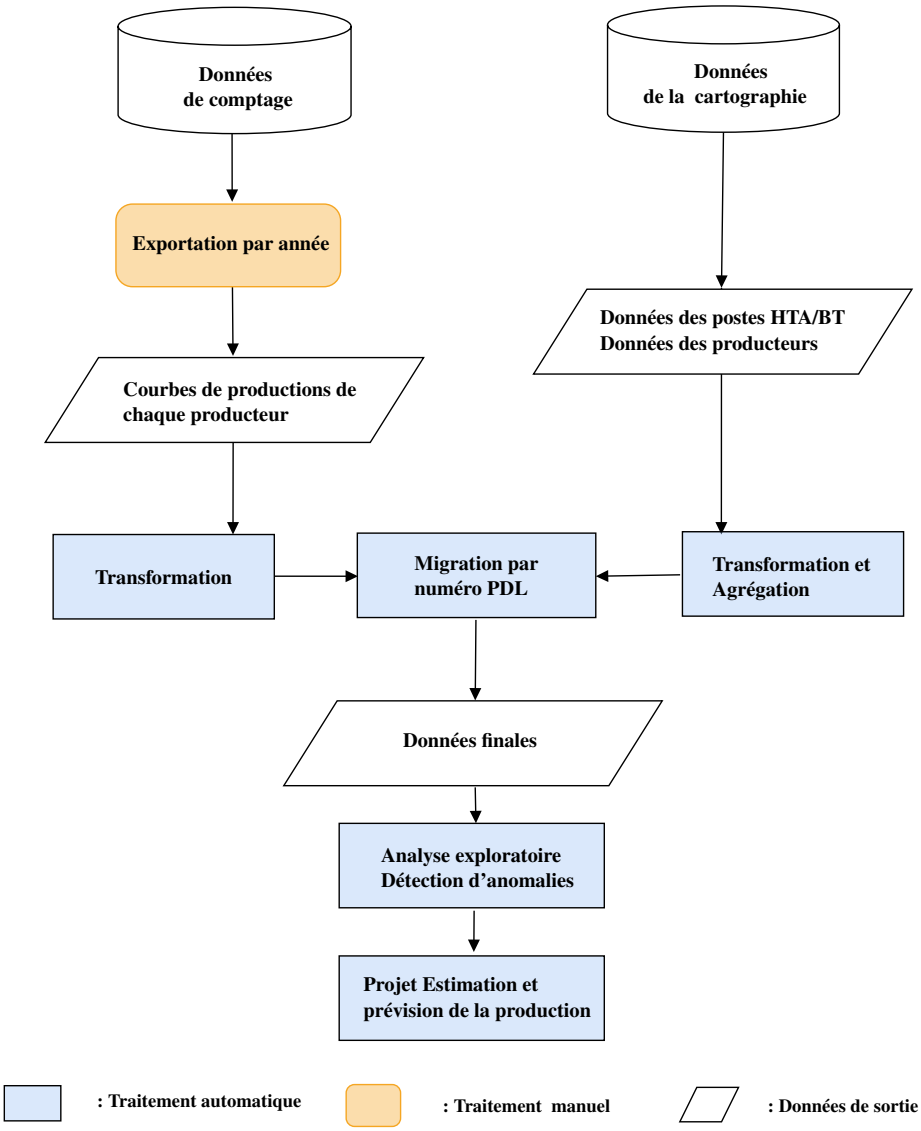


FIGURE 2.8 – Prétraitement de données de comptage

Nom des départs dans la base télémesures	Nom des départs dans la cartographie
C5 TROIS MOUTIER	LES TROIS MOUTIERS
C23 ST MARTIN RI	ST MARTIN
C26 BON MATOURS	BONNEUIL MATOURS
C3 CEINTURE CHAU	CEINTURE CHAUVIGNY
C5 CHA MONTREUIL	LA CHAPELLE MONTREUIL
C14 FONTAINE	FONTAINE LE COMTE
C24 ARCHAMBAULT	BOURG-ARCHAMBAULT

TABLEAU 2.1 – Exemple de données dans les deux bases télémesures et cartographie

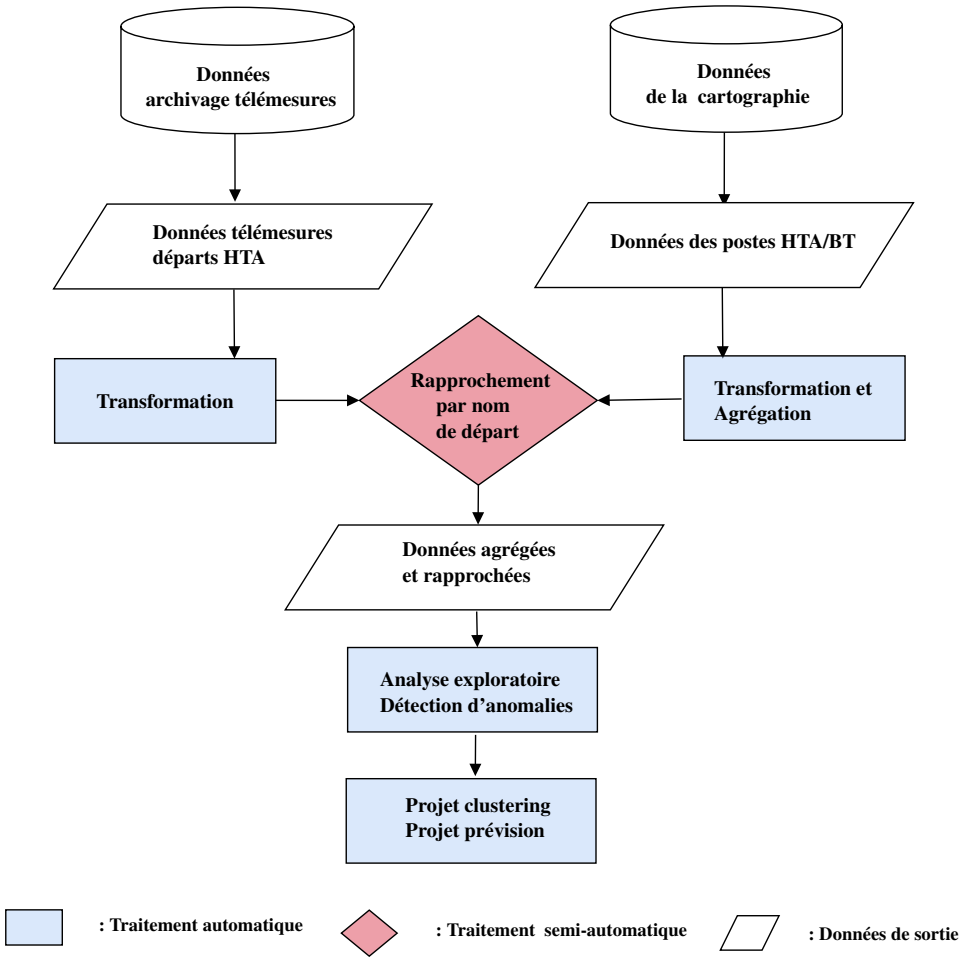


FIGURE 2.9 – Prétraitement de données dynamiques

— Détection de doublons : cette étape consiste à chercher les duplications dans les différentes données.

— Détection de données manquantes : cette étape consiste à identifier les informations manquantes et leur volume, ainsi que les raisons expliquant leur absence.

— Précision des données : cette étape fait référence à la fiabilité des données et leur précision comparées au phénomène observé.

— Détection des anomalies : cette étape identifie les données aberrantes et différentes anomalies comparées aux autres données.

La détection d'anomalies est une étape primordiale dans le processus de qualité de données. Dans la littérature, de nombreux travaux de recherche ont étudié cette notion d'anomalie. Aggarwal {2015} a élaboré un livre complet exposant les techniques et les méthodes permettant l'analyse et la détection des anomalies dans les données. Hawkins {1980} a défini une anomalie par « *une observation aberrante est une observation qui s'écarte tellement des autres observations au point de susciter des doutes sur le fait qu'elle a été générée par un processus différent ».* Chandola *et al.* {2009} ont présenté une synthèse approfondie des méthodes de détection des anomalies. Ils ont défini les anomalies comme « *des motifs dans les données qui ne sont pas conformes à une notion bien définie de comportement normal.».* Ils ont distingué trois types d'anomalies différents :

— Anomalie ponctuelle : c'est une observation individuelle de données considérée comme anormale par rapport au reste des observations. Il s'agit du type d'anomalie le plus traité dans la majorité des travaux de recherche sur la détection d'anomalies.

— Anomalie contextuelle : c'est une observation individuelle de données considérée comme anormale par rapport au reste des observations dans un contexte donné mais pas dans un autre contexte. Par exemple, les grandes valeurs de consommation d'électricité dans une région donnée peuvent être normales dans la saison d'hiver et anormales dans la saison d'été.

— Anomalies collectives : elles constituent une série de données anormales comparé à l'ensemble des données. Les données ponctuelles d'une anomalie collective peuvent être normales en elles-mêmes, mais leur combinaison en tant que collection est anormale.

Dans l'exploration des séries temporelles, la même catégorisation des anomalies est observée. La figure 2.10 illustre les types de détection d'anomalies dans les séries temporelles présentés par {Benkabou, 2018}. Nous distinguons trois types de détection :

— Détection locale : elle consiste à explorer les anomalies ponctuelles et contextuelles dans une série temporelle donnée.

— Détection des sous-séquences anormales : elle consiste à explorer les anomalies collectives présentes dans une série temporelle.

— Détection globale : elle permet d'identifier les séries anormales par rapport à l'ensemble des séries temporelles disponibles. Ce type de détection fait partie des objectifs du projet de regroupement de données présenté dans le chapitre 3.

Nous notons que les valeurs aberrantes sont considérées comme des anomalies contrairement aux valeurs extrêmes ou exceptionnelles qui font partie des données réelles. Dans la plupart des cas, la validation de l'anormalité d'une observation donnée est réalisée en exploitant les notions métier et en étudiant

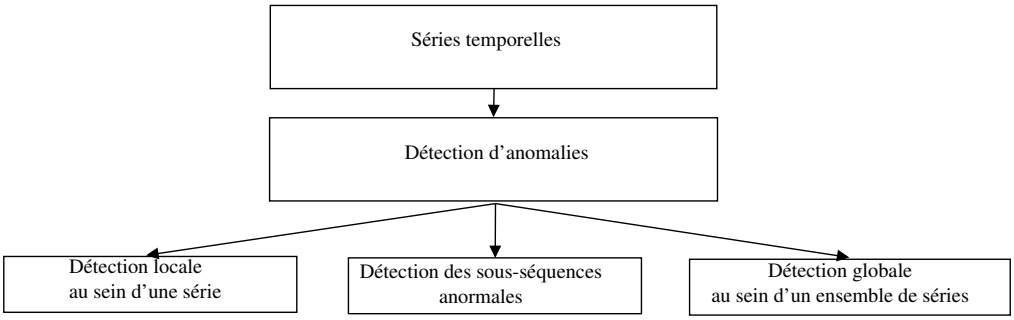


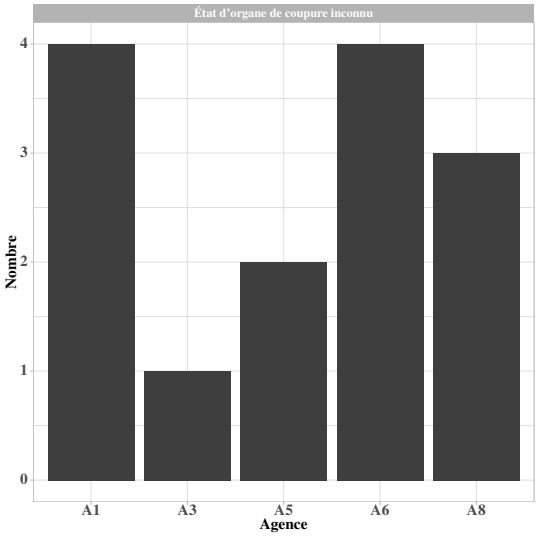
FIGURE 2.10 – Détection d'anomalies dans les séries temporelles (adapté de {Benkabou, 2018}).

les processus de génération de ces différentes données. Il existe de nombreuses techniques et méthodes permettant d'identifier les points aberrants et extrêmes. Pour un exposé détaillé de ces méthodes, nous pouvons nous référer au livre de Aggarwal {2015}.

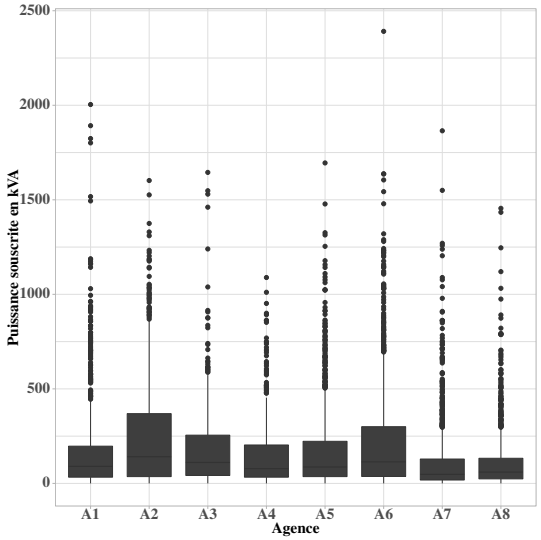
2.5.1 Exemple d'anomalies

Une étude de la qualité des différentes données exploitées dans cette thèse est réalisée. Cette exploration d'anomalies se fonde principalement sur les processus métier générant ces données. Dans le cas des données statiques de la topologie de réseau (voir figure 2.11), les anomalies peuvent fausser les différents calculs d'optimisation ou d'estimation du réseau. Par exemple, dans le cas des organes de coupures (voir figure 3.11a) certains de ces interrupteurs ont un état inconnu. Cependant, leur état est indispensable pour déterminer la configuration du réseau. D'autres anomalies dans les postes HTA/BT ont été aussi observées comme des puissances souscrites cumulées nulles (voir figure 3.11b). Ces puissances sont utilisées dans les différents calculs des foisonnements. Finalement, des longueurs des lignes électriques nulles ou infimes ont été identifiées (voir figures 2.11c et 2.11d). Les longueurs sont utilisées dans les calculs des résistances des lignes et qui sont essentielles pour l'optimisation et l'estimation d'état.

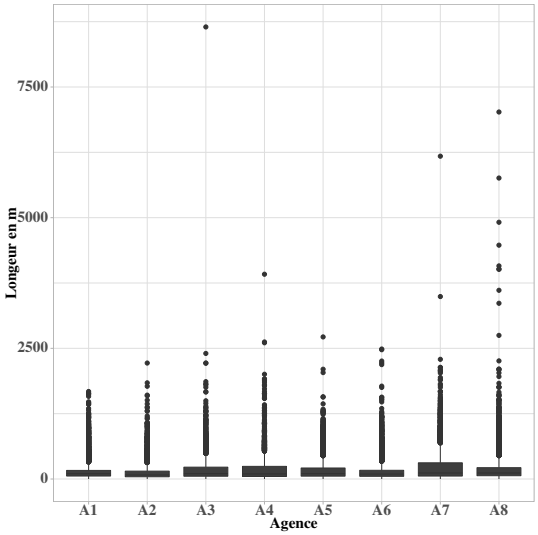
Concernant les données des courbes de charge, certaines données aberrantes ont été observées. La plupart de ces données aberrantes font partie de la réalité des données. Pour des raisons de la conduite et de la maintenance du réseau, certains départs ou des parties de ces départs sont ouverts et passés sur d'autres départs. Ceci crée des changements dans les courbes de charges. La figure 2.12 illustre certains de ces changements. Lors des périodes données, une consommation nulle est parfois observée au niveau d'un départ, parce qu'il a été ouvert. Au contraire, des faux pics de consommations sont aussi créés lorsque le départ est refermé. Dans le cas de la configuration partielle d'un départ en passant une partie sur un autre départ, ceci crée ainsi des changements dans les courbes de charge des deux départs. Finalement, certaines périodes observées anormales sont des données réelles de consommation, liées à la période estivale par exemple (voir figure 2.12).



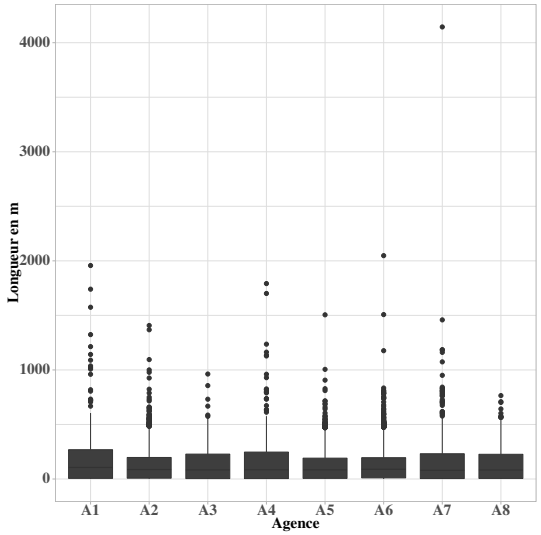
(a) Organes de coupure avec un état inconnu.



(b) Distributions des puissances souscrites des postes HTA/BT.



(c) Distributions des longueurs des lignes aériennes.



(d) Distributions des longueurs des lignes souterraines.

FIGURE 2.11 – Qualité des données de la cartographie.

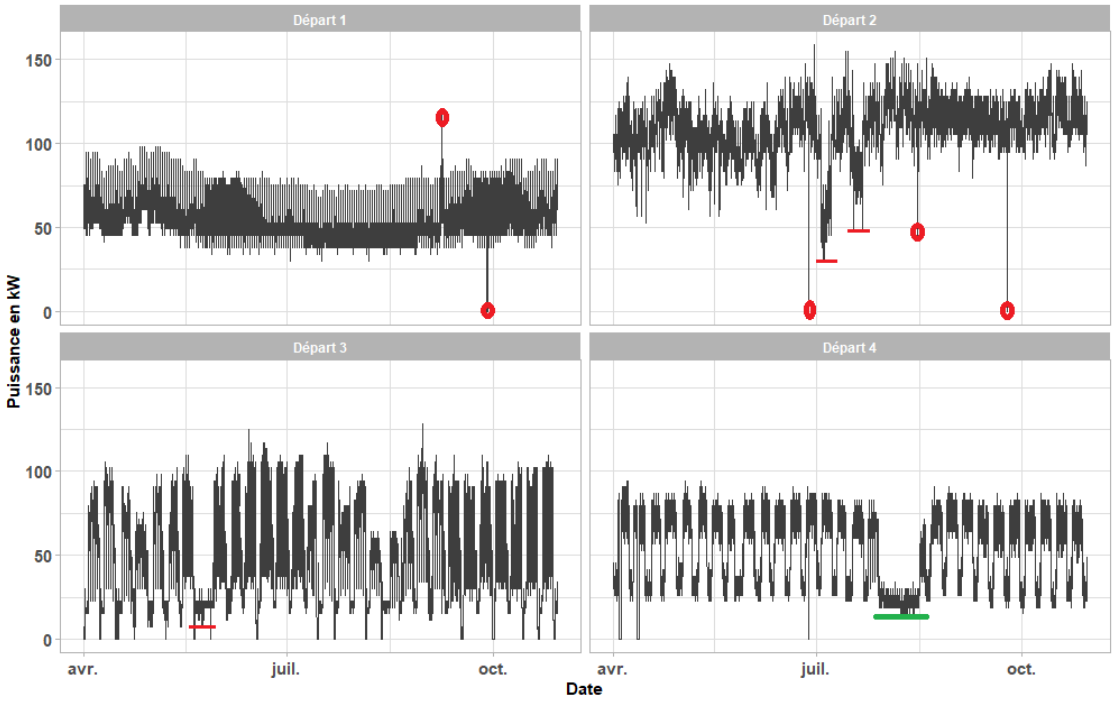


FIGURE 2.12 – Courbes de charge des départs HTA.

2.6 Conclusion du chapitre

Dans ce chapitre, nous avons présenté des notions métier du domaine de l'électrotechnique et des fondements de la théorie des séries temporelles, ainsi que des métriques de comparaison permettant la quantification de la différence entre les séries temporelles.

Par la suite, nous avons exposé les données utilisées et exploitées dans le cadre des différents projets de la thèse. Leur méthodologie de prétraitrement et de préparation est ainsi détaillée. Finalement, la qualité de ces données est présentée et des exemples d'anomalies sont illustrés. Le chapitre suivant exposera le projet de regroupement des données de charge des départs HTA.

Chapitre 3

Consommation électrique du réseau de SRD

Dans ce chapitre, nous proposons une approche de sélection des données de la consommation d'énergie les plus pertinentes en étudiant leur influence sur l'efficacité de l'optimisation. Cette méthodologie permet de détecter des similitudes (ou des dissimilitudes) dans les données afin de les regrouper dans des groupes homogènes. Cela permet d'identifier des paramètres représentatifs de chacun des consommateurs du réseau et de détecter des clients « exceptionnels » et « atypiques ». Nous avons évalué trois algorithmes de regroupement (clustering) de données, dont un algorithme original, en exploitant les données mesurées de la consommation d'électricité agrégée dans un niveau de tension supérieur de réseau. Cette étude a permis ainsi de confirmer ou d'infirmer des coefficients empiriques métier dans l'estimation de la consommation.

Sommaire

3.1 Introduction du chapitre

Pour faire face aux nouvelles mutations dans son système électrique, SRD a développé un outil d'optimisation de réseau permettant une gestion efficace et optimale des flux d'énergie en intégrant des moyens de production EnR décentralisés. Comme nous l'avons vu au chapitre 1, l'objectif de notre thèse est de résoudre certaines problématiques de modélisation de données en lien avec cet outil d'optimisation du réseau.

Dans ce chapitre, nous présentons notre étude de modélisation et de sélection de données pertinentes parmi toutes les données disponibles et leur influence sur l'efficacité de l'optimisation. L'objectif détaillé est présenté dans la première section. Par la suite, nous présenterons les méthodes utilisées dans ce projet et un exemple de travaux de recherche. Finalement les expérimentations et les différents résultats obtenus sont présentés dans la dernière partie.

3.2 Objectif détaillé

Nous avons vu dans le chapitre 1 que la gestion du réseau de distribution et l'utilisation de l'énergie en général ont connu ces dernières années des mutations profondes, notamment avec le développement des microgrids, des technologies de stockage d'énergie et de mobilité électrique ainsi que de l'intégration de l'énergie renouvelable décentralisée. L'objectif de cette transformation est de diminuer l'empreinte carbone de tous les usages d'énergie et de bien réussir leur transition énergétique vers un mode de consommation respectant plus l'environnement.

Le gestionnaire de réseau est impliqué dans ces transformations de l'usage de l'énergie. Il doit permettre, avec son réseau de distribution, de réussir le transit des différentes énergies vers les consommateurs en exploitant une énergie renouvelable produite localement. Avec ces nouveaux besoins de gestion, le réseau électrique intelligent a émergé pour résoudre certaines problématiques de gestion en exploitant des nouvelles technologies de communication, de monitoring du réseau et de collecte de données. Ceci permet au GRD de développer de nouveaux outils d'aide à la décision facilitant la gestion en utilisant de nombreuses données sur la variabilité de l'acheminement d'énergie.

Dans ce contexte, SRD a élaboré un outil innovant d'optimisation de réseau de distribution HTA en exploitant plusieurs données d'entrée, en modélisant plusieurs phénomènes électriques et en respectant les contraintes de dimensionnement du réseau. Cet outil a montré une efficacité d'optimisation en proposant une reconfiguration optimale des schémas d'exploitation en modifiant les états des interrupteurs du réseau {Ali Zazou, 2017}. Cependant, l'efficacité et l'efficience de cette optimisation résulte des structures de données d'entrée, de leur volume et de leur niveau de détail. L'efficacité dans ce cas se traduit par un temps de calcul réduit et une minimisation des coûts d'acheminement d'énergie en intégrant un maximum de production EnR locale et décentralisée.

L'objectif de cette présente étude est de sélectionner, parmi un ensemble de données archivées, les données les plus pertinentes en étudiant leur influence sur l'efficacité de l'optimisation. Nous proposons une approche de réduction de dimensionnalité de données, notamment les données des courbes des charges au niveau des départs HTA des postes sources. Cette méthodologie permet de détecter des similitudes (ou des dissimilitudes) dans ces données de charge afin de les regrouper dans certains groupes de ressemblance homogène. Ces groupes permettent une sélection de données en éliminant certaines redondances et en réduisant le volume de données d'entrée. C'est-à-dire, au lieu d'utiliser plusieurs courbes de charge dans l'optimisation, le regroupement ne permet d'utiliser que certaines courbes représentatives

des groupes. Par exemple, dans le cas des courbes de charge mesurées par les compteurs communicants, un grand volume de données sera stocké selon le nombre total des clients instrumentés par ces compteurs. Aujourd'hui, avec le déploiement des compteurs communicants sur tout le réseau, le GRD collecte les courbes de charges réelles d'une partie de ses clients autorisant l'exploitation de leurs données. Le projet de regroupement permettra de regrouper toutes ces courbes dans des groupes similaires afin de réaliser des études d'estimation (de profilage), de prévision ou d'optimisation. Dans le cas de SRD, avec plus de 140 000 clients appartenant majoritairement au segment C5 (90%), les compteurs communicants permettent l'enregistrement de nombreuses courbes de charge avec une granularité temporelle fine allant jusqu'à 30 minutes. Le regroupement permet dans ce cas de réduire le nombre total de ces courbes en sélectionnant des courbes représentatives qui seront utilisables par les différents outils de modélisation. Cela permet d'identifier ou d'estimer des paramètres représentatifs de chacun des consommateurs du réseau. De plus, le regroupement permet ainsi de détecter des clients « exceptionnels » et «atypiques». Nous appelons ce projet, projet « clustering » (regroupement).

Nous avons évalué cette approche de clustering par les seules données de charge des départs HTA disponibles à ce jour et mesurées au niveau des postes sources. Ces courbes de charge ont été normalisées par la puissance souscrite totale des départs. Ce processus statistique classique de normalisation de données permet d'introduire une notion métier connue dans le domaine de GRD, le foisonnement (voir chapitre 2). Cette donnée de foisonnement est essentielle pour l'optimisation ou l'estimation d'état du réseau. Elle permet d'estimer des courbes de charge avec une granularité temporelle fine au niveau des postes HTA/BT en utilisant les foisonnements calculés au niveau des départs HTA. Le clustering permet ainsi de confirmer ou infirmer ces coefficients empiriques de foisonnement (voir section 2.2.1) dans la modélisation de l'optimiseur.

3.2.1 Panorama de la consommation d'électricité du réseau de SRD

Le réseau de SRD est majoritairement un réseau rural. Les îlots IRIS du réseau sont de type H (Habitat) avec une part de 11% [1] et le reste des IRIS (89%) sont de type Z (non divisé) [2]. Les cartographies de la figure 3.1 illustrent la répartition spatiale de la consommation annuelle d'électricité en 2019. Les données sont disponibles dans la plateforme open data de SRD {SRD, 2021}. Les données de consommation des îlots IRIS contenant moins de 10 clients résidentiels sont secrétisées (valeurs vides). Nous observons une consommation assez élevée dans le centre Est du département. La commune de Mirebeau est la plus consommatrice avec une énergie annuelle totale de 32.5 GWh, suivie par la commune de Vivonne avec une énergie de 28 GWh. Nous notons que SRD ne couvre pas tout le territoire de la Vienne et qu'en particulier, cela n'intègre pas les communes de Poitiers et de Châtellerault qui sont desservies par le réseau d'ENEDIS.

Le résidentiel est le segment le plus consommateur dans le réseau de SRD avec une énergie annuelle totale de 758 GWh, soit une part de 59% de l'énergie totale consommée (1276 GWh), suivi par le secteur « Industrie manufacturière » avec une énergie totale de 117 GWh (9%). Les résidentiels consomment en moyenne 6 MWh dans l'année et les industriels consomment 370 MWh. La figure 3.2 illustre la consommation annuelle totale en GWh de chaque secteur d'activité. Nous notons que la production EnR annuelle représente 48% (616 GWh en 2019) de l'énergie consommée totale, soit 81% de l'énergie totale consommée par le résidentiel.

1. Statistiques calculées en comptant les IRIS avec une énergie totale non nulle
2. Définition d'Insee : « les IRIS d'habitat (code H) : leur population se situe en général entre 1 800 et 5 000 habitants. Ils sont homogènes quant au type d'habitat et leurs limites s'appuient sur les grandes coupures du tissu urbain (voies principales, voies ferrées, cours d'eau, ...). Pour les communes non découpées en IRIS, le type de l'IRIS est codé à Z. »

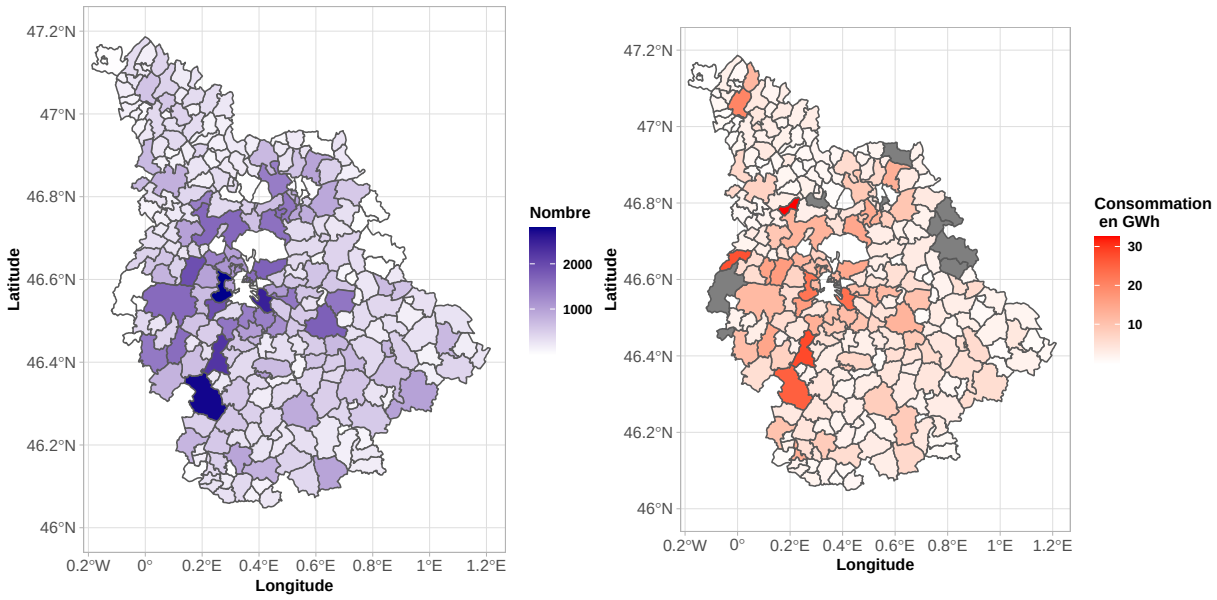


FIGURE 3.1 – Répartition spatiale de la consommation du réseau de SRD par maille IRIS
Lecture : Les deux cartographies illustrent la répartition spatiale de la consommation d'électricité du réseau de SRD en 2019. La carte à gauche illustre le nombre total de consommateurs par maille IRIS et la carte à droite représente la consommation annuelle totale par maille IRIS en GWh. Les valeurs manquantes sont représentées par la couleur grise.
Source : Données SRD disponibles dans la base open data {SRD, 2021} sous la licence ouverte v2.0 (Etalab) {Etalab, 2021}.

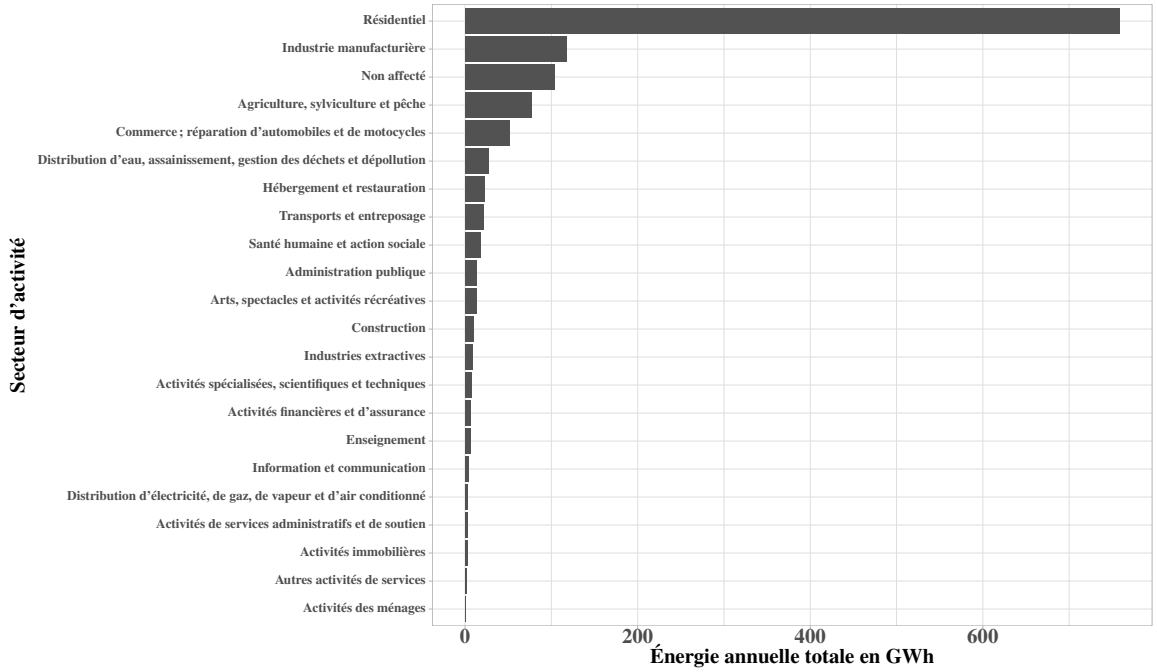


FIGURE 3.2 – Énergie annuelle consommée en GWH selon les secteurs d'activité.
Lecture : L'axe des abscisses représente l'énergie annuelle consommée en GWh et l'axe des ordonnées représente le secteur d'activité des consommateurs.
Source : Données SRD disponibles dans la base open data {SRD, 2021} sous la licence ouverte v2.0 (Etalab) {Etalab, 2021}.

3.3 Généralités

Le clustering est une approche de partitionnement de données multidimensionnelles. Il est connu dans le domaine de l'apprentissage automatique sous l'appellation d'apprentissage non supervisé. Il s'agit d'une démarche aveugle d'analyse de données non étiquetées en les segmentant en plusieurs sous-ensembles homogènes appelés clusters (groupes). Diday {1971} a défini le clustering ainsi : « *il s'agit de générer une partition à partir d'un corps de données sur lequel on ne demande pas nécessairement de faire d'hypothèse a priori. Cette partition doit réaliser au mieux les deux propriétés suivantes :*

— *les individus de chaque partie se ressemblent le plus possible,*

— *les individus de deux parties différentes se ressemblent le moins possible.»*

Cette ressemblance se caractérise par une mesure de similarité (ou dissimilarité) décrivant chaque objet et sa relation avec tous les autres objets. Une méthode de clustering cherche à regrouper les objets en se fondant sur la définition de similarité qui lui est fournie {Hastie *et al.*, 2009}. De point de vue statistique, l'objet à classer peut être une observation ou une variable. Dans notre cas, une observation est un instant t et une variable représente la courbe de charge d'un départ HTA. Nous cherchons donc à regrouper ces courbes de charge.

Avant de présenter les méthodes de clustering utilisées dans cette étude, nous allons définir quelques éléments de vocabulaire.

3.3.1 Éléments de vocabulaire

Soit $y_{i,t}$ une courbe de charge d'un départ HTA i. $y_{i,t}$ est une série temporelle avec $i = 1,2,...,p$ et $t = 1,2,...,T$. p est le nombre total de séries disponibles et T la taille de série sur une période donnée. Nous cherchons dans cette étude à regrouper les courbes de charge $y_{i,t}$ dans des clusters C_k homogènes. La plupart des méthodes de clustering utilisent comme entrée une matrice appelée matrice de distance ou matrice de similarité.

Pour une simplification des notations, nous notons $y_{i,t}$ par y_i et $\Omega = \{y_1, y_2, ..., y_p\}$ l'ensemble des courbes de charge à classer.

Indice de similarité, indice de dissimilarité et distance

Soit s une application de $\Omega \times \Omega$ dans $\mathbb{R}^+$, nous appelons s un indice de similarité lorsqu'il vérifie

— La symétrie : $s(y_i, y_j) = s(y_j, y_i)$ pour tous y_i et y_j dans Ω,

— La similarité d'une série avec elle-même : $s(y_i, y_i) = S > 0$, pour tout y_i dans Ω,

— La similarité est majorée par S : $s(y_i, y_j) \leqslant S$, pour tous y_i et y_j dans Ω.

Un indice de similarité normé s^* est défini à partir de s par

$$s^*(y_i, y_j) = \frac{s(y_i, y_j)}{S}$$

pour tous y_i et y_j dans Ω

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

L'indice de dissimilarité est une application d de $\Omega \times \Omega$ dans $\mathbb{R}^+$ défini par

$$d(i,j) = S - s(y_i, y_j)$$

pour tous y_i et y_j dans $\Omega \times \Omega$

Une distance sur Ω est, par définition, une dissimilarité vérifiant en plus la propriété d'inégalité triangulaire. Autrement dit, une distance d est une application de $\Omega \times \Omega$ dans $\mathbb{R}^+$ vérifiant

— $d(y_i, y_j) = d(y_j, y_i)$ pour tous i et j dans $\Omega \times \Omega$,

— $d(y_i, y_j) = 0 \iff y_i = y_j$,

— $d(y_i, y_j) \leqslant d(y_i, y_k) + d(y_j, y_k)$, pour tous y_i, y_j, y_k dans Ω^3.

Si Ω est fini, la distance peut être normée.

Une matrice de distance D **(ou de dissimilarité)** est une matrice carrée de taille p^2, tel que $d_{ij} = d(y_i, y_j)$. Cette matrice est fournie comme entrée dans la plupart des algorithmes de clustering. La majorité des algorithmes supposent une matrice de dissimilarité positive avec une diagonale nulle ($d_{y_i,y_i} = 0$ pour tout i dans $\{1, ..., p\}$).

La matrice de distance D peut être transformée dans un espace réduit en utilisant la méthode MDS (Multidimensional scaling) {Carroll et Arabie, 1998; Kruskal, 1964}. La méthode MDS cherche les vecteurs $z_1, z_2, ..., z_p$ dans un espace $\mathbb{R}^m$ de dimension m inférieure à T en minimisant la fonction dite de "stress" {Hastie *et al.*, 2009}

$$S_M(z_1, z_2, ..., z_p) = \sum_{j \neq k} (d_{jk} - ||z_j - z_k||)^2$$

Où $d_{jk} = d(y_i, y_j)$ est une mesure de dissimilarité.

La MDS permet de trouver une représentation dans un espace de dimension inférieure en préservant le mieux possible les distances entre les paires des séries.

Distance Euclidienne

La distance la plus utilisée dans le cas de variables quantitatives est la distance euclidienne. Elle est définie par

$$d(y_i, y_j) = \sqrt{\sum_{t=1}^{T} (y_{i,t} - y_{j,t})^2}$$

La figure 3.3 illustre un exemple de distance euclidienne entre deux courbes de charge. L'inconvénient de cette distance dans le cas des séries temporelles est qu'elle ne prend pas en compte la dimension temporelle de données. Puisque dans le cas de ces séries, parfois l'évolution temporelle est décalée d'une certaine période à cause de multiples raisons comme la variabilité spatiale des données, les retards de communication de données par les compteurs, etc. En effet, dans le cas d'un réseau électrique, les différents capteurs ne sont pas nécessairement synchronisés entre eux et peuvent réaliser leur mesure sur des périodes de temps présentant un décalage. De ce fait, un alignement entre les séries est créé. La distance euclidienne qui est une distance rigide ne permet pas de comparer les valeurs avec les valeurs décalées. Il nécessite aussi des séries temporelles de même taille. Son avantage est toutefois sa rapidité et sa complexité linéaire $O(n)$.

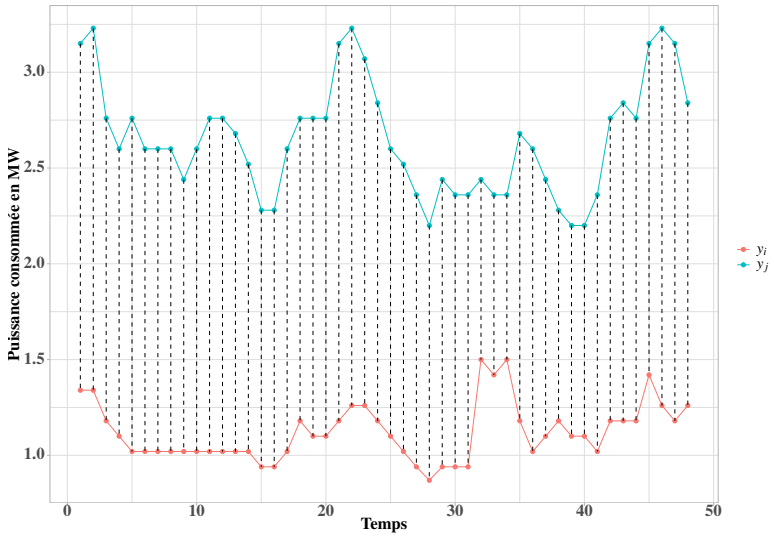


FIGURE 3.3 – Exemple de distance euclidienne entre deux courbes de charge sur une période de deux jours avec une granularité d'une heure.

Dynamic time warping (Déformation temporelle dynamique)

Pour résoudre la rigidité de la distance euclidienne, de nombreuses mesures de dissimilarité ont été élaborées dans le cas de séries temporelles. La plus connue est la distance Dynamic Time Warping (DTW) {Berndt et Clifford, 1994; Sakoe et Chiba, 1978}. Elle est appropriée dans la comparaison entre les séries en traitant les différentes transformations qui peuvent être présentes dans les données telles que les déplacements et les décalages temporels. De plus, elle permet de comparer des séries de tailles différentes. Son inconvénient est qu'elle est coûteuse en ressources informatiques. Par exemple, dans le cas de deux séries y_i et y_j de longueur n et m, sa complexité est de l'ordre $O(n \times m)$ {Keogh et Ratanamahatana, 2005; Ratanamahatana et Keogh, 2005}.

Supposons deux séries temporelles y_i et y_j avec

$$y_i = \{y_{i,1}, y_{i,2}, y_{i,3}, ..., y_{i,n}\}$$

$$y_j = \{y_{j,1}, y_{j,2}, y_{j,3}, ..., y_{i,m}\}$$

L'alignement entre y_i et y_j est calculé via une matrice A de taille $n \times m$. La matrice A définit la correspondance entre les deux séries. Chaque élément de la matrice A est donné par le carré de la distance euclidienne entre une paire de points $(y_{i,k}, y_{j,l})$, $A(k,l) = (y_{i,k} - y_{j,l})^2$ {Keogh et Ratanamahatana, 2005}.

La déformation temporelle est un chemin W continu de points de la matrice A et qui définit une correspondance entre y_i et y_j. W est donné par

$$W = w_1, w_2, w_3, ..., w_K$$

Où $w_u = (k,l)_u$ et $max(n,m) \leq K < n+m-1$

Ce chemin fait l'objet de plusieurs contraintes {Keogh et Ratanamahatana, 2005} (voir figure 3.4)

— **Conditions au limite :** $w_1 = (1,1)$ et $w_K = (m,n)$

— **Continuité :** Si $w_k = (a,b)$ Alors, $w_{k-1} = (a',b')$ Avec $|a-a'| \leq 1$ et $|b-b'| \leq 1$

— **Monotonie :** Si $w_k = (a,b)$ Alors, $w_{k-1} = (a',b')$ Avec $a-a' \geq 0$ et $b-b' \geq 0$

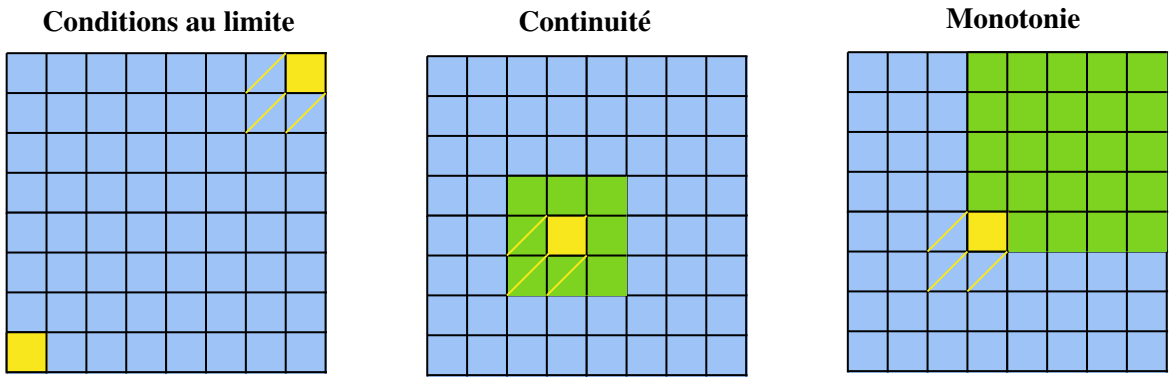


FIGURE 3.4 – Les contraintes de DTW

Il y a plusieurs chemins possibles qui vérifient ces contraintes. Dans le calcul de la distance DTW, nous nous intéressons seulement au chemin de poids minimal.

$$DTW(y_{i,n}, y_{j,m}) = \min\left(\sqrt{\sum_{u=1}^{K} A(w_u)}\right)$$

En utilisant la programmation dynamique, nous pouvons trouver ce chemin avec une fonction récursive {Keogh et Ratanamahatana, 2005}.

$$DTW(y_{i,k}, y_{j,l}) = d(y_{i,k}, y_{j,l}) + \min\left[DTW(y_{i,k-1}, y_{j,l-1}), DTW(y_{i,k-1}, y_{j,l}), DTW(y_{i,k}, y_{j,l-1})\right]$$

Cette fonction récursive détermine le parcours de la matrice A pour trouver le chemin W. Il existe dans la littérature d'autres manières définissant ce parcours. La distance DTW peut être normalisée dans certains cas de parcours par $n+m$.

La limite de DTW est de comparer toutes les valeurs des deux séries sur la période totale, même sur les points très éloignés dans le temps. Dans notre cas des courbes de charge, cela peut introduire certains défauts de comparaison. Il ne serait pas judicieux, par exemple, de comparer une valeur mesurée en janvier avec une valeur mesurée en juillet. De plus, l'algorithme DTW est sensible au pics de consommation, le calcul de l'alignement est sensible à ces points de pointe. Pour remédier à cela, les chercheurs ont introduit la notion de fenêtrage en construisant une bande symétrique autour de la diagonale de A {Mori *et al.*, 2016}. La bande la plus connue est la bande de Sakoe-Chiba {Sakoe et Chiba, 1978}. La figure 3.5 illustre un exemple de cette bande, elle oblige le chemin W à rester à l'intérieur. Elle évite donc la correspondance entre les points éloignés dans le temps. Les figures 3.5 et 3.6 illustrent un exemple de la bande de Sakoe-Chiba. Dans la figure 3.6, nous observons l'effet de cette bande dans la comparaison de deux courbes de charge. La bande r évite la comparaison entre les points éloignés. Elle calcule l'alignement seulement entre les instants les plus proches. Il a été démontré dans les travaux de {Wang *et al.*, 2013} que cette bande réduit ainsi le temps de calcul. La bande de Sakoe-Chiba fixe une fenêtre r dans la matrice A

$$l - r \leq k \leq l + r$$

Le cas $r = 0$ est équivalent à la distance euclidienne classique.

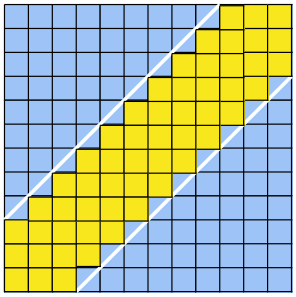


FIGURE 3.5 – Bande de Sakoe-Chiba.

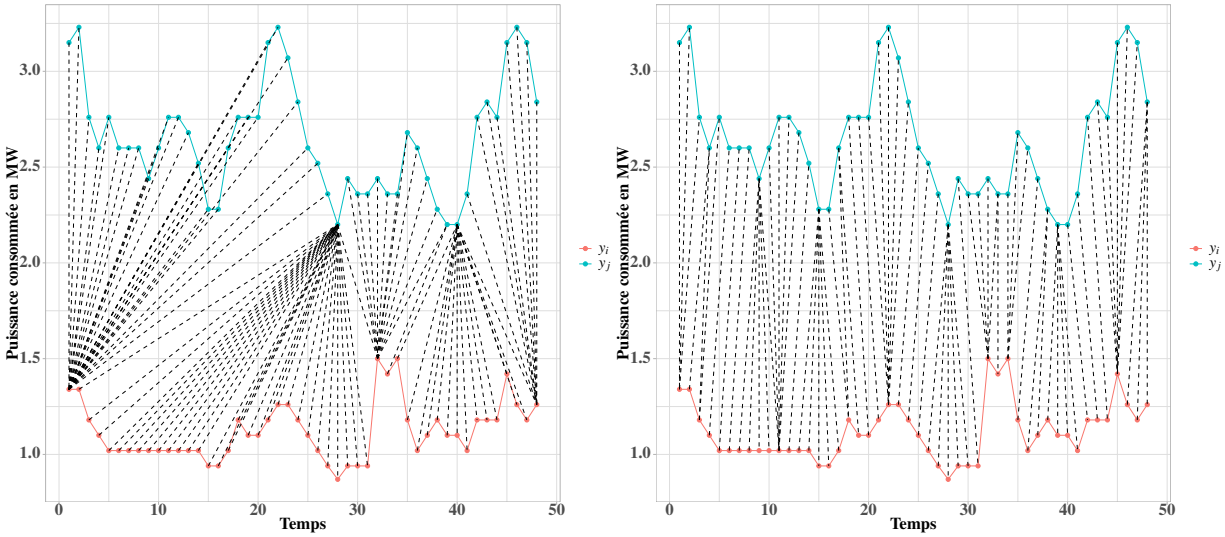


FIGURE 3.6 – Exemple de distance DTW entre deux courbes de charge.
Lecture : La figure à droite illustre la distance DTW entre deux courbes de charge sans fenêtrage Sakoe-Chiba. La figure à gauche illustre cette distance avec une bande de Sakoe-Chiba égale à 2.

Dans notre cas d'étude de clustering des courbes de charge, il faut comparer entre les p séries temporelles disponibles. Autrement dit, il faut créer $p(p-1)$ matrices A pour pouvoir trouver les chemins W entre chaque série et toutes les autres séries. Avec une complexité de $O(n^2)$ (les courbes de charge ont de même longueur n), la complexité totale du calcul de la matrice de distance D est de l'ordre $O(n^2 \times p^2)$.

3.3.2 Les algorithmes de clustering

Dans la littérature, de multiples méthodes du clustering de données ont été développées. Rokach et Maimon {2005} ont présenté une synthèse des principales méthodes de clustering. Ils ont divisé ces méthodes en plusieurs catégories : les méthodes hiérarchiques, les méthodes de partitionnement, les méthodes fondées sur une densité, les méthodes fondées sur un modèle, les méthodes fondées sur une grille et les méthodes de calcul souple (soft computing).

Dans notre étude de clustering, nous avons utilisé une méthode classique hiérarchique, une méthode de partitionnement et une autre méthode originale de partitionnement sous contrainte.

Nous rappelons qu'une partition P d'un ensemble Ω est un sous-ensemble de l'ensemble des parties de Ω $\mathscr{P}(\Omega)$ formé de parties disjointes $\{P_1, P_2, ..., P_k\}$ tel que

(i) Pour tout i dans $\{1,2,...,k\}$, $P_i \neq \emptyset$

(ii) Pour tous i et j dans $\{1,2,...,k\}$, $i \neq j$ implique $P_i \cap P_j = \emptyset$.

(iii) $\bigcup_{i=1}^{k} P_i = \Omega$

Une hiérarchie H sur Ω est un ensemble de parties $\{H_1, H_2, ..., H_k\}$ tel que

(i) $\emptyset \in H$

(ii) $\Omega \in H$

(iii) Pour tout élément y_i de Ω, $\{y_i\} \in H$

(iv) Pour tous i et j dans $\{1, 2, ..., k\}$, $H_i \cap H_j = \emptyset$ ou $H_i \subset H_j$ ou $H_j \subset H_i$

Agrégation autour des centres mobiles (k-means)

L'algorithme d'agrégation autour des centres mobiles, appelé aussi méthode de k-means, est une technique classique de partitionnement de données {Ball et Hall, 1967; Diday, 1971; Forgy, 1965; MacQueen *et al.*, 1967}. L'avantage de cette méthode est qu'elle est simple, intuitive, rapide et efficace dans le regroupement de grandes masses de données {Lebart *et al.*, 1995}.

k-means permet de calculer une partition P de Ω munie d'une distance d (souvent la distance euclidienne) en répétant plusieurs étapes. Le cardinal de P, noté k, est le nombre de clusters.

Premièrement, nous fixons un nombre k de clusters. Nous sélectionnons alors aléatoirement k centres provisoires de clusters appelés centroïdes c_l. A partir de ces centroïdes, nous construisons une première partition en affectant chaque élément de Ω au centroïde le plus proche. Par la suite, nous recalculons les centroïdes de la partition obtenue pour recalculer une nouvelle partition en affectant chaque élément au nouveau centroïde. Nous répétons cette opération jusqu'à ce que les centroïdes ne bougent plus. La figure 3.7 illustre un exemple de fonctionnement de cet algorithme.

L'inconvénient de k-means est que la solution de clustering proposée dépend en grande partie du choix aléatoire des premiers centroïdes. De plus, dans le cas de clustering des séries temporelles, les centroïdes sont des moyennes (centres de gravité) des objets des clusters. La limite de cette approche est que ces moyennes sont calculées sur les séries non alignées. Une alternative est de calculer l'alignement entre les membres de chaque clusters pour ensuite calculer le centroïde en supprimant cet alignement. Un exemple de cette approche est proposé dans les travaux de {Niennattrakul et Ratanamahatana, 2007}. Dans notre cas, nous avons utilisé une autre variante de k-means appelée k-medoids (k-médoïdes) {Park et Jun, 2009} qui utilise un médoïde comme centre de cluster au lieu de centroïde. Au contraire du centroïde qui est un élément artificiel calculé par une moyenne des éléments, le médoïde est un élément réel du cluster calculé en minimisant les distances entre l'ensemble des autres éléments.

$$m_l = arg\min_j \sum_{i=1}^{m} d(y_i, y_j)$$

pour tout j dans $\{1, ..., m\}$. Où m_l est le médoïde d'un cluster l et m le cardinal du cluster.

Classification ascendante hiérarchique (CAH)

L'algorithme de classification ascendante hiérarchique est un autre algorithme classique de regroupement de données. Les travaux les plus anciens sur cet algorithme montent à l'année 1963 avec Sokal *et al.* {1963}. Pour d'autres détails sur les fondements de CAH, on peut se référer aux livres {Hastie *et al.*, 2009; Lebart *et al.*, 1995}.

L'algorithme consiste à créer une hiérarchie en partant de la matrice de distance (ou de dissimilarité) D entre les éléments à regrouper. Il démarre de la partition triviale des singletons $\{y_i\}$, chaque élément

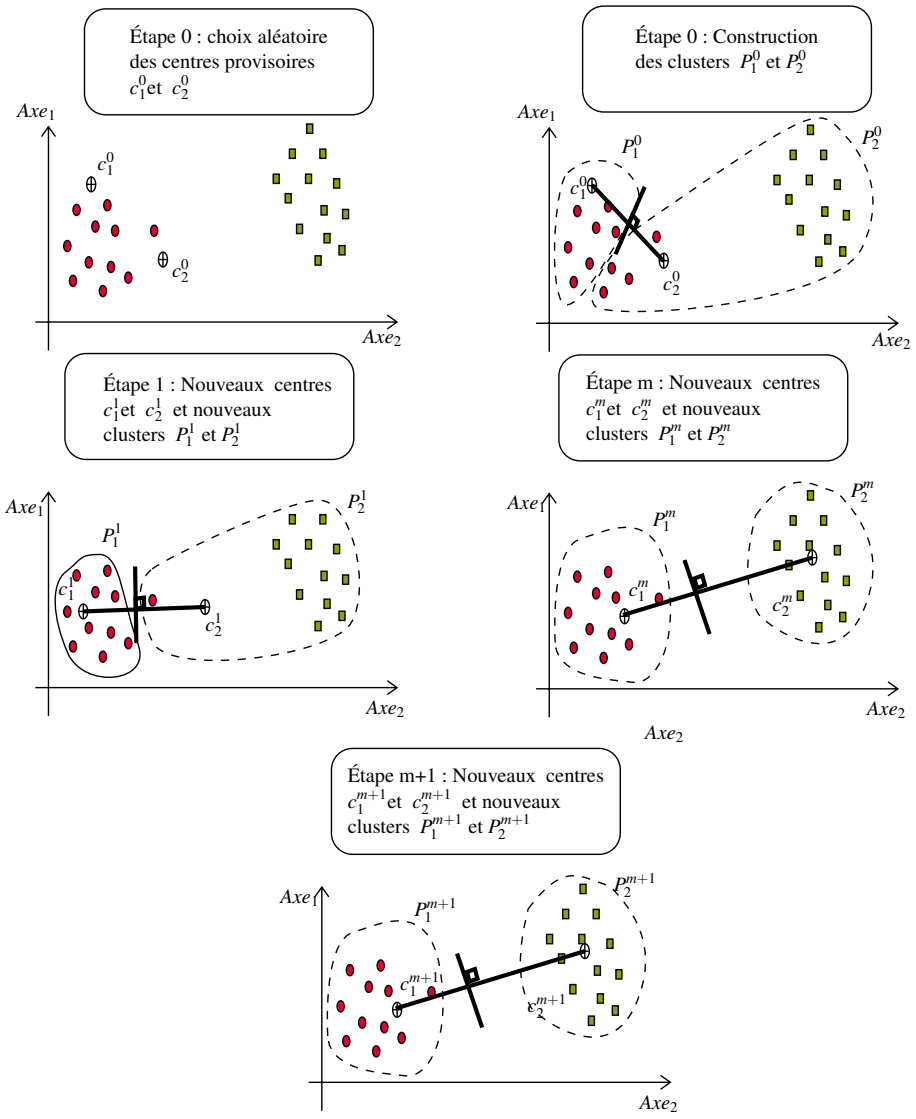


FIGURE 3.7 – Exemple d'étapes de l'algorithme de k-means (adapté de {Lebart *et al.*, 1995}).

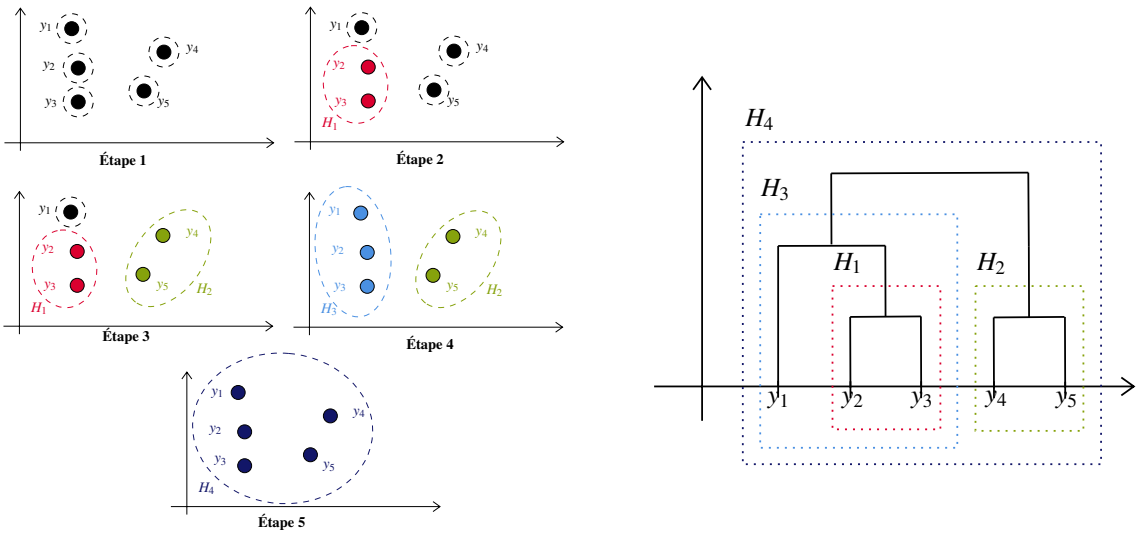


FIGURE 3.8 – Exemple d'étapes de l'algorithme de classification ascendante hiérarchique (adapté de {Lebart *et al.*, 1995}).

constituant un cluster. Ensuite, à chaque étape, il constitue des clusters formés par l'agrégation des deux éléments les plus proches de la partition précédente. Il s'arrête lorsqu'il obtient un cluster formé de l'ensemble des éléments Ω. L'algorithme peut être présenté sous forme d'un arbre. La figure 3.8 illustre un exemple de la méthode CAH.

L'algorithme met à jour à chaque étape la matrice de distance en calculant la distance entre les éléments regroupés dans un cluster et d'autres éléments dans un autre cluster. Il nécessite donc une stratégie de regroupement d'éléments qui consiste à calculer la distance entre les clusters disjoints. Cette stratégie est appelée critère de regroupement. Par exemple, dans le cas de trois éléments x, y et z, où x et y sont déjà regroupés dans une classe Z. Le critère de regroupement est une distance entre le groupe X et l'élément z appelée aussi un saut. Il existe dans la littérature plusieurs méthodes définissant ce critère de regroupement. Les plus connues sont les sauts minimal, maximal ou de Ward {Murtagh et Legendre, 2014} :

— Saut minimal : $d(X,z) = min\{d(x,z), d(y,z)\}$

— Saut maximal : $d(X,z) = max\{d(x,z), d(y,z)\}$

— Saut de Ward : $d(X,z) = \dfrac{2 \times d(x,z) + 2 \times d(y,z) + d(x,y)}{3}$

Nous avons utilisé dans notre cas d'étude, le saut de Ward {Murtagh et Legendre, 2014} qui est défini d'une manière générale entre trois groupes X, Y et Z par

$$d(X \cup Y, Z) = \frac{(n_X + n_Z) \times d(X,Z) + (n_Y + n_Z) \times d(Y,Z) + n_Z \times d(X,Y)}{n_X + n_Y + n_Z}$$

Où n_X, n_Y et n_Z sont les cardinaux des ensembles X, Y, et Z.

Le saut de Ward définit une inertie inter-groupes entre toutes les hiérarchies obtenues. Il constitue la hauteur de l'arbre hiérarchique. Nous cherchons donc dans cet arbre un découpage optimal qui maximise cette inertie inter-groupes.

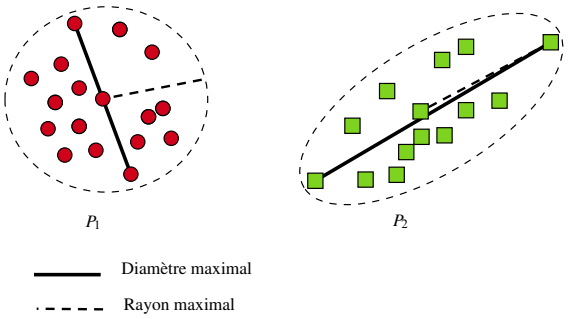


FIGURE 3.9 – Exemple de diamètre et de rayon maximaux

Equiwide clustering (EQW)

Equiwide clustering est un algorithme original élaboré au LIAS dans le cadre de notre thèse. En effet, les décideurs métier de SRD s'intéressaient à une solution de clustering fondée sur une erreur maximale. C'est-à-dire, l'algorithme de clustering des courbes de charge doit permettre un regroupement en respectant une erreur maximale entre les membres d'un cluster. Cette erreur se traduit dans le cas de clustering par une distance maximale à respecter entre les membres d'un groupe. Géométriquement, la distance maximale est équivalente au diamètre d'un cluster. Il peut être aussi représenté par un rayon maximal. L'algorithme Equiwide est un algorithme de partitionnement de données paramétré par cette contrainte de diamètre maximum. Il a été étudié et évalué dans le cadre du stage de fin d'étude de ANDERSEN {2020} au sein du LIAS. Ces travaux de stage ont pour objectifs la finalisation et l'implémentation de l'algorithme, ainsi que son évaluation et sa comparaison avec d'autres approches de clustering en exploitant les données de littérature et les données de SRD. Les résultats de ce protocole expérimental seront détaillés dans la partie résultats.

Mathématiquement, le diamètre d'un cluster P_k d'un ensemble Ω muni d'une distance d est donné par

$$diamètre(P_k) = \max_{y_i, y_j \in P_k} \{d(y_i, y_j)\}$$

De même le rayon est donné par

$$rayon(P_k) = \min_{y_i \in P_k} \{\max_{y_j \in P_k} \{d(y_i, y_j)\}\}$$

Un cluster P_k est dit homogène en diamètre si, et seulement si, son diamètre est inférieur à un seuil maximal D_{max}. De même, P_k est dit homogène en rayon si, et seulement si, son rayon est inférieur à un seuil maximal R_{max}. Une partition P de Ω est homogène si, et seulement si, tous les sous-ensembles P_k de P sont homogènes.

L'algorithme Equiwide permet d'identifier une partition P de Ω pour un seuil maximal M ($M = D_{max}$ ou $M = R_{max}$) vérifiant

(i) P est homogène sous M,

(ii) Pour toute partition homogène $P^{'}$ de Ω, le cardinal de $P^{'}$ est supérieur ou égal au cardinal de P. Autrement dit, le cardinal de P est minimal.

De plus, un sous-ensemble A de $\mathscr{P}(\Omega)$ est dit maximal compatible si, et seulement si, il vérifie

(i) A est homogène (compatible),

(ii) Il n'existe pas un sous-ensemble compatible B de $\mathscr{P}(\Omega)$ tel que $A \subseteq B$. Autrement dit, A est un ensemble maximal parmi tous les sous-ensembles compatibles de $\mathscr{P}(\Omega)$.

Pour un sous-ensemble A de $\mathscr{P}(\Omega)$, une couverture minimale de Ω par A est le plus petit sous-ensemble de B en cardinalité vérifiant

(i) $B \subset A$

(ii) $\bigcup_{B_k \in B} B_k = \Omega$

(iii) Il n'existe pas de sous-ensemble $B^{'}$ de A, tel que $|B^{'}| < |B|$ et $\bigcup_{B_k^{'} \in B^{'}} B_k^{'} = \Omega$

La couverture minimale permet de trouver une couverture complète de Ω avec un minimum d'ensembles compatibles possibles. Le nombre de ces ensembles est le nombre de clusters obtenus à la fin de l'algorithme.

L'algorithme Equiwide utilise en entrée une matrice de distance (ou de dissimilarité) D et un seuil maximal M pour trouver une partition homogène de Ω. Il est divisé en trois grandes étapes de calcul. Tout d'abord, il liste tous les sous-ensembles compatibles de Ω respectant le seuil fixé. Par la suite, il détermine la couverture minimale. Finalement, le clustering final est obtenu en choisissant une partition parmi l'ensemble des partitions homogènes obtenues dans les calculs de la couverture minimale. Ce choix se fonde sur une minimisation ou une maximisation de certains indicateurs décrivant le clustering, comme la minimisation des distances intra-clusters ou la maximisation du nombre d'éléments des plus gros clusters.

3.3.3 Évaluation du clustering

L'approche de clustering est une démarche aveugle de fouille de données non étiquetées, au contraire de l'apprentissage supervisé où une variable cible de classification est connue permettant une quantification de l'erreur de classement. Le clustering dépend tout d'abord du nombre de clusters à construire. Ce paramètre de cardinalité de la partition doit être fixé a priori dans plusieurs algorithmes de partitionnement comme la méthode k-means. Souvent, il est déterminé par les connaissances métier de la problématique étudiée ou par le niveau de réduction de dimensionnalité que nous souhaitons avoir dans les données. Pourtant, dans notre cas d'étude, l'expertise métier n'a pas pu déterminer ce nombre de clusters. Les agents de SRD sont uniquement en capacité de spécifier une mesure de dissimilarité et un seuil maximal en dessous duquel deux courbes de charge pourraient être appartenir au même groupe. Pour répondre à cette problématique, l'algorithme EQW a été développé pour permettre d'identifier, à partir d'un seuil maximal, le plus petit nombre k de clusters.

De plus, le clustering peut être résumé par plusieurs indicateurs décrivant la structure interne des groupes et les relations de chaque groupe avec les autres groupes, ces indicateurs sont appelés indicateurs internes. Il existe d'autres indicateurs dits externes permettant une comparaison entre plusieurs solutions de clustering. Avec ces indicateurs, nous cherchons dans un clustering un compromis entre une maximisation des distances inter-groupes et une minimisation des distances intra-groupes. Vendramin *et al.* {2010} ont présenté un examen approfondi des métriques utilisées dans l'évaluation de la qualité du clustering. Nous avons utilisé dans notre étude deux indicateurs, un indicateur externe et un indicateur interne.

Indicateurs externes

La plupart des indicateurs externes sont fondés sur un tableau de contingence (ou matrice de confusion) entre deux solutions de clustering P et $P^{'}$, de taille 2×2 composé de quatre cellules contenant

— n_{11} : le nombre de paires d'objets appartenant au même cluster dans P et dans $P^{'}$,

— n_{10} : le nombre de paires d'objets appartenant au même cluster dans P mais n'appartenant pas au même cluster dans $P^{'}$,

— n_{01} : le nombre de paires d'objets n'appartenant pas au même cluster dans P mais appartenant au même cluster dans $P^{'}$,

— n_{00} : le nombre de paires d'objets n'appartenant pas au même cluster dans P et dans $P^{'}$.

L'indicateur Rand Index (RI) {Rand, 1971} est donné par

$$\omega = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

Cet indicateur présente le taux de cohérence entre deux solutions de clustering. ω est dans $[0,1]$, $\omega = 0$ signifie que les deux partitions sont totalement différentes ($n_{11} = n_{00} = 0$), $\omega = 1$ signifie que les deux partitions sont égales.

L'inconvénient de l'indicateur RI est qu'il n'est pas adapté pour le hasard. En effet, il peut donner une valeur non nulle pour une comparaison entre deux partitions aléatoires. Pour remédier à cela, Hubert et Arabie {1985} ont proposé l'indicateur Adjusted Rand Index (ARI) normalisant RI pour que sa valeur soit nulle lorsque les deux partitions sont choisies par hasard et 1 lorsqu'une correspondance parfaite est atteinte {Vendramin *et al.*, 2010}.

$$\omega_A = \frac{n_{11} - \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{11}+n_{10}+n_{01}+n_{00}}}{\frac{2n_{11}+n_{10}+n_{01}}{2} - \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{11}+n_{10}+n_{01}+n_{00}}}$$

Indicateurs internes

Dans l'évaluation interne du clustering, nous avons utilisé la méthode de **silhouette**. Cette méthode consiste à attribuer à chaque objet y_i de Ω un indice {Rousseeuw, 1987}

$$s(i) = \frac{b(i) - a(i)}{max(b(i), a(i))} \tag{3.1}$$

Où $a(i)$ est la distance moyenne entre y_i et tous les éléments de son groupe d'appartenance. $b(i)$ est la distance minimale entre y_i et tous les éléments d'un autre groupe.

$$a(i) = \frac{1}{|P_k| - 1} \sum_{y_j \in P_k, i \neq j} d(y_i, y_j) \text{ et } b(i) = \min_{l \neq k} \frac{1}{|P_l|} \sum_{y_j \in P_l} d(y_i, y_j)$$

$s(i)$ varie entre -1 et 1, lorsqu'elle est proche de 1 cela signifie que y_i est bien classé. $s(i)$ proche de 0 indique que y_i peut aussi bien être affecté à un cluster voisin. $s(i)$ proche de -1 signifie que l'objet y_i est mal classé et il doit être déplacé dans un autre groupe. La figure 3.10 illustre un exemple de silhouette dans un clustering de trois groupes.

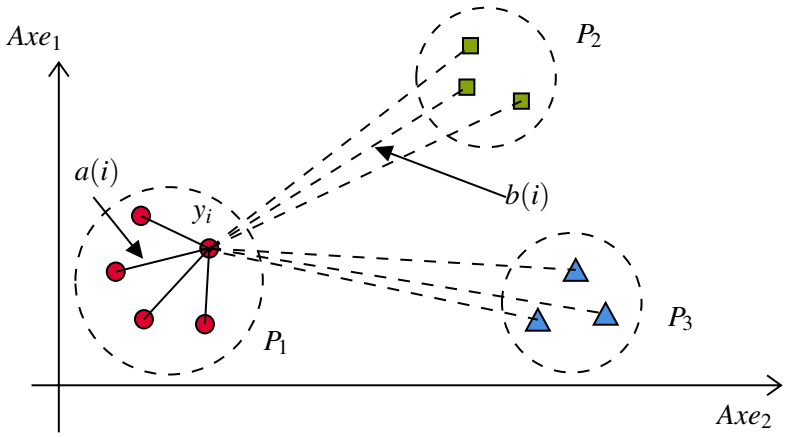


FIGURE 3.10 – Exemple de calcul de l'indice de silhouette.

L'indice de silhouette global pour un clustering est donné par la moyenne de tous les indices de silhouette $s(i)$ des objets y_i de Ω. Plus l'indice de silhouette global est proche de 1, plus le clustering est optimal.

3.3.4 État de l'art

Ces dernières années de nombreux travaux de recherche ont été développés dans le domaine de clustering des courbes de charge. Cet intérêt croissant est expliqué par l'évolution et la disponibilité d'un grand volume de données grâce aux compteurs communicants. L'exploration de ces données par les différents outils de modélisation ou de gestion de réseau peut devenir coûteuse en ressources informatiques (stockage et calcul).

Yang *et al.* {2013} ont présenté une synthèse approfondie des travaux de recherche sur le regroupement des données de charge dans un réseau intelligent. Le clustering permet une exploration de ces données de charge contenant des connaissances intéressantes pour les GRD. Ils ont présenté les méthodes de clustering fréquemment appliquées dans le regroupement des données de consommation ainsi que les indicateurs d'évaluation utilisés. Par la suite, ils ont présenté des applications du regroupement des courbes de charge comme l'identification des données atypiques ou aberrantes, la prévision de la charge et la fixation des prix d'électricité. Ce clustering des courbes de charge est influencé par l'environnement complexe du réseau intelligent et ses données de plus en plus massives et hautement dimensionnelles. Si *et al.* {2021} ont présenté un examen des algorithmes de clustering utilisés dans le clustering de la consommation électrique dans un réseau électrique intelligent. Ils ont résumé la méthodologie générale de regroupement des courbes de charge ainsi que les mesures de similarité, les algorithmes de clusterings et les indicateurs de validation utilisés avec leurs inconvénients et leurs avantages. Finalement, ils ont exposé des applications réelles du clustering dans le réseau électrique et les tendances de recherche future dans ce domaine. Les principales perspectives sont la détection des anomalies, la prévision de la charge, la sécurité des données et la réduction de dimensionnalité des données dans un environnement big data.

Tureczek *et al.* {2018} ont développé une méthodologie de regroupement de données en utilisant l'algorithme k-means et en exploitant des données réelles de 32 241 compteurs domestiques dans la ville danoise d'Esbjerg. Leur approche prenant en compte les autocorrélations dans les séries temporelles améliore les résultats de clustering obtenus par k-means en ajoutant la composante temporelle de données.

Granell *et al.* {2014} ont étudié l'effet de la granularité et de la taille des séries temporelles sur la qualité des résultats de clustering. Ils ont testé plusieurs algorithmes de clustering en exploitant des données réelles de consommations des résidentiels en variant la résolution temporelle de 60 secondes jusqu'à 240 minutes. Ils ont trouvé que les courbes de charge avec une granularité inférieure à 30 minutes, de préférence entre 8 minutes et 15 minutes, ont fourni un clustering de qualité comparé aux autres granularités. Les données avec une résolution supérieure ou égale à 60 minutes ne sont pas assez pertinentes pour un regroupement fiable des courbes de charge.

Bouveyron *et al.* {2018} ont élaboré une méthode de co-clustering [3] pour les données fonctionnelles. L'algorithme développé permet de résumer un grand ensemble de données de consommation afin de permettre son utilisation efficacement par les fournisseurs d'énergie. L'approche a été évaluée sur des données réelles collectées par les compteurs communicants (Linky) d'EDF avec une résolution de 30 minutes. Les résultats ont montré la pertinence de cette méthodologie.

La plupart des travaux de recherche dans le domaine du clustering de courbes de charge traitent les données des compteurs communicants individuels et résidentiels dans un niveau plus bas de réseau. Leur objectif est de réaliser une segmentation des différents consommateurs afin de regrouper les clients en étudiant leur comportement de consommation. Notre besoin de clustering, orienté par les différentes modélisations de l'optimiseur, consiste à étudier et à regrouper les courbes de charge au niveau des départs HTA selon les grandeurs et les niveaux des consommations mesurées, ce qui en général différent d'un regroupement par comportement de consommation.

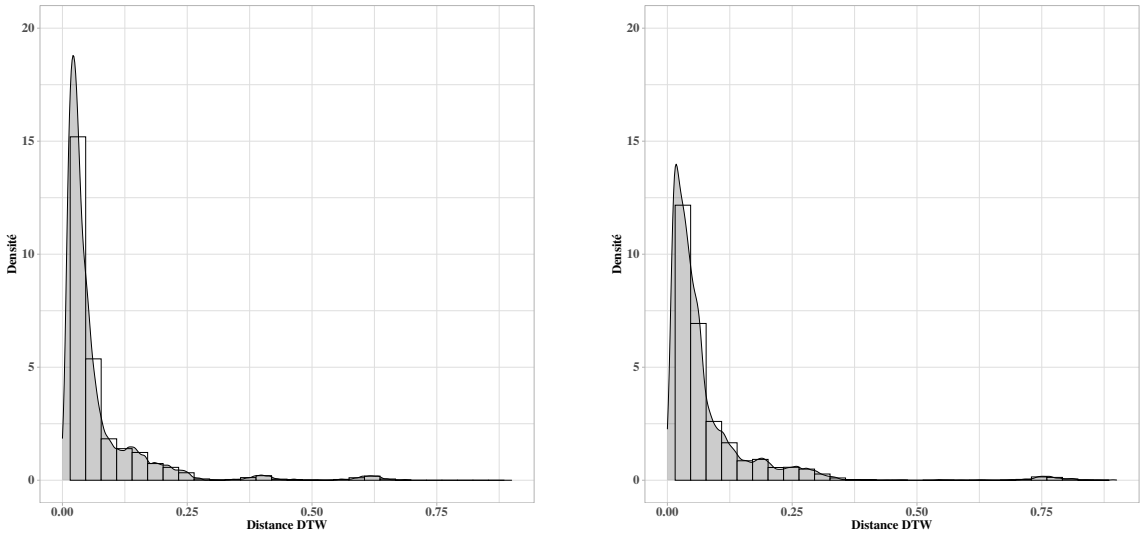
3.4 Données et expérimentation

Dans ce projet de clustering, nous avons utilisé les données des courbes de charge présentées dans la section 2.3.1. Ces données sont collectées en temps réel au niveau des départs HTA des postes sources et archivées avec une granularité temporelle de dix minutes au niveau de l'outil de conduite SCADA de SRD. Certains départs sont instrumentés par des technologies anciennes ne permettant la mesure que des données du courant en Ampère. D'autres départs, ainsi que les arrivées HTA, sont équipés par des technologies récentes permettant la mesure de la puissance active, la puissance réactive, la tension et le courant. Dans l'évaluation des méthodes de clustering, nous n'avons utilisé que les données du courant. Les puissances actives sont obtenues ensuite en supposant, pour une simplification des calculs, que le facteur de puissance $cos(\phi)$ égal à 1.

Nous avons au total les courbes de charge de 180 départs HTA au niveau de 16 postes sources du réseau de SRD. Nous avons éliminé certains départs dédiés vers des grands producteurs et les départs vers des postes d'étoilement. Nous avons gardé au final 145 départs que nous avons réussi à associer avec la base de données de la cartographie. Cette association permet de récupérer les informations sur la puissance souscrite totale cumulée au niveau du départ afin de normaliser les courbes de charge par cette puissance souscrite. Ces courbes normalisées sont les foisonnements instantanés au niveaux des départs HTA (voir section 2.2.1 chapitre 2). Pour un départ i, nous notons le foisonnement instantané par y_i. Nous avons utilisé les données mesurées entre le premier janvier 2017 et le 31 décembre 2017 avec une granularité de dix minutes. y_i est donc une série temporelle de taille $6 \times 24 \times 365 = 25\,560$.

Par la suite, nous avons calculé une matrice de distance entre toutes ces courbes de charge en utilisant la distance DTW avec une bande de Sakoe-Chiba égale à $r = 4$. Nous avons supposé que cette plage

3. Le co-clustering ou le clustering double est un clustering permettant de regrouper simultanément les individus et les variables au sens statistique.



(a) Distribution des distances DTW en été.

(b) Distribution des distances DTW en hiver.

FIGURE 3.11 – Distributions des distances DTW calculées entre les départs HTA.

de 80-minutes (40 minutes à droite et 40 minutes à gauche) est suffisante dans la comparaison. Vu le volume de données, nous n'avons pas réussi à calculer une matrice de distance sur toute la période annuelle. Nous avons donc divisé ces données sur des périodes mensuelles, ce qui nous a permis de comparer le clustering entre chaque mois sur deux saisons différents, l'été et l'hiver. Dans l'étude de la consommation, la saison d'été commence d'avril jusqu'à octobre. Un clustering sur la saison est ainsi calculé. Nous avons appliqué ensuite trois algorithmes de clustering : un algorithme hiérarchique CAH, un algorithme de partitionnement k-médoïdes et un algorithme original Equiwide. Finalement, une comparaison par l'indicateur ARI est réalisée entre les différents clustering obtenus par les trois algorithmes sur toutes les périodes temporelles. Une évaluation de la structure interne est aussi élaborée par l'indicateur de la silhouette. La figure 3.11 illustre les distributions des distances DTW calculées entre les différentes courbes de charge. Ces distances normalisées par T^2 (T est la taille de la série) ont la même dimension que les foisonnements.

Les résultats de cette étude de clustering ont été validés par une simulation d'estimation d'état pour une validation métier.

L'évaluation expérimentale a été réalisée dans les conditions réelles d'application sur un ordinateur équipé d'un processeur Intel Core i3-6006U à 3,00 GHz, de 2 cœurs, de 12 Go de RAM et d'un système d'exploitation Windows 10.

3.5 Résultats

Nous avons évalué les trois algorithmes de clustering en exploitant les données de consommation d'électricité collectées au niveau des départs HTA des postes sources. Les différents résultats sont présentés dans cette section. Nous avons fixé l'algorithme CAH comme algorithme de référence afin de comparer les résultats des deux autres algorithmes. L'algorithme CAH est évalué dans une première partie en variant le nombre k de clusters afin d'étudier la structure interne et externe des clusterings obtenus pour chaque période mensuelle et saisonnière. Par la suite, nous avons comparé le clustering proposé par CAH avec les deux clustering obtenus par EQW et k-médiodes. Nous notons que la détermination du

seuil maximal pour l'algorithme EQW est empirique en variant le rayon maximal pour obtenir un nombre de cluster égal à celui fixé pour CAH et k-médiodes. Ceci est différent du fondement de l'algorithme qui consiste à identifier un clustering à partir d'un seuil fixé a priori par le décideur. Cette expérimentation permet de le comparer avec les différents résultats obtenus par les deux autres algorithmes.

La figure 3.12 illustre les résultats de comparaison en utilisant l'indicateur externe ARI entre chaque clustering obtenu pour chaque période. Pour 2 ou 3 clusters, une ressemblance entre les différents clusterings est observée pour toutes les périodes mensuelles et saisonnières. A partir de 4 clusters, nous perdons cette stabilité de regroupement. Pour 3 clusters, nous avons une ARI de 95 % entre été et hiver, pour 4 et 5 clusters nous avons des ARI de 50 % et 40 %. Dans les cas de 4 et 5 clusters nous observons également une ressemblance entre les clusterings obtenus en août et octobre et la période totale d'été. Nous notons que le mois d'août est la période de moindre activité pour la plupart des industries.

Afin d'évaluer la structure interne de clustering, nous avons calculé l'indicateur interne présenté dans la section 3.3.3 pour les deux clusterings d'été et d'hiver en variant le nombre de clusters entre 2 et 20. La figure 3.13 illustre les résultats obtenus. Dans le cas de 3 clusters, nous avons une silhouette moyenne de 0.72 pour la période d'été et 0.73 pour la période d'hiver. Cette silhouette passe à 0.49 pour l'été et à 0.28 pour l'hiver dans le cas de 4 clusters. La hauteur des arbres hiérarchiques définissant une distance entre les hiérarchies selon le saut de Ward est égale à 0.57 pour un découpage en 2 clusters et à 0.34 pour 3 clusters dans le cas d'été. Cette hauteur, dans le cas de l'hiver, est égale à 0.73 pour 2 clusters et 0.37 pour 3 clusters. Nous remarquons donc une forte différence entre les niveaux d'agrégations 3, 2 et 1 groupes. Cela indique une modification de la structure interne des clustering en maximisant les distances inter-groupes. Le tableau 3.1 résume la structure interne de chaque clustering selon le nombre de clusters. Avec 3 clusters, nous avons obtenu un grand cluster de 134 objets, dans le cas de l'été suivi d'un cluster de 9 objets et un petit cluster de 2 éléments. Une répartition équivalente est conservée dans le cas de l'hiver avec 133 objets pour le grand cluster, 11 et 1 objets dans les autres clusters. Avec un clustering de 4 clusters, le grand cluster est divisé en deux clusters de 115 et 19 éléments dans le cas d'été. Dans le cas d'hiver, il est divisé en deux clusters de 79 et 54 éléments.

Les indicateurs ARI et silhouette montrent qu'à partir d'un nombre de clusters supérieur ou égal à 4 la pertinence de culstering diminue largement. Ceci s'explique par la division d'un grand cluster, trouvé dans les cas de deux et de trois clusters, en plusieurs clusters très proches pour un k supérieur ou égal à 4. En effet, la silhouette des éléments frontaliers des clusters très proches diminue car ils peuvent être déplacés dans des clusters voisins. Le rayon maximal obtenu pour le grand cluster est de 0.113 dans les deux cas d'été et d'hiver, ce qui représente 16% de la distance maximale calculée en été et 12% de la distance maximale calculée en hiver (voir 3.11). En regardant ces différents résultats obtenus par CAH, le culstering est pertinent dans les cas de deux et de trois clusters. Les silhouettes sont maximaux dans ces deux cas tout en conservant la même structure de clustering sur toutes les périodes de l'année.

3.5.1 Résultats d'évaluation de l'algorithme EQW

L'algorithme EQW a été évalué en exploitant les données de SRD sur les périodes d'été et d'hiver et en comparant avec les solutions de clusterings obtenues par les deux algorithmes CAH et k-médoïdes.

Avec des rayons maximaux de 0.168 dans l'hiver et 0.163 dans l'été, nous avons obtenu trois clusters en utilisant l'algorithme EQW. Quatre clusters sont obtenus avec des rayons maximaux de 0.167 dans l'hiver et 0.162 dans l'été. Nous notons que dans le cas de l'algorithme CAH les rayons maximaux sont égaux à 0.206 pour l'hiver et à 0.266 pour l'été (voir tableau 3.1). La figure 3.14 illustre les résultats

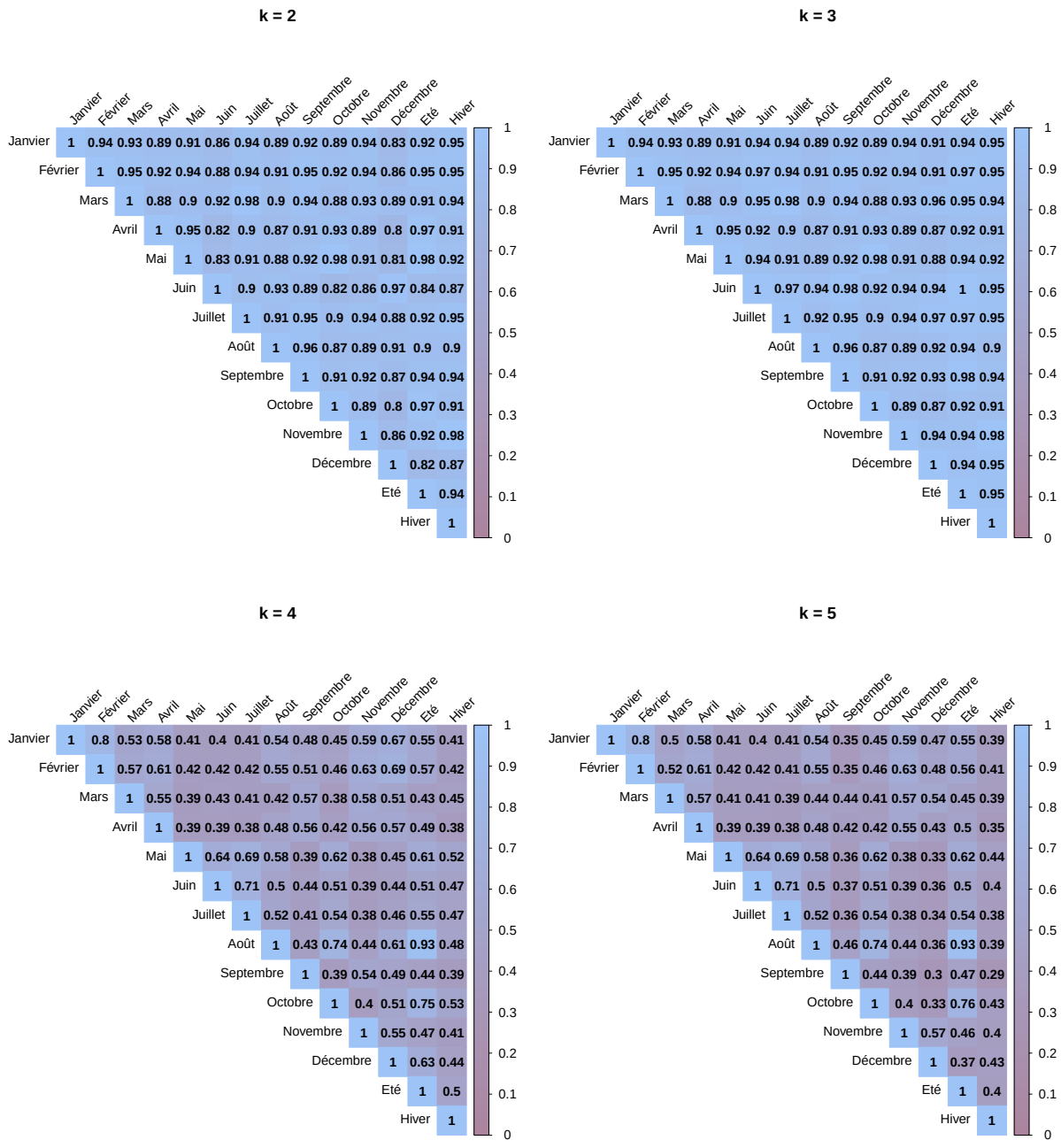
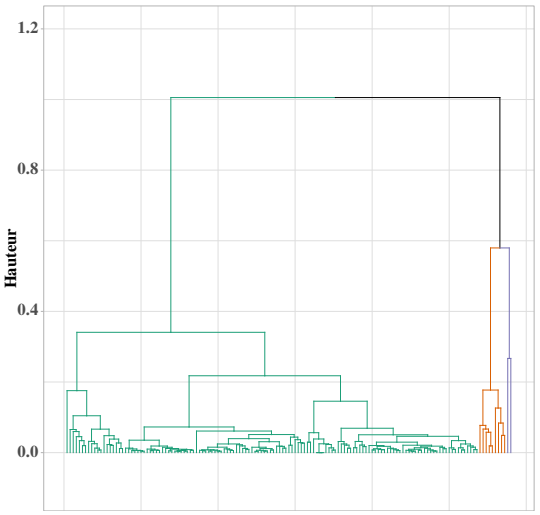
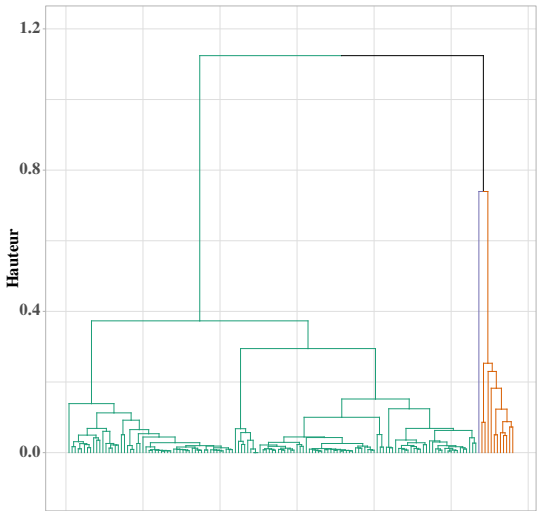


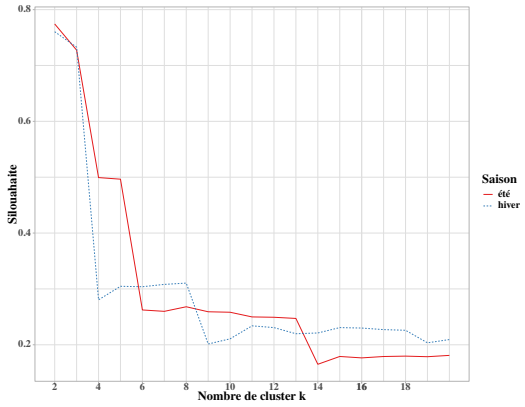
FIGURE 3.12 – Résultats de comparaison entre les clustering de chaque période en utilisant ARI. Lecture : Chaque figure illustre les résultats de comparaison en utilisant l'indicateur externe ARI entre les clusterings obtenus par CAH pour chaque période selon le nombre de cluster.



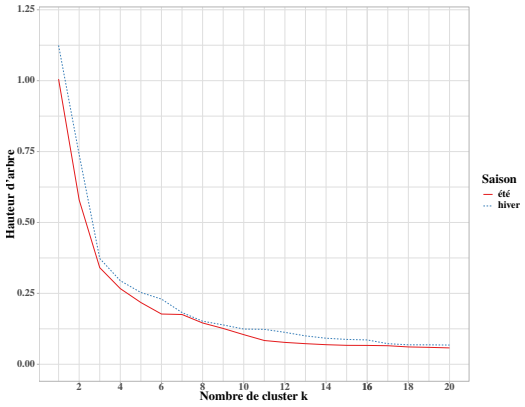
(a) Arbre hiérarchique clustering été.



(b) Arbre hiérarchique clustering hiver.



(c) Silhouette moyenne obtenue dans le cas des deux saisons été et hiver.



(d) Hauteur des arbres dans le cas des deux saisons été et hiver.

FIGURE 3.13 – Résultats de clustering des départs HTA du réseau de SRD.
Lecture : Les premières figures (en haut) 3.13a et 3.13b illustrent les arbres hiérarchiques obtenus dans le cas des départs de SRD avec un découpage en 3 clusters sur les deux périodes été et hiver. Les figures en bas 3.13d et 3.13c illustrent la hauteur des arbres et la silhouette moyenne selon le nombre de clusters k dans les deux périodes d'été et d'hiver.

Période	k	Cluster	Nombre d'élément	Diamètre max	Rayon max	Distance moyenne Cluster	Silhouette Moyenne cluster
Été	2	1	134	0.205	0.113	0.038	0.823
		2	11	0.486	0.247	0.172	0.178
Été	3	1	134	0.204	0.113	0.038	0.753
		2	9	0.178	0.103	0.087	0.465
		3	2	0.266	0.266	0.266	0.148
Été	4	1	115	0.101	0.051	0.028	0.569
		2	19	0.151	0.113	0.050	0.222
		3	9	0.178	0.103	0.087	0.266
		4	2	0.266	0.266	0.266	0.148
Hiver	2	1	133	0.192	0.113	0.044	0.818
		2	12	0.640	0.356	0.198	0.122
Hiver	3	1	133	0.192	0.113	0.044	0.772
		2	11	0.257	0.206	0.128	0.315
		3	1	-	-	-	-
Hiver	4	1	79	0.136	0.114	0.039	0.204
		2	54	0.115	0.089	0.031	0.417
		3	11	0.257	0.206	0.128	0.182
		4	1	-	-	-	-

TABLEAU 3.1 – Résultats des indicateurs internes obtenus dans le cas de clustering par CAH sur les deux périodes d'été et d'hiver. Le diamètre, le rayon maximal et la distance moyenne des clusters ont la même dimension que les foisonnements.

de comparaison par l'indicateur externe ARI entre les trois algorithmes sur les deux périodes. Nous observons une ressemblance pour les algorithmes CAH et EQW dans le cas de 3 clusters, 88% pour la période d'été et 95% pour la période d'hiver. Une différence est observée entre ces deux algorithmes et l'algorithme k-médoïdes, 43% entre EQW et k-médoïdes sur la période d'été et 59% pour la période d'hiver. Dans le cas de 4 clusters, nous avons une dissemblance sur l'ensemble des algorithmes sauf pour EQW et CAH dans la période d'été (78%) et entre EQW et k-médoïdes dans la période d'hiver. L'algorithme EQW reste stable entre les deux périodes hiver et été avec une ARI de 76%.

3.5.2 Discussion

Dans ce projet, nous avons développé une méthodologie de réduction de dimensionnalité des données de consommation en utilisant des méthodes d'apprentissage automatique non supervisé. L'évaluation des ces algorithmes en utilisant les données réelles de SRD sur plusieurs périodes de l'année a montré que ces données peuvent être regroupées en plusieurs clusters. Avec une partition en trois clusters, nous avons un grand cluster et deux autres petits clusters. La figure 3.15 illustre les clusterings obtenus par CAH, EQW et k-médoïdes en les représentant par la méthode MDS. Les départs limitrophes d'un cluster avec un cluster voisin ont une silhouette faible parfois négative. Ils peuvent être déplacés sur le cluster voisin. L'algorithme EQW a regroupé l'ensemble des départs sur un grand cluster et deux départs dans deux clusters disjoints.

L'exploration de ces départs éloignés des autres (regroupés dans les deux petits clusters) a montré que ce sont des départs exceptionnels (extrêmes) ou aberrants. La figure 3.17 représente les courbes de ces départs selon les clusters obtenus par CAH. Le départ du cluster 1 est le départ médoïde (représentant). Les départs regroupés dans le cluster 3 par le CAH sont des départs particuliers de SRD. Ils sont configurés avec un schéma d'exploitation différent comparé aux autres départs pour des raisons métiers

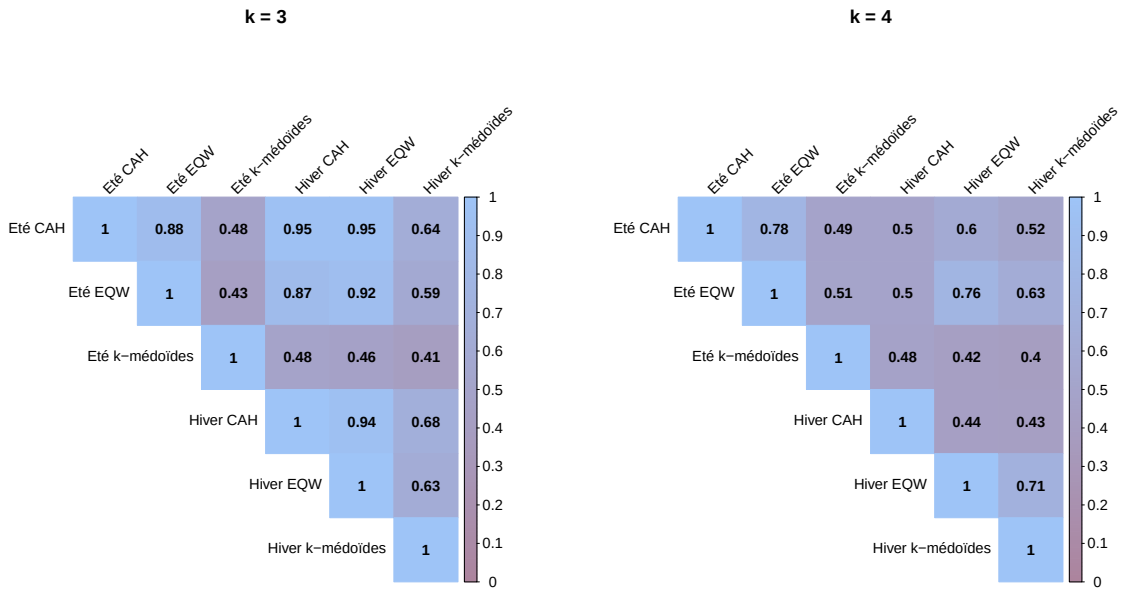


FIGURE 3.14 – Résultats de clustering par l'algorithme EQW.
Lecture : les deux figures illustrent les résultats de comparaison en utilisant l'indicateur externe ARI entre les clusterings obtenus par CAH, EQW et k-médoïdes selon le nombre de cluster.

de distribution. Les départs regroupés dans le cluster 2 par CAH desservent des grands clients industriels (consommateur HTA) qui ont en général une puissance souscrite supérieure à 1 MW. Certains de ces départs contiennent des périodes de consommation différentes surtout pendant la période estivale (mois août) et les week-ends. D'autres départs contiennent des périodes de consommation aberrantes en raison des défaillances de compteurs collectant les données. Le grand cluster 1 contient des départs desservant majoritairement les petits industriels et les résidentiels. Nous rappelons que 90% des clients de SRD sont des résidentiels. Cela peut expliquer la ressemblance entre les différents départs du cluster 1. La figure 3.16 montre la distribution des foisonnements des départs de chaque cluster obtenus par CAH sur une période d'été. Nous observons une différence entre les foisonnements des deux clusters 2 et 3 et les foisonnements du cluster 1. Les foisonnements des clusters 2 et 3 sont supérieurs aux foisonnements du cluster 1.

Ces résultats montrent que pour une réduction du nombre des courbes de charge, nous devons traiter les départs extrêmes séparément des autres départs. Le grand cluster regroupant la majorité des départs peut être divisé ensuite en sous-clusters selon les besoins de modélisation et de réduction de dimensionnalité. L'avantage de l'algorithme EQW pour un regroupement en lien avec un besoin métier est de s'assurer que les objets des sous-clusters respectent une certaine distance maximale intra-groupe. Cependant, la distance DTW est difficile à interpréter par l'expertise métier pour fixer le seuil maximal. Le choix du seuil reste donc empirique en étudiant les distributions de toutes les distances calculées entre tous les éléments. Pour remédier à cela, la distance euclidienne ou la distance de Manhatan ($d(y_i, y_j) = \sum_{t=1}^{T} |y_{i,t} - y_{j,t}|$) normalisées peuvent être appliquées, car elles sont plus faciles à expliquer pour l'expertise métier en introduisant la notion d'erreur absolue maximale. En revanche, elles restent des distances rigides pour les données temporelles.

Afin d'évaluer les hypothèses de foisonnement utilisées dans l'estimation de la charge au niveau de l'optimiseur, nous avons élaboré une simulation par l'estimateur d'état en utilisant plusieurs scénarios

sur les niveaux de foisonnement et en exploitant la méthodologie de clustering. Les différents résultats sont présentés dans la section suivante.

3.5.3 Validation par l'estimateur d'état

Nous rappelons que l'un des objectifs de clustering pour SRD est de confirmer ou infirmer les coefficients empiriques de foisonnement dans la modélisation de l'optimiseur. Cela permet ainsi une validation métier du clustering des départs HTA. Afin de réaliser cette expérimentation des foisonnements regroupés, nous avons élaboré une simulation d'estimation d'état du réseau pour chaque scénario de foisonnement en comparant avec les foisonnements individuels de chaque départ. Pour cela, nous avons utilisé le réseau de distribution de l'agence 6 de SRD en exploitant les courbes de charge mesurées en 2020. L'avantage de cette agence est qu'elle est desservie par quatre postes sources différents. Tous ces départs HTA (26 départs) font partie du grand cluster obtenu avant, aucun départ extrême n'existe sur cette agence. La figure 3.18 illustre le clustering obtenu par la méthode CAH sur une période annuelle. Pour les 26 départs de l'agence 6, nous pouvons découper l'arbre en 2 ou 4 clusters. Nous avons donc testé 4 scénarios différents :

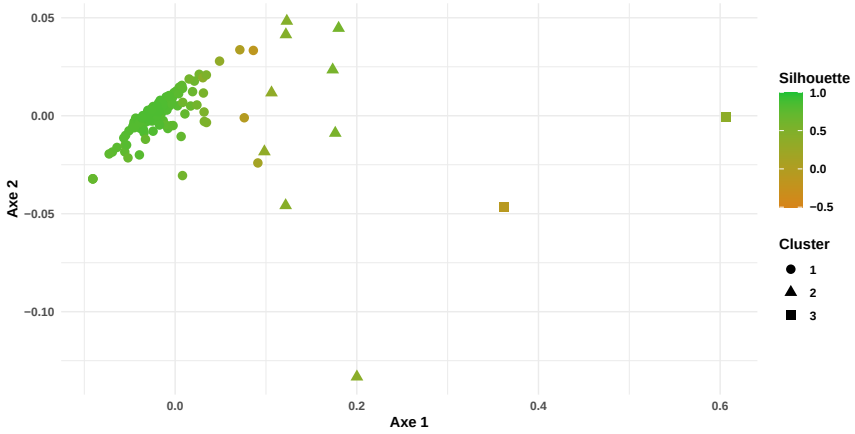
— Estimation de la charge par le foisonnement au niveau de l'agence.

— Estimation de la charge par le foisonnement au niveau du poste source.

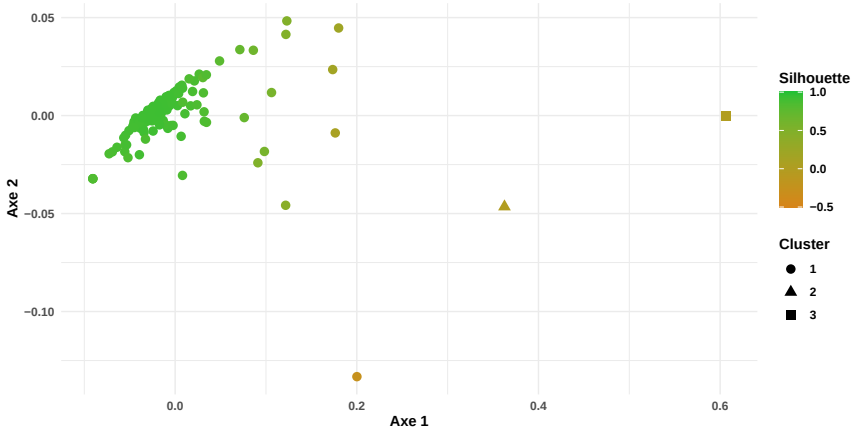
— Estimation de la charge par le foisonnement au niveau du cluster.

— Estimation de la charge par le foisonnement au niveau du départ HTA.

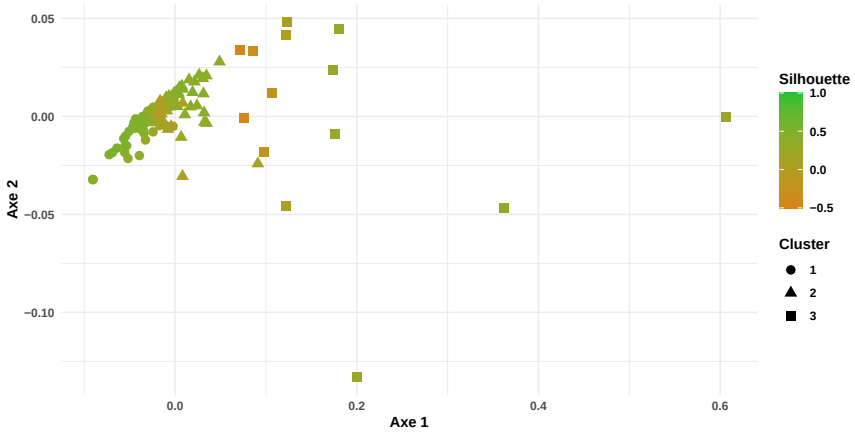
Nous avons comparé les résultats d'estimation d'état obtenus par les différents niveaux de foisonnement avec les résultats d'estimation obtenus en utilisant les foisonnements des départs. Concernant le clustering, nous avons testé deux possibilités : un clustering de deux clusters (clustering 2) et un clustering de 4 clusters (clsutering 4) obtenus par CAH. L'algorithme EQW a fourni un regroupement en 4 clusters avec un rayon maximal de 0.29 et un regroupement en 2 clusters avec un rayon de 0.389. La comparaison avec CAH en utilisant ARI est de 62% pour le custering 4 et 100% pour le clustering 2. L'estimation d'état est réalisée sur une période annuelle en utilisant deux schémas différents : un schéma normal et un schéma optimal obtenu par un foisonnement fixe de 15% au niveau de l'agence. Le tableau 3.2 résume les résultats de comparaison avec les pertes estimées dans chaque scénario de foisonnement en utilisant les métriques présentées dans la section 2.2.4. Le scénario clustering 4 est plus proche du scénario départ avec une MAPE de 5.9 %, suivi des scénarios clustering 2 et agence avec des MAPE de 6.5% et 6.6%. Le scénario foisonnement poste source est le moins pertinent par rapport aux autres scénarios avec une MAPE de 7.9%. En effet, les départs d'un poste source peuvent appartenir à des clusters différents. Donc, le regroupement par poste source peut regrouper deux départs éloignés et moins homogènes. Par conséquent, le foisonnement du poste source dans ce cas peut être inapproprié pour estimer la charge, car certains départs sont différents aux autres départs (présence d'industriels, schéma modifié, etc). Nous notons que le scénario de foisonnement utilisé aujourd'hui est ce scénario poste source. Le clustering permet donc de mieux estimer la charge au niveau de l'optimiseur par un foisonnement calculé au niveau d'un cluster. Le foisonnement de l'agence ici est le foisonnement de cluster le plus haut dans l'arbre hiérarchique, l'espace Ω de tous les départs.



(a) Clustering CAH



(b) Clustering EQW



(c) Clustering k-médoïdes

FIGURE 3.15 – Clustering obtenu dans le cas de la période d'été visualisé par la méthode MDS. Lecture : L'axe des abscisses et l'axe des ordonnées représentent les dimensions obtenues par la méthode MDS.

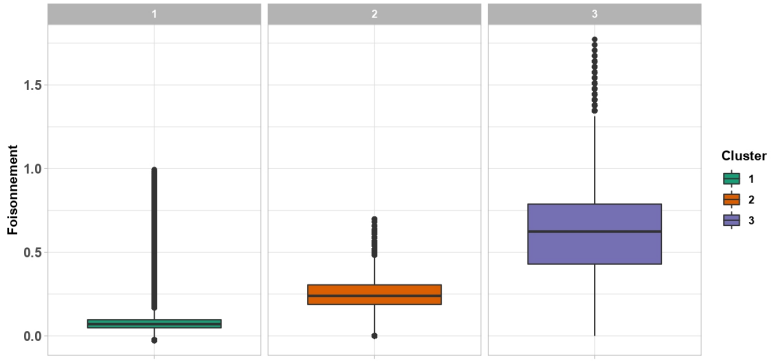


FIGURE 3.16 – Distributions des foisonnements de chaque cluster sur une période d'été.
Lecture : L'axe des ordonnées représente le foisonnement et l'axe des abscisses représente les boîtes à moustaches de chaque cluster sur l'ensemble des départs.

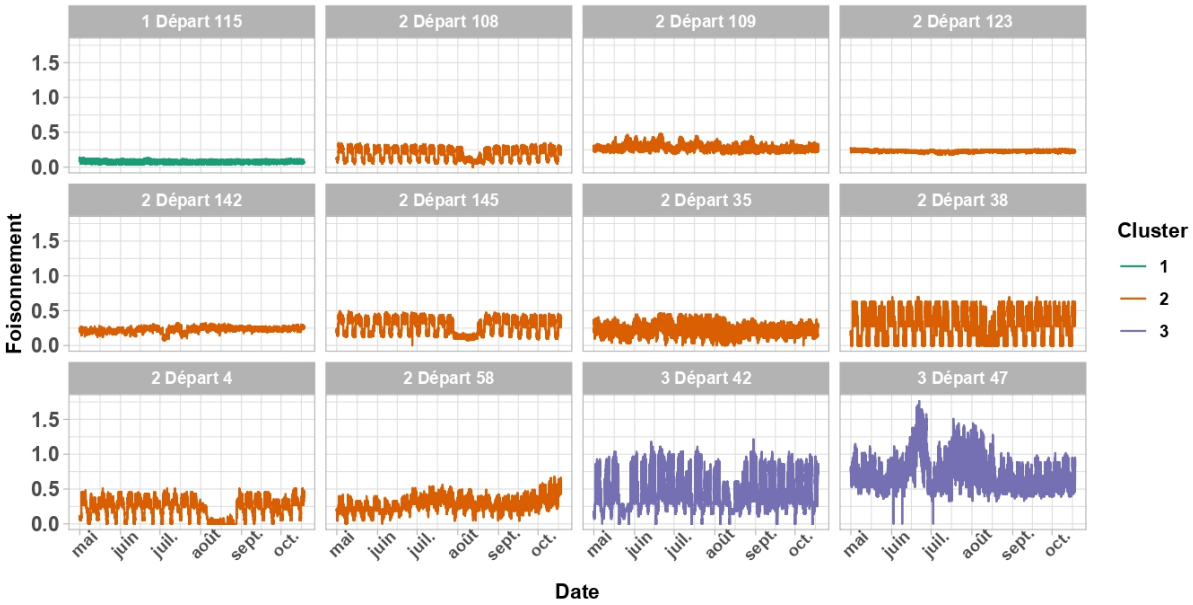


FIGURE 3.17 – Exemple de courbes de charge des départs selon le cluster.
Lecture : L'axe des abscisses représente la date sur une période d'été et l'axe des ordonnées représente le foisonnement. Les blocs illustrent les courbes des départs en fonction des clusters obtenus par CAH sur la période d'été. Le cluster 1 est représenté par le départ médoïde.

Schéma	Type de foisonnement	MBE^+ (MWh)	MBE^- (MWh)	MBE (MWh)	MAE (MWh)	RMSE (MWh)	MAPE (%)	MPE (%)
Normal	Clustering 4	0.026	-0.029	0.009	0.027	0.037	5.912	1.271
Normal	Clustering 2	0.033	-0.027	0.008	0.030	0.042	6.559	0.729
Normal	Agence	0.032	-0.028	0.004	0.030	0.041	6.671	-0.044
Normal	Poste source	0.043	-0.023	0.026	0.038	0.051	7.972	4.260
Optimal	Clustering 4	0.025	-0.025	0.009	0.025	0.033	6.273	1.711
Optimal	Clustering 2	0.029	-0.025	0.012	0.028	0.036	6.825	1.936
Optimal	Agence	0.027	-0.025	0.009	0.026	0.035	6.682	1.305
Optimal	Poste source	0.036	-0.024	0.023	0.033	0.044	8.082	4.529

TABLEAU 3.2 – Comparaison entre les pertes estimées par le foisonnement des départs et les pertes estimées par d'autres types de foisonnement.

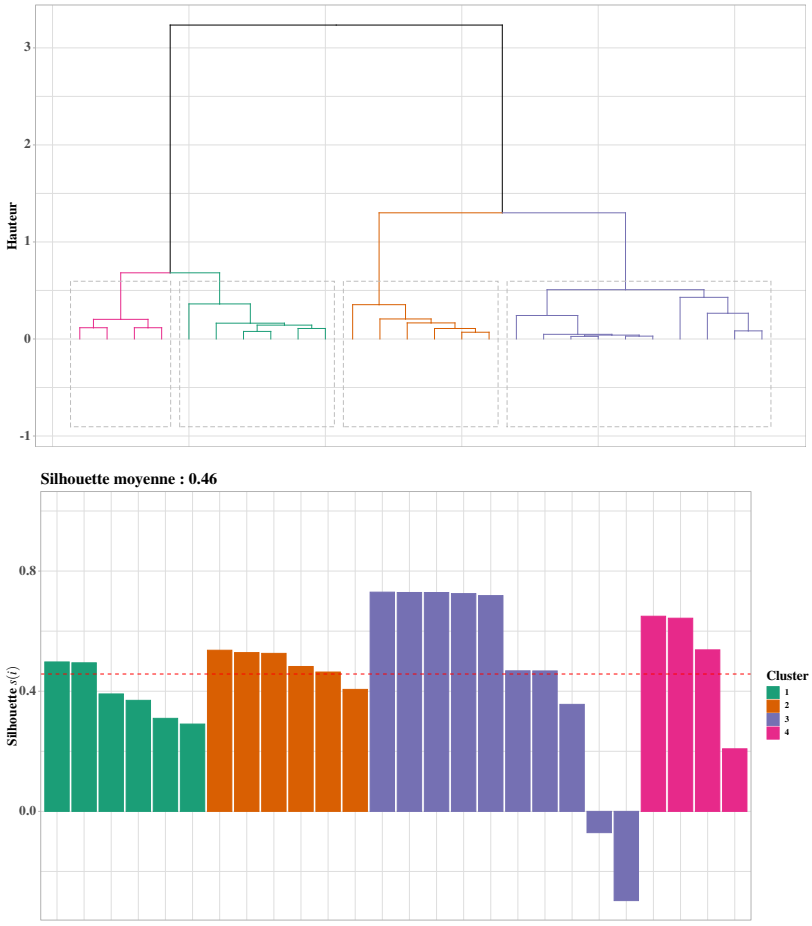


FIGURE 3.18 – Résultats de clustering des départs HTA de l'agence 6.
Lecture : La première figure (en haut) illustre l'arbre hiérarchique obtenu dans le cas de l'agence 6 avec un découpage en 4 clusters. La deuxième figure (en bas) illustre les indices de silhouette calculés pour chaque départ y_i.

3.6 Conclusion du chapitre

Dans cette étude, nous proposons une méthodologie de réduction de dimensionnalité des données de consommation d'un réseau de distribution de l'électricité. Cette approche fondée sur les techniques de regroupement de données permet au GRD de sélectionner les données les plus pertinentes en éliminant certaines redondances en détectant les similitudes existant dans ces données. Ceci permet de regrouper l'ensemble des courbes de charge disponibles dans des groupes homogènes. Cette homogénéité permet d'identifier des paramètres représentatifs de chacun des consommateurs du réseau et de détecter des consommateurs atypiques ou aberrants.

Ce projet permet ainsi d'évaluer des hypothèses de modélisation de la charge au niveau de l'optimiseur en utilisant les résultats de clustering. En effet, comparé aux foisonnements de chaque départ, l'estimation de la charge par les foisonnements au niveau des postes sources est moins pertinente que l'estimation par les foisonnements calculés au niveau des clusters.

Il sera judicieux de calculer un clustering sur les courbes de charge réelles au niveau des postes HTA/BT. Ceci permet de trouver le scénario d'estimation le plus proche de la réalité au niveau de chaque poste HTA/BT, car le foisonnement d'un départ HTA se situe dans un niveau plus haut du réseau. Cependant, ces courbes de charge dans un niveau plus bas de réseau ne sont pas disponibles à ce jour. Avec le déploiement des compteurs communicants sur tout le réseau, une grande partie des courbes de charge individuelles seront disponibles pour tous les clients autorisant la communication de leurs données. Une étude de clustering sur ces courbes de charge est intéressante pour valider au mieux les hypothèses de foisonnements et les résultats de clustering.

L'approche de clustering en utilisant la distance DTW est une stratégie pertinente de la réduction de dimensionnalité des données temporelles collectées au niveau d'un réseau de distribution de l'électricité. De plus, l'algorithme EQW élaboré dans cette étude permet de regrouper les données selon une erreur maximale définie par le décideur métier.

Chapitre 4

Estimation de la production PV

Dans ce chapitre, nous présenterons des méthodes d'interpolation spatiale pour estimer la production d'énergie des installations photovoltaïques (PV) distribuées dans un réseau de distribution d'électricité. L'objectif de cette étude est d'estimer la production des producteurs PV pour lesquels nous ne disposons pas de mesures des puissances avec une granularité fine. Nous avons utilisé les données de 3692 producteurs répartis sur une zone d'environ 7000 km^2 *pour évaluer les différentes méthodes d'interpolation. Sur cet ensemble de données, nous montrons que la méthode IDW (Inverse distance weighting) offre une meilleure précision d'estimation en comparant avec des méthodes naïves d'interpolation.*

Sommaire

4.1 Introduction

Pour faire face au réchauffement climatique, le développement des EnR dans le monde, et en France en particulier, a connu des changements majeurs ces dernières années. Le réseau français de distribution de l'électricité (HTA et BT) a été témoin d'une intégration importante de ce type de production.

SRD, comme les autres GRD, a connu une croissance considérable du nombre de producteurs PV et éoliens dans son réseau. La capacité de production EnR raccordée sur le réseau exploité par SRD est passée de 33 MW en 2010 à 366 MW en 2020. Ce parc EnR a injecté en 2020 une énergie de 725 GWh dans le réseau.

L'augmentation de ces moyens de production EnR génère plusieurs contraintes dans le réseau comme des fluctuations de tension, l'inversion du flux de puissance, etc {Agüero et Steffel, 2011}. Ils sont généralement intermittents, en particulier les énergies PV et éolienne dont la production dépend de plusieurs facteurs externes météorologiques. L'estimation et la prévision de ce type de production sont essentielles pour une gestion efficace et optimale du réseau. Elles nous permettent à la fois de minimiser les pertes d'énergie et les coûts de transmission, de contrôler les plages de tension sur l'ensemble du réseau et d'intégrer correctement les énergies renouvelables.

Dans ce chapitre, nous présenterons une méthodologie d'estimation des courbes de productions PV avec une granularité fine dans le cas de producteurs non instrumentés par des compteurs communicants. Cette estimation est fondée sur des méthodes d'interpolation spatiale en utilisant les données disponibles pour d'autres producteurs. Une méthodologie d'évaluation et de validation de ces méthodes est aussi présentée.

4.2 Objectif détaillé

Nous avons vu dans le chapitre 1 que dans le cadre de la thèse d'Ali Zazou {2017}, SRD a développé un outil d'optimisation de réseau de distribution en minimisant les pertes et en respectant les contraintes de dimensionnement du réseau. Au travers de la présente thèse, SRD cherche à s'équiper d'un outil de prévision de la production et de la consommation d'électricité, dans le but de d'alimenter un outil d'optimisation du réseau.

Pour élaborer un tel outil de prévision nous avons besoin de plusieurs données d'entrée, en particulier les données des courbes de la consommation et des courbes de la production de toutes les charges et les productions du réseau. Ces données sont généralement des séries temporelles, c'est-à-dire des suites de valeurs décrivant l'évolution de la quantité de la puissance consommée ou produite au cours du temps. Elles sont mesurées dans des endroits et des niveaux différents du réseau avec une granularité temporelle de dix minutes. Les courbes de charge sont collectées au niveau des départs HTA et les courbes de production sont mesurées au niveau des producteurs dont la capacité installée est supérieur à 36 kVA.

Comme nous avons vu dans la section 2.3.1 sur les données de télémesures, aujourd'hui à SRD les courbes de charges avec une granularité fine sont disponibles seulement au niveau des départs HTA. Ces courbes de consommation représentent la somme de toutes les consommations moins la somme de toutes les productions des clients et des producteurs connectés au départ HTA. C'est-à-dire, qu'une courbe de charge mesurée au niveau d'un départ HTA est une agrégation de toutes les courbes de charges des consommateurs moins les productions produites par les producteurs décentralisés de ce départ. Cependant, dans la modélisation de l'optimiseur, il est impératif de séparer le flux des consommations des flux

des productions. En effet, sans cette différenciation il n'est pas possible de modéliser les instants où la production égale la consommation. Sans séparation, nous considérons qu'il n'existe aucun transit dans le réseau, pourtant il y a un flux des productions décentralisées vers les consommateurs. L'énergie produite par ces producteurs décentralisés est consommée localement par les consommateurs voisins.

En revanche, comme nous avons pu le voir dans la section 2.3.1 certains producteurs ne sont pas équipés de compteurs permettant un enregistrement d'une courbe de production fine (granularité de dix minutes). Conséquemment, pour différencier la consommation de la production, il est indispensable d'estimer la part de la production manquante de ces producteurs avec une même granularité que celle des courbes de consommation. Généralement, ces producteurs sont des petits producteurs PV raccordés au niveau BT du réseau. Les autres producteurs comme les producteurs éoliens sont raccordés directement au réseau HTA et leur courbe est disponible dans l'outil de télérelève. Concernant les producteurs PV, nous distinguons trois catégories de producteurs :

1. Les petits producteurs ayant une capacité installée de moins de 36 kVA. Cette production est généralement une production en toiture des maisons des particuliers. Dans le cas de SRD, cette production est appelée PV diffus.
2. Les moyens producteurs ayant une capacité installée comprise entre 36 kVA et 250 kVA.
3. Les grands producteurs ayant une capacité installée dépassant 250 kVA. Ces producteurs sont raccordés au réseau HTA.

Les petits et moyens producteurs sont raccordés au réseau BT (400 V/230 V), tandis que les grands producteurs sont connectés directement au réseau HTA (20 kV). Les grands producteurs sont généralement surveillés et commandables en temps réel, les données étant archivées avec un pas de dix minutes. Les moyens producteurs sont équipés de compteurs permettant la télérelève a posteriori d'une courbe de charge avec un pas de dix minutes. Pour les petits producteurs, nous ne disposons pas aujourd'hui d'une courbe de charge, mais seulement d'un index annuel (production d'énergie cumulée sur une année).

Dans cette étude, nous avons développé une approche d'estimation de la production PV de ces petits producteurs en utilisant des méthodes d'interpolation spatiale en exploitant les données disponibles des autres producteurs. La production des moyens et grands producteurs sera utilisée comme référence dans l'estimation de la production des petits producteurs.

Avant la présentation de notre approche d'estimation, nous exposons des généralités sur la production PV et un panorama du parc PV de SRD.

4.2.1 Production photovoltaïque

L'énergie photovoltaïque est produite à partir du rayonnement solaire reçu par la terre. Ce rayonnement est une onde électromagnétique diffusée par la surface du soleil suite aux réactions de la fusion de l'hydrogène en hélium dans le noyau. Chaque seconde, le soleil libère environ 3.89 MJ d'énergie nucléaire qui est rapidement convertie en énergie thermique {Islam *et al.*, 2011}. Ce flux d'énergie est transporté ensuite vers la surface et diffusé sous forme de rayonnement électromagnétique. La quantité d'énergie émise par le soleil est de l'ordre de 64 MW/m^2. Après avoir voyagé plus de 150×10^6 km, seule une petite partie de cette énergie (1.7 MW/m^2) atteint le sommet de l'atmosphère terrestre sans absorption dans l'espace. Cette quantité reçue par la terre est appelée la constante solaire. Cependant, elle n'est pas totalement utilisable, en raison de la réflexion de l'onde reçue et donc de l'énergie qu'elle transporte {Islam *et al.*, 2011}. Ce phénomène de réflexion est appelé l'albédo. L'albédo représente la

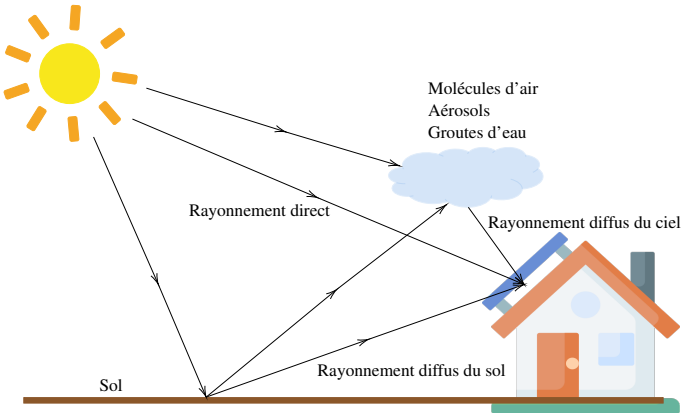


FIGURE 4.1 – Schéma illustrant les trois types de rayonnement reçus par les panneaux solaires (adapté de {Islam *et al.*, 2011})

part du rayonnement solaire renvoyé vers l'atmosphère, il est exprimé par un facteur entre 0 et 1 sans unité. L'albédo dépend de la nature de surface de réflexion. Par exemple, une surface sombre a un albédo de 0.15, tandis que la neige ou les nuages ont un albédo de 0.40 jusqu'à 0.90. Généralement, l'albédo global de la terre est d'environ 0.3, c'est-à-dire que 30% de l'énergie solaire reçue par la terre sera réfléchie {AUTIXIER et RONDEAU, 2021}.

Concernant les panneaux solaires, le rayonnement solaire incident est constitué de trois catégories de rayonnement {Islam *et al.*, 2011} (voir figure 4.1) :

— Le rayonnement direct : il s'agit du rayonnement non diffusé, non réfléchi qui atteint directement la surface des panneaux solaires en ligne droite.

— Le rayonnement diffus du ciel : il s'agit du rayonnement diffusé par la poussière, les aérosols, etc.

— Le rayonnement diffus du sol : il s'agit du rayonnement réfléchi par la surface de la terre (albédo)

Nous appelons le rayonnement global la combinaison de ces trois types de rayonnements {Islam *et al.*, 2011}. En France métropolitaine, la moyenne du rayonnement solaire global horizontal varie entre 1.1 MWh/m^2 dans le nord et 1.7 MWh/m^2 dans le sud. Dans la Vienne, la moyenne journalière égale environ 3.5 kWh/m^2 et la moyenne annuelle est de l'ordre de $1.3MWh/m^2$. La carte 4.2 illustre la répartition de cette irradiation sur tout le territoire français métropolitain. [1]

La transformation de ce rayonnement solaire en énergie électrique passe par des cellules PV. Une cellule PV est un composant électrique composé de matériaux semi-conducteurs absorbant une grande partie de rayonnement solaire. Les cellules les plus courantes sont composées de deux couches de silicium, une couche de type N dopée au phosphore qui possède plus d'électrons que le silicium, et une couche de type P dopée au bore qui possède moins d'électrons que le silicium. Par conséquence, quand un photon arrive sur la cellule, son énergie est transférée aux électrons du silicium. La libération de ces électrons crée une différence de potentiel entre les deux couches P et N et génère donc une tension électrique. Ce principe s'appelle l'effet photovoltaïque. Il a été découvert par le physicien français Antoine Becquerel en 1839. En assemblant plusieurs cellules, nous obtenons un panneau solaire capable de produire de l'électricité à partir de l'énergie solaire, sans pollution et sans composants mécaniques {ADEME, 2021}.

1. Carte disponible sur le site de Solargis {Solargis, 2021} sous la licence Creative Commons Attribution license (CC BY 4.0)

FIGURE 4.2 – Carte de l'irradiation globale horizontale en France
Source : Solargis
Lecture : Répartition de la moyenne d'irradiation globale horizontale dans la France métropolitaine (moyenne calculée sur la période 1994-2018)

Le rendement [2] d'un parc PV dépend de plusieurs facteurs internes et externes. D'une part, ce rendement résulte de la nature et de la composition des panneaux PV. En effet, il y a une différence de rendement selon les marques et les types des cellules solaires. En outre, le rendement d'un producteur PV dépend aussi de l'orientation et de l'inclinaison de ses panneaux. La figure 4.3 illustre l'orientation et l'inclinaison des panneaux d'un producteur PV. En France, pour tous les types de producteurs PV, la majorité de production est obtenue avec une orientation sud est une inclinaison de 30 degrés. Une production avec une orientation vers le nord est quant à elle inférieure à 30% de la capacité maximale de production. L'orientation et l'inclinaison sont souvent liées à des contraintes architecturales des bâtiments.

D'autre part, la production instantanée d'un parc PV dépend de plusieurs paramètres externes issus des phénomènes météorologiques comme le mouvement des nuages et des aérosols. De plus, l'existence d'obstacles comme les arbres ou les bâtiments à côté d'un parc PV peut empêcher les rayonnements solaires directs ou diffus d'arriver sur le panneau solaire. Ce phénomène est appelé l'ombrage. Il est soit partiel dans le cas d'un arbre par exemple, soit total dans le cas d'une couverture sur l'ensemble de panneaux solaires comme une poussière ou une branche tombée sur les panneaux {EdfEnR, 2021}. Une courbe de production d'un producteur PV résulte donc de tous ces facteurs externes et internes.

La puissance d'un panneau solaire est donnée par

$$P_R = \eta S I[1 - 0.05(t_0 - 25)]$$

Où η est le rendement de la cellule PV en pourcentage, S est la surface du panneau en m^2, I est l'ir-

2. Le rendement d'un panneau PV est la part du rayonnement solaire converti en électricité à partir de la totalité des rayonnements reçus. Il est exprimé en pourcentage.

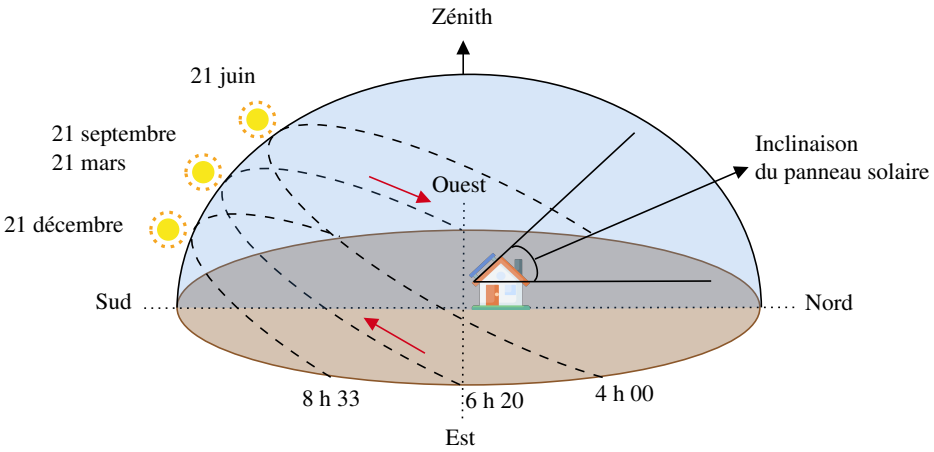


FIGURE 4.3 – L'orientation et l'inclinaison des panneaux d'un producteur PV (adapté de {Prinsloo et Dobson, 2015})

Type Producteur	Nombre total	Capacité totale en MVA	Moyenne en kVA	Écart type en kVA
Petit producteur	3048	15.01 (11.6%)	4.92	6.30
Moyen producteur	624	72.76 (56.4%)	116.60	52.10
Grand producteur	21	41.18 (32%)	1961.14	2318.50

TABLEAU 4.1 – Statistiques sur la capacité de production totale de parc PV de SRD selon les types des producteurs

radiation solaire en kW/m^2, t_0 représente la température extérieure en Celsius {Wan *et al.*, 2015}. La puissance crête est la puissance de production maximale d'un panneau solaire. Elle est calculée dans des conditions optimales de production lorsque l'irradiation solaire I est égale à 1 000 W/m^2 et la température t_0 est de 25 degrés Celsius.

4.2.2 Panorama du parc PV de SRD

Le réseau de SRD est constitué au total de 3693 producteurs dont 645 moyens et grands producteurs [3]. Ces producteurs, raccordés aux réseaux HTA et BT, sont répartis sur tout le territoire de SRD. Les figures 4.4 et 4.5 illustrent leur emplacement géographique, leur nombre total et la somme de leurs capacités par maille IRIS. Nous remarquons dans les deux cartes que les trois catégories de producteurs PV couvrent l'ensemble de département. Les petits producteurs sont plus présents dans le centre et le côté ouest du département. Les moyens et les grands producteurs sont plus présents au sud et au sud-ouest. Nous observons aussi que la proportion de petits producteurs est plus importante dans le nord comparé aux autres producteurs.

La capacité de production PV totale du réseau de SRD est de 129 *MVA* (données 2019). Les petits producteurs représentent 12 % de cette capacité, soit 15 *MVA*, les moyens producteurs représentent une part de 56 % et les grands producteurs ont une part de 32 % (voir tableau 4.1 et figure 4.6). Toutefois, nous avons seulement 21 grands producteurs contre 3048 petits producteurs. Nous notons que dans certains endroits du réseau, nous avons une présence de petits producteurs plus importante que des autres producteurs. Pour certaines communes, nous n'avons même que des petits producteurs.

Les petits producteurs répartis sur tout le territoire de SRD génèrent plusieurs contraintes sur le réseau. En particulier, leur production instantanée est aujourd'hui difficile à évaluer par rapport aux autres catégories de producteurs. En attendant le déploiement des compteurs communicants sur l'ensemble du

3. Données internes de SRD sur l'année 2019.

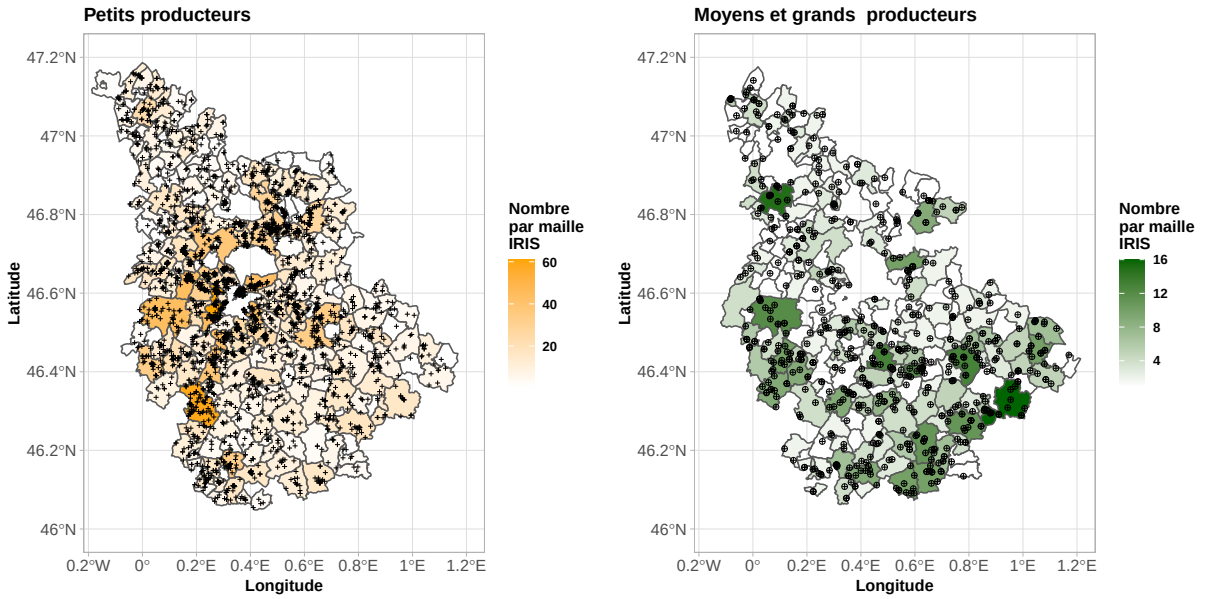


FIGURE 4.4 – Répartition spatiale des producteurs PV de SRD
Dans les deux cartographies, nous avons à gauche la répartition spatiale des petits producteurs PV sur le territoire de SRD dans le département de la Vienne, caractérisés par leur nombre total par maille IRIS. Dans la cartographie de droite, nous trouvons la répartition des moyens et grands producteurs et leur nombre. Les points représentés par des cercles et des croix dans les cartographies, montrent l'emplacement géographique des producteurs calculé à partir de leurs coordonnées GPS.

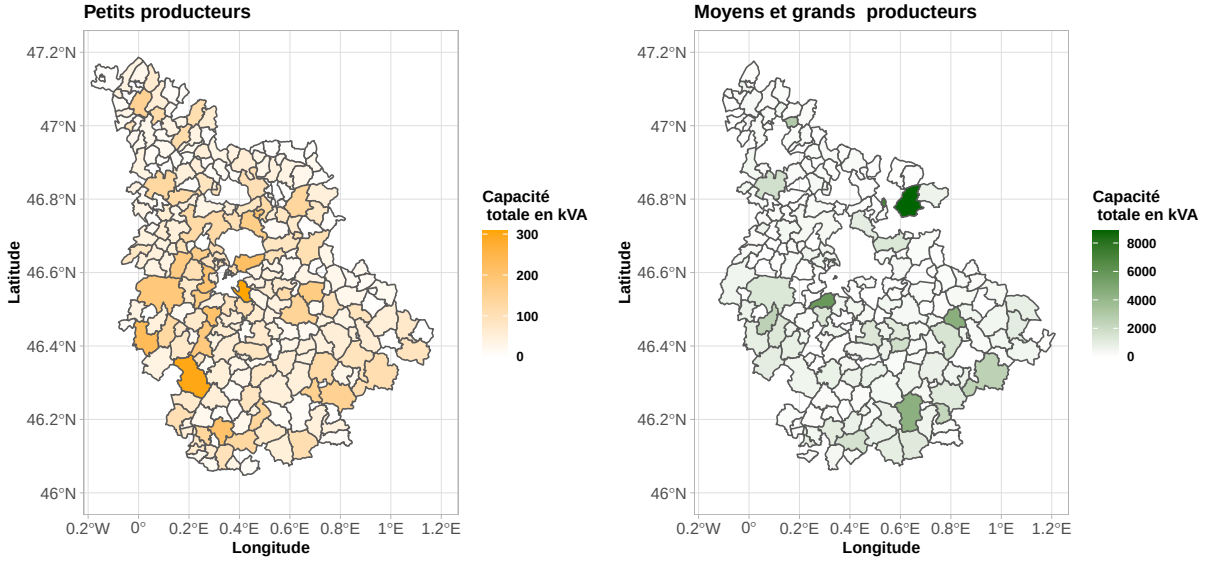


FIGURE 4.5 – Parc PV de SRD selon les types de producteurs
Dans les deux cartographies, nous avons à gauche la répartition spatiale de la capacité totale de production des petits producteurs PV. Dans la cartographie de droite, nous trouvons la répartition de la capacité totale des moyens et grands producteurs.

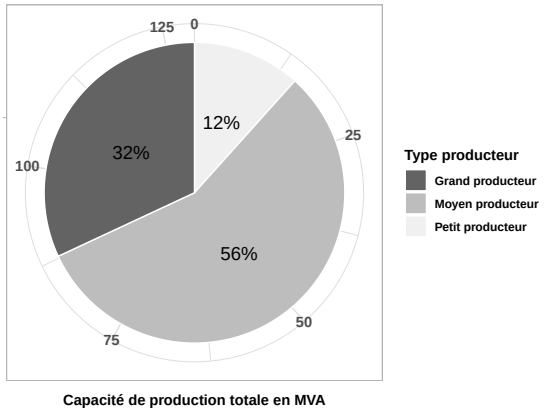


FIGURE 4.6 – La capacité totale de production selon les catégories de producteurs

réseau, nous avons développé une méthodologie d'estimation de cette production en utilisant des méthodes d'interpolation spatiale.

Nous avons vu précédemment que les différentes catégories de producteurs couvrent l'ensemble du département. De plus, la production PV dépend de plusieurs facteurs et phénomènes naturels météorologiques. Ces phénomènes peuvent être modélisés par des structures spatiales de par leur nature géographique. Pour ces raisons, nous avons appliqué des méthodes issues du domaine de l'interpolation spatiale dans cette approche d'estimation.

Dans cette présente étude, notre objectif est d'estimer le plus précisément possible les courbes de production PV fines non mesurables aujourd'hui à SRD, avec un temps de calcul acceptable, compte tenu du volume de données à estimer (3048 producteurs). Ces deux contraintes de rapidité et de précision ont conduit à tester les méthodes d'interpolation spatiale déterministes.

Ainsi, nous cherchons à trouver la production dans un point donné de l'espace en supposant que toutes les courbes de production sont disponibles aux autres points de référence. Ainsi, nous avons décidé de traiter l'espace indépendamment du temps, ce qui nous permet de simplifier les calculs en supposant que les courbes temporelles sont disponibles sur toute la période étudiée. Notre étude de l'état de l'art nous a permis de déterminer les méthodes les plus couramment utilisées dans des études de cas connexes dans la modélisation des phénomènes naturels comme les précipitations, la pollution, etc, afin d'évaluer leur efficacité sur notre ensemble de données de la production solaire.

4.3 Généralités

4.3.1 Méthodes d'interpolation spatiale

Dans cette section, nous présentons les notations, le vocabulaire et les méthodes d'interpolation spatiale

Le domaine de l'estimation et de l'interpolation spatiale intervient dans l'étude et la modélisation des phénomènes naturels dans un espace géographique donné. Il permet de modéliser par des équations mathématiques déterministes ou probabilistes la répartition des valeurs de ce phénomène sur l'espace étudié en modélisant les différentes corrélations entre les valeurs prises dans les sites de référence. Cette

structure spatiale présentée par le modèle permet d'estimer ou de prédire des valeurs du phénomène dans d'autres sites inconnus . Nous parlons d'estimation ou d'interpolation spatiale lorsque nous cherchons à estimer des valeurs en un point donné à partir des valeurs observées en d'autres points de l'espace. Les différentes notations et méthodes présentées dans ce chapitre sont détaillées dans le livre d'Arnaud et Emery {2000}.

Vocabulaire et notations

Soit z une fonction numérique prenant ses valeurs dans un espace limité, z est appelée une variable régionalisée. Nous notons

— z : la variable régionalisée.

— D : le champ de la régionalisation, le domaine de définition de z.

— $s \in D$: un site dans l'espace D, de coordonnées (x,y), $D \subseteq \mathbb{R}^2$.

— $z(s)$: la valeur observée sur le site $s \in D$. En général, cette valeur n'est pas connue.

— S : un sous-ensemble fini de D pour lequel la valeur observée $z(s), s \in S$ est connue, $S \subseteq D$ et $|S| = n$.

— $\hat{z}(s)$: la valeur estimée de $z(s)$ sur le site $s \in D$.

Lorsque z prend des valeurs dans l'espace D pour chaque instant t pour une période T, on parle alors d'une variable spatio-temporelle, z devient une fonction de $D \times T$ dans $\mathbb{R}^3$. Dans notre cas, les temps de mesure étant discrets, une variable spatio-temporelle sera définie en $\mathbb{R}^2 \times \mathbb{N}$.

Les notations deviennent alors :

— $z(s,t)$: la valeur observée sur le site $s \in D$ à l'instant t.

— $\hat{z}(s,t)$: la valeur estimée de $z(s)$ sur le site $s \in D$ à l'instant t.

Dans toutes les méthodes d'interpolation spatiale déterministes ou géostatistiques, nous cherchons à estimer une valeur $\hat{z}(s)$ sur un site s en utilisant les données de référence $z(s)$. Généralement, ces méthodes s'écrivent de la forme

$$\hat{z}(s) = \sum_{i=1}^{n} \lambda_i z_i(s) \tag{4.1}$$

Avec $\sum_{i=1}^{n} \lambda_i = 1$

L'approche naïve : l'approche d'interpolation naïve est une méthode d'estimation qui consiste simplement à calculer la moyenne de tous les $z(s)$ pour les sites connus.

$$\hat{z}(s) = \frac{1}{n} \sum_{s_i \in S} z(s_i) \tag{4.2}$$

Cette méthode sera utilisée comme référence afin d'évaluer la qualité des autres approches. En effet, notre motivation derrière le choix de cette méthode est de comparer son erreur avec l'erreur d'autres méthodes d'estimation. Elle nous permet donc de quantifier les gains en qualité de précision des autres méthodes d'estimation par rapport à cette méthode basique.

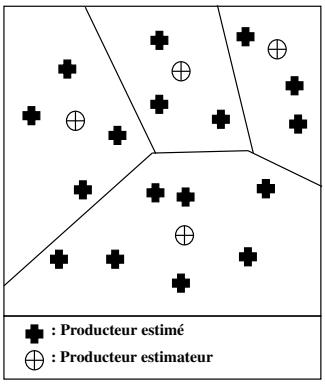


FIGURE 4.7 – Diagramme de Voronoï

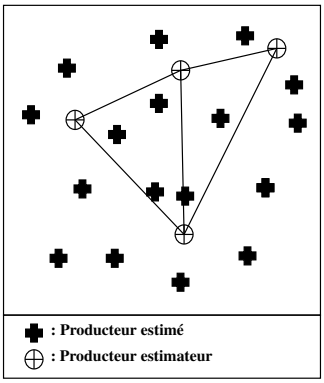


FIGURE 4.8 – Méthode TIN

Diagramme de Voronoï : nous appelons diagramme de Voronoï (ou diagramme de Thiessen ou cellules de Dirichlet) une partition de D telle que chaque partie de cette partition contient un site d'observation unique s_i de S appelé site représentatif, et telle que chaque point d'une partition est plus proche de son site représentatif que de tout autre représentant. Cette partition est notée $P = \{P_{s_1}, P_{s_2}, ..., P_{s_n}\}$.

$$\forall s_i \in S, P_{s_i} = \{s \in D \mid \forall s_j \in S, ||s - s_i|| \leq ||s - s_j||\} \tag{4.3}$$

Cette partition définit plusieurs polygones convexes sur l'espace D (voir Figure 4.7).

Cette méthode doit son nom au mathématicien russe Gueorgui Voronoi (1868-1908). Elle consiste à attribuer à tous les points d'un polygone la valeur mesurée dans son site représentatif, on parle aussi de l'interpolation par le plus proche voisin ou le voisin naturel.

Interpolation par triangulation TIN (Triangulated Irregular Network) : La méthode de triangulation consiste à diviser l'espace D en plusieurs triangles dont les sommets sont les sites d'observation (voir figure 4.8). Les valeurs d'un point sont alors estimées à partir des valeurs des sommets du triangle auquel il appartient. Il existe plusieurs méthodes de triangulation de l'espace, la plus utilisée étant la méthode de triangulation de Delaunay. Elle consiste simplement à prendre les points représentatifs des polygones de Voronoï comme sommets du triangle.

Il existe plusieurs méthodes d'interpolation à partir d'une triangulation, les plus connues sont l'interpolation linéaire et la méthode d'Akima {Akima, 1978}. Dans cette étude, nous n'avons testé que la méthode d'interpolation linéaire.

Nous essayons d'estimer la valeur de $\hat{z}(s)$ pour un site s appartenant au triangle (s_1, s_2, s_3) où $s_i =$

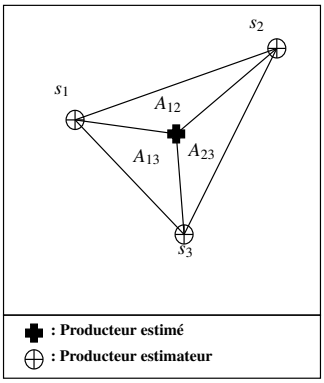


FIGURE 4.9 – Méthode TIN : zones de pondération

(x_i, y_i) sont les points de référence à partir de S (voir Figure 4.9).

L'estimation au site s dans l'espace $D \subset \mathbb{R}^2$ est définie comme suit :

$$\hat{z}(s) = \hat{z}(x,y) = \alpha x + \beta y + \lambda$$

Les équations des sommets du triangle s'écrivent comme suit :

$$\begin{cases} \alpha x_1 + \beta y_1 + \lambda = z(s_1) \\ \alpha x_2 + \beta y_2 + \lambda = z(s_2) \\ \alpha x_3 + \beta y_3 + \lambda = z(s_3) \end{cases}$$

La solution de ce système est alors :

$$\hat{z}(s) = \frac{|s_1, s, s_2| z(s_3) + |s_1, s, s_3| z(s_2) + |s_2, s, s_3| z(s_1)}{|s_1, s_2, s_3|} \tag{4.4}$$

où $|s_i, s_j, s_k|$ représente l'aire du triangle formé par (s_i, s_j, s_k).

La méthode TIN consiste donc à diviser l'espace en plusieurs triangles irréguliers. Ces triangles sont utilisés pour l'estimation d'un point à l'intérieur d'un triangle donné en pondérant par les surfaces des petits triangles formés entre les sommets.

Interpolation par la moyenne des k voisins les plus proches KNN : cette méthode est une généralisation de la méthode de Voronoï où, au lieu de prendre la valeur du plus proche voisin, nous calculons la moyenne des k plus proches voisins.

Soit k-NN(s) les k plus proches voisins de s en S. En cas d'égalité, les sites de référence les plus proches sont choisis arbitrairement.

$$\hat{z}(s) = \frac{1}{k} \sum_{s_i \in k\text{-NN}(s)} z(s_i) \tag{4.5}$$

Dans cette méthode, nous devons déterminer le nombre de voisins k optimal à utiliser. Généralement, le nombre k est déterminé empiriquement. A partir des estimations des différents k, le k optimal est celui qui minimise l'erreur d'estimation.

Pondération par l'inverse de la distance IDW (Inverse Distance Weighting) la méthode IDW (voir figure 4.10) a été développée par le service météorologique national américain en 1972 ; elle est simple, intuitive et rapide.

La méthode IDW est une variation de la méthode de la moyenne des k plus proches voisins, en calculant une moyenne pondérée où les poids sont l'inverse de la distance entre chaque site.

Des poids plus élevés sont attribués aux sites les plus proches, tandis que des poids plus faibles sont attribués aux sites les plus éloignés.

$$\hat{z}(s) = \frac{\sum_{s_i \in k\text{-NN}(s)} w_i^p \cdot z(s_i)}{\sum_{s_i \in k\text{-NN}(s)} w_i^p} \tag{4.6}$$

Le poids $w_i = \frac{1}{d(s_i,s)}$ est l'inverse de la distance euclidienne $d(s_i,s)$ entre les sites s_i et s. p est un scalaire dans $\mathbb{R}$ appelé la puissance.

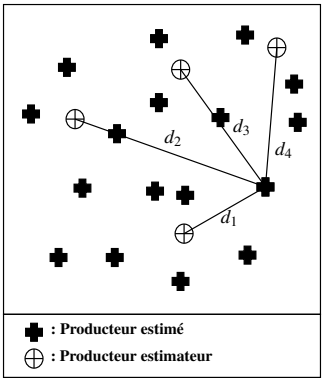


FIGURE 4.10 – Pondération par l'inverse de la distance dans le cas de quatre voisins

La limite de cette méthode est sa sensibilité aux valeurs aberrantes. En outre, la détermination du paramètre p, du nombre de voisins ou du rayon maximum, est empirique. Dans le cas ou $p = 0$, nous obtenons la méthode KNN.

Krigeage ordinaire OK (Ordinary kriging) la méthode OK {Krige, 1951} fait partie des méthodes du domaine de la géostatistique. Dans cette catégorie de modèles, nous considérons que la variable régionalisée z est une fonction aléatoire notée Z.

Dans le krigeage ordinaire, nous supposons que la moyenne de la variable Z est inconnue

L'estimation de Z, avec les poids w_i, est :

$$\hat{Z}(s) = a + \sum_{s_i \in S} w_i Z(s_i)$$

L'espérance de l'erreur d'estimation s'écrit comme suit

$$E[\hat{Z}(s) - Z(s)] = a + \sum_{s_i \in S} w_i E[Z(s)] - E[Z(s)] = a + \left(\sum_{s_i \in S} w_i - 1 \right) m = 0$$

.

Comme m est inconnue, nous fixons $a = 0$ et $\sum_{s_i \in S} w_i = 1$ pour avoir un estimateur sans biais. Pour plus d'informations sur cette méthode et d'autres méthodes géostatistiques, nous pouvons nous référer au livre d'Arnaud et Emery {2000}.

Méthodologie de test

Afin de choisir la méthode la plus performante, nous devons quantifier la qualité et la précision des différents modèles en utilisant plusieurs métriques. Ces métriques permettent de comparer les différentes méthodes entre elles et de choisir au final la méthode la plus performante. Zhang *et al.* {2015} ont procédé à un examen approfondi des métriques utilisées dans la production solaire.

Erreurs d'estimation Dans cette étude, nous avons utilisé plusieurs métriques statistiques présentées dans la section 2.2.4 du chapitre 2 comme MBE_s, MAE_s et $RMSE_s$

Dans cette étude d'estimation, nous avons normalisé la métrique $RMSE_s$ par l'écart-type et la moyenne de $z(s,t)$, cette dernière étant notée $\overline{z}(s,t)$.

$$nRMSE_{\sigma}(s) = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(z(s,t) - \hat{z}(s,t))^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(z(s,t) - \overline{z}(s,t))^2}}$$

$$nRMSE_{m}(s) = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(z(s,t) - \hat{z}(s,t))^2}}{\overline{z}(s,t)}$$

Méthodologie

Afin de choisir la méthode d'interpolation la plus efficace, nous devons calculer les erreurs d'estimation de chaque méthode, mais à ce jour, nous ne disposons pas de données complètes sur la production PV des petits producteurs. C'est pourquoi nous avons adopté une méthodologie de test utilisant uniquement les données observées sur les sites des moyens et grands producteurs.

Nous avons utilisé la méthode de validation croisée {Kohavi *et al.*, 1995} (Cross validation) pour sélectionner la méthode d'interpolation la plus performante. La figure 4.11 illustre le fonctionnement de cette stratégie. Nous avons pris un échantillon aléatoire de 80 % des sites pour l'estimation et de 20 % pour l'évaluation des méthodes, la méthode la plus performante étant celle qui minimise l'erreur moyenne.

La méthode de validation croisée permet d'évaluer les différents modèles en calculant la précision d'estimation sur plusieurs échantillons de test. Elle partitionne l'ensemble des sites de référence en K sous-ensembles de même taille. La partition i est utilisée pour évaluer la qualité des modèles estimés en utilisant les autres $K-1$ partitions. En répétant cette opération K fois, les sites sont estimés dans une seule partition i et servent de points de référence (estimateurs) pour les autres partitions.

Enfin, pour optimiser les paramètres des algorithmes IDW, nous avons appliqué la méthode de validation croisée emboîtée (Nested Cross-Validation), qui permet de diviser la partie d'estimation en sous-partitions de test et de validation.

4.3.2 État de l'art

Li et Heap {2014} ont présenté une revue complète des méthodes d'interpolation spatiale les plus

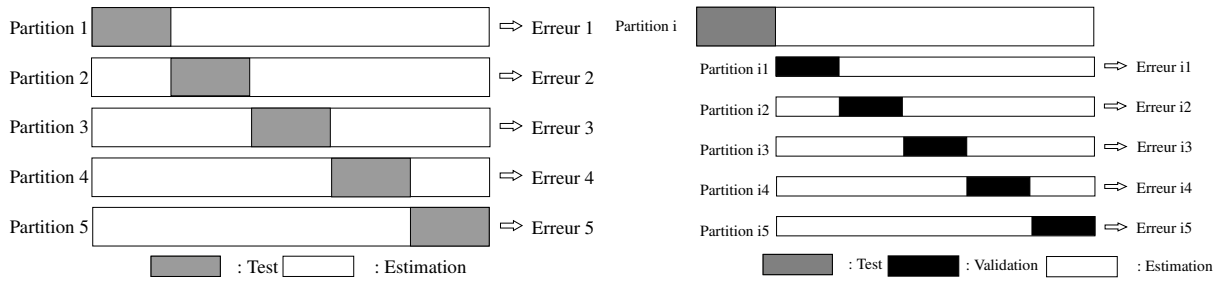


FIGURE 4.11 – La méthode de validation croisée

largement utilisées dans le domaine de l'environnement. Ces méthodes sont utilisées dans plusieurs domaines tels que l'exploitation minière, l'ingénierie, l'agriculture et la météorologie. Elles sont divisées en trois catégories de méthodes : les méthodes déterministes (non géostatistiques), les méthodes stochastiques (géostatistiques) et les méthodes hybrides. Une classification de 25 méthodes utilisées dans la littérature est proposée en fonction de leurs caractéristiques pour montrer les similitudes entre elles.

Susanto *et al.* {2016} ont comparé trois des méthodes d'interpolation les plus largement utilisées, dans lesquelles le temps est traité indépendamment de la dimension spatiale. Ces méthodes sont le Krigeage ordinaire OK, la pondération par l'inverse des distances IDW et la triangulation TIN. Ils ont comparé la qualité des estimations en utilisant les métriques RMSE et MAE ainsi que le temps de calcul requis par chaque méthode. Leurs résultats montrent que la méthode OK est la technique la plus efficace quantitativement dans l'ensemble, mais son temps de calcul augmente considérablement avec le nombre de points observés. Les auteurs ont donné un poids élevé à l'efficacité du calcul. Selon leur méthode de classement, ils ont affirmé que la méthode OK est la moins bonne et que la technique IDW est la méthode d'interpolation spatiale la plus appropriée dans l'ensemble.

Chen et Liu {2012} ont utilisé la méthode IDW pour estimer les précipitations au milieu de Taïwan. Ils optimisent ensuite la précision de l'interpolation en se fondant sur les paramètres de la méthode IDW tels que la puissance et le rayon de recherche en utilisant la méthode de validation croisée. Leurs résultats ont montré que IDW est meilleure en période sèche qu'en période d'inondation. L'approche IDW a des performances médiocres dans les cas extrêmes.

Qiao *et al.* {2018} ont analysé la pollution des métaux lourds dans les sols, en particulier l'Arsenic, en utilisant les méthodes d'interpolation spatiale OK et IDW. La précision de l'interpolation est élevée dans les régions où la corrélation spatiale est élevée et la variabilité spatiale de la concentration d'Arsenic est faible. Pour certaines régions, ils montrent que les deux méthodes OK et IDW sont similaires, mais la méthode OK est plus efficace dans le cas de l'estimation de valeurs extrêmes.

Lu et Wong {2008} ont développé un algorithme de pondération adaptative à distance inverse fondé sur la méthode IDW qui permet d'optimiser et d'ajuster ces paramètres (puissance et rayon) en fonction de la distribution des points observés. Leurs résultats ont montré que IDW adaptative est plus efficace que la méthode classique avec des paramètres constants et qu'elle est meilleure que OK dans certains cas où la distribution spatiale des points est difficile à modéliser par un variogramme.

Joseph et Kang {2011} ont développé une nouvelle méthode d'interpolation fondée sur IDW et la régression linéaire. Leurs résultats ont montré que cette méthode a la même précision que OK et qu'elle est plus rapide. Ceci permet d'interpoler facilement de grands volumes de données. Ils ont également développé une heuristique pour calculer les intervalles de confiance. Ils montrent que ces intervalles sont plus efficaces que les intervalles de confiance fournis par la méthode OK.

Travaux sur l'estimation de l'énergie solaire

La plupart des travaux qui ont été trouvés dans la littérature dans le domaine de l'interpolation spatiale de l'énergie solaire estiment l'irradiation solaire dans un site à partir des points observés, et non la production réelle d'une installation photovoltaïque.

SEN et SAHIN {2001} ont appliqué les méthodes d'interpolation spatiale pour estimer l'irradiation solaire sur des sites où les mesures d'irradiation n'existent pas. Ils ont testé la méthode du Semi-Variogramme Cumulé (SVC) et la méthode IDW. Ils ont validé les méthodes sur des mesures mensuelles d'irradiation de 29 stations de mesure par une approche de validation croisée. Ils ont trouvé que la méthode SVC est meilleure que la méthode IDW.

Alsamamra *et al.* {2009} ont analysé les méthodes de krigeage résiduel (Residual Kriging RK) et OK, pour estimer la moyenne mensuelle du rayonnement solaire global dans le sud de l'Espagne. Ils ont évalué les deux méthodes sur des données de 4 ans provenant de 166 stations (112 stations pour l'apprentissage et 54 pour les tests). Ils ont constaté que la méthode OK est la plus précise, avec une RMSE de 5 % pour les mois d'été et de 11 % pour les mois d'automne et d'hiver. Ils ont proposé une amélioration de la méthode RK en intégrant des données externes sur les ombres portées par les caractéristiques topographiques. La méthode RK a ainsi été améliorée avec une RMSE de 5,5 % en été et de 10,5 % en automne et en hiver.

Jamaly et Kleissl {2017} ont analysé les corrélations spatiales et temporelles dans les données d'irradiation, puis ils ont appliqué la méthode OK spatiale et spatio-temporelle. Ils proposent une fonction de covariance paramétrique anisotropique pour modéliser les nuages transitoires. Leur méthode est validée avec un ensemble de données simulées et réelles. Ils ont montré que le modèle anisotrope est le plus efficace avec une erreur nRMSE de 7,92 %. Ce modèle a amélioré de 66 % l'erreur obtenue par le modèle de persistance.

Rodríguez-Amigo *et al.* {2017} ont appliqué quatre méthodes d'interpolation spatiale pour cartographier l'irradiation horizontale globale (GHI). Ils ont utilisé le logiciel ArcGis en exploitant les données de 71 stations météorologiques au sol réparties sur une zone de 94 226 km^2 en Espagne. Ces données sont constituées des mesures de GHI avec un historique de 7 ans et une granularité de 30 minutes. Elles ont été transformées pour calculer le GHI journalier et pour obtenir au final la moyenne annuelle. Ils ont évalué leurs résultats à l'aide des métriques MBE, MAE, RMSE et MAPE en utilisant la méthode de validation croisée. Les résultats ont été testés aussi avec les données de 4 stations de l'agence météorologique nationale espagnole. La contribution de leurs travaux est d'évaluer les méthodes d'interpolation spatiale dans l'estimation des données climatologiques sur de grandes zones avec un faible nombre de points de mesures. Ils ont trouvé que la méthode de Krigeage Universel avec semi-variogramme quadratique était la plus juste en précision d'estimation en utilisant les indicateurs MAE et RMSE.

4.4 Données et expérimentation

Dans cette étude, nous avons utilisé les données de comptage et les données de cartographie (voir section 2.3 chapitre 2). Pour 3693 producteurs PV raccordés au réseau de SRD (voir tableau 4.1), nous avons calculé leur emplacement géographique grâce aux coordonnées GPS. Ces points représentent les sites dans la modélisation spatiale. Nous avons aussi la capacité installée des producteurs PV et leur capacité crête. En plus de ces données statiques, nous avons les courbes de productions de tous les moyens et grands producteurs. Nous avons au total plus de 645 courbes de production mesurées. Cependant

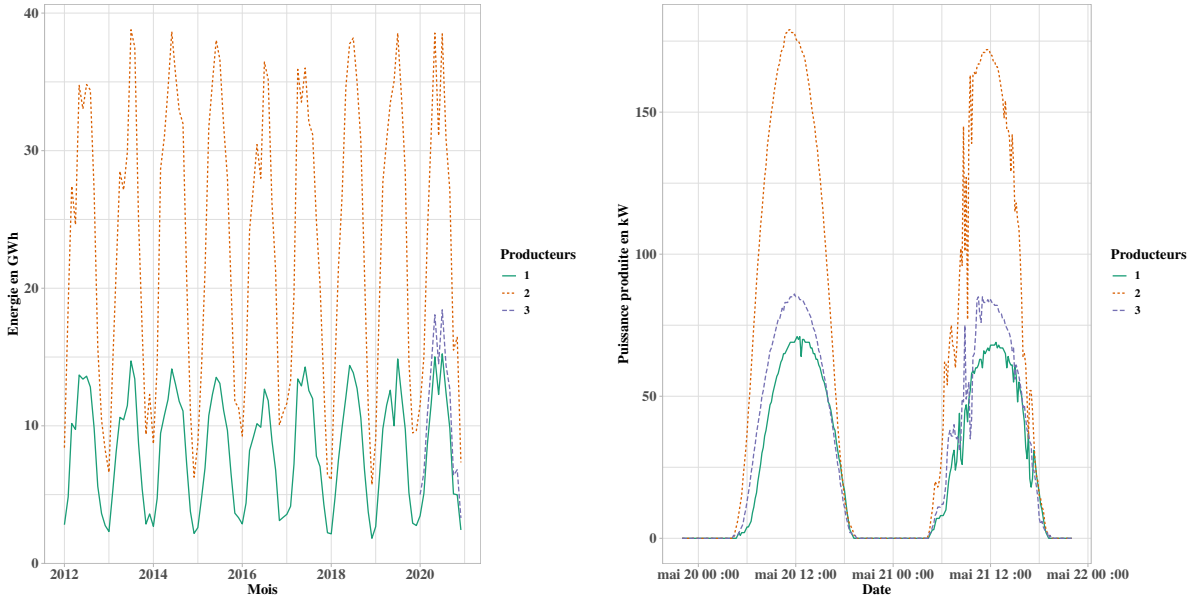


FIGURE 4.12 – Exemples de courbes de productions de trois producteurs
Lecture : Dans la figure de gauche, l'axe des abscisses présente le temps avec un pas d'un mois et l'axe des ordonnées présente l'énergie mensuelle de ces trois producteurs. Dans la figure de droite, l'axe des abscisses présente le temps avec un pas de 10 minutes, l'axe des ordonnées présente la puissance produite en kW des trois producteurs. Nous pouvons remarquer la fluctuation de la production en raison du passage de nuages.

nous avons retiré les producteurs qui avaient une période manquante en 2019 en raison d'opérations de maintenance ainsi que les producteurs que nous n'avons pas réussis à associer à la base de données de la cartographie afin d'ajouter les informations statiques (coordonnées GPS, capacité, etc.) Nous avons gardé au final les courbes de 621 moyens et grands producteurs sur l'année 2019. Pour chaque producteur, nous avons une courbe de production avec une granularité de 10 minutes de janvier à décembre 2019. Chaque courbe est une série temporelle de $6 \times 24 \times 365 = 52\,560$ valeurs. En combinant l'information spatiale des producteurs avec leurs courbes de production, nous obtenons les données $z(s,t)$ de chaque producteur au site s. Les z sont donc des séries spatio-temporelles.

La figure de gauche dans 4.12 illustre les énergies mensuelles produites par trois producteurs depuis janvier 2012. Nous observons que pendant l'été, ils produisent beaucoup plus qu'en hiver. La figure de droite (voir figure 4.12) illustre les courbes de productions des trois producteurs, le producteur 1 ayant une capacité de 89 kVA, le producteur 2 une capacité de 250 kVA et le producteur 3 une capacité de 95 kVA. Les fluctuations dans les courbes de production sont liées au passage de nuages au-dessus des panneaux solaires. Ces installations étant exploitées par SRD pour le compte du producteur dans le cadre d'une convention de prestation, nous connaissons leurs orientations et inclinaisons. La majorité des panneaux solaires du producteur 1 sont orientés à 234 degrés par rapport au nord avec une inclinaison de 14 degrés. Pour le producteur 2, ils sont orientés à 180 degrés par rapport au nord (plein sud) avec une inclinaison de 20 degrés. Concernant le producteur 3, ils sont orientés à 8 degrés par rapport au nord avec une inclinaison de 12 degrés. Nous remarquons dans les trois courbes le décalage de production lié à la configuration des producteurs.

Ces courbes présentent la production instantanée de producteurs de taille différente. En effet nous avons des producteurs qui ont une capacité de production supérieure à 6000 kVA, tandis que d'autres ont une capacité égale à 36 kVA. Il est indispensable pour l'estimation de normaliser ces courbes afin de travailler avec des données du même ordre de grandeur. Dans cette étude, nous avons expérimenté

deux scénarios de normalisation. Le premier scénario consiste à normaliser les courbes de production en fonction de leur capacité installée maximale de production (voir figure 4.13) puisque cette information est disponible pour tous les producteurs (petits, moyens et grands).

$$z_n(s,t) = \frac{z(s,t)}{C_s}$$

C_s est la capacité installée du producteur s.

Ici $z_n(s,t)$ représente la part de la production instantanée à un instant t en fonction de la capacité installée du producteur. Généralement, ce facteur est entre 0 et 1 (0%-100%), cependant certains producteurs dépassent parfois leur capacité maximale de production.

Dans le deuxième scénario, nous avons normalisé les données par l'énergie annuelle produite par le producteur. Ce scénario de normalisation est acceptable puisque la production annuelle d'énergie est une information disponible pour tous les producteurs. Les petits producteurs communiquent leurs index de comptage chaque année.

$$z_n(s,t) = \frac{z(s,t)}{E_{T,s}}$$

$E_{T,s}$ est l'énergie totale produite par le producteur s sur une période T (dans notre cas, la période T est l'année entière).

$$E_{T,s} = \int_T z(s,t)dt$$

$z_n(s,t)$ représente dans ce scénario un pourcentage de la puissance instantanée produite en fonction de l'énergie totale produite sur l'année. $z_n(s,t)$ est donc une courbe de très petites valeurs qui ne dépassent pas 0.001 sur la journée.

La figure 4.13 montre un exemple de normalisation de données de trois producteurs. La normalisation par l'énergie annuelle permet de mieux ajuster les grandeurs des productions.

L'évaluation expérimentale a été réalisée dans les conditions réelles d'application sur un ordinateur équipé d'un processeur Intel Core i3-6006U à 3,00 GHz, de 2 cœurs, de 12 Go de RAM et d'un système d'exploitation Windows 10.

Le calcul de la matrice de distance entre les producteurs est réalisé par les packages sf {Pebesma, 2018} et sp {Bivand *et al.*, 2013} en langage R {R Core Team, 2020} avec une projection CRS 4326 (WGS84 - World Geodetic System 1984).

4.5 Résultats

Dans ce projet d'estimation des courbes de production des petits producteurs PV, nous avons testé plusieurs méthodes d'interpolation spatiale. Nous avons fixé dans cette étude un temps de calcul maximum d'une heure pour la partie interpolation. En effet, cette estimation spatiale sera utilisée dans d'autres outils de prévision, nous cherchons donc à minimiser le temps de calcul de cette phase d'interpolation. Les algorithmes géostatistiques ont dépassé ce délai en raison de la quantité de données à traiter. Leurs résultats ne sont donc pas détaillés dans ce document.

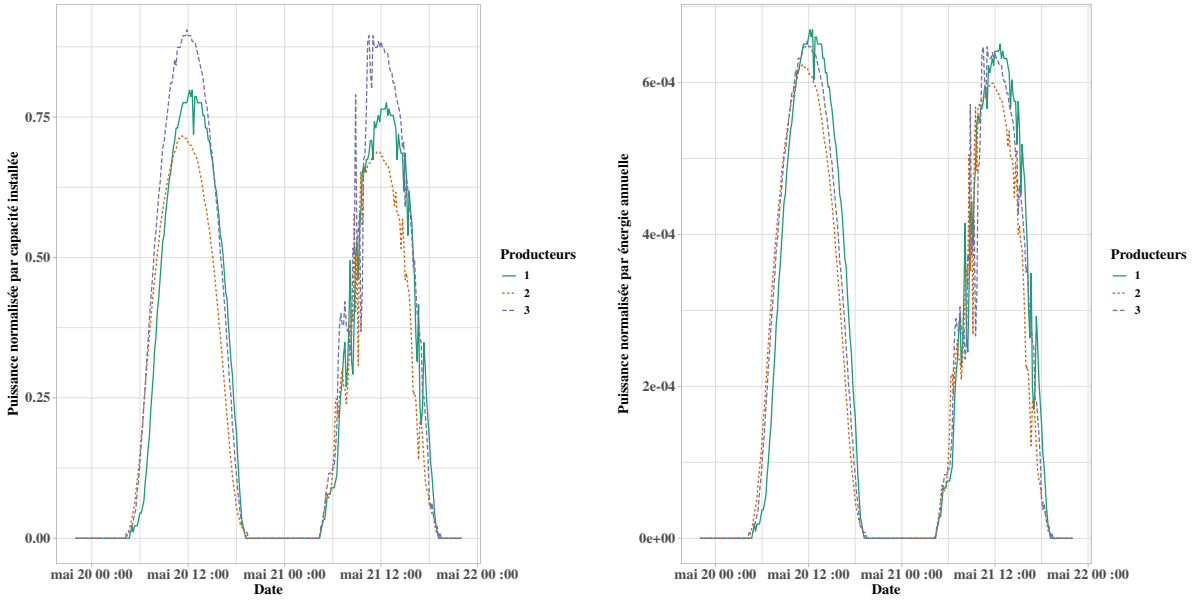


FIGURE 4.13 – Exemples de normalisation de trois courbes de productions.
Lecture : Dans la figure de gauche, l'axe des abscisses présente le temps avec un pas de 10 min et l'axe des ordonnées présente la puissance produite normalisée par la capacité installée des producteurs. Nous pouvons remarquer la fluctuation de la production en raison du passage de nuages. Dans la figure de droite, l'axe des abscisses présente le temps, l'axe des ordonnées présente la puissance produite normalisée par l'énergie annuelle des trois producteurs.

Méthode	**Nombre de voisins**	**Temps écoulé (s)**
Calcul de la matrice de distance	620	657.14
Méthode naïve	-	34,67
TIN	3	31.41
Voronoi	1	5.78
IDW	5	7.15
IDW	10	8.39
IDW	30	12.29
KNN	5	6.16
KNN	10	6.93
KNN	30	7.67

TABLEAU 4.2 – Temps de calcul de chaque méthode

Nous présentons au final les résultats de cinq méthodes d'interpolation spatiale déterministe. Le temps de calcul de chaque algorithme est présenté dans le tableau 4.2. Ce temps est divisé en deux parties. Le premier temps de calcul donné est celui nécessaire au calcul d'une matrice de distance entre tous les sites de la partition de test et les sites de la partition d'estimation. Ensuite, nous appliquons les algorithmes d'interpolation en utilisant cette matrice de distance. Notons que cette matrice est statique et peut être calculée une seule fois.

Nous avons appliqué les cinq algorithmes en utilisant la méthode de validation croisée présentée dans la section 4.3.1. Pour la méthode IDW, nous avons fixé $p = 1$ et un nombre maximum de voisins de 10. Pour l'algorithme KNN, nous avons fixé le nombre de voisins à 5. Au début, ces paramètres sont fixés d'une manière arbitraire.

Avant de commencer notre expérimentation d'estimation, nous avons remarqué dans les données issues de la cartographie que certains parcs PV sont divisés en plusieurs points de livraison (PDL).

FIGURE 4.14 – Exemple de sites de production
Les cercles rouges présentent les producteurs identifiés par PDL, leur emplacement est calculé en utilisant les coordonnées GPS.
Licence : fond de carte OpenStreetMap hébergé par Esri, sous la licence Creative Commons by Attribution (CC BY 4.0)

En effet, ces parcs sont équipés de plusieurs compteurs et donc plusieurs PDL. L'image satellite de la figure 4.14 montre un exemple de ces parcs. Les sites calculés par les coordonnées GPS (présentés par les cercles dans l'image) correspondent aux points de connexion avec le réseau. Ils sont identifiés par le numéro PDL. Cependant, ces points appartiennent au même parc PV. Cet artefact dans les données peut fausser les résultats d'estimation, du fait que nous estimons la courbe de production d'un producteur avec les données de ses voisins qui font partie en réalité de ce même site de production. En outre, la distance entre les deux sites sera très petite donc un grand poids sera donné au plus proche voisin. Pour remédier à ce problème, nous avons fixé une distance minimale de séparation entre les sites de la partition test et les sites de la partition estimation. Il faut néanmoins noter que nous n'aurons pas ce problème entre les petits et les moyens/grands producteurs. Nous avons déterminé cette distance minimale de séparation empiriquement. La figure 4.15 présente les résultats de qualité de l'estimation via la métrique *RMSE* des cinq algorithmes en fonction de la distance minimale de séparation dans le cas de la normalisation par la capacité installée. Nous remarquons l'effet de cet artefact sur la précision de l'estimation. Nous passons d'une RMSE de 0.045 avec une distance nulle à une *RMSE* de 0.051 avec une distance de 50 m. Après une distance de séparation de 50 m l'erreur croit linéairement en fonction de la distance, puisque nous nous éloignons du plus proche voisin.

Après avoir fixé la distance minimale de séparation, nous avons effectué notre expérimentation d'estimation en comparant les cinq méthodes. La qualité de l'estimation est mesurée par les métriques présentées dans la section 2.2.4. Les résultats de l'interpolation sont présentés dans la figure 4.16 en utilisant la métrique $nRMSE_\sigma$, permettant la normalisation de la RMSE par l'écart-type des valeurs observées afin de comparer les résultats des deux scénarios de normalisation. La RMSE qui pénalise les erreurs importantes dans l'estimation nous permet de mesurer la précision de la méthode IDW qui est générale-

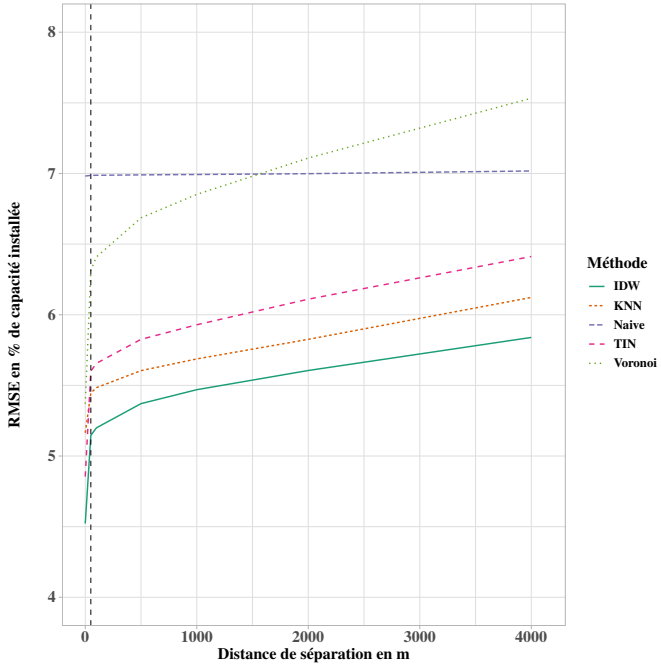


FIGURE 4.15 – L'effet de la distance de séparation sur la qualité d'estimation
L'axe des abscisses est la distance de séparation en mètres, et l'axe des ordonnées est l'erreur *RMSE* d'estimation de chaque méthode. La droite verticale présente la distance de 50 *m*

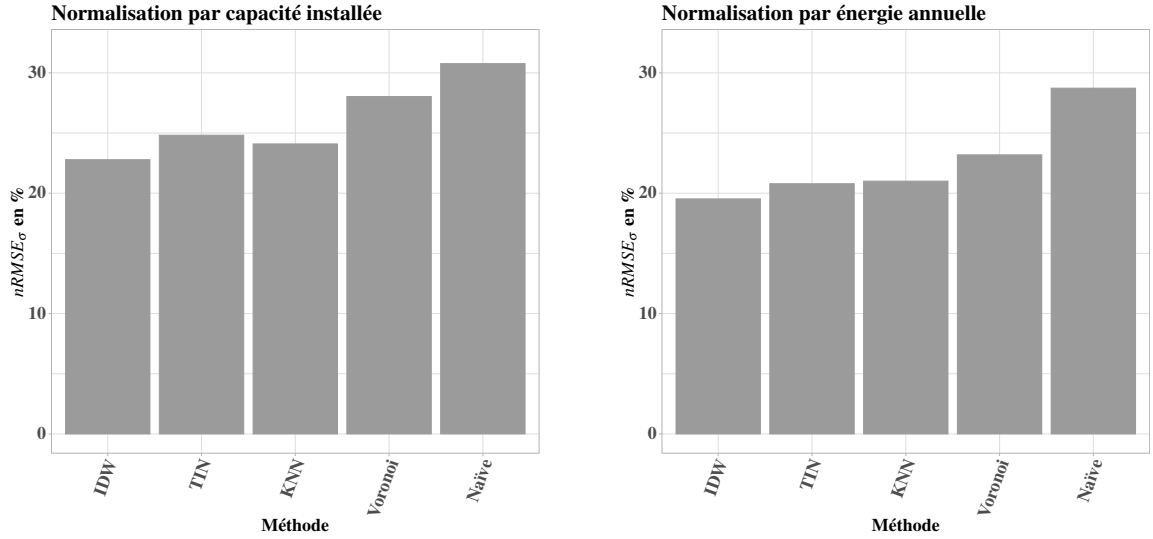


FIGURE 4.16 – Résultats obtenus pour les 5 méthodes dans les deux cas de normalisation.
Lecture : L'axe des abscisses représente les cinq méthodes et l'axe des ordonnées est l'erreur $nRMSE_\sigma$ d'estimation de chaque méthode . Le graphique de gauche montre ces résultats dans le cas de la normalisation par la puissance installée, et le graphique de droite montre les résultats dans le second cas de normalisation par l'énergie annuelle.

Méthode	MAE (% de capacité)	RMSE (% de capacité)	$nRMSE_\sigma$ (%)	$nRMSE_m$ (%)
IDW	2.3	5.1	22.784	35.931
KNN	2.4	5.4	24.093	37.992
TIN	2.5	5.6	24.817	39.130
Voronoi	2.8	6.3	28,030	44,191
Naïve	3.2	6.9	30,766	48,487

TABLEAU 4.3 – Résultats obtenus en cas de normalisation par la capacité installée. Les métriques MAE et RMSE présentent un pourcentage de la capacité maximale de production des producteurs.

Méthode	MAE (% d'énergie)	RMSE (% d'énergie)	$nRMSE_\sigma$ (%)	$nRMSE_m$ (%)
IDW	0.00149	0.00352	19.527	30.798
KNN	0.00157	0.00378	21.001	33.116
TIN	0.00155	0.00374	20.787	32.776
Voronoi	0.00171	0.00417	23.186	36.556
Naïve	0.00235	0.00517	28.727	45.264

TABLEAU 4.4 – Résultats obtenus dans le cas de normalisation par l'énergie annuelle

ment très sensible aux valeurs aberrantes et extrêmes. Dans le cas des deux scénarios de normalisation, l'algorithme IDW a offert la meilleure performance parmi les algorithmes évalués.

Le tableau 4.3 résume l'erreur d'estimation pour chaque méthode dans le cas de la normalisation par la capacité installée. Avec les mêmes paramètres que le deuxième scénario de normalisation, IDW surpasse les autres méthodes tant pour la métrique MAE que celle de RMSE. Elle est suivie de KNN et TIN, dont la précision est similaire, puis de Voronoï, et enfin de l'approche Naïve. IDW fournit une amélioration de 26 % de la RMSE par rapport à l'approche naïve.

Dans le second scénario (tableau 4.4), les valeurs brutes de MAE et RMSE sont nettement inférieures puisque la normalisation par l'énergie annuelle représente un pourcentage instantané de la production en fonction de l'énergie totale annuelle du producteur. IDW reste la méthode la plus précise. La stratégie de normalisation n'a pas d'impact majeur sur l'efficacité de chaque méthode. La précision des deux scénarios est similaire, la seule différence notable étant que TIN surpasse ici KNN.

La seconde stratégie de normalisation est toutefois plus précise pour toutes les méthodes. En particulier pour IDW, la $nRMSE_\sigma$ est améliorée de 3 points, passant de 22,78 % dans le premier scénario à 19,53 % dans le deuxième.

Dans le cas de l'algorithme IDW, l'erreur résiduelle de l'estimation MBE est répartie de la même manière sur l'ensemble de territoire. Cette erreur est symétrique et centrée au zéro. Cependant, certains producteurs sont soit surestimés (erreur négative), soit sous-estimés (erreur positive) du fait de la différence de configuration de ces producteurs avec leurs voisins référents. La figure 4.17 illustre la répartition spatiale et la distribution de cette erreur dans le cas de la normalisation par la capacité installée.

4.5.1 Optimisation des paramètres

Comme nous l'avons vu dans la section 4.3.1, les algorithmes IDW et KNN nécessitent la détermination de certains paramètres de synthèse. Dans notre étude, nous les avons déterminés empiriquement en les faisant varier sur l'ensemble de données. Nous avons utilisé la méthodologie présentée dans la section 4.3.1, à savoir la validation croisée emboîtée qui permet d'isoler la partition de test de la partition de validation. Cette dernière partition permet donc d'optimiser les paramètres des algorithmes sans influer sur les résultats de comparaison avec les autres méthodes dans la partie test.

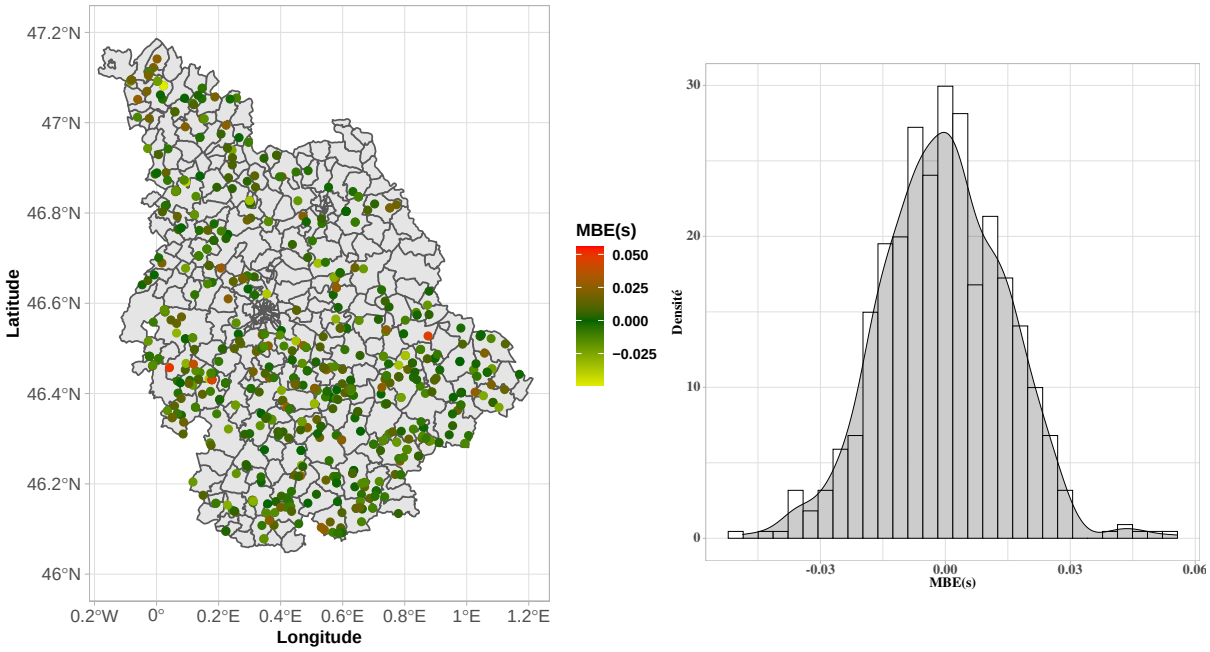


FIGURE 4.17 – Distribution de l'erreur résiduelle de l'estimation
La carte à gauche présente la répartition spatiale de l'erreur MBE(s) de chaque producteur. L'histogramme à droite illustre la distribution de cette erreur.

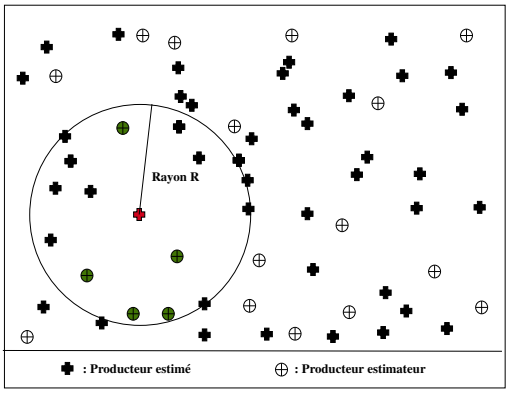


FIGURE 4.18 – Détermination des voisins les plus proches par un rayon maximum

Le paramètre à optimiser dans le cas de l'algorithme KNN est le nombre maximum de voisins à prendre en compte dans l'estimation. Nous pouvons également, au lieu de fixer ce nombre maximum, déterminer un rayon maximum. Ce rayon définit un cercle dans lequel seuls les producteurs de référence qui se trouvent à l'intérieur sont considérés (voir figure 4.18). De même, concernant l'algorithme IDW, nous pouvons déterminer le nombre de voisins ou le rayon maximum à prendre en compte dans l'estimation. De plus, nous pouvons optimiser la puissance p dans IDW utilisée dans la pondération. Dans le cas où $p = 0$, IDW est équivalent à la méthode KNN. Le rayon ou le nombre de voisins permettent d'éliminer les estimateurs qui ne contribuent pas à l'estimation {Shepard, 1968}.

Dans cette expérimentation, nous avons fait varier le rayon maximum entre 6.5 *km* et 125 *km*, la puissance de 0 à 5 avec un pas de 0.1, et le nombre de voisins entre 1 et 60. La borne inférieure du rayon maximum (6.5 *km*) est choisie pour avoir au minimum un voisin pour tous les producteurs. Les résultats de cette étude empirique sont présentés dans la figure 4.19. Nous avons gardé le même découpage des parties test et estimation élaboré dans la comparaison des méthodes. Seul un nouveau découpage de la partie estimation est réalisé pour optimiser les différents paramètres. Les résultats d'évaluation obtenus

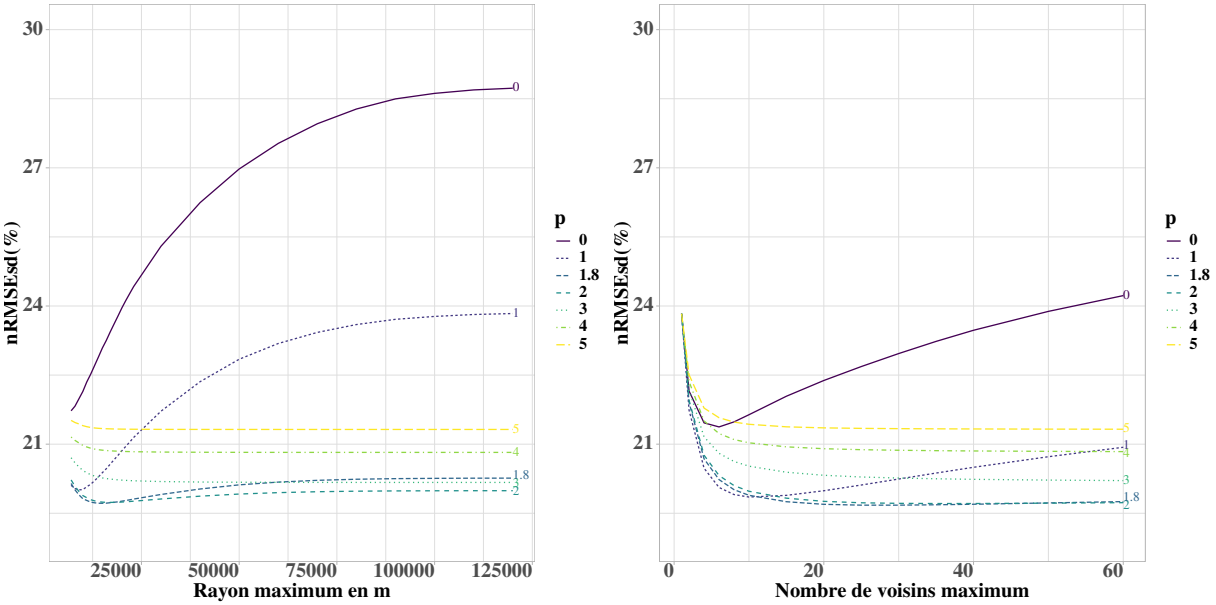


FIGURE 4.19 – Optimisation des paramètres IDW et KNN. Lecture : sur la figure de gauche, nous avons sur l'axe des abscisses le rayon maximum, et sur l'axe des ordonnées la qualité de l'estimation par la métrique $nRMSE_\sigma$ pour chaque puissance p. Dans le graphique de droite, nous avons le nombre de voisins sur l'axe des abscisses et sur l'axe des ordonnées, nous trouvons les $nRMSE_\sigma$ pour chaque p. La méthode KNN est équivalente à $p = 0$ dans IDW

sont réalisés sur la partie test en utilisant les données normalisées par l'énergie annuelle.

Dans le cas de l'algorithme IDW, nous avons obtenu une erreur $nRMSE_\sigma(s)$ minimale de 19.72% pour un rayon maximum optimal de 15 *km* et une puissance optimale p égale 1.8. Dans le cas d'optimisation du nombre de voisins maximum, nous avons obtenu une $nRMSE_\sigma(s)$ minimale de 19.68% dans le cas d'un nombre maximum de 25 voisins et une puissance optimale de 1.8. Dans le cas de l'algorithme KNN (voir $p = 0$ dans la figure 4.19) nous avons obtenu un résultat optimal pour un rayon maximum de 6.5 *km* avec une $nRMSE_\sigma(s)$ de 21.72%. Dans le cas du nombre de voisins maximum, nous avons eu une $nRMSE_\sigma(s)$ de 21.37% pour six voisins.

Dans les deux cas, la définition du voisinage sur la base d'un nombre de voisins au lieu d'une distance maximale a toujours offert une meilleure précision.

4.5.2 Évaluation pour les petits producteurs

Dans cette étude, nous avons élaboré une approche d'estimation de la production solaire des petits producteurs avec une granularité temporelle fine. Pour évaluer les différentes méthodes d'estimation, nous avons utilisé les données des moyens et grands producteurs en utilisant la méthodologie de validation croisée présentée dans la section 4.3.1. Cependant, nous ne disposons d'aucune donnée complète pour valider notre approche dans le cas des petits producteurs. Les seules données disponibles sont les données de comptage d'énergie annuelle et les courbes de production journalière de neuf petits producteurs disponibles dans la base de données Epices 2.3.1. Nous avons utilisé ces deux jeux de données pour quantifier les différences entre les énergies réelles et les énergies estimées par l'interpolation spatiale.

Concernant les données de comptage des petits producteurs, les énergies sont communiquées par les producteurs dans des intervalles et des périodes irréguliers. Nous avons sélectionné un échantillon de 348 petits producteurs qui ont communiqué leurs index entre le premier décembre 2018 et le 31 janvier 2020 sur une période de comptage totale entre 360 et 370 jours afin de calculer l'énergie annuelle produite

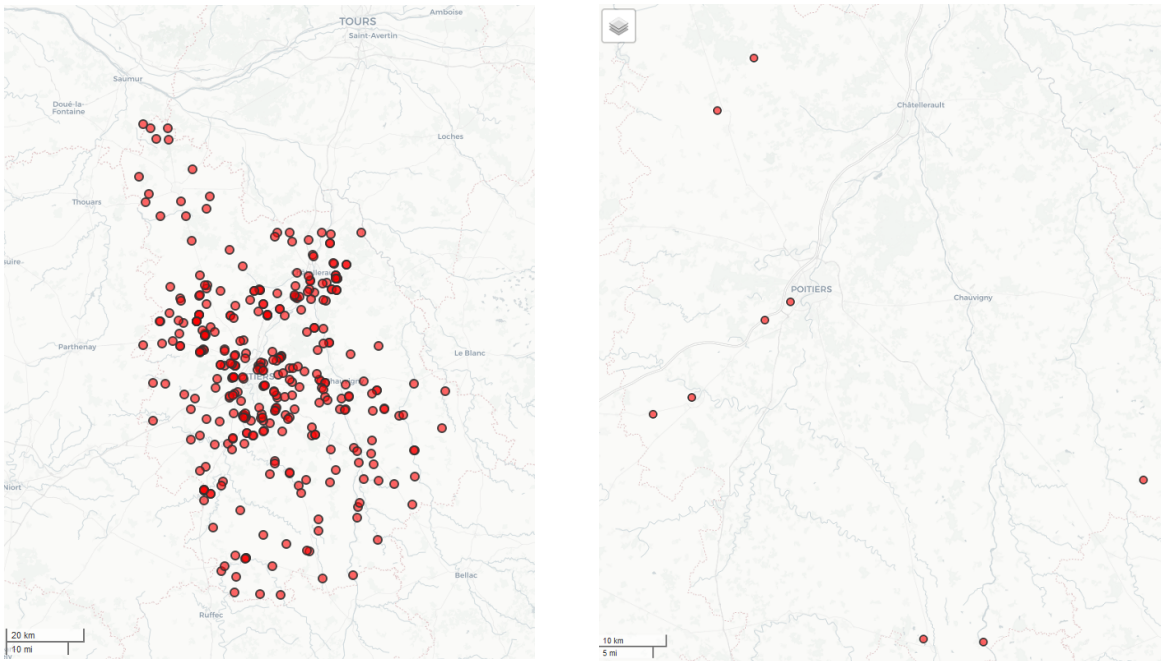


FIGURE 4.20 – Répartition spatiale des petits producteurs de comparaison.
Lecture : la figure à gauche illustre l'emplacement de 348 petits producteurs dans la base de comptage (l'échelle de la carte est de 20 km). La figure à droite illustre l'emplacement des 9 producteurs dans la base Epices (l'échelle de la carte est de 10 km).
Licence carte : fond de carte par OpenStreetMap France, sous la licence Creative Commons CC BY-SA 2.0.

en 2019. Les courbes disponibles dans Epices contiennent de nombreuses périodes manquantes (intra et inter jour). Ceci fausse le calcul des énergies journalières disponibles dans l'outil puisqu'il manque une partie d'énergie sur la journée. Les courbes avec une granularité intra jour ne sont pas disponibles pour tous les producteurs, seules les données avec une granularité journalière et mensuelle sont disponibles.

Nous avons estimé les courbes de production des petits producteurs avec une granularité de dix minutes en exploitant les données des producteurs de référence (612 moyens et grands producteurs) normalisées par la capacité installée. Nous avons utilisé l'algorithme IDW avec 25 voisins et une puissance de 1.8 (paramètres optimaux). Par la suite, nous avons calculé les énergies annuelles à partir des courbes estimées pour les comparer avec les énergies annuelles réelles de comptage. Dans le cas des données Epices, les courbes de productions fines (pas 10 minutes) des neuf producteurs sont estimées en exploitant les données des moyens et grands producteurs normalisées par l'énergie annuelle. Les énergies produites journalières sont calculées par la suite. La figure 4.21 illustre les résultats obtenus dans le cas des données annuelles des comptages. La majorité des énergies annuelles sont surestimées.

La figure 4.21 illustre les résultats obtenus dans le cas des données annuelles des comptages. La figure à gauche présente l'énergie réelle communiquée par les producteurs en fonction de l'énergie estimée par IDW. La figure de droite présente la distribution de l'erreur MPE de chaque producteur. Nous remarquons que la majorité des énergies annuelles sont surestimées (points au-dessus de la droite), la MPE médiane égale à -14 % avec une MAE médiane de 502 kWh. Nous notons que certains producteurs sont beaucoup trop surestimés avec une MPE allant jusq'à -230 %. Ceci peut être expliqué par plusieurs raisons. Premièrement, nous supposons dans l'estimation spatiale que le producteur a produit en continuité sur toute l'année, comme ses voisins estimateurs. Des problèmes sur une installation peuvent fausser l'estimation et entraîner une surestimation de l'énergie. Deuxièmement, sur les petits producteurs, l'orientation, l'in-

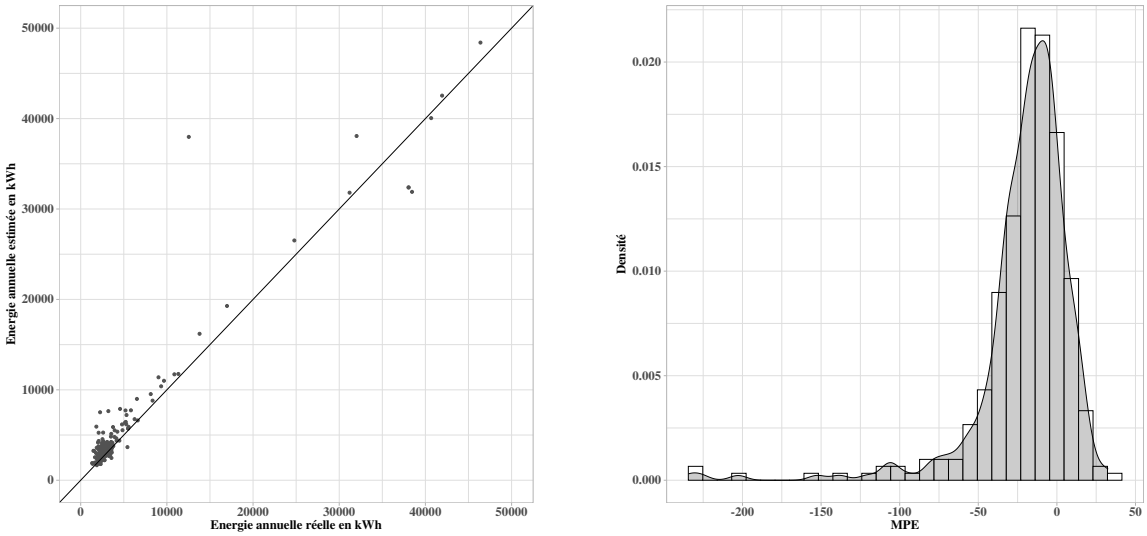


FIGURE 4.21 – Résultats obtenus dans le cas de données annuelles de comptage.

clinaison ou encore les ombrages sont probablement moins optimisés que pour les grandes centrales (les panneaux sont posés sur une toiture existante). La normalisation par énergie annuelle est plus pertinente que la normalisation par capacité installée, car elle permet de mieux estimer certains producteurs bien ou mal dimensionnés par rapport à leur capacité comparé aux capacités de leurs voisins.

La figure 4.22 et le tableau 4.5 présentent les résultats obtenus dans le cas d'estimation des énergies journalières des neuf producteurs Epices. L'orientation des panneaux solaires est par rapport au nord (orientation plein nord égale à 0°). L'erreur MAPE varie entre 4.38% et 26.58%, nous n'avons pas pu calculer la MAPE dans le cas du producteur 8, car il y a des valeurs nulles dans la courbe de production Epices. Les courbes de la figure 4.22 illustrent les différences entre les énergies estimées et les énergies calculées dans la base Epices. Nous remarquons pour les producteurs 8 et 9 que certaines périodes de production manquantes influent sur la qualité de l'estimation. Le producteur le mieux estimé est le producteur 6 avec une erreur MAPE annuelle de 4.38%. Le tableau 4.6 présente les résultats de l'estimation selon les saisons. Dans le cas des producteurs 1 et 2, nous observons une surestimation de l'énergie journalière sur l'année. Concernant les producteurs 3, 4, 5, 6 et 7, une sous-estimation est remarquée dans le cas d'hiver contre une surestimation dans le cas d'été. Cela est dû aux différences d'orientation et d'inclinaison entre les producteurs de référence et les producteurs estimés. En effet, les grands producteurs sont dimensionnés généralement auteur de l'optimum été (orientation 180° et inclinaison 30°), or certains producteurs inclinés au-delà des 30° produisent moins l'été mais davantage l'hiver, cela implique une surestimation ou une sous-estimation selon chaque saison. L'ajout des informations sur les périodes de maintenance et d'arrêt de production, ainsi que l'orientation et l'inclinaison des panneaux solaires peuvent éviter la surestimation ou la sous-estimation de l'énergie pour les différents producteurs estimés.

4.5.3 Discussion

Sur la base des résultats obtenus, nous avons développé une méthodologie d'estimation de la production PV des producteurs non instrumentés par un compteur communicant. Cette estimation est fondée sur les méthodes d'interpolation spatiales déterministes en utilisant les données des producteurs instrumentés comme référence. Nous avons testé deux scénarios de normalisation de données pour mieux rapprocher les courbes de production. Dans cette normalisation, nous n'avons utilisé que les informations dispo-

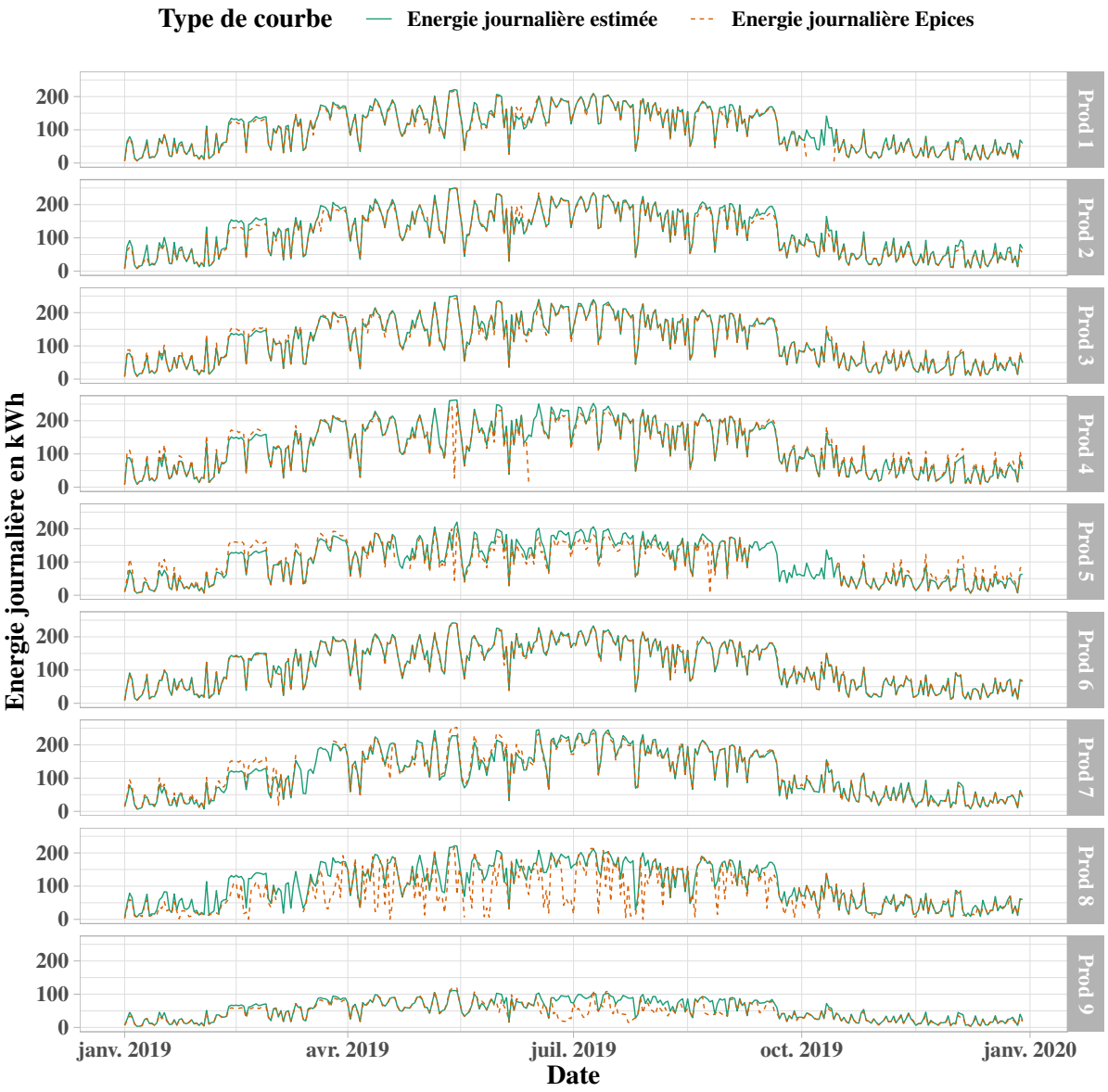


FIGURE 4.22 – Comparaison entre les courbes de production journalière Epices et la courbe journalière estimée.
Lecture : L'axe des abscisses représente la date et l'axe des ordonnées représente l'énergie journalière en kWh. Chaque bloc représente les courbes de production journalière estimées et mesurées dans Epices de chaque producteur.

Producteur	Capacité (kVA)	Orientation (°)	Inclinaison (°)	MBE^+ (kWh)	MBE^- (kWh)	MBE (kWh)	MAE (kWh)	RMSE (kWh)	MAPE (%)
Prod 1	27.00	186.00	16.00	4.09	-5.25	-3.30	5.01	8.46	11.71
Prod 2	32.00	146.00	12.00	3.29	-8.82	-4.60	6.89	10.17	7.65
Prod 3	30.00	145.00	20.00	4.96	-5.68	0.19	5.28	7.33	4.83
Prod 4	33.00	165.00	25.00	7.62	-8.92	0.24	8.20	17.62	15.14
Prod 5	30.00	165.00	25.00	13.49	-13.30	-2.65	13.37	20.76	26.58
Prod 6	34.00	165.00	25.00	3.68	-4.39	-0.25	4.02	5.52	4.38
Prod 7	30.00	190.00	17.00	9.56	-8.30	1.52	8.99	15.92	16.30
Prod 8	30.00	190.00	20.00	5.67	-40.65	-24.25	28.26	49.60	–
Prod 9	16.00	180.00	17.00	2.00	-13.35	-7.83	9.27	18.48	26.53

TABLEAU 4.5 – Résultats de comparaison obtenus dans le cas des 9 producteurs disponibles dans la base Epices.

Producteur	Saison	MBE (kWh)	MAE (kWh)	MAPE (%)
Prod 1	Été	-2.31	5.09	3.90
Prod 1	Hiver	-4.36	4.93	20.07
Prod 2	Été	-2.35	5.96	3.88
Prod 2	Hiver	-6.86	7.83	11.46
Prod 3	Été	-3.30	5.98	4.07
Prod 3	Hiver	3.72	4.57	5.59
Prod 4	Été	-6.11	8.81	21.22
Prod 4	Hiver	6.59	7.58	9.06
Prod 5	Été	-14.68	16.67	40.93
Prod 5	Hiver	8.88	10.22	12.83
Prod 6	Été	-0.67	4.91	3.39
Prod 6	Hiver	0.19	3.12	5.38
Prod 7	Été	-0.07	8.83	5.73
Prod 7	Hiver	3.28	9.17	27.96
Prod 8	Été	-33.37	38.36	–
Prod 8	Hiver	-14.55	17.52	–
Prod 9	Été	-12.93	15.27	42.73
Prod 9	Hiver	-2.67	3.20	10.15

TABLEAU 4.6 – Résultats obtenus dans le cas des 9 producteurs selon les saisons été et hiver.

nibles des petits producteurs comme leur capacité installée et leur énergie totale produite annuellement. Il existe d'autres méthodes de normalisation comme la normalisation par le maximum de production sur l'année mais ces informations ne sont pas disponibles pour les petits producteurs. D'après les résultats obtenus, la normalisation par l'énergie annuelle nous permet d'améliorer la qualité de l'estimation, car certains producteurs ne sont pas correctement configurés et leur capacité installée ne tient pas compte des défauts environnementaux comme l'ombrage, l'orientation et l'inclinaison sous-optimales. De ce fait, ils produisent moins d'énergie que celle prévue par leur capacité installée, certains producteurs ne dépassent pas 60% de leur capacité sur toute l'année même sur des périodes d'été. Par conséquent, la normalisation par l'énergie annuelle est plus pertinente et permet de mieux saisir l'hétérogénéité entre les producteurs. La meilleure approche de normalisation dans notre cas d'étude est la normalisation par l'énergie annuelle.

Dans les deux figures 4.23 et 4.24, les valeurs de production réelles sur deux jours d'été sont représentées pour deux producteurs. Le premier jour est un jour avec un ciel clair, et le deuxième jour est nuageux. Le premier producteur (figure 4.23) est estimé avec une erreur $nRMSE_{\sigma}$ de 10% et le second (Figure 4.24) avec une erreur de 30%. Nous constatons clairement pour le premier producteur, par ciel clair, les avantages de la normalisation par la production annuelle d'énergie, car elle rapproche mieux chaque estimation de la production réelle.

Les méthodes d'estimation spatiale permettent de mieux saisir le mouvement des nuages, car ce phénomène est très local. L'estimation spatiale est pertinente dans ce cas car elle permet de saisir ces mouvements des nuages à partir de la production des producteurs de référence. Dans cette approche nous avons utilisé seulement les données endogènes des courbes de production des voisins référents. Aucune autre donnée externe n'est exploitée.

Toutefois, la limite de cette approche est qu'elle ne prend pas en compte les informations sur l'orientation, l'inclinaison et l'ombrage des panneaux. En effet, nous ne disposons d'aucune information aujourd'hui sur les caractéristiques physiques des panneaux solaires et la structure interne des producteurs.

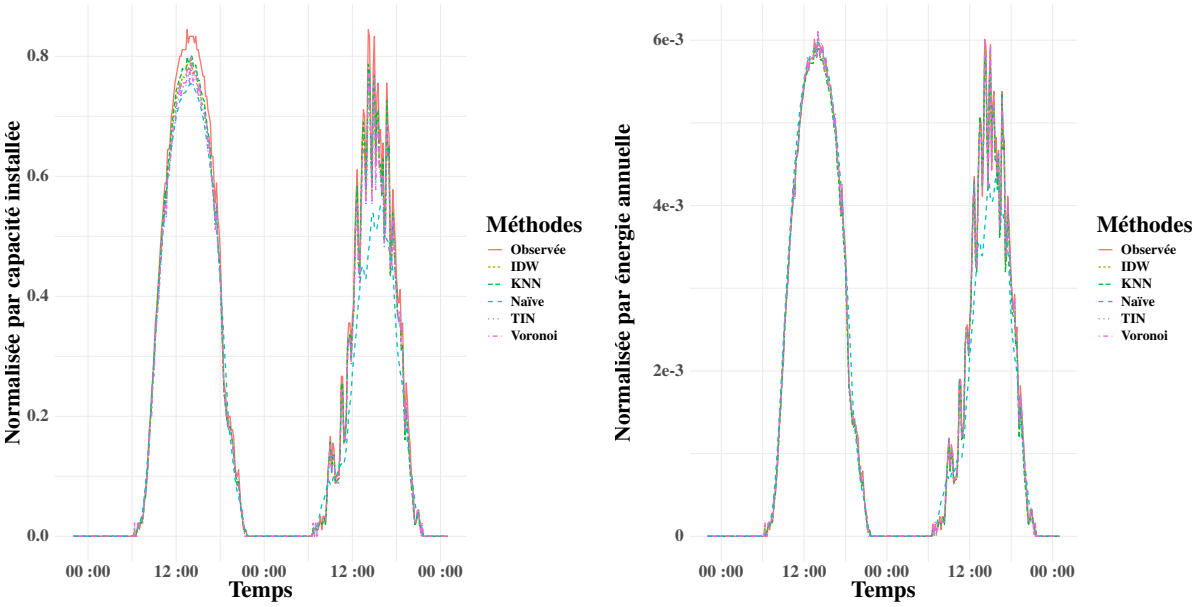


FIGURE 4.23 – Exemple d'un producteur bien estimé
Sur l'axe des abscisses, nous avons les heures de deux jours d'été, et sur l'axe des ordonnées, à gauche, nous avons la production toutes les 10 minutes normalisée par la capacité installée des producteurs et à droite, nous avons la production normalisée par l'énergie annuelle

Par conséquent, l'estimation spatiale dans le cas d'un producteur différent en termes d'orientation et d'inclinaison est moins pertinente, même si nous saisissons bien les informations sur le mouvement des nuages. La figure 4.24 illustre l'exemple d'un producteur mal estimé à cause de cette différence d'orientation par rapport à ses voisins. Sa courbe de production est décalée par rapport aux autres courbes à cause de la configuration de ses panneaux solaires. Cependant, nous captons bien les mouvements de nuages. Pour améliorer la qualité d'estimation, il faut ajouter ces informations internes (ombrage, orientation et inclinaison) dans la méthodologie d'estimation. Malheureusement, nous ne disposons pas de ces informations aujourd'hui dans les bases de données de SRD, à part pour certains moyens et grands producteurs dans la base de données EnR (voir section 2.3.1).

La figure 4.25 illustre la méthodologie finale suivie dans l'estimation d'une courbe de production d'un producteur donnée. Après une étape d'importation des données statiques de la cartographie et des courbes de production des producteurs de référence, une matrice de distance est calculée. Si l'énergie annuelle du producteur estimée est disponible, nous normalisons les données par cette énergie, sinon nous les normalisons par sa capacité installée. Ensuite, nous appliquons la méthode IDW avec 25 voisins et une puissance p égale à 1.8 (paramètres optimaux). Finalement, nous exportons la courbe de production estimée du producteur.

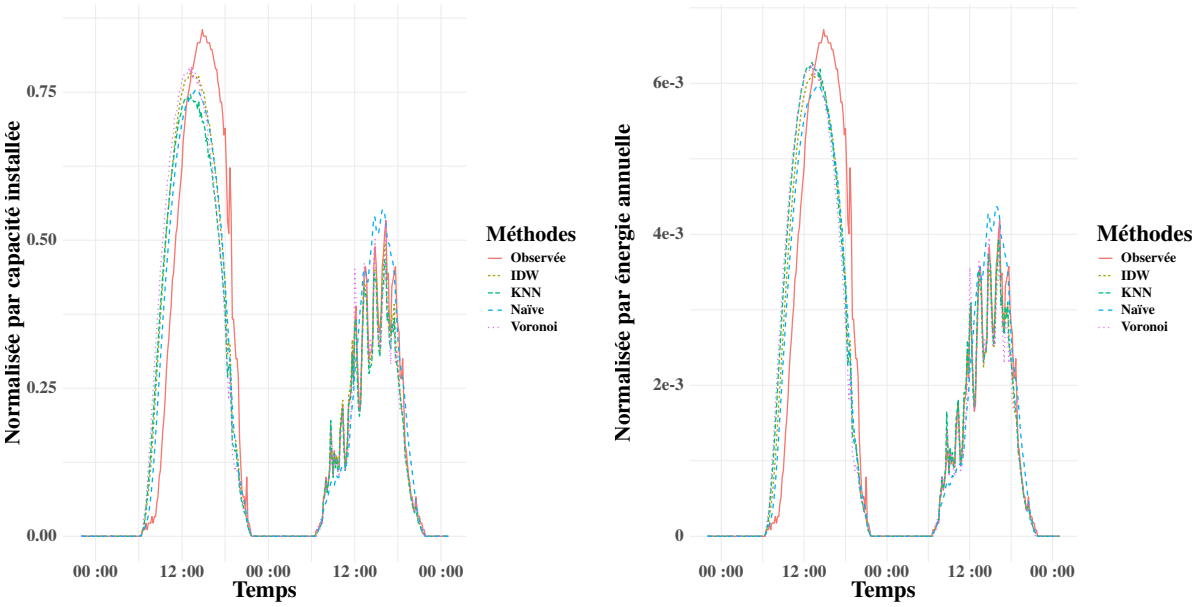


FIGURE 4.24 – Exemple d'un producteur mal estimé
Sur l'axe des abscisses, nous avons les heures de deux jours d'été, et sur l'axe des ordonnées, à gauche, nous avons la production toutes les 10 minutes normalisée par la capacité installée des producteurs et à droite, nous avons la production normalisée par l'énergie annuelle

4.6 Conclusion du chapitre

Dans ce chapitre, nous avons développé une méthodologie d'estimation de la production PV dans un réseau de distribution d'électricité en utilisant des méthodes d'interpolation spatiale. L'objectif de cette étude est d'estimer en granularité fine la production des petits producteurs raccordés au réseau BT pour lesquels nous ne disposons que de données agrégées (biannuelles). Ces estimations de production PV sont ensuite utilisées pour séparer le flux des consommations des flux des productions, afin de calculer les charges réellement consommées nécessaires pour l'optimiseur.

L'évaluation des différentes méthodes a montré que la méthode IDW est la plus précise et transmet assez bien les informations sur les mouvements des nuages captées par les variations de puissance des producteurs voisins. Les résultats montrent un $nRMSE_\sigma$ de 19 % obtenu par la méthode IDW, contre 29 % obtenu par la méthode naïve. En outre, nous pouvons optimiser cette méthode en optimisant plusieurs paramètres tels que le rayon ou le nombre de voisins à prendre en compte dans l'interpolation. Sur cet ensemble de données réelles, les paramètres optimaux pour p et le nombre de voisins étaient respectivement de 1,8 et 25.

D'autre part, la configuration des panneaux solaires des producteurs peut dégrader la qualité de l'estimation si l'on tente d'estimer un producteur dont la configuration diffère sensiblement de celle de ses voisins. Dans ce cas, la production journalière a été déplacée de manière inappropriée, même si nous avons capturé les mouvements des nuages. Ce décalage pourrait être en grande partie causé par l'orientation des panneaux. L'ajout de cette information dans les méthodes d'estimation pourrait améliorer la qualité de l'estimation. Cependant, cette information n'est actuellement pas disponible pour tous les producteurs. A noter que le territoire de SRD (département de la Vienne) est peu vallonné. Le calcul des distances entre les producteurs fonctionne bien et le mouvement des nuages n'est pas influencé par les collines qui peuvent bloquer la circulation des masses d'air. La présente étude d'estimation sera donc différente dans des régions montagneuses.

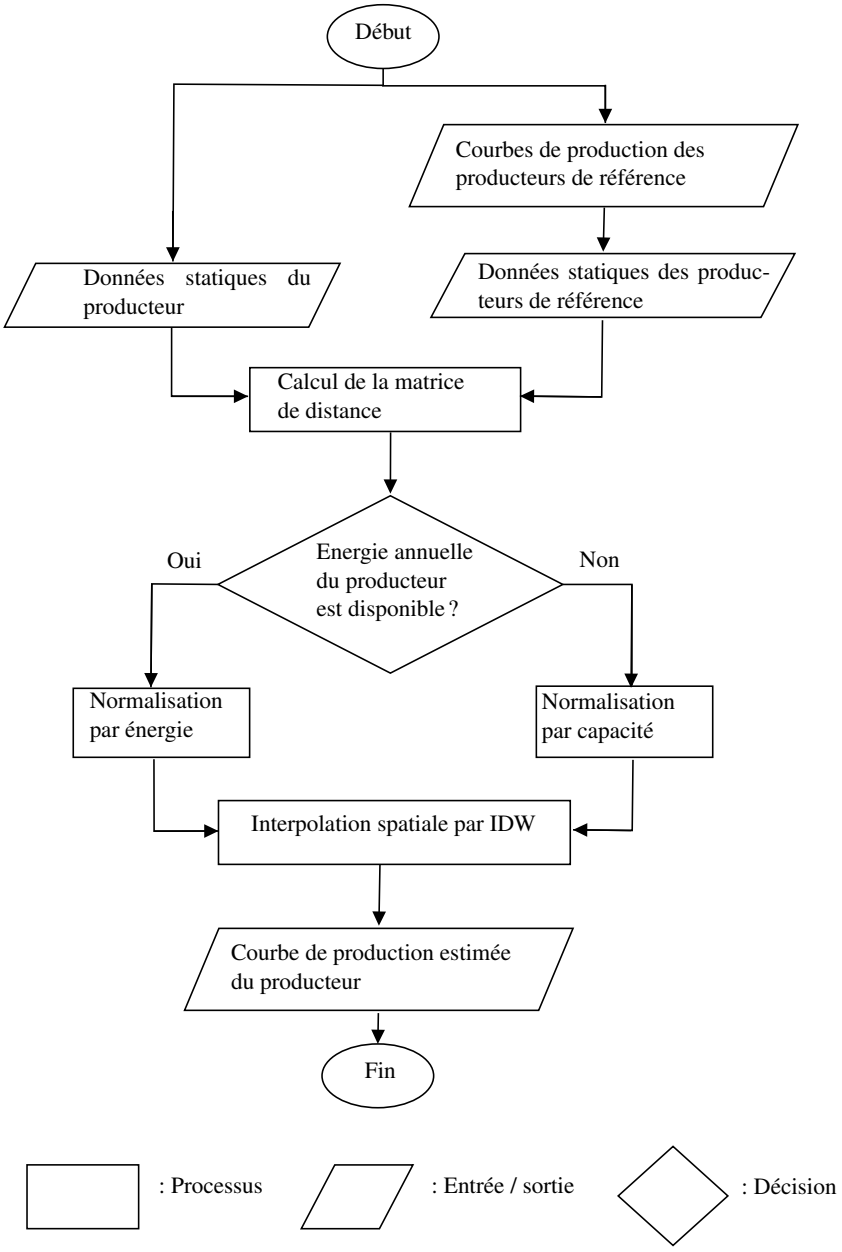


FIGURE 4.25 – Méthodologie finale d'estimation spatiale

Ces méthodes d'interpolation déterministe sont toutefois limitées car nous ne prenons pas en compte les caractéristiques aléatoires du phénomène étudié. Nous n'utilisons que les informations géométriques de l'espace. Par exemple, dans la méthode IDW, nous ne nous intéressons qu'à la distance qui sépare les producteurs estimés de leurs voisins estimateurs. Cette méthode ne tient pas compte de la variation spatio-temporelle de la production. Cette production dépend de plusieurs paramètres internes liés à la configuration du parc PV tels que son orientation, son inclinaison, et d'autres paramètres externes tels que les aléas de la météo (rayonnement, direction du nuage et sa vitesse, etc.).

Généralement, les méthodes déterministes ne permettent pas d'étudier ces variations. Les méthodes d'interpolation probabilistes peuvent améliorer la précision d'estimation. Ces méthodes font partie du domaine de la géostatistique mais elles sont coûteuses en ressources informatiques. Par exemple, nous avons testé la méthode d'interpolation spatiale probabiliste OK sur les données de production d'énergie PV. Celle-ci a dépassé la limite d'une heure de calcul que nous avions fixée comme limite pour les différentes méthodes.

Dans le cas de grands volumes de données, la méthode IDW peut être une méthode appropriée pour estimer la production des producteurs PV en n'utilisant que les données endogènes d'historique de production.

Chapitre 5

Prévision de la production PV

Dans ce chapitre, nous présentons une méthodologie de prévision de la production solaire des producteurs photovoltaïques distribués dans un réseau de distribution d'électricité (HTA et BT). Le but de cette étude est d'élaborer une prévision ponctuelle court terme d'un horizon d'une heure pour gérer l'intermittence de la production solaire et une prévision probabiliste long terme pour planifier et optimiser le réseau sur un horizon d'un mois à trois mois. Nous avons évalué les différentes approches de prévision en exploitant les données de 621 producteurs instrumentés par des compteurs communicants. Sur ces données, nous montrons que les algorithmes d'apprentissage automatique avec une approche globale améliorent les prévisions fournies par des méthodes naïves. Les prévisions ont été validées par un estimateur d'état du réseau pour quantifier les différences de pertes, de chutes de tension et d'élévation de tension entre l'état estimé prévu de réseau et l'état estimé réel.

Sommaire

5.1 Introduction du chapitre

Dans le chapitre précédent, nous avons présenté notre étude sur l'estimation de la production PV en utilisant les méthodes d'interpolation spatiale. L'objectif de ce travail est l'élaboration d'une approche d'estimation d'une production non mesurable aujourd'hui avec une granularité fine. Cette production estimée doit être utilisée pour la prévision de la production solaire du réseau de SRD.

Nous avons vu précédemment que la France a été mobilisée ces dernières années pour réduire ses émissions de gaz à effets de serre. Avec la loi du 17 août 2015 relative à la transition énergétique [1], la France s'est engagée à renforcer son indépendance énergétique en augmentant sa capacité de production en énergie renouvelable de 50% en horizon 2030. Cette transition énergétique favorise la diversification des moyens de production de l'énergie en valorisant les productions des EnR.

Dans ce chapitre, nous présentons notre travail sur la prévision de la production solaire de SRD. Nous exposons dans une première partie l'état de l'art et les différents modèles utilisés dans cette étude. Dans une deuxième partie, la méthodologie de prévision et d'évaluation est présentée. Finalement, nous exposons les différents résultats obtenus en exploitant les données réelles de SRD.

5.2 Objectif détaillé

Nous avons présenté précédemment dans le chapitre 4 que SRD a connu entre 2010 et 2019 une croissance de 700 % de sa capacité totale de production liée à l'énergie solaire. Cette augmentation s'explique par les politiques étatiques favorisant le développement de ces moyens de production et également le coût des panneaux solaires qui a chuté de 85% depuis 2010 {IRENA, 2021}. Toutefois, cette énergie solaire est intermittente et variable puisqu'elle dépend fortement des aléas de la météo. C'est-à-dire, la puissance de sortie d'un panneau PV est très incertaine en raison de plusieurs facteurs météorologiques comme la température, la vitesse du vent, la couverture nuageuse, les niveaux d'aérosols atmosphériques, le taux d'humidité, etc. Conséquemment, sa dépendance à ces conditions rend la prévision de sa production complexe et difficile.

L'intégration de cette énergie intermittente dans le réseau génère plusieurs contraintes de gestion. Par exemple, elle peut créer des variations de tensions et donc dégrader la qualité de fourniture de l'électricité. Pour faire face à ces changements, une prévision de cette production est indispensable pour bien gérer cette intermittence. Cette prévision nous permet à la fois d'optimiser le réseau sur un horizon temporel donné et d'identifier les contraintes de dimensionnement affectant le réseau dans un état de fonctionnement futur. D'une part, la prévision nous permet d'anticiper les problèmes de gestion de réseau, comme les variations de tensions, afin d'assurer une qualité de fourniture de l'électricité aux différents clients. D'autre part, elle nous permet de calculer le schéma d'exploitation optimal de réseau en considérant les différentes courbes prédictives de charges et de production.

Dans ce projet de prévision de la production solaire, nous avons développé des modèles de prévision permettant le maintien de la sécurité du réseau, le contrôle des variations des tensions et l'optimisation des flux de l'énergie décentralisée. Pour gérer l'intermittence de l'énergie PV, elle peut être stockée. Cependant, avec un coût de stockage très élevé aujourd'hui cette solution reste chère et difficile à appliquer. De plus, même la solution du stockage d'énergie nécessite des prévisions pour bien gérer et optimiser les différentes ressources.

1. LOI n° 2015-992 du 17 août 2015 relative à la transition énergétique pour la croissance verte

Notre méthodologie de prévision est orientée vers l'optimisation du réseau en intégrant les courbes prédictives de production et de consommation. Notre projet de prévision s'est déroulé en trois grandes étapes :

— Étape 1 : Estimation des courbes de production des petits producteurs (voir chapitre 4).

— Étape 2 : Séparation entre les courbes de consommation et les courbes de production.

— Étape 3 : Prévision de la production et de la consommation d'énergie sur plusieurs horizons temporels.

Dans l'étape 1, nous estimons la production non mesurable avec une granularité fine, cette étape fait l'objet du chapitre 4. Dans la deuxième étape, nous différencions la courbe de consommation totale au niveau des départs HTA et la courbe de production totale de ces départs. Cette étape nous permet de calculer la charge réelle consommée au niveau d'un départ HTA sans l'énergie produite localement par les producteurs décentralisés. Dans l'étape finale, nous développons des modèles de prévision de la charge réelle consommée et de la production des EnR. Cette prévision est sur plusieurs horizons temporels selon les cas d'usage. Globalement, nous distinguons trois catégories de prévision en fonction de l'horizon temporel.

Premièrement, nous avons besoin d'une prévision à court terme avec un horizon de prévision d'une heure et avec un pas de dix minutes. Le but de cette prévision est d'anticiper les contraintes arrivant sur le réseau sur un horizon court terme, pour permettre aux agents d'en réaliser la conduite. Deuxièmement, une prévision à moyen terme d'un horizon de quinze jours est nécessaire pour planifier les différents travaux sur le réseau. En effet, les travaux sont planifiés quinze jours avant leur réalisation. Cette prévision permet donc de bien gérer ces différentes maintenances en configurant le réseau avec un schéma de secours optimal. Finalement, nous avons besoin d'une prévision d'un mois à trois mois afin de configurer le réseau par un schéma d'exploitation optimal prédictif à long terme.

En outre, SRD voudra expérimenter dans le futur une solution d'optimisation stochastique en plus de la solution d'optimisation déterministe existante aujourd'hui. Cette solution d'optimisation stochastique sera utilisée dans un horizon temporel à long terme en intégrant les différentes informations probabilistes dans la modélisation. En réalité, la gestion du réseau est très complexe puisque l'environnement extérieur du réseau est dominé par une stochasticité liée aux aléas de la météo et aux différents équipements du réseau. Une solution d'optimisation stochastique est donc intéressante pour une gestion simultanée de la charge réelle, des nouveaux usages d'électricité comme la mobilité électrique, de l'énergie solaire, de l'énergie éolienne, des différentes formes de stockage de l'énergie et finalement de la gestion des micro-réseaux {Anderson *et al.*, 2011}. Pour ces raisons, nous avons développé dans le cas de l'horizon long terme une approche de prévision probabiliste qui sera utilisée comme entrée dans l'outil d'optimisation probabiliste.

De nombreux travaux de la littérature considérant la prévision de la production solaire. Nous pouvons les classer en plusieurs familles d'approches de prévision. D'une part, nous distinguons ces approches selon l'horizon temporel : très court terme, court terme, moyen terme et long terme. D'autre part, nous les distinguons par leur nature, ponctuelle ou probabiliste, et par les différentes données d'entrée (endogènes ou exogènes). Les modèles de prévision ponctuelle permettent une prévision d'une valeur pour un horizon donné et les modèles probabilistes permettent une prévision d'une distribution de valeurs. Dans les travaux sur la prévision de la production PV, peu de chercheurs ont travaillé sur la deuxième famille de modèles probabilistes comparativement avec les travaux sur la prévision de la consommation

ou de l'énergie éolienne {Antonanzas *et al.*, 2016}. Cette catégorie de prévision probabiliste est intéressante pour un GRD, du fait qu'elle mesure les incertitudes sur les maximums et les minimums des prévisions et donc les risques liés de sur-production ou sur-consommation. En modélisant ces différents besoins de SRD en matière de prévision, nous avons choisi trois catégories de prévision résumées dans le tableau 5.1.

Type de prévision	**Horizon**	**Pas de prévision**	**Données**	**Nature de prévision**	**Besoin métier**
Court terme	1 h	10 min	Endogènes	Ponctuelle	Gérer l'intermittence
Moyen terme	15 jours	1 h	Endogènes et exogènes	Ponctuelle et probabiliste	Planification des travaux
Long terme	1 mois-3 mois	1 h	Endogènes	Probabiliste	Optimisation long terme

TABLEAU 5.1 – Besoins de SRD en matière de prévision de la production solaire

Dans cette présente étude, nous avons développé deux approches de prévision. Premièrement une prévision court terme est étudiée en utilisant l'historique de la production des moyens et grands producteurs PV (données endogènes). Deuxièmement, une approche de prévision probabiliste à long terme est développée en utilisant les mêmes données. Finalement nous proposons une méthodologie d'évaluation et de validation des différents modèles en comparant avec des méthodes de prévision naïves et en utilisant un estimateur d'état de réseau. Ce calculateur d'état nous permet de valider les différents modèles avec un aspect métier d'un gestionnaire de réseau de distribution en identifiant les pertes et les variations de tension. Dans le temps imparti de cette thèse, nous n'avons pas considéré la prévision à moyen terme.

Dans la section suivante, nous présentons les fondements des méthodes utilisées dans l'élaboration de ces méthodes de prévision.

5.3 État de l'art

Avec les évolutions des recherches dans le domaine des mathématiques et de la statistique ainsi qu'avec le progrès de l'informatique et des capacités de calcul machine, la prévision est devenue un domaine mature de recherche scientifique avec diverses applications dans plusieurs domaines d'activités scientifiques et économiques. La prévision est devenue une tâche statistique courante et essentielle dans la vie des entreprises. Elle constitue un outil d'aide à la décision important éclairant la planification et la gestion des ressources.

Dans les entreprises, elle est souvent confondue avec la planification et les objectifs. Les objectifs sont ce que nous souhaitons dans le futur, tandis que la prévision consiste à prévoir le futur en exploitant toutes les données historiques et les connaissances métier de l'entreprise. La planification est une application de la prévision et des objectifs. Une bonne planification se fonde sur une détermination appropriée des actions à mener pour que la prévision soit proche des objectifs {Hyndman et Athanasopoulos, 2018}.

Dans le cas d'un GRD, la prévision est un outil primordial pour la gestion optimale et performante de réseau. Elle permet une détermination et une identification des meilleures actions à appliquer dans le réseau selon l'évolution et la variabilité de la consommation et de la production.

5.3.1 Présentation des modèles de prévision

Il existe de multiples familles de méthodes et de modèles dans le domaine de la prévision. Elles dépendent majoritairement de la nature des données disponibles, de leur qualité, de leur volume et de leur granularité. Nous distinguons deux grandes familles de méthodes : les méthodes qualitatives lorsque nous n'avons aucune donnée disponible ou les données existantes ne sont pas pertinentes et les méthodes quantitatives dans le cas contraire. Les méthodes quantitatives sont applicables si deux conditions sont remplies. D'une part, nous avons des informations numériques suffisantes et pertinentes sur le passé du phénomène étudié. D'autre part, nous supposons que certaines caractéristiques et aspects sur les tendances du phénomène se répètent et se poursuivent dans le futur {Hyndman et Athanasopoulos, 2018}. Les études sur les problèmes de la prévision quantitative se fondent sur le domaine d'analyse de séries chronologiques (présentées dans le chapitre 2). Dans notre cas d'étude, nous utilisons les données historiques de la production des parcs PV de SRD. Ces données constituent des séries temporelles mesurées avec un pas de 10 minutes.

La prévision de la production solaire dépend en grande partie du rayonnement solaire global qui atteint les panneaux solaires. Ce rayonnement solaire est un phénomène météorologique dépendant d'autres phénomènes physiques liés à l'atmosphère terrestre et au soleil. Il possède donc des tendances qui se poursuivent dans un futur proche ou lointain. Par exemple, il dépend de la saison, la production en été est plus grande qu'en hiver (voir figure 4.3 chapitre 4).

Dans cette section, nous allons présenter les modèles de prévision utilisés dans notre présent travail sur la prévision de la production solaire. Les modèles sont distingués en trois catégories :

— les modèles naïfs,

— les modèles statistiques classiques,

— les modèles d'apprentissage automatique.

Dans la modélisation générale des séries temporelles, soit par les méthodes naïves, statistiques ou apprentissage machine, l'objectif est de trouver un modèle décrivant le mieux la réalité du phénomène étudié. Nous cherchons à trouver et à sélectionner un modèle qui, à partir de plusieurs données d'entrée endogènes ou exogènes, fournisse une prévision proche de la réalité observée. La modélisation générale du problème s'écrit de la façon suivante :

$$Y_t = f(Y_{t-1}, Y_{t-2}, ..., X_{t,1}, X_{t,2}, .., X_{t,m}) + \varepsilon_t \tag{5.1}$$

avec f une fonction à estimer par le modèle, Y_{t-i} représente l'historique des valeurs de la donnée endogène à prédire, $X_{t,i}$ sont les données exogènes, m est le nombre de variables exogènes et ε_t un bruit blanc.

Par la suite, nous notons cette modélisation par l'équation suivante

$$Y_t = f(X_t) + \varepsilon_t \tag{5.2}$$

Nous appelons Y_t la variable cible à prévoir et X_t les variables explicatives du modèle. Un modèle est une estimation $\hat{f}$ de la fonction f. Dans l'apprentissage statistique ou automatique le but est d'approcher le mieux la fonction f en utilisant et en sélectionnant les données les plus pertinentes.

Modèles naïfs

L'approche naïve de prévision appelée aussi modèle de persistance est largement utilisée dans la prévision météorologique. Dans ce modèle simple, nous supposons que la quantité de la production PV ou du rayonnement solaire à un instant $t+1$ égale tout simplement la quantité à l'instant t.

$$Y_{t+1} = Y_t \tag{5.3}$$

Cette simplicité de prévision est non négligeable. L'approche naïve est difficile à devancer par les autres méthodes plus complexes surtout dans des horizons de prévision court terme. Une généralisation de cette méthode naïve est la méthode de moyenne mobile ou de moyenne glissante. Ce modèle considère que la valeur à l'instant $t+1$ est une moyenne glissante des anciennes valeurs sur une période k dans le passé. Ce filtrage permet d'atténuer les fluctuations transitoires, atténuant ainsi les erreurs de mesure, ou les fluctuations importantes.

$$Y_{t+1} = \frac{1}{k}\sum_{i=0}^{k-1} Y_{t-i} \tag{5.4}$$

Malgré la simplicité fondamentale de ces modèles, ils font partie des modèles de référence les plus utilisés dans la prévision de la production solaire. Dans le développement de toute nouvelle approche de prévision, il est important de comparer la précision de prévision avec celle de l'approche naïve. Il a été démontré que la qualité de prévision de cette méthode naïve diminue nettement avec l'horizon temporel de prévision {Huang *et al.*, 2012; Madsen *et al.*, 2005; Wan *et al.*, 2015}.

Dans notre étude, nous avons utilisé les deux méthodes présentées par les équations (5.3) et (5.4) comme méthodes de référence de comparaison des autres approches développées.

Modèles statistiques

Nous avons vu dans la section 2.2.3 chapitre 2 qu'une série temporelle Y_t peut se décomposer dans le cas additif par

$$Y_t = m_t + s_t + \varepsilon_t$$

Ou dans le cas multiplicatif par

$$Y_t = m_t \times s_t \times \varepsilon_t$$

où m_t représente la composante tendancielle, s_t représente la composante saisonnière et ε_t est un bruit aléatoire.

Dans la prévision statistique, nous supprimons les deux composantes tendancielles et saisonnières. Nous appliquons ensuite le modèle sur les résidus stationnaires. Parfois, il est nécessaire d'appliquer une transformation avant de réaliser la modélisation. Par exemple, si nous cherchons à prévoir une quantité positive comme un prix ou une quantité d'énergie produite, certains modèles prévoient des valeurs négatives. Pour remédier à ce problème, nous pouvons appliquer une transformation logarithmique sur la série. Elle permet ainsi de réduire les fluctuations dans les données. S'il y a des valeurs nulles, comme dans notre cas par exemple où à certaines heures nous avons une production nulle, nous pouvons ajouter

une constante positive sur les données avant d'appliquer la transformation logarithmique. Concernant la suppression de la tendance et de la saisonnalité, il existe plusieurs méthodes permettant cela (voir chapitre 2).

Dans la phase finale, nous choisissons le modèle qui correspond le mieux aux résidus en utilisant plusieurs métriques statistiques permettant de mesurer la qualité de prévision en regardant les différences entre les valeurs fournies par le modèle et les valeurs réelles observées dans une période du passé. Dans la phase finale, nous choisissons le meilleur modèle en précision afin de prévoir les résidus. Par la suite, nous inversons toutes les transformations que nous avons appliquées sur la série (transformation logarithme, ajout de constante) pour obtenir la prévision finale.

Modèles machine learning

Dans l'apprentissage machine supervisé, nous cherchons à ajuster une fonction à partir d'un ensemble de données d'entrée. D'une manière générale, supposons que notre modèle est de la forme $Y = f(X) + \varepsilon$. Dans cette opération d'apprentissage, nous observons le phénomène étudié via les données disponibles à la fois d'entrée X et de sortie Y et nous les assemblons ensuite sous forme d'un ensemble d'apprentissage $A = \{(y_i, x_i)\}$ pour tout i dans $[1, N]$. Dans cet ensemble, les données d'entrée x_i observées alimentent un système artificiel produisant les sorties y_i via une fonction approximative $\hat{f}$ tel que $y_i = \hat{f}(x_i)$. Nous appelons ce système artificiel un algorithme d'apprentissage. L'algorithme d'apprentissage a la capacité d'ajuster la fonction $\hat{f}$ en fonction de l'erreur $e_i = y_i - \hat{f}(x_i)$ entre les données réelles observées et les données générées par l'algorithme. L'erreur e_i est appelée erreur d'apprentissage ou résidu. Dans l'apprentissage supervisé, nous distinguons deux familles de modélisations, une modélisation d'un problème de classification et une modélisation d'un problème de régression. Dans le cas de la classification, la variable cible Y est discrète, comme dans la classification binaire où $Y \in \{0, 1\}$. Dans le cas de la régression, la variable sortie Y est continue $Y \in \mathbb{R}$, comme notre cas d'étude où Y est la puissance produite par un producteur PV. L'apprentissage machine supervisé est un processus d'apprentissage par l'exemple, c'est-à-dire qu'une fois le modèle créé, nous espérons que les paires d'entrées-sorties produites lors de sa mise en pratique seront proches des données utilisées dans la phase d'apprentissage {Hastie *et al.*, 2009}.

Dans cette étude nous avons utilisé un algorithme statistique Arima et deux algorithmes d'apprentissage machine Random forest (RF) et Gradient boosting machine (GBM). Un exposé théorique des ces algorithmes est présenté dans la partie annexe A.1.

5.3.2 État de l'art sur la prévision de la production solaire

L'histoire de la prévision du rayonnement solaire remonte à la fin du XIX^e siècle avec les avancés dans les approches de prévision météorologique. L'arrivée des ordinateurs et des calculateurs centraux a favorisé le développement de cette branche de prévision et il a réduit les temps de calcul des différentes simulations. Aujourd'hui, la prévision de la production solaire est traitée avec une attention sans précédent de la part des différentes communautés scientifiques dans le domaine de la prévision de l'énergie. Ceci est expliqué par l'importance de la gestion de l'intermittence de cette énergie pour réussir son intégration dans le réseau électrique. Cette intégration constitue un défi majeur pour la transition énergétique du réseau en déployant une production majoritairement renouvelable {Yang *et al.*, 2018}.

Pour donner un ordre de grandeur dans la littérature de la prévision de la production solaire, nous

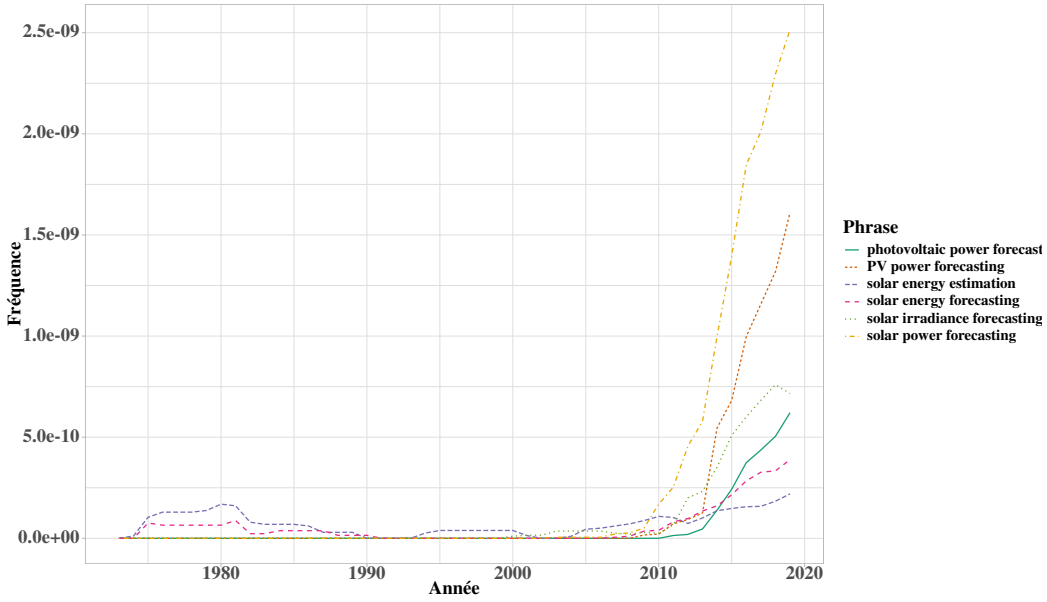


FIGURE 5.1 – Exemple de fréquence des mots dans des livres.
Source : Google Books Ngram Viewer
Lecture : l'axe des abscisses représente les années depuis 1975, et l'axe des ordonnées représente les fréquences d'apparition des phrases dans des livres calculées par l'outil Google Books Ngram.

avons utilisé l'outil Google Books Ngram Viewer[2] permettant de quantifier la fréquence d'apparition de phrases dans des livres sur une période donnée. Nous avons remarqué que les chercheurs se sont intéressés à partir de 1975 à la prévision de la production solaire. Ensuite à partir de l'année 2009 (voir figure 5.1), plusieurs livres sont apparus sur la prévision de cette production. Le livre de Kleissl {2013} a présenté un état de l'art complet sur les différentes approches de prévision. En outre, une recherche sur Google Scholar par exemple de la phrase « PV power forecasting » renvoie plus de 100 000 résultats de recherche et la recherche de la phrase « solar energy estimation » renvoie plus d'un million de résultats.

Dans cette littérature riche et volumineuse, plusieurs articles de synthèse ont été rédigés pour résumer une partie de cette littérature. Yang *et al.* {2018} ont présenté un article de synthèse en utilisant les techniques de fouille de données textuelles (text miniming) sur les données de 249 articles issues du site de ScienceDirect. Ils ont utilisé ainsi les métadonnées de 1000 articles renvoyés par Google Scholar pour extraire les données sur les auteurs, les revues, les conférences, le nombre de citations, etc. Leur étude a pour objectif l'identification des dernières avancées et tendances de recherche dans le domaine de la prévision solaire.

Wan *et al.* {2015} ont présenté un examen approfondi des méthodes de prévision de l'énergie solaire et ses applications dans le domaine des réseaux électriques. Ils ont classé ces méthodes en quatre familles d'approches, une approche statistique, une approche d'intelligence artificielle (IA), une approche physique et une approche hybride. Les approches statistiques sont fondées sur les techniques statistiques dans le domaine de l'analyse de séries temporelles en utilisant les données historiques, comme la méthode de moyenne mobile autorégressive intégrée (ARIMA). Les approches d'IA utilisent les dernières avancées dans le domaine d'apprentissage machine telles que les réseaux de neurones artificiels (artificial neural networks ANNs), les forêt d'arbres décisionnels (Random Forest RF), les machines à vecteurs de support (support vector machine SVM), etc. Cette catégorie de méthodes peut être classée dans la famille des approches statistiques. Les méthodes physiques font partie du domaine de prévision numérique du

2. La compilation est soumise à une licence Creative Commons Attribution 3.0 Unported.

temps (numerical weather prediction NWP) en utilisant les données des prévisions météorologiques et des images satellites. Les approches hybrides combinent des approches des trois familles de méthodes. Le choix d'une catégorie de prévision résulte de plusieurs facteurs comme les données disponibles, l'horizon temporel de prévision, le domaine d'utilisation de la prévision, etc. Les auteurs ont présenté ainsi des applications de la prévision solaire dans le domaine de gestion des réseaux électriques intelligents. Ces applications sont en fonction de l'horizon temporel de prévision. La prévision très court terme (d'une seconde à une minute) et la prévision court terme (jusqu'à 48-72 heures) sont utilisées particulièrement dans la gestion des activités d'exploitation des centrales PV, la régulation automatique de la production, le contrôle du stockage, le marché de l'énergie, etc. Les prévisions moyen et long termes sont utilisées essentiellement dans la gestion des travaux et des maintenances du réseau et l'évaluation et la planification des installations des parcs PV dans le réseau.

Antonanzas *et al.* {2016} ont présenté un examen approfondi et complet des dernières avancées de la recherche sur la prévision solaire. Ils ont étudié plusieurs articles pour rassembler une grande partie des connaissances sur la prévision de la production PV et dégager les dernières tendances dans le domaine. Ces travaux se distinguent par l'horizon temporel allant d'une seconde à des semaines, ainsi que des horizons spatiaux d'un site unique à une région. Premièrement, ils ont présenté le but et les différentes motivations derrière la prévision solaire. Deuxièmement, ils ont présenté un résumé des différentes méthodes et techniques utilisées dans ces travaux. Troisièmement, ils ont présenté les différences entre les prévisions ponctuelles et probabilistes et les avantages de chaque catégorie de prévision. Ensuite, ils ont classifié et résumé les différents articles selon les horizons temporels et les données d'entrée. Finalement, les métriques utilisées dans la mesure de performance de prévision sont présentées. Les auteurs ont conclu dans ce travail plusieurs synthèses. D'une part, plus l'horizon temporel est éloigné, plus la part des travaux utilisant les modèles NWP augmentent. Les approches statistiques et IA sont non seulement utilisées dans la plupart des travaux, mais elles ont montré de bonnes performances en qualité de prévision par rapport aux méthodes physiques. La plupart des derniers articles utilisent les méthodes d'apprentissage machine, car elles permettent une facilité de modélisation sans connaître les différentes caractéristiques des parcs PV, comme l'orientation, l'inclinaison, l'ombrage, etc. D'autre part, la prévision d'une courbe agrégée sur un territoire grand est plus précise qu'une prévision pour chaque producteur dans un site donné, ceci est expliqué par l'effet de lissage qui annule certains bruits dans les courbes individuelles des producteurs. Finalement, la plupart des anciens travaux traitent la prévision ponctuelle, cependant les articles récents introduisent des méthodes de prévision probabilistes. La prévision probabiliste permet une meilleure gestion des risques liés à variabilité de la production et donc permet une meilleure prise de décision. En comparaison, avec la prévision de la charge ou la prévision de la production éolienne, la prévision probabiliste dans le cas de la production solaire est encore inachevée, plusieurs défis doivent donc encore être résolus (voir figure 5.2).

van der Meer *et al.* {2018} ont présenté une synthèse des travaux récents dans le domaine de la prévision probabiliste de l'énergie solaire et de la charge. L'objectif de cette prévision probabiliste est de fournir une distribution prédictive complète d'un état futur ou de prévoir que la quantité prévue tombera dans un intervalle de confiance prévu. Ils ont analysé plusieurs travaux dans ce domaine et les performances de chaque méthode utilisée. Ils ont conclu qu'il n'existe pas un modèle unique applicable dans tous les cas d'usages. Un modèle performant dans un jeu de données n'est pas nécessairement performant dans d'autres jeux de données. Le but de leur étude était de trouver un lien entre la prévision probabiliste de la production solaire et la prévision de la charge afin de prévoir la charge réelle nette consommée. Ils ont montré que ces deux types de prévision se rejoignent en termes de résolution spatiale et temporelle

mais que la similarité en termes de variabilité entre la production et la charge peut différer d'une manière significative.

Hong *et al.* {2016} ont organisé une compétition de prévision de l'énergie intitulée « the Global Energy Forecasting Competition 2014 » (GEFCom2014). Dans cette compétition, les participants ont développé des approches de prévision probabiliste de la charge, du prix de l'énergie, de l'énergie éolienne et de l'énergie solaire. Cette compétition a attiré 581 participants de 61 pays différents. L'avantage de ce type de travaux est la normalisation de la métrique et des données utilisées. Les participants ont travaillé sur les mêmes jeux de données et leurs modèles ont été évalués par la même métrique. Dans la partie de la prévision solaire, les organisateurs ont utilisé les données de trois stations PV en Australie avec une granularité d'une heure et les données de prévision de 12 variables météorologiques. L'objectif de cette compétition était de fournir une prévision de la production solaire de ces trois parcs PV sur un horizon de 24 *h*. Les prévisions devaient être exprimées sous forme de 99 quantiles. Les cinq premières équipes gagnantes ont utilisé des algorithmes issus de la catégorie statistique et apprentissage machine comme Gradient Boosting (GB), k-Nearest Neighbour (k-NN), Quantile Regression Forest (QRF), Multiple Quantile Regression (MQR), RF et SVM. L'équipe gagnante a utilisé GB et k-NN en exploitant toutes les informations fournies.

Russo *et al.* {2017} ont exposé la problématique de prévision de l'énergie solaire et un état de l'art sur les méthodes de prévision. Ensuite, ils ont présenté leur approche de prévision court terme d'un horizon d'une heure avec un pas de quinze minutes en utilisant plusieurs données d'entrée. Ils ont exploité un logiciel de modélisation de données appelé « The brain Project » utilisant un algorithme génétique. L'approche de prévision est évaluée sur les données d'un parc PV d'une capacité de 1.05 kW situé à Catane en Italie. Ils ont utilisé plusieurs données d'entrée, mais aucune donnée sur le mouvement des nuages n'a été utilisée. Les résultats ont montré que les algorithmes avec deux entrés seulement performent mieux que d'autres algorithmes naïfs.

Rana *et al.* {2016} ont utilisé les méthodes d'apprentissage automatique pour une prévision court terme de la production de l'énergie PV en exploitant les courbes de production et les données météorologiques. Ils ont appliqué les algorithmes ANN et SVM avec deux méthodologies d'apprentissages différentes. Une approche n'utilise que les données de la production (données endogènes). Cette approche est appelée approche univariée. La deuxième approche est une approche multivariée qui utilise les données endogènes et des données externes météorologiques (données exogènes). Les résultats ont montré que l'approche univariée est plus précise en qualité de prévision par rapport à l'approche multivariée, avec une erreur relative moyenne (MAPE) comprise entre 4.15% et 9.34%. Ils ont montré que l'énergie solaire pour des horizons à court terme peut être prévue avec précision en n'utilisant que les données endogènes de l'historique de la production.

Li *et al.* {2016} ont appliqué deux algorithmes d'apprentissage machine (ANN et SVR) pour prévoir l'énergie d'un parc solaire de 6 MW en Floride aux États-Unis sur plusieurs horizons temporels (15 min, 1 h et 24 h).Ils ont évalué leur approche en utilisant les données endogènes et exogènes (météorologiques). Les données endogènes sont constituées des historiques de mesures de 11 onduleurs du parc PV. Ils ont développé une approche de prévision hiérarchique en utilisant les données de ces onduleurs. C'est-à-dire qu'ils ont créé un modèle de prévision pour chaque onduleur. Par la suite, les prévisions des onduleurs sont agrégées au niveau du parc PV. Ils ont montré que les prévisions des onduleurs améliorent la qualité de prévision du parc PV par rapport aux approches classiques.

Golestaneh *et al.* {2019} ont présenté une méthodologie de prévision probabiliste de la production

solaire. Cette approche permet une quantification des incertitudes sur la prévision sous forme d'un intervalle de confiance. Leur méthodologie est applicable aussi dans le cas où les données des images satellites ou le mouvement des nuages ne sont pas disponibles. Ils ont trouvé que l'erreur de prévision ne suit aucune des densités paramétriques comme la gaussienne. Pour cette raison, ils ont généré les densités prédictives en utilisant les méthodes non-paramétriques. Ils ont utilisé Extreme learning machine (ELM), qui a fourni des prévisions ponctuelles et probabilistes de l'énergie solaire allant de quelques minutes à une heure. Ils ont évalué les différentes approches de prévision sur plusieurs données de production dans des régions avec des climats différents.

Bessa *et al.* {2015} ont développé un algorithme de prévision probabiliste de la production PV sur un horizon de six heures. Ils ont utilisé plusieurs courbes de production PV recueillies dans un réseau de distribution de l'électricité dans la ville d'Évora en Portugal. Leurs résultats ont montré que la combinaison de toutes ces données dans un cadre commun de prévision peut améliorer la précision (entre 8% et 12%) de la prévision par rapport à un modèle univarié. Ils ont trouvé que l'utilisation des données distribuées dans le réseau améliore la prévision ponctuelle et probabiliste au niveau des postes HTA/BT (postes de distribution public). L'approche multivariée spatio-temporelle exploite donc pleinement toutes les informations collectées par les compteurs communicants dans le réseau.

Hossain *et al.* {2017} ont développé une approche de prévision de la production solaire sur un horizon d'une heure et d'un jour en utilisant l'algorithme ELM. Ils ont évalué leur modèle en utilisant des données endogènes sur l'historique de trois parcs PV en Malaisie et des données exogènes météorologiques. Leurs résultats ont montré, en comparant avec d'autres algorithmes (SVR et ANN), que l'algorithme ELM fournit une meilleure qualité de prévision dans un temps de calcul réduit.

Lonij *et al.* {2013} ont élaboré une approche de prévision solaire sous un ciel nuageux pour 80 producteurs PV en toiture répartis sur une zone de 50 $km \times 50$ km à Tucson. Ils ont utilisé les données de plusieurs capteurs installés dans la zone d'étude avec une granularité de 15 minutes. Leur approche de prévision utilisant les corrélations entre les producteurs PV était plus précise en prévision par rapport au modèle naïf et les approches utilisant les données des images satellites.

Zamo *et al.* {2014a} et Zamo *et al.* {2014b} ont élaboré un travail complet sur la prévision de la production solaire en utilisant les données météorologiques et des modèles NWP. Les auteurs ont appliqué plusieurs algorithmes d'apprentissage statistique pour développer une prévision sur un horizon d'une heure dans le premier article et sur un horizon de deux jours dans le deuxième article. Dans le premier travail, ils ont élaboré une approche de prévision ponctuelle en n'utilisant que les données endogènes des parcs PV et les données exogènes de la prévision météorologique fournie par Météo France (modèle ARPEGE). Les résultats ont montré que l'algorithme RF était le plus performant en qualité de prévision par rapport aux autres algorithmes. Dans le deuxième article, ils ont développé une approche de prévision probabiliste sur un horizon de deux jours. Le but de ce travail est de fournir une distribution de probabilité de la production prévue présentée par des quantiles. Ils ont montré que les prévisions sont toujours plus précises en comparant avec les prévisions météorologiques. Cependant, aucun modèle de prévision n'a dominé les autres modèles sur les données utilisées.

Vaz *et al.* {2016} ont développé une approche de prévision jusqu'à un mois en utilisant l'algorithme NARX(Nonlinear Autoregressive with eXogenous inputs). Ils ont évalué leur méthodologie sur les données météorologiques et les données historiques des parcs PV voisins. Ils ont montré que l'utilisation des données historiques des parcs voisins améliore la qualité de prévision pour les deux saisons hiver et été. Leurs résultats ont montré que l'algorithme NARX était plus performant par rapport au modèle naïf pour

les horizons de prévision supérieurs à 15 minutes.

Lin et Pai {2016} ont développé un modèle de prévision mensuelle de la production solaire pour un ensemble de producteurs basés à Taiwan en n'utilisant que les données historiques de production. Les auteurs ont expérimenté des techniques statistiques utilisables dans le domaine de séries temporelles. Ils ont appliqué au début la technique de la décomposition saisonnière (SD) pour supprimer l'effet de la saisonnalité dans la courbe de production. Ensuite, ils ont élaboré la prévision en utilisant l'algorithme Least-Square Support Vector Regression (LS-SVR) en optimisant les paramètres par l'algorithme génétique. Les résultats ont montré que cet algorithme est le plus précis en prévision avec une MAPE de 7.84% en comparant avec les autres algorithmes comme ARIMA, SARIMA, ANN et LS-SVR sans SD. Ils ont montré l'importance d'utilisation de SD dans les données avant de développer la prévision. Au final, ils ont recommandé l'élaboration d'une prévision pour chaque saison (été et hiver).

Résumé

Nous récapitulons dans le tableau 5.2 les travaux présentés précédemment par ordre de l'horizon temporel de prévision. Nous avons constaté après notre recherche bibliographique sur la prévision de la production solaire que ce domaine est encore immature en comparant avec la prévision de la charge et de la production éolienne (voir figure 5.2). Ceci est expliqué par l'absence d'une base de données commune de comparaison entre les différentes approches et le manque d'une métrique de précision de prévision identique permettant une mesure des qualités des modèles. Aujourd'hui, dans ces travaux le seul moyen de comparaison est l'utilisation d'un modèle naïf de prévision permettant une quantification de gain de prévision des différentes approches développées. Cependant, les données utilisées dans l'évaluation sont de nature différente, puisque les producteurs se situent dans des sites avec des climats différents et ils ont des capacités installées éloignées (de 1 kW jusqu'à 6 MW). Nous ajoutons à cela l'absence d'informations sur la configuration et les caractéristiques des parcs solaires comme l'orientation, inclinaisons, ombrage, marque, etc. La normalisation entre les travaux de recherche permet d'identifier et de sélectionner les meilleures approches de prévision utilisées dans l'état de l'art. La compétition GEFCom2014 {Hong *et al.*, 2016} est un bon exemple des avantages de cette normalisation dans la comparaison entre les approches et les travaux de recherche. En outre, nous notons que dans ces travaux, nous n'avons trouvé aucun article traitant un grand nombre de producteurs raccordés dans un réseau de distribution de l'électricité. Le maximum de nombre de producteurs que nous avons trouvé est 80 dans les travaux de {Lonij *et al.*, 2013} suivi par les travaux de {Bessa *et al.*, 2015} qui ont utilisé 44 producteurs raccordés au réseau HTA et BT.

Finalement, nous notons que nous n'avons pas trouvé beaucoup de travaux sur la prévision long terme d'un mois et plus. Ceci est expliqué par l'absence des prévisions météorologiques pour ces horizons lointains. En outre, la plupart des travaux sur la prévision de la production solaire sont focalisés sur des horizons très court terme et court terme, car les marchés de l'énergie en général sont plus actifs sur ces horizons proches.

Prévision probabiliste

Nous avons vu précédemment dans la sous-section 5.3.1 qu'il existe deux catégories de prévision en général. D'une part, nous avons une prévision ponctuelle produisant une seule valeur pour chaque pas de prévision sur l'horizon temporel. D'autre part, nous avons une prévision probabiliste prévoyant un intervalle ou une distribution de valeurs. Les avantages de l'approche probabiliste pour le GRD sont

Article	Année	Horizon temporel	Pas de prévision	Méthodes	Résultats	Données	Nombre de producteurs	Capacité installée	Pays
{Russo *et al.*, 2017}	2014	1 h	15 min	Genetic Programing, ANN	ANN avec deux entrées	Données production, variables calendaires, pas de données sur les nuages.	Un producteur	1.05 kW	Catane, Italie
{Rana *et al.*, 2016}	2016	1 h	5 min	ANN, SVM	ANN plus performent que SVR pour les horizons grands	Endogènes, rayonnment solaire , température, humidité, vitesse de vent	Un producteur	1.22 MW	Brisbane, Australie
{Li *et al.*, 2016}	2016	15 min, 1 h et 24 h	15 min	ANN, SVR, Approche hiérarchique	Utilisation des données des onduleurs améliore la prévision	Endogènes	Un producteur avec 11 onduleurs	6 MW	Floride, États-Unis
{Golestaneh *et al.*, 2019}	2016	1 h	10 min	ELM, QR	ELM probabiliste	Endogènes, données météo, pas de données sur les nuages	Deux producteurs	1 kW et 433 kW	Deux sites avec climat différent Singapore et Australie
{Bessa *et al.*, 2015}	2015	1-6 h	1 h	VAR et VARX. Modèles univariés et mutlivariés	VARX utilisation des données météo améliore la prévision	Endogènes, Informations recueillies dans le réseau, données météo	44 Producteurs répartis sur 10 postes HTA/BT	-	Évora, Portugal
{Hossain *et al.*, 2017}	2017	1 h et 1 jour	1 h et 1 jour	ELM, SVR, ANN	ELM avec un temps de calcul réduit	Endogènes et Exogènes	Trois producteurs	2, 1.8 et 2.7 kW	Malaisie
{Lonij *et al.*, 2013}	2013	6 h	15 min	Modèles naïve, Approche fondée sur les corrélations entre les sites	L'approche proposée est plus performent que l'approche naïve	Endogènes	80 producteurs distribués sur les toits des maisons	-	Tucson, États-Unis
{Zamo *et al.*, 2014a}	2014	28 45 h	1 h	Modèle naïf, LM, tree, Bagging, Boostin, RF, SVM, GAM	RF	Endogènes et exgogènes NWP, Modèle ARPEGE	28 producteurs dans deux sites différents	-	Deux régions en France
{Zamo *et al.*, 2014b}	2014	66 72 h	1 h	QR, QRF, LMQR	QR	Endogènes et exgogènes NWP, Modèle PEARP	28 producteurs dans deux sites différents	-	Deux régions en France
{Vaz *et al.*, 2016}	2016	15 min - 1 mois	15 min	NARX, Modèle naive	NARX avec les données des producteurs voisins	Endogènes et exogènes météo	5 producteurs	-	Utrecht, Pays-Bas
{Lin et Pai, 2016}	2016	1 mois	1 mois	ESDLS-SVR LS-SVR, GA, ARIMA, SARIMA GRNN	LS-SVR	Endogènes	16 producteurs	-	Taiwan

TABLEAU 5.2 – Récapitulatif des travaux de la prévision de la production solaire

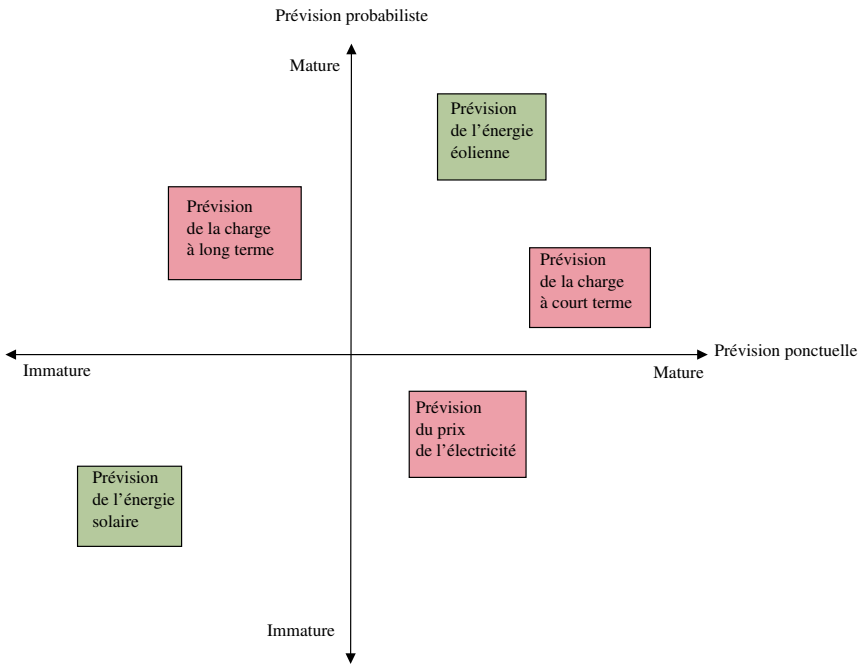


FIGURE 5.2 – Maturité des travaux dans le domaine de prévision de l'énergie.
Source : Hong *et al.* {2016}

multiples. Elle permet par exemple la quantification des limites supérieures ou inférieures des prévisions possibles, l'élaboration des scénarios de prévision de la production en exploitant les distributions prévues ou le calcul des probabilités de confiance pour chaque valeur prévue. Cela permet au GRD d'anticiper les contraintes arrivant sur le réseau ou d'optimiser et de planifier les différentes ressources de gestion de réseau. La figure 5.3 présente des exemples de type de prévision. Pour une prévision probabiliste, nous avons une prévision sous forme d'intervalle avec une borne supérieure et inférieure ou sous forme d'une distribution de valeurs. Dans notre cas de prévision long terme, nous avons développé une approche de prévision probabiliste d'une distribution de valeurs sur un mois.

En général, nous définissons une prévision probabiliste par une estimation d'une distribution statistique dans un état futur. Cette prévision probabiliste peut être présentée par une fonction de densité de probabilité (Probability density function PDF) ou par une fonction de répartition (Cumulative distribution function CDF). L'avantage de la représentation par la fonction CDF est que nous pouvons la résumer par un ensemble de quantiles. Nous pouvons donc prévoir une fonction CDF en prévoyant un ensemble de quantiles.

Nous rappelons que pour une variable aléatoire Y, la fonction de répartition F s'écrit de la façon suivante

$$F_Y(y) = \int_{-\infty}^{y} f(t)\mathrm{d}t = P(Y \leq y) \tag{5.5}$$

où f est la fonction de densité de probabilité de Y.

Le quantile d'ordre τ est par définition égale à $q_\tau = inf(y : F_Y(Y \geq \tau))$. Dans le cas où F_Y est continue et strictement croissante, nous avons $P(Y < q_\tau) = \tau$ avec τ dans $[0,1]$. Le quantile q_τ nous indique la probabilité τ à partir de laquelle un évènement y se situe en dessous de ce quantile q_τ. La figure 5.4 illustre un exemple d'une fonction CDF, pour un τ égal à 0.5, le quantile $q_{0.5}$ (la médiane) égale à 60 kW.

Un intervalle contenant Y avec une probabilité p peut se définir par un quantile inférieur q_{τ_1} et un

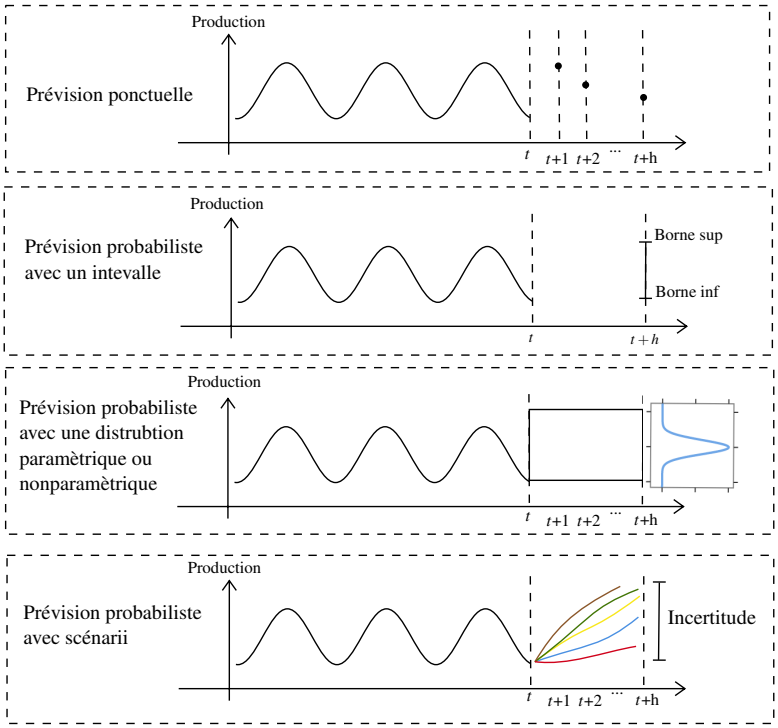


FIGURE 5.3 – Exemple de type de prévision (adapté de {Ordiano, 2019} et {Rana *et al.*, 2015}

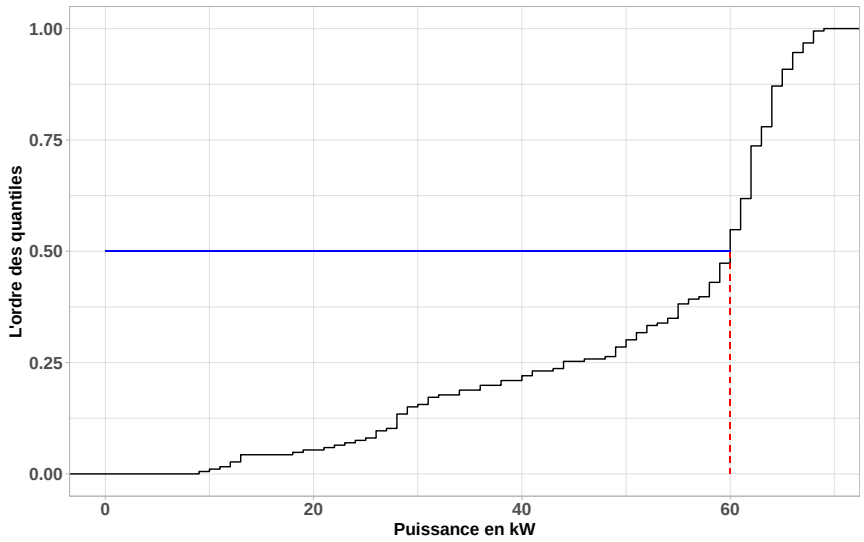


FIGURE 5.4 – Exemple d'une fonction CDF

quantile supérieur q_{τ_2}.

$$P(q_{\tau_1} \leq Y \leq q_{\tau_2}) = q_{\tau_2} - q_{\tau_1} \tag{5.6}$$

avec $q_{\tau_1} < q_{\tau_2}$

L'équation (5.6) nous permet de calculer la borne supérieure et la borne inférieure de l'intervalle prévu dans le cas d'une prévision probabiliste d'un intervalle.

Dans le cas où F est continue et strictement croissante, nous pouvons l'utiliser pour transformer un échantillon d'une variable aléatoire U uniformément distribuée sur $[0,1]$ en un échantillon de Y et inversement. Cette transformation est l'inverse de la fonction de répartition {Ordiano, 2019}

$$y = F^{-1}(u) \tag{5.7}$$

L'équation 5.6 nous assure la simulation des échantillons d'une distribution présentée par une fonction CDF.

Nous disons que deux variables aléatoires Y_1 et Y_2 sont dépendantes si Y_1, en prenant des valeurs spécifiques, influence les résultats possibles de Y_2. Cette dépendance s'exprime par la fonction de distribution conditionnelle.

$$F(y_2|y_1) = P(Y_2 \leq y_2|Y_1 = y_1) \tag{5.8}$$

Dans le cas de plusieurs variables aléatoires $Y_1, Y_2, ..., Y_m$, nous avons

$$F(y_m|y_1, y_2, ..., y_{m-1}) = P(Y_m \leq y_m|Y_1 = y_1, Y_2 = y_2, ..., Y_{m-1} = y_{m-1}) \tag{5.9}$$

En outre, la distribution de probabilité conjointe de plusieurs variables aléatoire $Y_1, Y_2, ..., Y_m$ s'écrit de la manière suivante

$$\begin{aligned} F(Y_1, Y_2, ..., Y_m) &= \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} ... \int_{-\infty}^{y_m} f(t_1, t_2, ..., t_m) \mathrm{d}t_1 \mathrm{d}t_2 ... \mathrm{d}t_m \\ &= P(Y_1 \leq y_1, Y_2 \leq y_2, ..., Y_m \leq y_m) \end{aligned} \tag{5.10}$$

Un problème connu dans le domaine des probabilités est lorsque nous souhaitons estimer la fonction de répartition conjointe de plusieurs variables aléatoires dépendantes. Par exemple, dans le cas où nous souhaitons étudier en même temps la consommation de l'électricité, la production éolienne et la production solaire, nous devons estimer la fonction CDF conjointe de ces trois variables. Une alternative est l'utilisation des fonctions copules. Une fonction copule est une fonction mathématique capable de capturer la structure de dépendance de plusieurs variables aléatoires en se fondant uniquement sur leurs fonctions de répartition marginales. L'utilisation de ces copules est fondée sur le théorème de Sklar {Ordiano, 2019}.

Le théorème de Sklar (1959) dit que la distribution conjointe F_Y d'un vecteur aléatoire $Y = (Y_1, Y_2, ..., Y_m)$ peut être écrite comme une fonction de ces distributions marginales F_{Y_i}

$$F_Y(y) = C_Y(F_{Y_1}(y_1), ..., F_{Y_m}(y_m)) \tag{5.11}$$

avec $C_Y : [0,1] \times [0,1] \times ... \times [0,1]$ est la distribution conjointe des variables aléatoires $U_j = F_{Y_j}(Y_j)$ pour $j = 1, 2, ..., m$. Cette transformation nous assure que les U_j ont toujours des distributions marginales uniformes. De plus, si les U_j sont continues, la fonction copule est unique ({Nelsen, 2007; Schoelzel et Friederichs, 2008}). Le théorème de Sklar nous assure donc l'existence d'une fonction définissant une structure de dépendance entre plusieurs variables aléatoires.

Le théorème de Sklar peut être formulé dans le cas des CDF conjointes conditionnelles {Patton, 2006}.

$$F_{Y|Y_{m+1}}(y_1, y_2, ..., y_m|y_{m+1}) = C_{Y|Y_{m+1}}(F_{Y|Y_{m+1}}(y_1|y_{m+1}), ..., F_{Y|Y_{m+1}}(y_m|y_{m+1})|y_{m+1}) \tag{5.12}$$

Dans le domaine de la prévision météorologique en général, il existe trois manières pour définir la fonction CDF : par une fonction *CDF* paramétrique, par une fonction *CDF* non paramétrique ou par les prévisions d'ensemble. Dans le cas de la production solaire, il a été démontré dans la littérature qu'il est difficile de définir la CDF par une seule loi de probabilité. Dans les travaux de Golestaneh *et al.* {2019}, par exemple, ils ont trouvé que la distribution ne suit aucune densité paramétrique comme la gaussienne. La fonction CDF non paramétrique est une estimation discrète d'une fonction CDF. Elle permet de définir une CDF prédictive sans aucun axiome sur l'évènement futur. Dans ce cas, la prévision de la CDF est fournie par l'ensemble des quantiles prévus. Ce type de prévision est appelé la prévision par quantiles {Pinson *et al.*, 2007}. Il a été largement utilisé par exemple dans la compétition GEFCom2014. Dans la prévision par quantiles plusieurs méthodes statistiques et d'apprentissage machine ont été utilisées dans la littérature comme la régression quantile QR, GBM, ANN pour estimer les distributions prédictives. La troisième famille de méthodes est la méthode de prévision d'ensemble ou la prévision par scénario. Cette approche est largement utilisée dans la prévision probabiliste météorologique. Généralement cette prévision est générée par des modèles physiques de prévision (NWP). Dans le cas par exemple d'un modèle physique, la prévision par scénario est donnée par un ensemble de prévisions calculées en modifiant les conditions initiales des équations de modélisation du phénomène météorologique. Ce système de scénarios représente les différentes incertitudes sur l'événement futur {Lauret *et al.*, 2019}.

La méthode de prévision probabiliste de production solaire la plus utilisée dans la littérature est la régression quantile QR. Cette méthode permet d'étudier la variation des quantiles d'une distribution conditionnelle $F_{Y|X}$ en fonction des variables explicatives X {Givord et d'Haultfoeuille, 2013}. Nous supposons que les quantiles de $F_{Y|X}$ ont une forme linéaire

$$q_\tau(Y|X) = X^{'}\beta_\tau \tag{5.13}$$

Avec τ dans $[0,1]$ et $\beta_\tau = (\beta_{1,\tau}, \beta_{2,\tau}, ..., \beta_{m,\tau})$ un vecteur de $\mathbb{R}^m$ correspondant aux m variables explicatives.

L'expression (5.13) peut s'écrire de la manière équivalente {Givord et d'Haultfoeuille, 2013}

$$Y = X^{'}\beta_\tau + \varepsilon_\tau \tag{5.14}$$

Avec $q_\tau(\varepsilon_\tau|X) = 0$

Approche locale et globale

L'approche globale de prévision a été introduite par Januschowski *et al.* {2020}. Cette approche intervient dans la prévision de plusieurs séries temporelles de nature similaire, par exemple la prévision de la production PV de plusieurs producteurs ou la prévision de la charge au niveau de plusieurs départs HTA. Nous avons deux modélisations possibles dans ce cas de prévision. La première consiste à développer un modèle indépendamment pour chaque série temporelle, cette approche est appelée une approche locale. La deuxième approche consiste à élaborer conjointement un seul modèle de toutes les séries temporelles. Cette approche est appelée une approche globale.

La distinction entre ces deux approches est la manière dont le modèle est entraîné, c'est-à-dire la façon d'estimer les paramètres. Nous ne cherchons pas avec la méthode globale à trouver une structure de dépendance particulière entre les séries temporelles. Cette structure est déterminée principalement par des approches de prévision multivariées. Il est donc important de distinguer entre une approche multivariée et une approche globale. Une approche multivariée peut être une approche locale en étudiant les corrélations avec d'autres données exogènes comme les variables météorologiques. Une approche globale peut être univariée avec seulement des données endogènes ou multivariée avec d'autres données exogènes. La plupart des méthodes statistiques classiques sont des méthodes locales, tandis que les méthodes d'apprentissage machine peuvent être utilisées dans une approche locale ou une approche globale. Récemment, les approches globales ont été utilisées dans plusieurs travaux de recherche comme les travaux de Wen *et al.* {2017} et Salinas *et al.* {2020}. Ils ont montré une performance de prévision par rapport à d'autres approches dans plusieurs compétitions de prévision comme la compétition M5 {Makridakis *et al.*, 2020}. Le modèle ligthGBM avec une approche globale a gagné la première place dans cette compétition parmi plus de 5000 participants. Bojer et Meldgaard {2020} ont présenté un article de synthèse des compétitions de prévision déroulées sur la plateforme Kaggle. Tous les concours présentés dans cet article ont été remportés par des modèles globaux. Cette méthodologie d'approche globale a été utilisée ainsi dans des prévisions probabilistes comme les travaux de Wen *et al.* {2017}.

Dans l'approche locale, nous développons un modèle pour chaque série temporelle $Y_{s,t}$ dans l'ensemble de séries disponibles $\{Y_{s,t}\}_{s=1}^{S}$ où S est le nombre total de séries.

$$Y_{s,t} = f_s(X_{s,t}) + \varepsilon_{s,t}$$

C'est-à-dire, nous devons estimer un n nombre de paramètres pour chaque série. Pour les S séries, il faut estimer donc $n \times S$ paramètres différents. Ce nombre de paramètres croit avec le nombre de séries à prévoir.

Dans l'approche globale nous construisons un seul modèle pour l'ensemble de séries temporelles disponibles.

$$Y_t = f(X_t) + \varepsilon_t$$

Nous n'estimons que n paramètres de modèle pour l'ensemble de séries. Ce nombre reste fixe si nous augmentons le nombre de séries. L'avantage de cette approche est donc un gain de temps et de ressources matériels dans l'élaboration de la prévision. En outre, l'utilisation de toutes les données dans

Données d'entrée approche locale				
Date	**PDL**	**Puissance** t	**Puissance** t**-1**	**...**
t_1	PDL_1	$p_{1,1}$	-	...
t_2	PDL_1	$p_{1,2}$	$p_{1,1}$	...
t_3	PDL_1	$p_{1,3}$	$p_{1,2}$	...
t_4	PDL_1	$p_{1,4}$	$p_{1,3}$	...
t_5	PDL_1	$p_{1,5}$	$p_{1,4}$	...
t_6	PDL_1	$p_{1,6}$	$p_{1,5}$	...
t_7	PDL_1	$p_{1,7}$	$p_{1,6}$	...
t_8	PDL_1	$p_{1,8}$	$p_{1,7}$	...
t_9	PDL_1	$p_{1,9}$	$p_{1,8}$	...
...	PDL_1	...	...	...
t_i	PDL_1	$p_{1,i}$	$p_{1,i-1}$	...
t_{i+1}	PDL_1	$p_{1,i+1}$	$p_{1,i}$	...
t_{i+2}	PDL_1	$p_{1,i+2}$	$p_{1,i+1}$	...
...	...	...	...	...

Données d'entrée approche globale							
Date	**PDL**	**Puissance** t	**Capacitée**	**Coordonnées GPS**	**Energie Annuelle**	**Puissance t-1**	**...**
t_1	PDL_1	$p_{1,1}$	C_1	(x_1, y_1)	E_1	-	...
t_1	PDL_2	$p_{2,1}$	C_2	(x_2, y_2)	E_2	-	...
t_1	PDL_3	$p_{3,1}$	C_3	(x_3, y_3)	E_3	-	...
t_2	PDL_1	$p_{1,2}$	C_1	(x_1, y_1)	E_1	$p_{1,1}$	...
t_2	PDL_2	$p_{2,2}$	C_2	(x_2, y_2)	E_2	$p_{2,1}$	...
t_2	PDL_3	$p_{3,2}$	C_3	(x_3, y_3)	E_3	$p_{3,1}$	...
t_3	PDL_1	$p_{1,3}$	C_1	(x_1, y_1)	E_1	$p_{1,2}$	...
t_3	PDL_2	$p_{2,3}$	C_2	(x_2, y_2)	E_2	$p_{2,2}$	...
t_3	PDL_3	$p_{3,3}$	C_3	(x_3, y_3)	E_3	$p_{3,2}$	...
...	...	...	...	...	...	...	...
t_i	PDL_1	$p_{1,i}$	C_1	(x_1, y_1)	E_1	$p_{1,i-1}$	...
t_i	PDL_2	$p_{2,i}$	C_2	(x_2, y_2)	E_2	$p_{2,i-1}$	...
t_i	PDL_3	$p_{3,i}$	C_3	(x_3, y_3)	E_3	$p_{3,i-1}$	...
...	...	...	...	...	...	...	...

FIGURE 5.5 – Exemple de données d'entrée dans les deux approches locale et globale

un seul modèle améliore la qualité de prédiction du modèle, surtout pour les approches d'apprentissage machine qui s'améliorent selon le volume de données d'entrées.

Dans notre cas de prévision de la production solaire de réseau de SRD, nous avons plus de 600 courbes de production à prévoir. Donc une approche de prévision globale est applicable dans notre cas de prévision court terme et long terme. En utilisant cette approche, nous développons un seul modèle pour l'ensemble de producteurs présents dans le réseau. Toutefois, avec une approche locale de prévision nous avons besoin d'élaborer plus de 600 modèles différents. La figure 5.5 illustre un exemple de données d'entrée dans les deux approches locale et globale. Dans l'approche globale, nous pouvons ajouter d'autres informations statiques sur les séries temporelles permettant une différenciation entre chaque série. Par exemple, dans notre de cas de prévision, nous pouvons ajouter toutes les informations sur les producteurs comme leur capacité de production, les coordonnées GPS, etc. L'utilité de ces variables sera discutée dans la partie modélisation 5.5.

5.4 Méthodologie de prévision

Le travail de prévisionniste consiste à développer un modèle décrivant au mieux le phénomène étudié en utilisant les données historiques mesurées. Une prévision de qualité est une prévision proche de la réalité observée. Dans notre travail de prévision, nous avons suivi une méthodologie composée de plusieurs étapes (voir figure 5.6), afin de sélectionner le modèle le plus précis.

La première étape consiste à définir le problème de prévision à étudier. Dans cette étape, nous modélisons notre besoin de prévision en comprenant le phénomène étudié, la manière dont la prévision sera utilisée et qui sera l'utilisateur final de cette prévision. Le premier travail est donc de discuter avec tous les services intéressés par la prévision et les services chargés de la collecte et du stockage des données. La définition du besoin de prévision est une étape souvent difficile, mais elle est très importante dans le

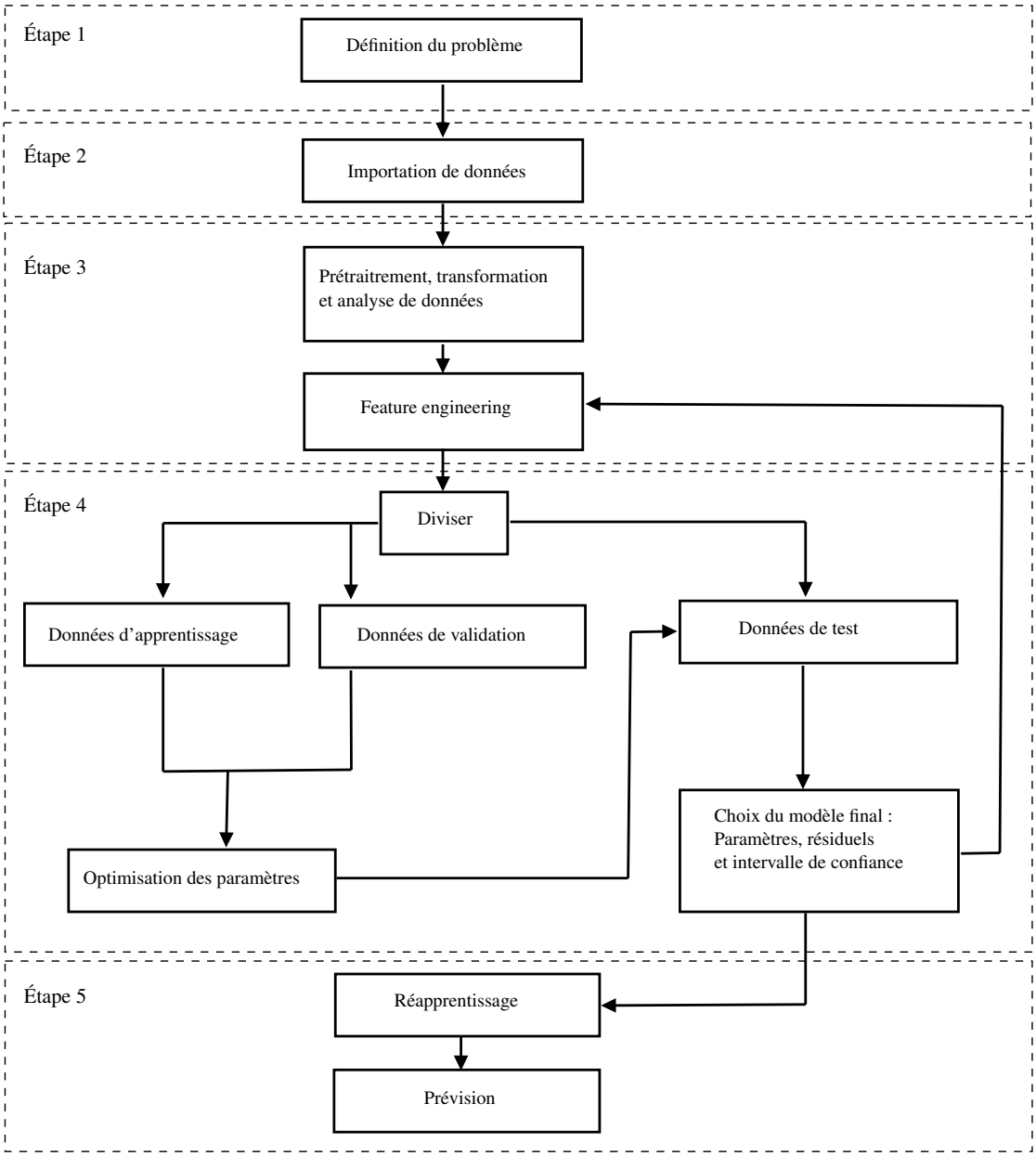


FIGURE 5.6 – Méthodologie suivie dans le développement des modèles de prévision

processus de prévision {Hyndman et Athanasopoulos, 2018}.

Après l'étape de compréhension du besoin de la prévision vient l'étape de collecte d'informations. Dans cette étape nous cherchons toutes les données nécessaires pour la prévision. Ces données sont toutes les informations statiques et dynamiques stockées dans les bases de données internes de l'entreprise ou des données externes disponibles dans l'open data. Nous ajoutons aussi toutes les informations métier des personnes travaillant directement ou indirectement dans la prévision {Hyndman et Athanasopoulos, 2018}.

Après la récupération de toutes les données, nous avons besoin de les préparer et de les prétraiter pour les rendre exploitables par les outils d'analyse de données. Cette étape de prétraitement et d'analyse est une étape exploratoire de données. Nous commençons toujours par tracer les différentes séries temporelles pour dégager visuellement les tendances, la saisonnalité, des changements du comportement et toute valeur aberrante ou extrême. Nous regardons ainsi les corrélations entre les différentes variables explicatives disponibles pour la prévision {Brockwell et Davis, 2016}. Nous notons que pour une prévision à un horizon temporel donné, il faut toujours vérifier que les variables explicatives seront aussi disponibles pour cet horizon.

Ensuite vient l'étape de développement de modèles. Dans cette étape, nous choisissons le type de famille de modèles à tester selon l'historique de données disponibles, les différentes corrélations entre les variables explicatives et la variable cible et la manière dont les prévisions doivent être utilisées {Hyndman et Athanasopoulos, 2018}. Dans cette étape, nous comparons plusieurs modèles de prévision de natures différentes. Comme dans l'état de l'art, la qualité de la prédiction est comparée à la qualité de prédiction obtenue par les modèles naïfs. Le développement d'un modèle consiste à estimer ses paramètres en fonction de données d'apprentissage. Généralement dans cette étape de sélection de modèles, nous utilisons une méthodologie de validation et d'évaluation pour choisir le modèle le plus performant en qualité de prévision. La méthodologie de validation et d'évaluation sera détaillée dans la partie 5.4.1.

Une fois que le modèle est sélectionné et que ses paramètres sont estimés, nous l'utilisons pour élaborer la prévision finale. Dans cette étape de prévision, nous devons construire notre outil final de prévision en vérifiant la disponibilité et la qualité de données d'entrée et en estimant la performance et la qualité de prévision en temps réel.

5.4.1 Méthodologie d'évaluation

L'élaboration d'une prévision de qualité nécessite une méthodologie de validation et d'évaluation des différents modèles développés afin de sélectionner le modèle le plus précis. Pour pouvoir comparer les différents modèles et notamment avec les méthodes naïves, nous avons besoin d'une stratégie d'évaluation objective permettant une quantification de la précision de prévision.

Dans notre étude, nous avons divisé l'ensemble de données observées en trois parties. Une première partie d'apprentissage est utilisée pour entraîner les différents modèles. Dans cette partie, nous ajustons le modèle sur les données d'apprentissage mesurées du phénomène étudié en estimant ses différents paramètres.

La partie de validation de modèle sert à contrôler l'apprentissage en évitant le sur-ajustement du modèle. Cette problématique de sur-apprentissage est connue dans le domaine de machine learning sous le nom d'« Overfiting ». C'est-à-dire que le modèle capture toutes les informations dans les données d'apprentissage, mais qu'il n'a pas la capacité de généraliser sur d'autres données. Avec la méthode de

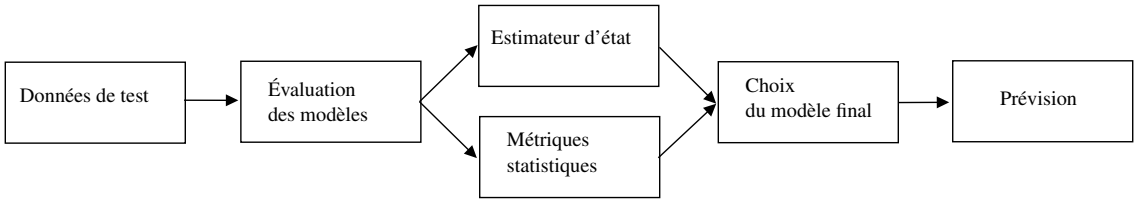


FIGURE 5.7 – Méthodologie d'évaluation et de sélection des modèles

validation, le modèle est entraîné sur la partie d'apprentissage et contrôlé sur la partie de validation. Ce contrôle consiste à estimer et à optimiser les hyper-paramètres en calculant les différentes erreurs de prévision en utilisant les données de validation. Une autre stratégie de validation est la méthode de validation croisée présentée dans le chapitre 4. Cependant dans le cas des séries temporelles, son utilisation est compliquée, car il faut diviser les parties selon une période donnée en respectant l'axe temporel.

Finalement, nous avons une partie de test servant à l'évaluation des différents modèles développées en utilisant plusieurs métriques et mesures de qualité en comparant avec la réalité observée. Nous notons l'importance de vérifier que les données explicatives du modèle seront ainsi disponibles dans l'horizon temporel de prévision. Ce problème est connu dans le domaine d'apprentissage machine sous le nom de « Leakage », il consiste à entraîner le modèle sur des données qui ne seraient pas disponibles au moment de la prévision. Par exemple, nous souhaitons réaliser une prévision de la production solaire sur un horizon mensuel et nous utilisons le rayonnement solaire comme variable explicative du modèle. Nous aurons un très bon modèle en qualité de prévision, mais lorsque nous testons le modèle dans la réalité nous n'aurons pas l'information de la prévision du rayonnement solaire sur un horizon d'un mois. Conséquemment, le modèle ne sera pas utilisable dans la réalité.

Le leakage désigne ainsi la fuite de données de la partie apprentissage vers la partie de test consciemment ou inconsciemment. Cette fuite passe généralement lorsque nous souhaitons remplacer les valeurs manquantes dans l'ensemble de données. Par exemple, si nous utilisons la méthode de $k-NN$ pour estimer les données manquantes avant la division de la partie apprentissage et de la partie test, nous aurons des informations qui vont passer vers la partie de test via l'estimateur $k-NN$. Cela implique une surestimation de la précision du modèle sur la partie test, car il connaît bien une partie de ces données.

Métriques de performance

Nous proposons dans ce projet de prévision une méthodologie d'évaluation des modèles (voir figure 5.7) en utilisant à la fois des métriques statistiques classiques et des simulations du réseau électrique via un calculateur d'état. Après l'apprentissage des différents modèles, nous souhaitons en effet quantifier la qualité de prévision de chaque modèle développé. Le calculateur d'état nous permet d'estimer la qualité de modèle dans un fonctionnement réel du réseau. Dans la partie évaluation statistique, nous avons utilisé les métriques déjà présentées dans la section 2.2.4 du chapitre 2, $nMBE^{+}$+, $nMBE^{-}$, nMBE, nMAE et nRMSE normalisées par la capacité installée des producteurs, ainsi que la métrique MAPE.

Nous avons utilisé le Skill Score présenté dans beaucoup de travaux de recherche permettant une quantification de la qualité d'un modèle par rapport au modèle naïf.

$$ss = 1 - \frac{nRMSE_{\text{modèle}}}{nRMSE_{\text{naïf}}}$$

L'évaluation par l'estimation du réseau consiste à réaliser la simulation du calculateur d'état présentée dans le chapitre 3. L'estimateur nous permet de calculer les pertes totales des lignes électriques, les élévations et les chutes de tension au niveau des postes HTA/BT. Nous avons élaboré une stratégie d'évaluation des modèles de prévision en utilisant ces différents résultats. Cette évaluation nous permet de quantifier la capacité d'un modèle à capturer certaines contraintes liées au dimensionnement du réseau.

Dans cette simulation, nous élaborons une prévision pour chaque modèle développé sur les données de test. Par la suite, nous obtenons une courbe de valeurs prévues $\{\hat{y}_t\}$ par le modèle. Cette courbe sera utilisée ensuite comme entrée dans l'estimateur d'état. Concernant les courbes de charge et les courbes de production des autres producteurs comme les éoliennes, nous avons utilisé les données réelles mesurées sur la période de test. Cela nous permet d'étudier et de comparer la prévision du modèle avec la réalité observée et avec les autres modèles développés. Les courbes de production des petits producteurs sont estimées par la méthode d'interpolation spatiale présentée dans le chapitre 4 en utilisant les courbes prévues des moyens et grands producteurs.

Après la simulation, nous obtenons plusieurs courbes décrivant l'état du réseau. D'une part, nous avons une courbe de pertes totales générées par lignes électriques dans le réseau pour chaque prévision. D'autre part, nous avons les courbes de tension estimées dans les postes HTA/BT. Nous calculons ensuite le maximum et le minimum de tension estimés dans un instant t. Au final, nous avons trois courbes différentes pour chaque prévision fournie. Une première estimation de qualité des modèles est réalisée par les métriques statistiques déjà utilisées sur les courbes de production réelles comme MBE, MAE, RMSE et MAPE. En outre, nous avons utilisé d'autres métriques décrivant le respect des contraintes de tension pour chaque modèle.

Nous commençons par transformer les deux courbes de maximum $v_{max,t}$ et de minimum $v_{min,t}$ de tension en deux courbes binaires $b_{max,t}$ et $b_{min,t}$. Ces courbes sont obtenues en vérifiant que les bornes min et max restent dans un intervalle de variations de tension. Cet intervalle de tension est pour un GRD une norme de qualité de fourniture à respecter pour tous les clients. Il se définit pour une tension V de réseau par

$$0.95 \times V_0 \leq V \leq 1.05 \times V_0$$

Avec la tension V_0 égale la tension normale de réseau, dans le réseau HTA par exemple $V_0 = 20\ kV$. Dans les résultats calculés par l'estimateur d'état la tension V_0 est présentée par 1 $v.u$ (voltage magnitude). C'est-à-dire pour une tension de 20 kV par exemple, la tension doit rester entre 19 kV (0.95 $p.u$) et 21 kV (1.05 $p.u$). Les courbes binaires des bornes maximales $v_{max,t}$ et minimales $v_{min,t}$ estimées dans le réseau s'écrivent donc

$$b_{max,t} = \begin{cases} 1 & \text{si } v_{max,t} \leq 1.05 \\ 0 & \text{si } v_{max,t} > 1.05 \end{cases} \tag{5.15}$$

$$b_{min,t} = \begin{cases} 1 & \text{si } v_{min,t} \geq 0.95 \\ 0 & \text{si } v_{min,t} < 0.95 \end{cases} \tag{5.16}$$

A partir de ces courbes, nous construisons une matrice appelée la matrice de confusion présentée dans la figure 5.8.

		Valeurs prévues : 1	Valeurs prévues : 0
Valeurs réelles	1	VP Vrais Positifs	FN Faux Négatifs
	0	FP Faux Positifs	VN Vrais Négatifs

FIGURE 5.8 – Matrice de confusion

Avec

$$VP = \sum_{t=1}^{T} \mathbb{1}\{b_{max,t} = 1 \wedge \hat{b}_{max,t} = 1\}$$

$$VN = \sum_{t=1}^{T} \mathbb{1}\{b_{max,t} = 0 \wedge \hat{b}_{max,t} = 0\}$$

$$FP = \sum_{t=1}^{T} \mathbb{1}\{b_{max,t} = 0 \wedge \hat{b}_{max,t} = 1\}$$

$$FN = \sum_{t=1}^{T} \mathbb{1}\{b_{max,t} = 1 \wedge \hat{b}_{max,t} = 0\}$$

À partir de cette matrice nous calculons deux métriques appelées sensibilité et spécificité. La sensibilité indique le taux de vrais positifs, c'est-à-dire la proportion des 1 qui sont correctement prévus par le modèle. La spécificité mesure la proportion des vrais négatifs 0 qui sont correctement prévus.

$$\text{Sensibilité} = \frac{VP}{VP+FN}$$

$$\text{Spécificité} = \frac{VN}{VN+FP}$$

Dans notre cas d'étude, la sensibilité indique la capacité de respect général du modèle des contraintes des variations de tensions, tandis que la spécificité mesure la capacité du modèle à prévoir les contraintes affectant le réseau dans le futur.

A partir de ces deux métriques, nous obtenons la métrique Prévision équilibrée (Balanced Accuracy BA) qui est égale à

$$BA = \frac{\text{Sensibilité} + \text{Spécificité}}{2}$$

La métrique BA nous indique la moyenne de la sensibilité et de la spécificité. Une BA égale 1 signifie que le modèle prévoit 100% des variations de tensions estimées dans la réalité en respectant et en ne respectant pas les contraintes.

La figure 5.9 illustre un exemple d'élévation de tension estimée par le calculateur d'état. Nous remarquons des élévations de tension dans certains nœuds du réseau dans une journée de production solaire. Le modèle 1 est le plus précis en prévision de cette contrainte comparé aux modèles 2 et 3 car il indique plus précisément à quel moment la configuration de réseau considérée deviendra invalide. Cette mesure n'est pas toujours directement liée à la précision (par exemple mesurée avec la RMSE) de la valeur prédite

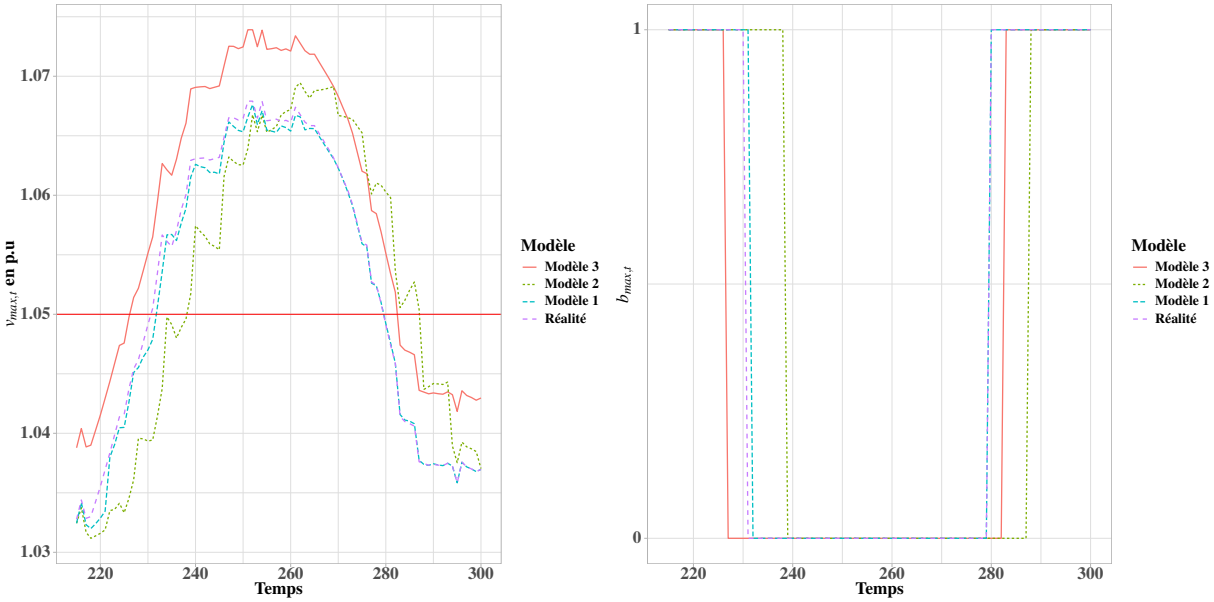


FIGURE 5.9 – Exemples de prévision d'élévation de tension.
Lecture : Dans la figure de gauche, l'axe des abscisses présente le temps et l'axe des ordonnées présente la tension maximale $v_{max,t}$ en p.u calculée par l'estimateur d'état. La droite horizontale représente la tension limite à respecter dans le réseau. Dans la figure de droite, l'axe des abscisses présente le temps, l'axe des ordonnées présente la courbe binaire $b_{max,t}$ calculée par l'équation (5.15).

même si, pour cet exemple, le modèle 1 est à la fois le plus précis et le plus représentatif du respect (ou non respect) des contraintes dimensionnantes du réseau. Les courbes binaires représentent le respect des modèles de la limite maximale (1.05) d'élévation de tension.

Prévision court terme

Dans le cas de la prévision court terme de 10 *min* à 60 *min*, nous avons développé une méthodologie d'évaluation des prévisions des modèles en divisant l'ensemble de données en trois parties, apprentissage, validation et test.

Nous avons étudié la qualité de prévision sur une année entière en la divisant en 12 mois. Pour chaque mois nous utilisons les deux derniers jours pour la validation et le test du modèle. Avec cette stratégie, nous testons la qualité de chaque modèle sur les différents mois de l'année. Sur le dernier jour d'un mois réservé au test du modèle, nous cherchons pour chaque heure à avoir une prévision sur un horizon d'une heure avec un pas de 10 *min*. Pour simplifier la démarche, nous avons choisi l'instant 0 de chaque heure pour réaliser une prévision sur un horizon de 60 *minutes*. L'erreur de prévision est donc la différence entre la prévision de modèle et les valeurs observées sur cette période horaire. La figure 5.10 illustre cette méthodologie d'évaluation. L'erreur statistique du modèle est la moyenne de toutes les erreurs horaires dans l'ensemble des périodes testées

$$\overline{err} = \sum_{m=1}^{M} \sum_{h=1}^{H} erreur_{h,i}$$

M égale le nombre total de mois et H le nombre total d'heures.

Dans le cas de l'estimateur d'état, nous simulons le réseau avec les courbes prévues par chaque modèle et les courbes réelles observées. La production des petits producteurs est estimée par l'interpolation spatiale en utilisant les courbes prévues. Ensuite, nous calculons les différences entre la prévision et la

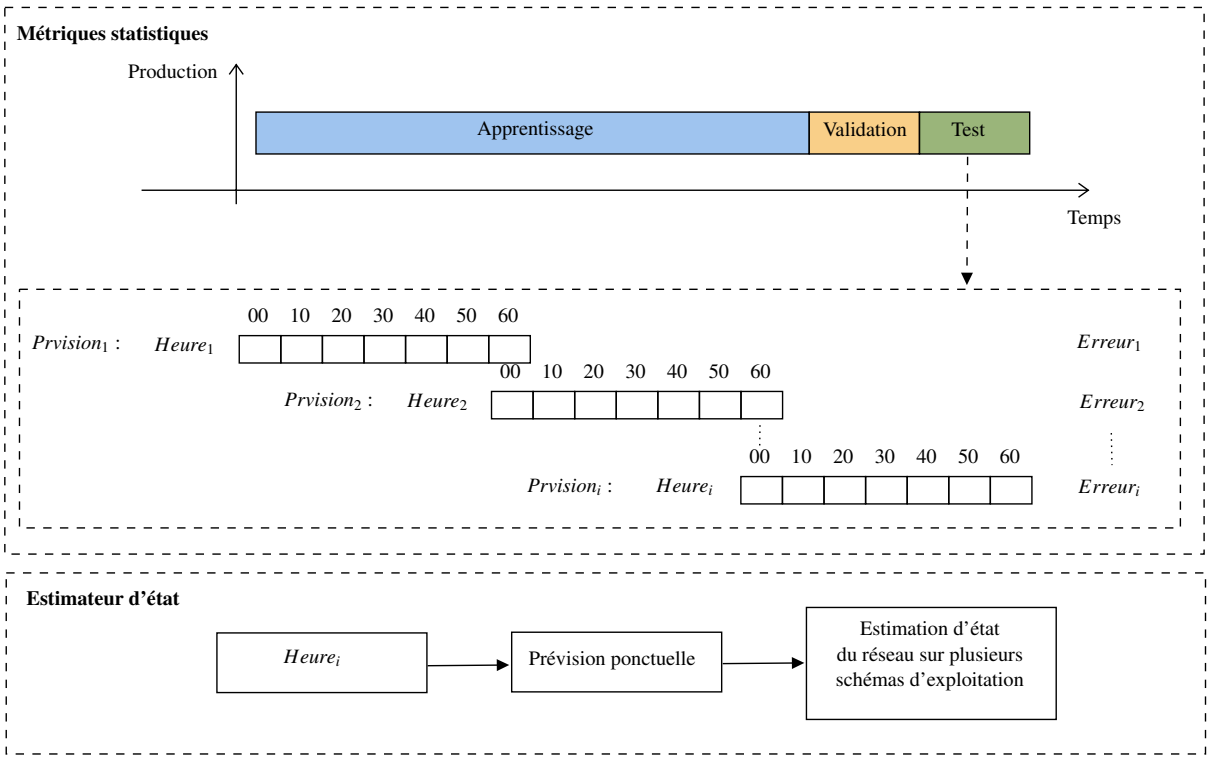


FIGURE 5.10 – Méthodologie d'évaluation dans le cas de la prévision court terme

réalité en termes de pertes et de variations de tension. Ce calcul nous permet à la fois d'identifier le modèle qui respecte le mieux le dimensionnement de réseau par rapport à la réalité observée et qui prévoit le mieux les contraintes affectant le réseau. Dans le cas des pertes, nous avons utilisé les métriques MBE, MAE, RMSE et MAPE entre les pertes estimées par le modèle et les pertes estimées par les données réelles. Les élévations et les chutes de tension sont étudiées par ces métriques plus les métriques binaires (Sensibilité, Spécificité et BA).

Prévision long terme

Dans ce cas de prévision, nous avons étudié la qualité des modèles sur une année entière en la divisant en quatre trimestres :

— Trimestre 1 noté Hiver 1 contenant les mois janvier, février et mars 2020,

— Trimestre 2 noté Été 1 contenant les mois avril, mai et juin 2020,

— Trimestre 3 noté Été 2 contenant les mois juillet, août et septembre 2020,

— Trimestre 4 noté Hiver 2 contenant les mois octobre, novembre et décembre 2020.

Nous cherchons dans chaque trimestre réservé au test du modèle à prévoir les distributions de chaque mois en prévoyant les quantiles. Par exemple, dans le cas du trimestre Hiver 1, nous utilisons les données de janvier 2012 à septembre 2019 dans l'apprentissage des modèles et les données de octobre, novembre et décembre pour la validation.

Les quantiles prévus seront utilisés pour définir la fonction CDF via la méthode d'interpolation linéaire. La figure 5.10 illustre la méthodologie d'évaluation suivie dans ce cas de prévision. L'erreur statistique égale la moyenne de toutes les erreurs calculées dans l'ensemble de périodes de test. Dans le cas de cette prévision, nous comparons la qualité du modèle par rapport aux quantiles réels calculés sur un mois. Nous notons qu'il faut calculer une erreur pour chaque mois, heure et quantile :

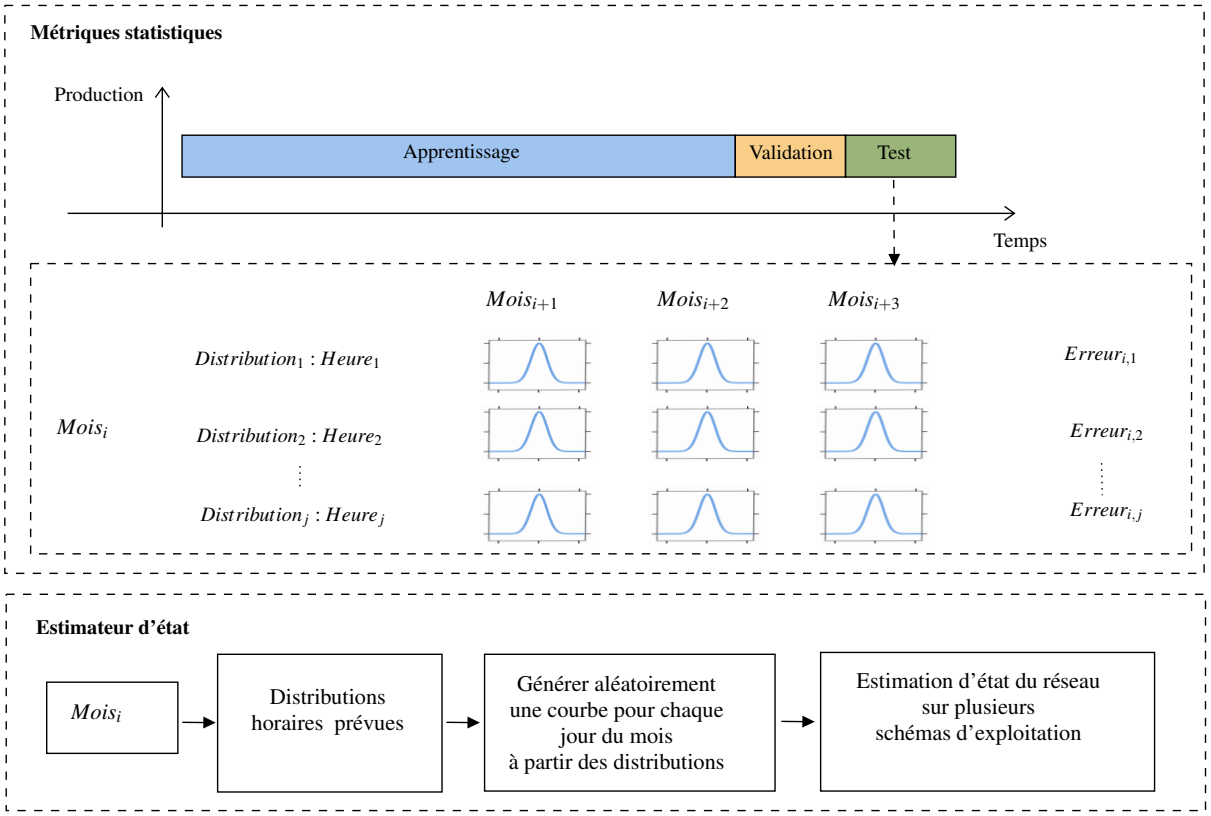


FIGURE 5.11 – Méthodologie d'évaluation dans le cas de la prévision long terme

$$\overline{err} = \sum_{m=1}^{M} \sum_{h=1}^{H} \sum_{q_\tau \in Q} erreur_{h,q_\tau,m}$$

M égale le nombre total de mois, H le nombre total d'heures et Q l'ensemble des quantiles utilisés dans la prévision. Dans notre cas τ est dans $\{0, 0.1, 0.2, ..., 0.9, 1\}$.

Dans le cas de l'estimateur d'état, nous avons utilisé les fonctions CDF prévues par les quantiles pour générer des échantillons de production. L'équation (5.7) ($y = F^{-1}(u)$) nous permet de réaliser cette simulation.

En réalisant des tirages aléatoires sur la probabilité u de la variable aléatoire uniforme U_i, nous obtenons des valeurs de production donc un scénario de production. Afin de simplifier cette simulation, nous supposons pour une journée que la probabilité u sera unique pour tous les producteurs et pour toutes les heures. Conséquemment en tirant un u dans $[0, 1]$ nous obtenons une courbe de production en interpolant par les quantiles prévus de la fonction CDF. Sur une période mensuelle, nous avons utilisé les données réelles de la charge et de la production des autres producteurs comme les éoliennes et les usines biogaz. La production solaire est obtenue par la courbe générée aléatoirement par la fonction CDF prévue. Nous notons que cette simulation n'est valable que si nous supposons que la production solaire est indépendante de la production éolienne et biogaz, ainsi que de la consommation. C'est une hypothèse assez forte, mais nous semble être une simplification raisonnable.

Nous avons réalisé un nombre de tirages équivalent au nombre de jours du mois. La production des petits producteurs est estimée en utilisant les courbes de production des moyens et grands producteurs générées aléatoirement. Au final, la qualité de modèle est obtenue en comparant avec les résultats de simulation des courbes générées par la fonction CDF réelle calculée sur le mois.

5.5 Données et modélisation

Dans cette étude de prévision de la production solaire de SRD, nous avons utilisé les mêmes données présentées dans la section 4.4 du chapitre 4 des moyens et grands producteurs PV raccordés au réseau HTA et BT. Pour supprimer l'effet du changement entre l'horaire d'été et celui d'hiver dans les données, nous avons converti la date en fuseau horaire GMT pour tous les producteurs. Nous n'avons gardé ainsi que les heures de production comprises entre 4h et 20h.

Dans la partie d'estimateur d'état, nous avons utilisé plusieurs schémas d'exploitation dans l'agence 6 de SRD. Les différents schémas sont calculés en utilisant l'optimiseur selon plusieurs scénarios sur le foisonnement global de toute l'agence (voir section 2.2.1). Les résultats d'estimation dans le cas des schémas optimaux avec un foisonnement de 15%, de 10% et de 5% ont été similaires. Nous avons ajouté d'autres limitations dans l'optimiseur comme la variation maximale de tension. Nous avons gardé au final 3 schémas dans la comparaison des modèles :

— Le schéma normal d'exploitation (Normal),

— Le schéma optimal avec un foisonnement global de 15% (Optimal1),

— Le schéma optimal avec un foisonnement de 5 % en limitant les variations de tension de ± 4 % (Optimal2).

5.5.1 Prévision court terme

Dans le cas de la prévision court terme, nous avons utilisé les courbes de production de 621 producteurs (moyens et grands). Pour chaque producteur, nous avons les puissances mesurées chaque 10 min en kW de janvier à décembre 2020. La courbe de production est une série temporelle d'une granularité de 10 min et de taille égale à $6 \times 17 \times 366 = 37\,332$ (17 représente le nombre d'heures de production). Cette série a été représentée par $z(s,t)$ dans le chapitre 4. Dans le cas de la prévision, nous notons cette série par

$$Y_{s,t} = z(s,t)$$

où s représente le producteur. Dans le cas de la prévision par une approche locale, nous modélisons notre problème par

$$Y_{s,t} = f_{s,t}(X_{s,t}) + \varepsilon_{s,t} \tag{5.17}$$

$Y_{s,t}$ est la variable cible à prédire, $X_{s,t}$ sont les variables explicatives, $\varepsilon_{s,t}$ est un bruit aléatoire. Dans le cas des données endogènes, les données explicatives sont l'ensemble $X_{s,t} = Y_{s,t-1}, Y_{s,t-2}, Y_{s,t-3}, ..., Y_{s,1}$. Dans le cas d'apprentissage machine nous pouvons ajouter d'autres variables explicatives calendaires comme le jour du mois $J_t \in \{1,2,3,...,31\}$, le jour de la semaine $J_t \in \{1,2,3,4,5,6,7\}$, l'heure $H_t \in \{4,5,...,20\}$, les minutes $M_t \in \{0,10,20,...,50\}$, etc.

Dans ce cas d'étude, il faut développer un modèle de prévision pour chaque producteur. C'est-à-dire dans notre cas, il faut entraîner 621 modèles différents, donc l'estimation $n \times 621$ paramètres différents (n égale le nombre de paramètres du modèle).

Dans le cas de prévision par approche globale, nous développons un seul modèle global regroupant toutes les courbes de production dans une seule matrice. Cette matrice en long format est obtenue en

empilant les courbes de productions ensemble (voir un exemple de matrice dans la figure 5.5). Nous modélisons notre problème de prévision ici par

$$Y_t = f(X_t) + \varepsilon_t \tag{5.18}$$

Les variables explicatives sont les données endogènes d'historique de prévision. Nous pouvons ajouter dans le cas d'approche globale les données statiques des producteurs comme la capacité de production et l'énergie annuelle calculée sur l'année précédente. Ces informations permettent au modèle global de capturer les informations sur la normalisation de données. Nous pouvons ajouter aussi les informations sur l'emplacement géographique des producteurs pour capturer l'information sur la répartition spatiale. Nous ajoutons les variables calendaires comme l'année, le mois, le jour, l'heure, etc. afin d'ajouter les informations sur les saisonnalités et les tendances globales. Finalement, nous notons qu'il est nécessaire d'ajouter dans ce type d'approche un identifiant unique de producteur, comme le numéro PDL. Le numéro PDL est utilisé comme une variable explicative dans le modèle afin d'identifier chaque producteur séparément.

5.5.2 Prévision long terme

Dans ce cas de prévision, nous avons utilisé les courbes de production des moyens et grands producteurs. Pour chaque producteur, nous avons les puissances mesurées chaque 10 min en kW de janvier 2012 à décembre 2020. Nous notons que pour les producteurs raccordés après janvier 2012, les valeurs de la période avant la date de mise en service sont considérées manquantes.

Nous avons modélisé notre prévision par une approche probabiliste. Le besoin de SRD dans cette prévision est d'estimer pour chaque producteur la distribution de la production pour une heure donnée sur un mois. C'est-à-dire, pour chaque producteur s et pour chaque heure de production solaire $h \in \{4,5,6,...,19,20\}$ sur un mois t, nous cherchons la fonction CDF représentant la distribution de ces valeurs de productions. Dans notre modélisation, nous avons transformé la série temporelle originale $Y_{s,t}$ en plusieurs séries temporelles de distributions horaires avec un pas mensuel. C'est-à-dire, nous déterminons pour chaque mois t la fonction F de la variable $Y_{s,h,t}$ avec $h \in \{4,5,6,...,19,20\}$.

$$F_{Y_{s,h,t}}(y) = P(Y_{s,h,t} \leq y) \tag{5.19}$$

La fonction F peut être résumée par les quantiles q_τ.

Dans notre méthodologie de prévision nous avons utilisé la méthode de prévision par quantile. Cette méthode non paramétrique, nous permet d'estimer la fonction CDF en prévoyant ces quantiles. Nous avons fixé 11 quantiles q_τ, avec τ dans $\{0,0.1,0.2,...,0.9,1\}$. Par exemple q_0 et q_1 représentent le minimum et le maximum de la production horaire sur un mois. Dans notre modélisation, nous avons développé une prévision pour chaque quantile à partir des quantiles calculés précédemment. Pour simplifier la modélisation, nous avons supposé que la prévision pour un quantile d'ordre τ donné est une fonction des anciens quantiles de ce même ordre τ. Nous obtenons donc pour chaque producteur, heure et quantile une série temporelle

$$Q_{\tau,s,h,t} = f[Q_{\tau,s,h,t-1}, Q_{\tau,s,h,t-2}, ...,] + \varepsilon_{\tau,s,h,t} \tag{5.20}$$

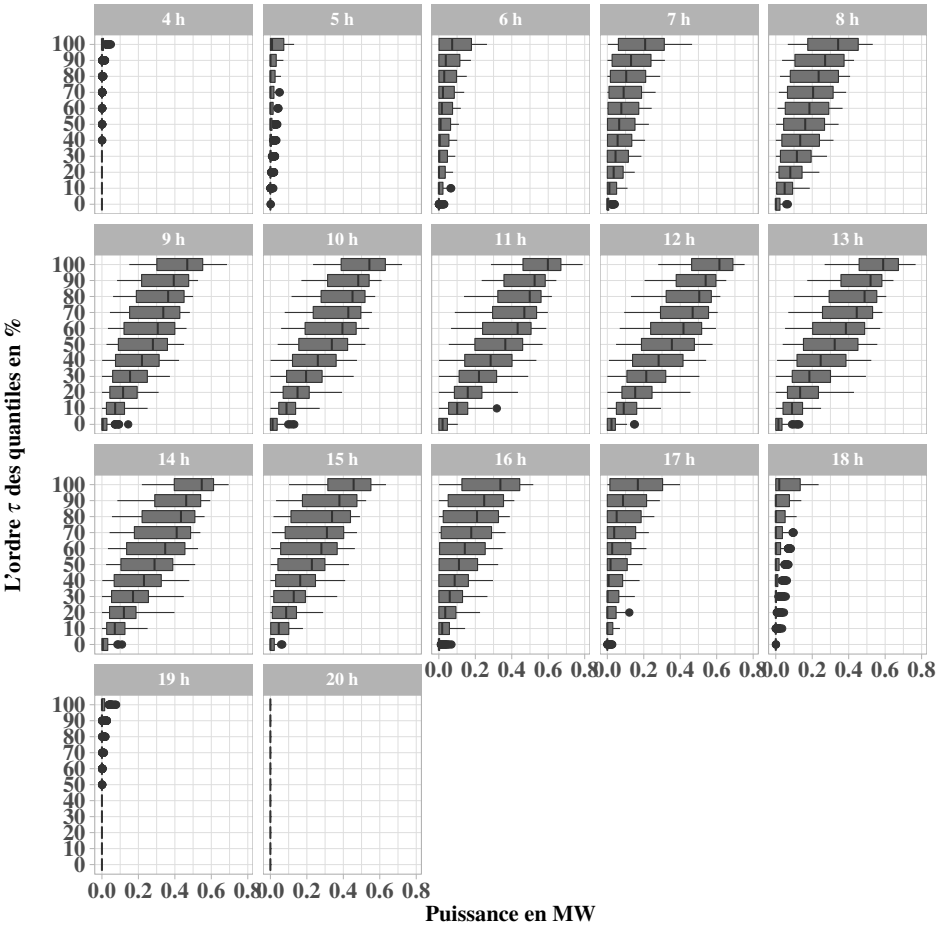


FIGURE 5.12 – Exemple des distributions des CDFs horaires dans le cas d'un producteur HTA d'une capacité de 0.8 *MVA* sur une période de 2012 à 2020.
Lecture : L'axe des abscisses représente les distributions des quantiles de production en MW (le quantile présente une puissance de production). L'axe des ordonnées représente l'ordre τ des quantiles. Chaque bloc représente l'heure de la production correspondante. Ces distributions des quantiles résument la distribution des fonctions CDF pour chaque heure de production sur la période 2012-2020.

Avec $Q_{\tau,s,h,t} = q_\tau(Y_{s,h,t})$ et $\varepsilon_{\tau,s,h,t}$ un bruit aléatoire.

Nous avons traité ces séries temporelles des quantiles par les méthodes de prévisions présentées avant. Dans le cas des modèles d'apprentissage machine, nous pouvons ajouter d'autres informations calendaires comme nous l'avons fait dans la prévision court terme. La figure 5.12 illustre les distributions des quantiles pour chaque heure de production sur la période 2012-2020 dans le cas d'un producteur PV d'une capacité installée de 800 kVA.

Dans l'approche globale nous empilons toutes les séries temporelles $Q_{\tau,s,h,t}$ ensemble dans une seule matrice en long format. Nous ajoutons dans cette approche les informations sur l'ordre des quantiles, $\tau \in \{0, 0.1, ..., 0.9, 1\}$. τ sera utilisé comme variable explicative du modèle.

La fonction CDF finale prévue est obtenue en réalisant une interpolation linéaire des différents quantiles prévus.

$$F_{Y_{s,h,t}}(y) = \sum_{i=1}^{N-1} \left[\tau_i + (y - q_{\tau_i}) \frac{\tau_{i+1} - \tau_i}{q_{\tau_{i+1}} - q_{\tau_i}} \right] \mathbb{1}(q_{q_{\tau_i}} \leq y \leq q_{\tau_{i+1}})$$

τ_i dans $\{0, 0.1, ..., 0.9, 1\}$ et i dans $\{1, 2, ..., N\}$ avec N le nombre total des quantiles prévus.

Dans le cas de l'évaluation par l'estimateur d'état, les courbes de production sont générées aléatoirement en calculant l'inverse de la fonction F et en tirant aléatoirement des u dans $\{0, 0.01, 0.02, ..., 0.99, 1\}$

$$y = F^{-1}_{Y_{s,h,t}}(u)$$

La valeur de u est globale pour l'ensemble des heures et des producteurs, mais elle est tirée aléatoirement pour chaque jour du mois.

5.6 Résultats

Dans cette section, nous présentons les différents résultats obtenus en appliquant la méthodologie de prévision et d'évaluation présentée précédemment et en exploitant les courbes de production réelles des moyens et grands producteurs PV.

Nous avons utilisé les méthodes d'apprentissage machine disponibles dans la bibliothèque *H2o* {LeDell *et al.*, 2020} (GBM et RF) permettant une parallélisation des calculs sur plusieurs cœurs. Le modèle ARIMA est disponible dans le package *forecast* {Hyndman et Khandakar, 2008} en R {R Core Team, 2020}. L'apprentissage des modèles est réalisé sur une machine virtuelle Ubuntu 18.04.4 fournie par le LIAS, avec un processeur Intel(R) Xeon(R) CPU E5-2630 v3 à 2.40GHz avec 6 cœurs, et une mémoire vive de 45Go.

5.6.1 Prévision court terme

Dans l'étude de la prévision court terme de la production solaire de SRD, nous avons testé cinq algorithmes de prévision. Dans l'approche locale, nous avons utilisé l'algorithme naïf de prévision présenté dans l'équation (5.3) et l'algorithme de la moyenne glissante présenté par l'équation (5.4) sur une heure MA1 et sur 2 heures MA2. Dans le cas de l'approche globale, nous avons utilisé deux algorithmes d'apprentissage machine GBM et RF, ainsi qu'un modèle Ensemble regroupant les deux prévisions de RF et de GBM.

L'évaluation des modèles est réalisée par les métriques statistiques et par les résultats d'estimation du réseau en utilisant un calculateur d'état. D'une part, la qualité d'un modèle est déterminée statistiquement en utilisant les métriques MBE^+, MBE^-, MBE, MAE et $RMSE$ entre les courbes prévues et les courbes réelles observées. Les métriques MBE^+, MBE^- nous indiquent les taux de la sur-prévision et de la sous-prévision d'un modèle. D'autre part, la précision d'un modèle est validée par un calculateur d'état de réseau en utilisant les courbes prévues. Les différences entre les résultats d'estimation du réseau de chaque modèle sont quantifiées par les métriques MBE, MAE, $RMSE$ et $MAPE$, ainsi que dans le cas des chutes et d'élévation de tension par les métriques binaires.

Dans la partie d'évaluation par les métriques statistiques, nous avons obtenu les résultats résumés dans le tableau 5.3, l'algorithme RF est le plus performant en qualité de prévision par rapport au GBM et aux méthodes naïves.

Sur l'ensemble de producteurs, l'algorithme RF en approche globale est le plus précis en qualité de prévision sur l'ensemble de périodes testées. La figure 5.13 présente les résultats obtenus en utilisant la métrique $nRMSE_s$. Les distributions présentées par les boîtes à moustaches illustrent la répartition des $nRMSE_s$ obtenues pour chaque producteur. Nous remarquons que les algorithmes d'apprentissage machine améliorent la qualité de prévision par rapport aux méthodes naïves sur l'ensemble des producteurs.

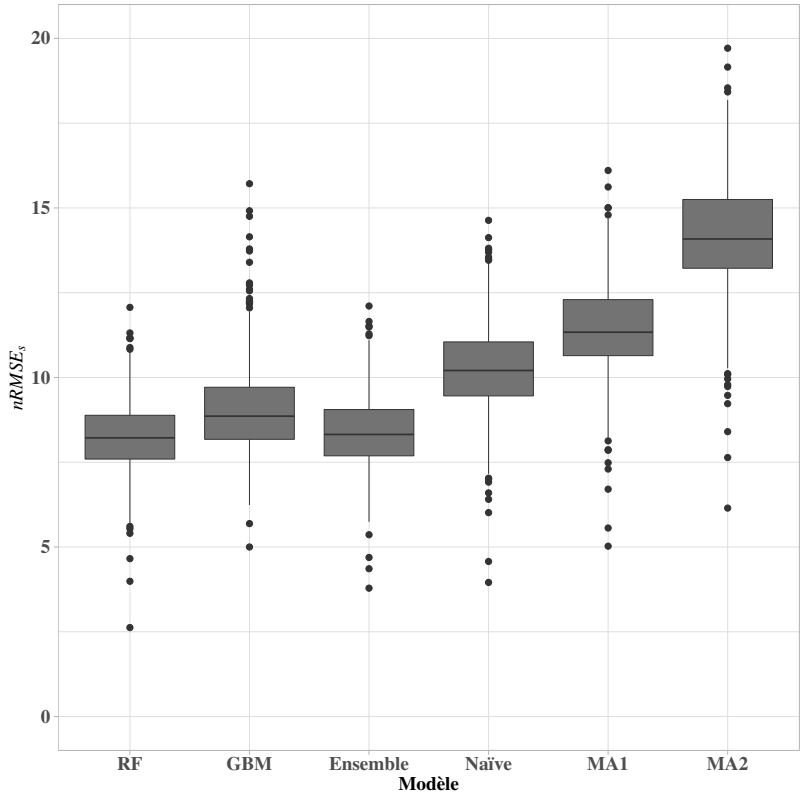


FIGURE 5.13 – Les distributions des erreurs $nRMSE_s$ calculées pour chaque producteur
Lecture : L'axe des abscisses représente les modèles et l'axe des ordonnées représente les distributions des $nRMSE_s$ en pourcentage de capacité installée.

La moyenne glissante sur une période plus grande dégrade la qualité de prévision. Nous notons que la méthode naïve a fourni une prévision assez juste vu sa simplicité fondamentale.

Globalement sur l'ensemble de prévisions fournies, la méthode RF est la plus performante en qualité de prévision avec une nRMSE égale 8.32% suivie par la méthode GBM 9.12%. La prévision moyenne de RF et GBM (Ensemble) n'a pas amélioré la qualité de prévision de RF. La méthode naïve a fourni une prévision de 10.34 % suivie par la méthode de la moyenne glissante 11.48 % pour une heure et 14.24 % pour 2 heures. Le Skill score ss de RF est de 19.19 % suivi par Ensemble 17.76 % et GBM 11.50 %. Le ss présente le pourcentage d'amélioration de la qualité de prévision par rapport au modèle naïf. Les résultats de MBE^+, MBE^- et MBE indiquent que les méthodes d'apprentissage machine ont tendance à sur-prévoir la production sur l'ensemble de producteurs.

Les résultats de la simulation par estimateur d'état sur l'agence 6 indiquent que l'algorithme RF est

Modèle	$nMBE^+(\%)$	$nMBE^-(\%)$	$nMBE(\%)$	$nMAE(\%)$	$nRMSE(\%)$	ss(%)
RF	5.07	-3.83	-0.64	4.27	8.32	19.19
GBM	6.84	-4.48	-0.94	5.22	9.12	11.50
Ensemble	5.77	-4.04	-0.79	4.62	8.47	17.76
Naïve	8.33	-8.36	-0.00	5.67	10.34	0
Naïve MA1	10.13	-8.75	0.00	7.22	11.48	-11.42
Naïve MA2	12.89	-10.48	0.00	9.51	14.24	-38.24

TABLEAU 5.3 – Précision des modèles dans le cas de la prévision court terme. Les métriques $nMBE^+(\%)$, $nMBE^-(\%)$, $nMBE(\%)$, $nMAE(\%)$ et $nRMSE(\%)$ représentent l'erreur de prévision en pourcentage de la capacité installée des producteurs. Le skill score ss présente un pourcentage de différence par rapport à la RMSE du modèle naïf

le plus performant sur l'ensemble des schémas d'exploitation testés comparé à l'estimation du réseau en utilisant les courbes réelles. Le tableau 5.4 résume les résultats obtenus au niveau des pertes, des chutes de tension et des élévations de tension. Sur le schéma de fonctionnement normal du réseau, l'algorithme RF a fourni une MAPE de 0.8% contre une MAPE de 4.66% obtenue dans l'estimation du réseau sans production. Dans l'ensemble des trois schémas, aucune chute de tension n'est calculée. Par contre, des élévations de tension ont été estimées, l'algorithme RF est le plus performant en qualité de prévision avec une BA de 95.58% dans le cas du schéma normal. La sensitivité est égale à 99.17% et la spécificité égale à 92%. La méthode naïve a fourni une BA de 90.62% avec une sensitivité de 98.71% et une spécificité égale à 82.53%. L'algorithme RF a prévu 10 FP, 9 FN et 115 VN contre 22 FP, 14 FN et 104 VN prévus par la méthode naïve. Nous notons que l'augmentation du nombre de FP par un modèle est plus contraignante pour le fonctionnement du réseau. En effet un modèle qui prévoit moins de FP est plus pertinent pour la conduite de réseau, car il prévoit assez bien les contraintes affectant le réseau. Il évite de prévoir des fonctionnements normaux du réseau pour des fonctionnements contraints réels. Le même ordre est respecté dans les deux autres schémas. Un faible nombre d'élévation de tension est estimé : 25 points dans le cas de RF et 31 points dans le cas de la méthode naïve. Ce faible nombre de dépassements s'explique par la configuration de ces schémas optimaux conçus pour ne pas engendrer d'élévations de tension. Nous notons que la différence du nombre total de points d'élévation de tension entre les différentes méthodes s'explique par la non prise en compte des points de divergence de l'estimateur d'état. L'algorithme RF permet donc de bien prévoir le fonctionnement normal de réseau et les contraintes affectant la qualité de distribution de l'électricité.

5.6.2 Prévision long terme

Dans le cas de la prévision long terme, nous avons testé six algorithmes : deux algorithmes d'apprentissage machine RF et GBM avec une approche globale, une méthode statistique classique ARIMA et quatre algorithmes naïfs. Dans l'évaluation des modèles, nous avons comparé les algorithmes RF, GBM et ARIMA avec plusieurs méthodes naïves : une méthode naïve utilisant la valeur du dernier mois, une méthode naïve (NaïveAnnée) utilisant les données de l'année dernière comme valeurs de prévision des prochains mois et deux méthodes de la moyenne glissante sur deux mois (MA2) et sur trois mois (MA3). Par exemple, dans le cas du trimestre Hiver 1 (janvier, février et mars 2020) la prévision du modèle naïf sur l'horizon de ces trois mois est la valeur observée dans le mois décembre 2019. Les valeurs prévues par la méthode NaïveAnnée sont les valeurs observées en janvier, février et mars de l'année dernière (2019).

En comparaison avec les courbes générées par les CDF réelles (voir tableau 5.5), l'algorithme RF a fourni une meilleure qualité de prévision avec une nRMSE de 5.91 %, suivi de GBM (6.10 %) et de NaïveAnnée (6.55%). L'algorithme RF a amélioré la précision de la prévision de la méthode NaïveAnnée de 9.74% et de la méthode naïve de 68%. Nous notons que l'algorithme NaïveAnnée a fourni une bonne qualité de prévision vu ses fondements très basiques. La figure 5.14 illustre les distributions des $nRMSE_s$ calculées pour chaque producteur. Les méthodes RF, GBM et NaïveAnnée améliorent largement l'ensemble des prévisions de chaque producteur comparé aux autres méthodes. La méthode statistique ARIMA avec une approche locale n'a pas devancé ces trois méthodes. Nous notons que la production de l'année dernière utilisée dans la méthode NaïveAnnée fait partie des variables explicatives des deux modèles RF et GBM. En regardant l'erreur dans chaque trimestre et en fonction de l'horizon de prévision (voir figure 5.15), nous remarquons dans le cas de deux trimestres «Hiver 1» et «Eté 2» que les méthodes naïves (Naïve, MA1 et MA3) ont fourni une bonne qualité de prévision comparées aux deux

Modèle	Schéma	Pertes			Chutes de tension			Élévation de tension									
		MAE (MW)	RMSE (MW)	MAPE (%)	MAE (v.u)	RMSE (v.u)	MAPE (%)	MAE (v.u)	RMSE (v.u)	MAPE (%)	FP	FN	VN	VP	BA (%)	Sensibilité (%)	Spécificité (%)
RF	Normal	0.0043	0.0235	0.8154	0.0001	0.0003	0.0098	0.0009	0.0020	0.0897	10	9	115	1081	95.58	99.17	92
RF	Optimal1	0.0040	0.0219	0.9213	0.0001	0.0004	0.0141	0.0005	0.0011	0.0460	3	2	25	1186	94.55	99.83	89.28
RF	Optimal2	0.0040	0.0262	0.9544	0.0001	0.0004	0.0133	0.0005	0.0011	0.0461	3	2	25	1186	94.55	99.83	89.28
GBM	Normal	0.0054	0.0237	1.1096	0.0001	0.0003	0.0119	0.0011	0.0022	0.1073	20	9	106	1084	91.65	99.17	84.12
GBM	Optimal1	0.0050	0.0222	1.2042	0.0002	0.0005	0.0176	0.0006	0.0012	0.0569	8	0	20	1186	85.71	100	71.42
GBM	Optimal2	0.0050	0.0265	1.2465	0.0002	0.0004	0.0166	0.0006	0.0012	0.0571	8	0	20	1191	85.71	100	71.42
Ensemble	Normal	0.0048	0.0235	0.9372	0.0001	0.0003	0.0106	0.0010	0.0020	0.0973	11	9	114	1081	95.18	99.17	91.20
Ensemble	Optimal1	0.0045	0.0220	1.0377	0.0002	0.0004	0.0155	0.0005	0.0012	0.0509	6	1	22	1188	89.24	99.91	78.57
Ensemble	Optimal2	0.0044	0.0263	1.0714	0.0001	0.0004	0.0147	0.0005	0.0012	0.0509	6	1	22	1191	89.24	99.91	78.57
Naïve	Normal	0.0054	0.0223	1.1684	0.0001	0.0004	0.0142	0.0012	0.0025	0.1181	22	14	104	1077	90.62	98.71	82.53
Naïve	Optimal1	0.0053	0.0229	1.2897	0.0002	0.0006	0.0200	0.0007	0.0014	0.0631	3	6	25	1183	94.39	99.49	89.28
Naïve	Optimal2	0.0049	0.0225	1.3397	0.0002	0.0005	0.0192	0.0007	0.0014	0.0635	3	6	24	1185	94.19	99.49	88.88
MA1	Normal	0.0075	0.0234	1.6821	0.0002	0.0005	0.0196	0.0016	0.0030	0.1578	18	11	107	1078	92.29	98.9899	85.6
MA1	Optimal1	0.0075	0.0244	1.9226	0.0003	0.0006	0.0252	0.0008	0.0016	0.0798	7	10	21	1171	87.07	99.15	75
MA1	Optimal2	0.0071	0.0238	1.9997	0.0002	0.0006	0.0246	0.0008	0.0016	0.0799	7	10	21	1182	87.08	99.16	75
MA2	Normal	0.0104	0.0266	2.3380	0.0003	0.0006	0.0269	0.0022	0.0039	0.2153	28	21	98	1071	87.92	98.07	77.77
MA2	Optimal1	0.0102	0.0264	2.7058	0.0003	0.0008	0.0345	0.0011	0.0020	0.1055	7	7	21	1182	87.2	99.41	75
MA2	Optimal2	0.0099	0.0257	2.8318	0.0003	0.0008	0.0333	0.0011	0.0020	0.1067	7	7	21	1185	87.2	99.41	75
Sans production	Normal	0.0212	0.0335	4.6684	0.0006	0.0014	0.0642	0.0070	0.0112	0.6640	126	0	0	1086	50	100	0
Sans production	Optimal1	0.0204	0.0324	5.2764	0.0008	0.0019	0.0852	0.0040	0.0067	0.3804	28	0	0	1188	50	100	0
Sans production	Optimal2	0.0205	0.0343	5.4867	0.0008	0.0019	0.0831	0.0040	0.0067	0.3824	28	0	0	1193	50	100	0

TABLEAU 5.4 – Résultats d'estimateur d'état dans le cas de l'agence 6 en utilisant les prévisions court termes

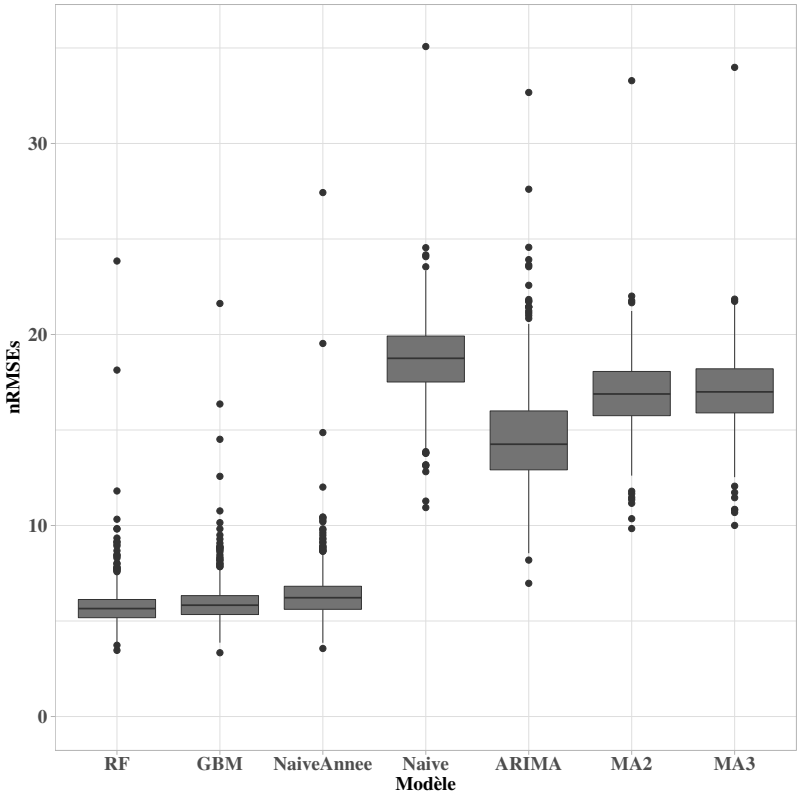


FIGURE 5.14 – Les distributions des erreurs $nRMSE_s$ calculées pour chaque producteur
Lecture : L'axe des abscisses représente les modèles et l'axe des ordonnées représente les distributions des $nRMSE_s$ en pourcentage de capacité installée.

autres trimestres «Hiver 2» et «Eté 1» . Cela montre que l'utilisation des distributions des derniers mois appartenant à la même saison (été ou hiver) est plus précise que les distributions d'une saison différente. Les trois méthodes RF, GBM et NaïveAnnée ont fourni une précision assez similaire sur l'ensemble des horizons, sauf pour le mois de novembre (horizon 2 dans Hiver 2), une nRMSE de 10% est obtenu contre 4% dans les deux autres mois (octobre et décembre). Ceci est expliqué par une augmentation de la production au mois de novembre 2020 comparé à la production produite en même mois de l'année dernière. Par exemple, la moyenne des médianes de toutes les heures et de tous les producteurs en 2020 est de 24.14 kW contre une moyenne de 9.57 kW en 2019.

Les résultats obtenus dans le cas d'estimateur d'état sur l'ensemble des mois testés (voir tableau 5.6) indique que l'algorithme RF a fourni une meilleure précision d'estimation des pertes avec une MAPE

Modèle	$nMBE^{+}(\%)$	$nMBE^{-}(\%)$	$nMBE(\%)$	$nMAE(\%)$	$nRMSE(\%)$	$ss_1(\%)$	$ss_2(\%)$
RF	4.73	-2.15	0.23	3.04	5.91	68.55	9.74
GBM	4.95	-2.46	-0.06	3.27	6.10	67.55	6.88
NaïveAnnée	5.21	-5.10	-0.70	3.30	6.55	65.18	0
Naive	17.98	-12.29	3.94	11.68	18.83	0	-187.56
ARIMA	13.79	-4.57	5.07	8.98	14.84	21.05	-126.60
MA2	15.92	-11.69	2.10	10.86	16.99	9.65	-159.32
MA3	14.92	-12.22	1.17	10.99	17.14	8.81	-161.75

TABLEAU 5.5 – Précision des modèles dans le cas de la prévision long terme. Les métriques $nMBE^{+}(\%)$, $nMBE^{-}(\%)$, $nMBE(\%)$, $nMAE(\%)$ et $nRMSE(\%)$ représentent l'erreur de prévision en pourcentage de la capacité installée des producteurs. Les skill score ss_1 et ss_2 représentent le pourcentage de différence par rapport à la RMSE des méthodes Naïve et NaïveAnnée

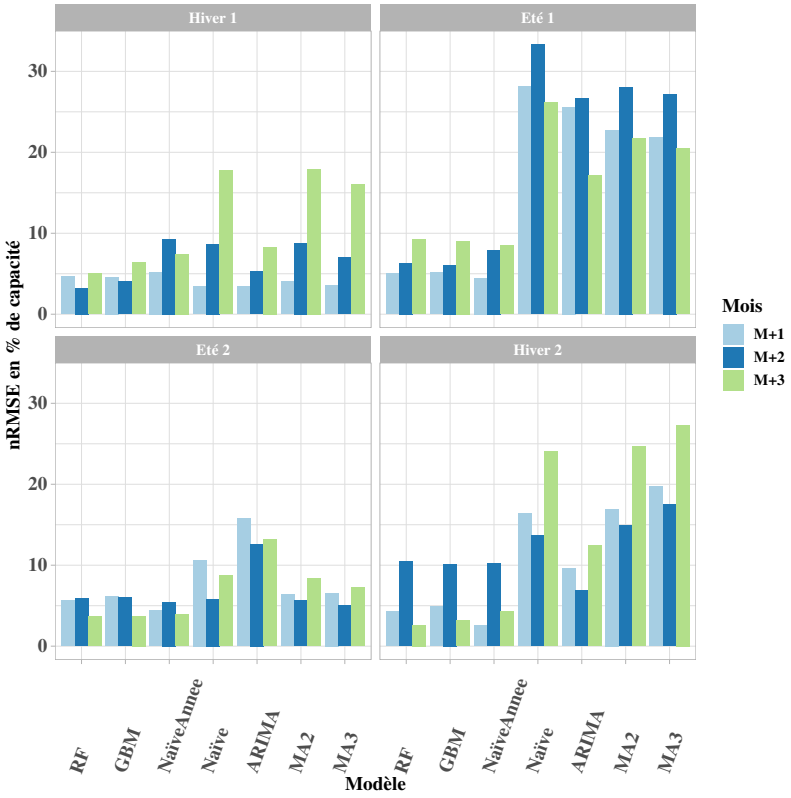


FIGURE 5.15 – L'erreur nRMSE en fonction des trimestres et l'horizon de prévision.
Lecture : les modèles sont présentés dans l'axe des abscisses et l'erreur nRMSE est présentée dans l'axe des ordonnées. Dans les différents blocs, nous avons les résultats obtenus dans le cas des trimestres en fonction de chaque mois de prévision sur les horizons d'un mois (M+1) jusqu'à trois mois (M+3).

de 0.62 % dans le schéma normal, suivi de GBM avec une MAPE de 0.67% et NaïveAnnée avec une MAPE de 0.75% contre une MAPE de 5.71 % obtenue dans le cas d'estimation sans production. Dans le cas des élévations de tension, l'algorithme NaïveAnnée a fourni une meilleure précision des prévisions des contraintes affectant le réseau avec une BA de 93.2%, une sensitivité de 98.65 % et une spécificité de 87.39, suivi de GBM avec une BA de 91.09% et de RF qui a fourni une BA de 90.72%. NaïveAnnée a prévu 30 FP, 78 FN et 208 VN contre 43 FP, 33 FN et 196 prévus par RF. En revanche, dans les deux schémas optimaux, il n'y a que 3 points de dépassements de tension.

5.6.3 Discussion

Nous avons développé dans cette étude une méthodologie de prévision de la production solaire du réseau électrique de SRD. La prévision est élaborée en utilisant plusieurs méthodes naïves, statistiques et apprentissages machine en exploitant les données disponibles des courbes de production des moyens et grands producteurs. Nous avons testé ces méthodes pour une prévision ponctuelle de 10 min à 60 min et pour une prévision probabiliste d'un mois à trois mois. L'évaluation des différentes méthodes est réalisée par de nombreuses métriques statistiques et par les résultats des simulations fournies par le calculateur d'état en utilisant plusieurs schémas d'exploitation dans le cas de l'agence 6 de SRD.

Dans le cas de la prévision court terme, les méthodes d'apprentissage machine avec une approche globale ont fourni une meilleure qualité de prévision comparées aux méthodes naïves. L'algorithme RF a amélioré la qualité de prévision de la méthode naïve de 19%. La figure 5.16 illustre un exemple de prévision d'un producteur prévu par RF avec une nRMSE de 8.32 % (proche de la nRMSE globale). Nous remarquons la différence entre la qualité de prévision dans une journée claire et une journée nuageuse.

Modèle	Schéma	Pertes			Chutes de tension			Élévation de tension									
		MAE (MW)	RMSE (MW)	MAPE (%)	MAE (v.u)	RMSE (v.u)	MAPE (%)	MAE (v.u)	RMSE (v.u)	MAPE (%)	FP	FN	VN	VP	BA (%)	Sensibilité (%)	Spécificité (%)
RF	Normal	0.0041	0.0288	0.6279	0.0001	0.0003	0.0113	0.0007	0.0013	0.0658	43	33	196	5773	90.72	99.43	82.00
RF	Optimal1	0.0043	0.0273	0.7772	0.0001	0.0003	0.0113	0.0003	0.0006	0.0250	2	0	1	6045	66.66	100.00	33.33
RF	Optimal2	0.0042	0.0249	0.8036	0.0001	0.0003	0.0115	0.0003	0.0006	0.0254	2	0	1	6048	66.66	100.00	33.33
GBM	Normal	0.0044	0.0288	0.6777	0.0001	0.0003	0.0120	0.0007	0.0014	0.0706	41	34	197	5773	91.09	99.41	82.77
GBM	Optimal1	0.0045	0.0273	0.8336	0.0001	0.0003	0.0118	0.0003	0.0006	0.0270	2	0	1	6041	66.66	100	33.33
GBM	Optimal2	0.0045	0.0249	0.8645	0.0001	0.0003	0.0121	0.0003	0.0006	0.0274	2	0	1	6051	66.66	100	33.33
Naïve Année	Normal	0.0045	0.0233	0.7526	0.0001	0.0003	0.0112	0.0007	0.0014	0.0703	30	78	208	5739	93.02	98.65	87.39
Naïve Année	Optimal1	0.0047	0.0242	0.8766	0.0001	0.0003	0.0109	0.0003	0.0007	0.0280	2	0	1	6044	66.66	100	33.33
Naïve Année	Optimal2	0.0047	0.0226	0.9139	0.0001	0.0003	0.0111	0.0003	0.0007	0.0284	2	0	1	6051	66.66	100	33.33
Naïve	Normal	0.0155	0.0324	3.2472	0.0006	0.0013	0.0583	0.0025	0.0046	0.2418	222	44	16	5778	52.98	99.24	6.72
Naïve	Optimal1	0.0159	0.0334	3.8567	0.0006	0.0014	0.0634	0.0010	0.0022	0.1000	3	0	0	6054	50	100	0
Naïve	Optimal2	0.0160	0.0319	3.9874	0.0006	0.0014	0.0644	0.0010	0.0022	0.0997	3	0	0	6046	50	100	0
ARIMA	Normal	0.0091	0.0301	2.0316	0.0004	0.0009	0.0394	0.0018	0.0034	0.1699	169	9	70	5801	64.56	99.84	29.28
ARIMA	Optimal1	0.0094	0.0295	2.4644	0.0004	0.0010	0.0403	0.0007	0.0016	0.0678	3	0	0	6046	50	100	0
ARIMA	Optimal2	0.0094	0.0272	2.5427	0.0004	0.0010	0.0413	0.0007	0.0016	0.0677	3	0	0	6050	50	100	0
MA2	Normal	0.0146	0.0316	2.9489	0.0005	0.0011	0.0529	0.0023	0.0041	0.2247	211	49	28	5765	55.43	99.15	11.71
MA2	Optimal1	0.0153	0.0338	3.4966	0.0006	0.0012	0.0582	0.0010	0.0020	0.0939	3	0	0	6047	50	100	0
MA2	Optimal2	0.0148	0.0305	3.7544	0.0006	0.0012	0.0592	0.0010	0.0020	0.0943	3	0	0	6050	50	100	0
MA3	Normal	0.0148	0.0332	2.9126	0.0005	0.0011	0.0521	0.0024	0.0042	0.2293	207	68	31	5743	55.92	98.82	13.02
MA3	Optimal1	0.0154	0.0342	3.4787	0.0006	0.0012	0.0582	0.0010	0.0020	0.0955	3	0	0	6043	50	100	0
MA3	Optimal2	0.0153	0.0315	3.5991	0.0006	0.0012	0.0586	0.0010	0.0020	0.0957	3	0	0	6040	50	100	0
Sans production	Normal	0.0259	0.0402	5.7109	0.0009	0.0015	0.0892	0.0047	0.0080	0.4551	239	0	0	5813	50	100	0
Sans production	Optimal1	0.0257	0.0389	6.6157	0.0009	0.0015	0.0921	0.0021	0.0041	0.1986	3	0	0	6041	50	100	0
Sans production	Optimal2	0.0258	0.0376	6.8524	0.0009	0.0016	0.0936	0.0021	0.0042	0.1993	3	0	0	6054	50	100	0

TABLEAU 5.6 – Résultats d'estimateur d'état dans le cas de l'agence 6 de SRD en utilisant les prévisions long terme

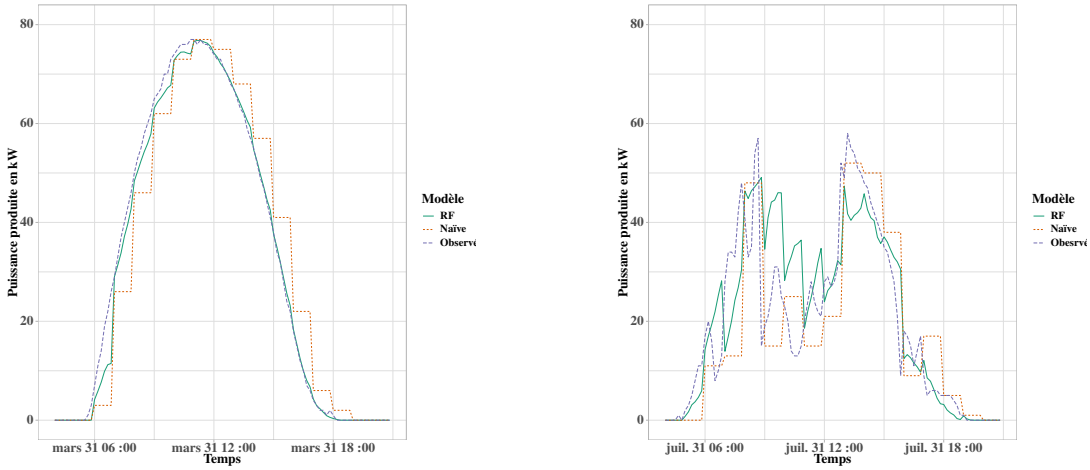


FIGURE 5.16 – Exemple de prévision dans une journée claire et une journée nuageuse.
Lecture : L'axe des abscisses est le temps avec un pas de 10 minutes et l'axe des ordonnées est la puissance produite en kW. La figure à gauche illustre un exemple de prévision dans une journée claire et la figure à droite illustre la prévision dans une journée nuageuse.

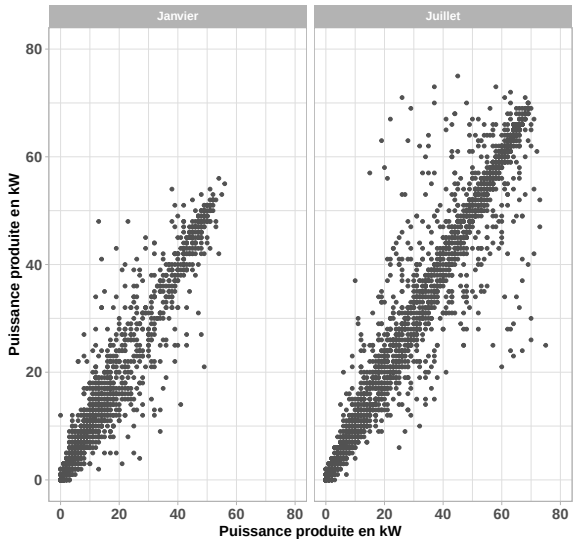


FIGURE 5.17 – Exemple de relation entre la production à l'instant t et la production à l'instant $t-1$.
Lecture : L'axe d'abscisses est la puissance produite à l'instant t et l'axe des ordonnées est la puissance produite à l'instant $t-1$. Les blocs représentent les mois janvier et juillet.

L'algorithme RF prévoit mieux la production solaire sur une journée claire comparé à l'algorithme naïf. La courbe prévue par l'algorithme naïf est une courbe décalée d'une heure de la courbe de production réelle. Cependant, dans une journée nuageuse les deux algorithmes ont des difficultés à prévoir le passage du nuage. Dans ce cas, la courbe prévue par RF est aussi décalée de la courbe réelle. Ceci est expliqué par la forte corrélation (97%) entre la production à l'instant t et la production à l'instant $t-1$. La variable explicative production à l'instant $t-1$ a une grande importance dans la décision de l'algorithme RF par rapport aux autres variables explicatives. La figure 5.17 illustre le lien entre la production à l'instant t et la production à l'instant $t-1$ dans deux mois de l'année (janvier et juillet), la corrélation moyenne est de l'ordre de 97%.

Concernant la prévision long terme l'approche naïve (NaïveAnnée) utilisant les données de l'année dernière comme valeurs de prévision est plus performant que l'approche naïve utilisant les valeurs du dernier mois. L'algorithme RF a amélioré la précision de prévision de 9% par rapport à la prévision

fournie par la méthode NaïveAnnée. Toutefois, la méthode NaïveAnnée est limitée pour les nouveaux producteurs raccordés récemment à cause d'absence de données historiques. L'algorithme RF est avantageux dans ce cas, car il prend en compte les périodes manquantes dans la décision de prévision. En effet, dans la construction des arbres de décision les valeurs manquantes sont interprétées comme une information distincte. De plus, dans l'approche globale l'algorithme permet de capter toutes les informations des périodes de production des producteurs raccordés précédemment. Une alternative pour l'approche NaïveAnnée est d'estimer une courbe historique d'un nouveau producteur en utilisant la méthodologie d'interpolation spatiale présentée dans le chapitre 4.

La figure 5.18 illustre un exemple de prévision des quantiles d'un producteur d'une capacité installée de 250 kVA orienté plein sud. Pour chaque heure de production solaire, nous observons les prévisions fournies par RF et NaïveAnnée en comparaison avec les quantiles réels. Les quantiles d'ordres 0 et 1 représentent les productions minimale et maximale d'une heure donnée dans un mois, le quantile d'ordre 0.5 représente la médiane de cette distribution de production horaire. Les courbes de la production médiane illustrent les différences qui peuvent apparaître entre les médianes d'un mois et les médianes du même mois de l'année dernière. Cette différence est illustrée par exemple dans les mois d'avril, juin et novembre. Elle est moins significative dans les prévisions des productions maximales. De plus, l'algorithme RF permet de mieux ajuster la prévision de la méthode NaïveAnnée en utilisant les informations disponibles de la production des derniers mois. La figure 5.19 illustre les relations entre la production d'un mois et la production des mois précédents. Nous remarquons que les relations sont fortes pour le mois $t-1$ avec une corrélation de 94% et les mois de l'année dernière $t-12$ et $t-13$ avec des corrélations de 97% et 94%. Ces deux variables explicatives ont beaucoup d'importance dans la décision de l'algorithme RF.

5.7 Conclusion du chapitre

Dans cette étude, nous avons élaboré des approches de prévision de la production solaire dans un réseau de distribution de l'électricité. La prévision permet au GRD de bien gérer l'intermittence des producteurs raccordés au réseau (HTA et BT) et de bien planifier les ressources. Nous avons développé deux types de prévision en n'utilisant que les données endogènes d'historique de production : une prévision court terme pour une anticipation des contraintes affectant le réseau d'un horizon d'une heure et une prévision long terme pour une optimisation prédictive du réseau.

Dans le cas de la prévision court terme, l'algorithme RF a amélioré la qualité de prévision de 19% par rapport à l'algorithme naïf en fournissant une prévision ponctuelle sur un horizon d'une heure avec un pas de dix minutes. Dans le cas de la prévision probabiliste long terme, l'algorithme RF a amélioré de 68% la précision de la prévision fournie par l'algorithme naïf et de 9.7% la qualité de prévision fournie par la méthode NaïveAnnée. Les prévisions fournies par les algorithmes naïfs ont été assez justes vu la simplicité fondamentale de ces algorithmes. Les algorithmes d'apprentissage machine avec une approche globale permettent de mieux ajuster les prévisions naïves en combinant plusieurs variables explicatives d'historique de production et d'autres informations statiques des producteurs. L'avantage de l'algorithme RF avec une approche globale pour le GRD est de créer un seul algorithme de prévision pour l'ensemble des producteurs raccordés dans le réseau et instrumentés par un compteur communicant, avec un temps de calcul réduit (en moyenne 12 minutes d'apprentissage) dans les deux cas de prévisions. De plus, la simulation par l'estimateur d'état a confirmé que globalement à l'échelle d'une agence l'algorithme RF a fourni une précision de prévision de l'état du réseau comparé à la simulation utilisant les données réelles.

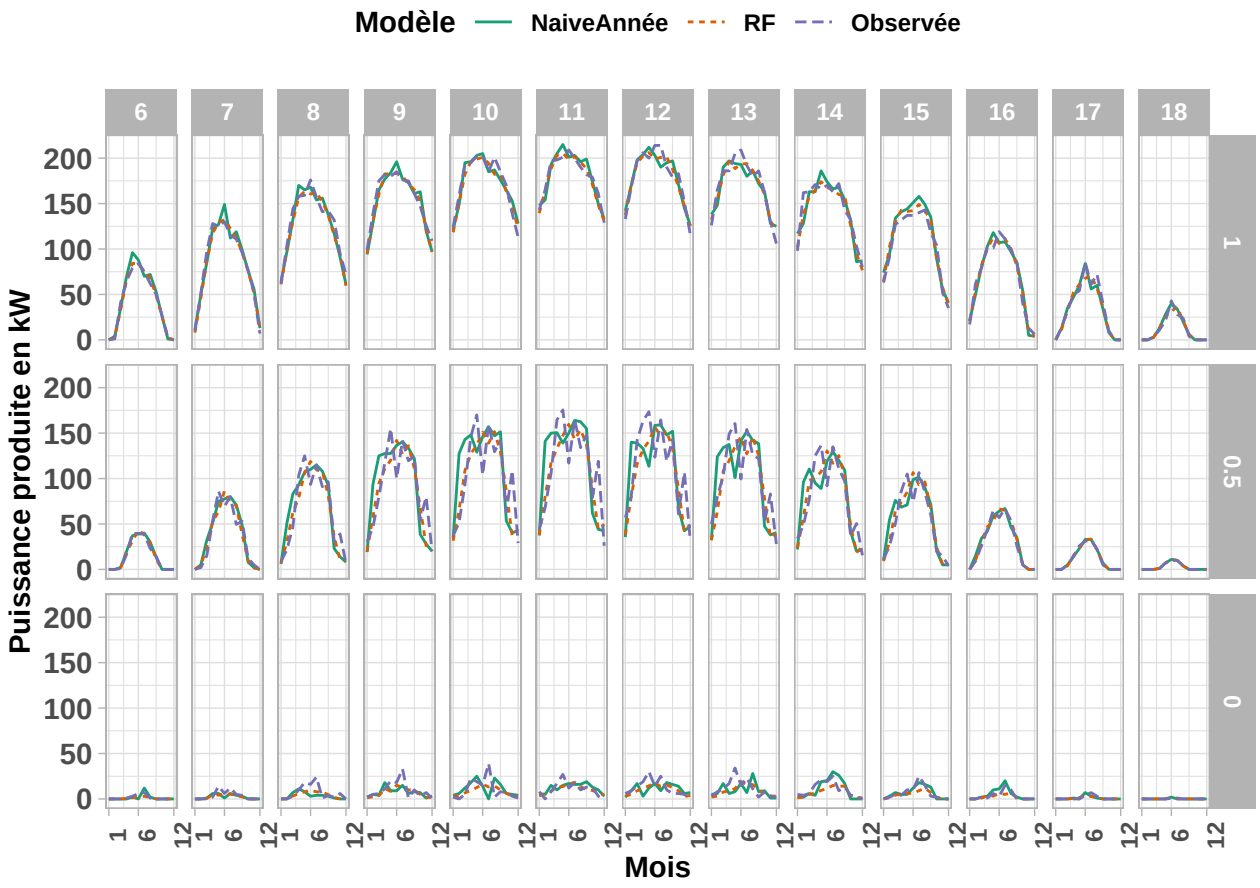


FIGURE 5.18 – Exemple de prévision long terme.
Lecture : L'axe des abscisses représente les mois et l'axe des ordonnées représente la puissance produite en kW. Les blocs représentent les quantiles d'ordre 0, 0.5 et 1 en fonction des heures de la journée de production.

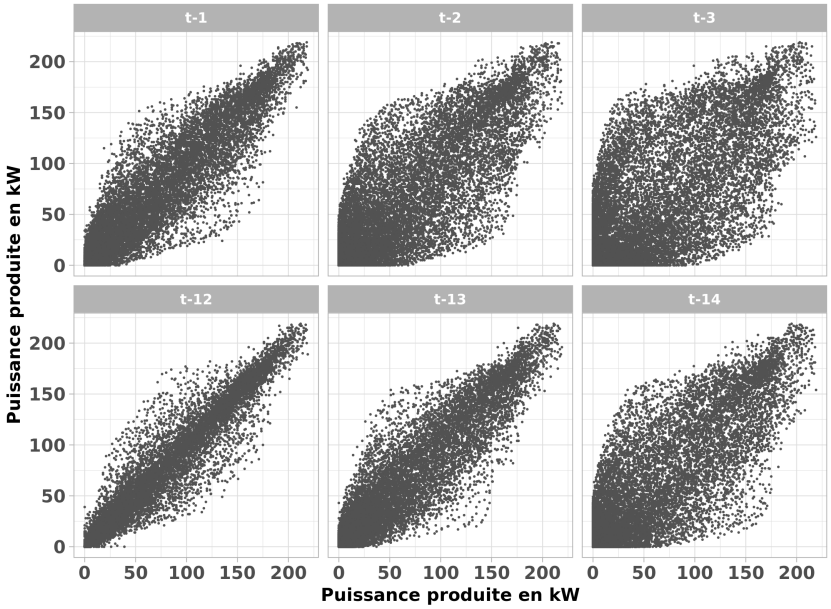


FIGURE 5.19 – Exemple de relation entre la production à l'instant t et la production aux instants précédents.
Lecture : L'axe des abscisses représente la puissance produite en kW dans un mois t et l'axe des ordonnées représente la puissance produite en kW dans les mois précédents. Les blocs représentent les relations entre la prodution du mois t et les mois $t-1$,$t-2$ et $t-3$ et les mois de l'année dernière $t-12$,$t-13$ et $t-14$.

La combinaison de la prévision et de l'estimation d'état permet d'anticiper les chutes et les élévations de tension et d'autres contraintes affectant le réseau dans un état de fonctionnement futur. De plus, la prévision probabiliste long terme permet au GRD de tester plusieurs scénarios sur la prévision de la production pour bien gérer et optimiser les ressources d'exploitation en conséquence. Ceci permet ainsi de bien intégrer de nouveaux producteurs PV décentralisés dans le réseau.

Dans le cas de nombreux producteurs PV distribués dans un réseau de distribution de l'électricité, les méthodes d'apprentissage automatique avec une approche globale peuvent être une stratégie appropriée de la prévision de la production solaire en n'exploitant que les données endogènes.

Conclusion générale et perspectives

Les travaux de recherche élaborés dans cette thèse ont été réalisés dans le cadre d'une convention CIFRE entre le gestionnaire de réseau de distribution de l'électricité SRD et le laboratoire LIAS. Ce projet de thèse développe des méthodes d'estimation et de prévision de la consommation de l'électricité et de la production d'énergie renouvelable du réseau de SRD. Il s'inscrit dans le contexte de recherche et de développement dans le domaine des smartgrids, particulièrement sur la détermination d'un schéma d'exploitation optimal du réseau électrique. Cela permet de mieux intégrer les moyens de production EnR décentralisés en respectant les différentes contraintes techniques et réglementaires de gestion du système électrique. Le but est d'optimiser les flux d'énergie acheminés dans le réseau afin, notamment, de minimiser les pertes par effet Joule dues au transport. La thèse est divisée en trois axes principaux.

Dans un premier axe, une méthodologie de réduction de dimensionnalité des données de consommation du réseau de distribution a été élaborée. Cette approche fondée sur les techniques de regroupement de données (clustering) permet de sélectionner les données les plus pertinentes en éliminant certaines redondances. Elle permet de regrouper l'ensemble des courbes de charge dans des groupes homogènes. Les différents groupes permettent d'identifier des paramètres représentatifs des consommateurs et de détecter des consommateurs atypiques. Ceci a permis ainsi d'évaluer des hypothèses de modélisation de la charge au niveau de l'optimiseur. En effet, l'estimation de la charge par les foisonnements au niveau des postes sources utilisée aujourd'hui dans l'optimisation est moins pertinente que l'estimation par les foisonnements calculés au niveau des groupes homogènes. De plus, un algorithme de clustering Equiwide original est développé dans le cadre de ce projet permettant un clustering fondé sur une erreur maximale autorisée entre les membres de chaque groupe. L'algorithme calcule, en utilisant cette erreur définie par un décideur métier, une solution de clustering respectant cette contrainte de ressemblance minimale entre les différentes données. Cette approche de clustering est une stratégie pertinente de réduction de dimensionnalité des courbes de charge collectées au niveau d'un réseau de distribution de l'électricité, afin de minimiser les temps de calcul et de développer des modèles de prévision pour l'optimisation des réseaux électriques.

Dans un deuxième axe, une méthodologie d'estimation de la production photovoltaïque est développée. Cette approche utilisant les méthodes d'interpolation spatiale permet d'estimer les courbes de production en granularité fine pour tous les producteurs non instrumentés par des compteurs communicants avec un temps de calcul raisonnable. Cette estimation exploite les données disponibles des moyens et grands producteurs et les différentes données de comptage agrégées (biannuelles). La méthode d'interpolation spatiale déterministe IDW (Inverse distance weighting) est la plus précise et transmet assez bien les informations sur les mouvements des nuages captées par les variations de puissance des producteurs voisins. Les résultats montrent un $nRMSE_{\sigma}$ de 19 % obtenu par la méthode IDW, contre 29 % obtenu par la méthode naïve. De plus, la normalisation des données avec la production annualisée offre la meilleure précision d'estimation parmi les approches de normalisation étudiées.

Finalement dans un troisième axe, des approches de prévision de la production solaire ont été proposées. La prévision permet au GRD de bien gérer l'intermittence des producteurs raccordés au réseau (HTA et BT). Les prévisions développées n'utilisent que les données endogènes d'historique de production. D'une part, une prévision court terme est élaborée pour anticiper les contraintes affectant le réseau d'un horizon d'une heure. D'autre part, une prévision long terme est proposée pour une optimisation prédictive du réseau d'un horizon de trois mois. Plusieurs méthodes de prévision ont été évaluées sur les données de production solaire de 621 producteurs raccordés au réseau de SRD avec des stratégies d'utilisation différentes (locales et globales). L'évaluation de ces méthodes a montré que l'algorithme d'apprentissage automatique RF (Random forest) avec une approche globale améliore les prévisions des méthodes naïves dans les deux cas de prévision. Finalement, une validation des différentes méthodes par un estimateur d'état du réseau est réalisée en quantifiant les différences entre les pertes, les chutes et les élévations de tension prévues par les méthodes et la réalité observée dans le réseau.

Perspectives

Les travaux de recherche de cette thèse font partie du projet IMAGE de SRD qui vise l'élaboration d'un schéma d'exploitation optimal du réseau électrique. Dans le court terme, il reste à finaliser la prévision moyen terme de la production photovoltaïque avec une approche ponctuelle et probabiliste d'un horizon de 15 jours. De plus, avec la disponibilité des données météorologiques dans le futur proche pour un horizon allant jusqu'à 15 jours, il sera intéressant de tester la qualité de prévision des modèles exogènes utilisant ces données externes et de les comparer avec les modèles endogènes élaborés dans cette thèse pour les deux horizons court terme et moyen terme.

Par la suite, il faut développer les prévisions de la consommation de l'électricité pour des horizons similaires aux horizons prévus dans le cas de la production solaire. En effet, la prévision de la consommation se fondera sur les résultats obtenus dans le projet clustering (voir chapitre 2). Le clustering a permis de développer des groupes homogènes de la consommation des différents départs HTA. Il reste donc à élaborer une prévision pour chaque cluster de consommation obtenu sur trois horizons temporels : court terme, moyen terme et long terme.

Pour ces mêmes horizons temporels, il faut développer également une prévision de la production éolienne. Les producteurs éoliens de SRD sont tous raccordés au niveau du réseau HTA et leurs données sont collectées avec une granularité temporelle fine. Il ne sera pas donc nécessaire de procéder à une estimation de données comme dans le cas de la production solaire. Les mêmes approches de prévision développées dans cette thèse peuvent être appliquées dans le cas de cette production. Pour une quinzaine de producteurs éoliens du réseau de SRD, l'approche de prévision globale permettra de développer un seul modèle de prévision pour tous les producteurs pour les différents horizons temporels.

Finalement, les différentes approches de clustering et de prévision ont été validées par un estimateur d'état du réseau selon plusieurs scénarios possibles. Il sera judicieux d'appliquer la même méthodologie de simulation en utilisant l'optimiseur. Cette évaluation permettra de déterminer les différences entre les schémas optimaux obtenus selon les scénarios étudiés.

Table des figures

Liste des tableaux

Bibliographie

Académie Française : Dictionnaire d'académie française en ligne, 9ème édition, 2021. URL https://www.dictionnaire-academie.fr/article/A9E0741. (consulté le 30/01/2021).

ADEME : L'électricité solaire, 2021. URL https://www.ademe.fr/sites/default/files/assets/documents/guide-pratique-electricite-solaire.pdf. (Consulté le 2021-02-14).

Aggarwal Charu C : Outlier analysis. *In Data mining*, pages 237–263. Springer, 2015.

Agüero Julio Romero et Steffel Steve J : Integration challenges of photovoltaic distributed generation on power distribution systems. *In 2011 IEEE Power and Energy Society General Meeting*, pages 1–6. IEEE, 2011.

Akima Hiroshi : A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical Software (TOMS)*, 4(2):148–159, 1978.

Ali Zazou Abdelkrim : *Conception d'un outil d'optimisation dynamique du schéma d'exploitation du réseau de distribution d'éléctricité de SRD*. Thèse de doctorat, IASE-ENSMA, Poitiers, avril 2017.

Alsamamra Husain, Ruiz-Arias Jose Antonio, Pozo-Vázquez David et Tovar-Pescador Joaquin : A comparative study of ordinary and residual kriging techniques for mapping global solar radiation over southern spain. *Agricultural and Forest meteorology*, 149(8):1343–1357, 2009.

ANDERSEN Jennie : Adaptation et évaluation d'un algorithme de partitionnement de données fondé sur un diamètre maximum. Stage de fin d'études, Ecole Nationale Supérieure de Mécanique et d'Aérotéchnique, 2020.

Anderson Roger N, Boulanger Albert, Powell Warren B et Scott Warren : Adaptive stochastic control for the smart grid. *Proceedings of the IEEE*, 99(6):1098–1115, 2011.

Antonanzas J., Osorio N., Escobar R., Urraca R., Martinez-de Pison F.J. et Antonanzas-Torres F. : Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016. ISSN 0038092X. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X1630250X.

Arnaud Michel et Emery Xavier : *Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatistiques*. Hermès, 2000.

AUTIXIER Laurène et RONDEAU Nathalie : Processus naturels, albedo, 2021. URL https://www.emse.fr/~bouchardon/enseignement/processus-naturels/up1/web/wiki/MC%20-%20Planeto%20-%20Albedo%20-%20Autixier%20&%20Rondeau.htm. (Consulté le 2021-02-14).

Ball Geoffrey H et Hall David J : A clustering technique for summarizing multivariate data. *Behavioral science*, 12(2):153–155, 1967.

Benkabou Seif-Eddine : *Détection d'anomalies dans les séries temporelles : application aux masses de données sur les pneumatiques*. Thèse de doctorat, Université de Lyon, 2018.

Berndt Donald J et Clifford James : Using dynamic time warping to find patterns in time series. *In KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA :, 1994.

Bessa R.J., Trindade A., Silva Cátia S.P. et Miranda V. : Probabilistic solar power forecasting in smart grids using distributed information. *International Journal of Electrical Power & Energy Systems*, 72:16–23, 2015. ISSN 01420615. URL https://linkinghub.elsevier.com/retrieve/pii/S0142061515000897.

Bivand Roger S., Pebesma Edzer et Gomez-Rubio Virgilio : *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. URL https://asdar-book.org/.

Boisnault C, Brisset J F, Pinget A, Hossenlopp L, Libens H, Souchère C et Electric Schneider : PCCN : de l'étude à la réalisation. page 6.

Bojer Casper Solheim et Meldgaard Jens Peder : Kaggle forecasting competitions : An overlooked learning opportunity. *International Journal of Forecasting*, 2020.

Bouneau Christophe, Jacques LECOUTURIER, Jean-Yves ARZUL, Christophe BOUNEAU, Richard CAZENAVE, Bernard DUCHÊNE, Claude FERNANDEZ, André LAURENT et Jacques PÉRÈS : *Le système nerveux du réseau français de transport d'électricité : 1946-2006 : 60 années de contrôle électrique*. Lavoisier, 2012.

Bouveyron Charles, Bozzi Laurent, Jacques Julien et Jollois François-Xavier : The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 67(4):897–915, 2018.

Breiman Leo : Random forests. *Machine learning*, 45(1):5–32, 2001.

Brockwell Peter J. et Davis Richard A. : *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing, 2016. ISBN 978-3-319-29852-8 978-3-319-29854-2. URL http://link.springer.com/10.1007/978-3-319-29854-2.

Carroll J Douglas et Arabie Phipps : Multidimensional scaling. *Measurement, judgment and decision making*, pages 179–250, 1998.

Chandola Varun, Banerjee Arindam et Kumar Vipin : Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Chen Feng-Wen et Liu Chen-Wuing : Estimation of the spatial rainfall distribution using inverse distance weighting (idw) in the middle of taiwan. *Paddy and Water Environment*, 10(3):209–222, 2012.

Cleveland Robert B et others : Stl : A seasonal-trend decomposition procedure based on loess. 1990. *DOI : citeulike-article-id*, 1435502, 1990.

CRE : Présentation des réseaux d'éléctricité, 2020. URL https://www.cre.fr/Electricite/Reseaux-d-electricite/Presentation-des-reseaux-d-electricite. (Consulté le 2021-01-23).

CRE : Introduction aux smart grids, 2021. URL https://www.smartgrids-cre.fr/introduction-aux-smart-grids. (Consulté le 2021-01-30).

Dagum Estela Bee et Bianconcini Silvia : *Seasonal adjustment methods and real time trend-cycle estimation*. Springer, 2016.

Diday Edwin : Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, 19(2):19–33, 1971.

Doulet Alain : Réseaux de distribution d'électricité - présentation. page 13, 2010.

DOULET Alain et HORSON Jean-Paul : Smart grids : contexte, acteurs et enjeux. (ref. article : d4963), 2019. URL https://www.techniques-ingenieur.fr/base-documentaire/energies-th4/problematiques-communes-des-reseaux-electriques-du-fonctionnement-au-comptage-42266210/smart-grids-contexte-acteurs-et-enjeux-d4963/.

EdfEnR : Lexique : Ombrages, 2021. URL https://www.edfenr.com/lexique/ombrages/. (Consulté le 2021-02-14).

Enedis : Enedis et les ELD | enedis, 2020. URL https://www.enedis.fr/relations-avec-les-eld. (Consulté le 2020-11-23).

Etalab : Licence ouverte / open licence, 2021. URL https://www.etalab.gouv.fr/licence-ouverte-open-licence. (Consulté le 2021-01-27).

Fan Wenfei et Geerts Floris : Foundations of data quality management. *Synthesis Lectures on Data Management*, 4(5):1–217, 2012.

Forgy Edward W : Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

Givord Pauline et d'Haultfoeuille Xavier : La régression quantile en pratique. 2013.

Golestaneh Faranak, Pinson Pierre et Gooi H. B. : Very short-term nonparametric probabilistic forecasting of renewable energy generation— with application to solar energy. *IEEE Transactions on Power Systems*, 31(5):3850–3863, 2019. ISSN 0885-8950, 1558-0679. URL http://ieeexplore.ieee.org/document/7372485/.

Gonzalez-Longatt Francisco M et Rueda José Luis : *PowerFactory applications for power system analysis*. Springer, 2014.

Granell Ramon, Axon Colin J et Wallom David CH : Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. *IEEE Transactions on Power Systems*, 30(6):3217–3224, 2014.

Hastie T., Tibshirani R. et Friedman J.H. : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer series in statistics. Springer series in statistics New York, 2009. ISBN 9780387848846.

Hawkins Douglas M : *Identification of outliers*, volume 11. Springer, 1980.

Hong Tao, Pinson Pierre, Fan Shu, Zareipour Hamidreza, Troccoli Alberto et Hyndman Rob J : Probabilistic energy forecasting : Global energy forecasting competition 2014 and beyond, 2016.

Hossain Monowar, Mekhilef Saad, Danesh Malihe, Olatomiwa Lanre et Shamshirband Shahaboddin : Application of extreme learning machine for short term output power forecasting of three grid-connected pv systems. *journal of Cleaner Production*, 167:395–405, 2017.

Huang Rui, Huang Tiana, Gadh Rajit et Li Na : Solar generation prediction using the arma model in a laboratory-level micro-grid. *In 2012 IEEE third international conference on smart grid communications (SmartGridComm)*, pages 528–533. IEEE, 2012.

Hubert Lawrence et Arabie Phipps : Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Hyndman Rob J et Athanasopoulos George : *Forecasting : principles and practice*. OTexts, 2018.

Hyndman Rob J et Khandakar Yeasmin : Automatic time series forecasting : the forecast package for R. *Journal of Statistical Software*, 26(3):1–22, 2008. URL `http://www.jstatsoft.org/article/view/v027i03`.

IGN : métadonnées de produit contours... IRIS®, 2021. URL `https://geoservices.ign.fr/ressources_documentaires/Espace_documentaire/BASES_VECTORIELLES/CONTOURS_IRIS/IGNF_CONTOURS-IRISr_2-1.html`. (Consulté le 2021-01-27).

IRENA : Renewable power generation costs in 2019, 2021. URL `/publications/2020/Jun/Renewable-Power-Costs-in-2019`. (Consulté le 2021-02-22).

Islam Md Kafiul, Ahammad Tanvir, Pathan Enamul Haq, Mushfiqul ANM et Khandokar Md Rezwanul Haque : Analysis of maximum possible utilization of solar radiation on a solar photovoltaic cell with a proposed model. *International journal of modeling and optimization*, 1(1):66, 2011.

Jamaly Mohammad et Kleissl Jan : Spatiotemporal interpolation and forecast of irradiance data using kriging. *Solar Energy*, 158:407–423, 2017.

Januschowski Tim, Gasthaus Jan, Wang Yuyang, Salinas David, Flunkert Valentin, Bohlke-Schneider Michael et Callot Laurent : Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1):167–177, 2020.

Jaray Olivier, Thomesse Jean-Pierre et Tavella Jean-Philippe : L'INTEROPERABILITE DANS LES POSTES PCCN. page 5.

Joseph V Roshan et Kang Lulu : Regression-based inverse distance weighting with applications to computer experiments. *Technometrics*, 53(3):254–265, 2011.

Keogh Eamonn et Ratanamahatana Chotirat Ann : Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

Kleissl Jan : *Solar energy forecasting and resource assessment*. Academic Press, 2013.

Kohavi Ron et others : A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

Krige Daniel G : A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

Kruskal Joseph B : Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Lauret Philippe, David Mathieu et Pinson Pierre : Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194:254–271, 2019.

Lebart Ludovic, Morineau Alain et Piron Marie : *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris, 1995.

LeDell Erin, Gill Navdeep, Aiello Spencer, Fu Anqi, Candel Arno, Click Cliff, Kraljevic Tom, Nykodym Tomas, Aboyoun Patrick, Kurka Michal et Malohlava Michal : *h2o : R Interface for the 'H2O' Scalable Machine Learning Platform*, 2020. URL `https://CRAN.R-project.org/package=h2o`. R package version 3.30.0.1.

Li Jin et Heap Andrew D : Spatial interpolation methods applied in the environmental sciences : A review. *Environmental Modelling & Software*, 53:173–189, 2014.

Li Zhaoxuan, Rahman Sm, Vega Rolando et Dong Bing : A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. *Energies*, 9(1):55, 2016. ISSN 1996-1073. URL `http://www.mdpi.com/1996-1073/9/1/55`.

Lin Kuo-Ping et Pai Ping-Feng : Solar power output forecasting using evolutionary seasonal decomposition least-square support vector regression. *Journal of Cleaner Production*, 134:456–462, 2016.

Lonij Vincent PA, Brooks Adria E, Cronin Alexander D, Leuthold Michael et Koch Kevin : Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. *Solar Energy*, 97:58–66, 2013.

Lu George Y et Wong David W : An adaptive inverse-distance weighting spatial interpolation technique. *Computers & geosciences*, 34(9):1044–1055, 2008.

MacQueen James et others : Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Madsen Henrik, Pinson Pierre, Kariniotakis George, Nielsen Henrik Aa et Nielsen Torben S : Standardizing the performance evaluation of short-term wind power prediction models. *Wind engineering*, 29 (6):475–489, 2005.

Makridakis S, Spiliotis E et Assimakopoulos V : The m5 accuracy competition : Results, findings and conclusions. *Int J Forecast*, 2020.

Marty Max, Dixneuf Daniel et Gilabert Delphine Garcia : *Principes d'électrotechnique : cours et exercices corrigés*. Dunod, 2005.

Mori Usue, Mendiburu Alexander et Lozano José Antonio : Distance measures for time series in r : The tsdist package. *R J.*, 8(2):451, 2016.

Murtagh Fionn et Legendre Pierre : Ward's hierarchical agglomerative clustering method : which algorithms implement ward's criterion ? *Journal of classification*, 31(3):274–295, 2014.

Nelsen Roger B : *An introduction to copulas*. Springer Science & Business Media, 2007.

Niennattrakul Vit et Ratanamahatana Chotirat Ann : On clustering multimedia time series data using k-means and dynamic time warping. *In 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pages 733–738. IEEE, 2007.

Ordiano Jorge Ángel González : *New Data-Driven Probabilistic Forecasting Methods with Applications in Energy Systems*. Thèse de doctorat, KIT-Bibliothek, 2019.

Park Hae-Sang et Jun Chi-Hyuck : A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

Patton Andrew J : Modelling asymmetric exchange rate dependence. *International economic review*, 47 (2):527–556, 2006.

Pebesma Edzer : Simple Features for R : Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. URL `https://doi.org/10.32614/RJ-2018-009`.

Pinson Pierre, Nielsen Henrik Aa, Møller Jan K, Madsen Henrik et Kariniotakis George N : Non-parametric probabilistic forecasts of wind power : required properties and evaluation. *Wind Energy : An International Journal for Progress and Applications in Wind Power Conversion Technology*, 10 (6):497–516, 2007.

Prinsloo G.J. et Dobson R.T. : *Solar Tracking*. SolarBooks, 2015. ISBN 97890365338671.

Qiao Pengwei, Lei Mei, Yang Sucai, Yang Jun, Guo Guanghui et Zhou Xiaoyong : Comparing ordinary kriging and inverse distance weighting for soil as pollution in beijing. *Environmental Science and Pollution Research*, 25(16):15597–15608, 2018.

R Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Rana Mashud, Koprinska Irena et Agelidis Vassilios G : 2d-interval forecasts for solar power production. *Solar Energy*, 122:191–203, 2015.

Rana Mashud, Koprinska Irena et Agelidis Vassilios G. : Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. *Energy Conversion and Management*, 121, 2016. ISSN 01968904. URL `https://linkinghub.elsevier.com/retrieve/pii/S0196890416303934`.

Rand William M : Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Ratanamahatana Chotirat Ann et Keogh Eamonn : Three myths about dynamic time warping data mining. *In Proceedings of the 2005 SIAM international conference on data mining*, pages 506–510. SIAM, 2005.

Rodríguez-Amigo María del Carmen, Díez-Mediavilla Montserrat, González-Peña David, Pérez-Burgos A et Alonso-Tristán Cristina : Mathematical interpolation methods for spatial estimation of global horizontal irradiation in castilla-león, spain : A case study. *Solar Energy*, 151:14–21, 2017.

Rokach Lior et Maimon Oded : Clustering methods. *In Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

Rousseeuw Peter J : Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

RTE : RTE en bref, 2020. URL `https://www.rte-france.com/rte-en-bref`. (Consulté le 2020-11-23).

Russo M., Leotta G., Pugliatti P.M. et Gigliucci G. : Genetic programming for photovoltaic plant output forecasting. *Solar Energy*, 105:264–273, 2017. ISSN 0038092X. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X14000991.

Sakoe Hiroaki et Chiba Seibi : Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Salinas David, Flunkert Valentin, Gasthaus Jan et Januschowski Tim : Deepar : Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Schoelzel Christian et Friederichs Petra : Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15(5):761–772, 2008.

SDES : Chiffres clés des énergies renouvelables - Édition 2020, 2021a. URL https://www.statistiques.developpement-durable.gouv.fr/chiffres-cles-des-energies-renouvelables-edition-2020. (Consulté le 2021-01-25).

SDES : Données locales de consommation d'énergie, 2021b. URL https://www.statistiques.developpement-durable.gouv.fr/donnees-locales-de-consommation-denergie. (Consulté le 2021-01-27).

SEN ZEKAI et SAHIN AHMET D : Spatial interpolation and estimation of solar irradiation by cumulative semivariograms. *Solar Energy*, 71(1):11–21, 2001.

Shepard Donald : A two-dimensional interpolation function for irregularly-spaced data. *In Proceedings of the 1968 23rd ACM national conference*, pages 517–524, 1968.

Shiskin Julius : *The X-11 variant of the census method II seasonal adjustment program*. Numéro 15. US Department of Commerce, Bureau of the Census, 1967.

Si Caomingzhe, Xu Shenglan, Wan Can, Chen Dawei, Cui Wenkang et Zhao Junhua : Electric load clustering in smart grid : Methodologies, applications, and future trends. *Journal of Modern Power Systems and Clean Energy*, 9(2):237–252, 2021.

SmartGrid.gov : Smart grid : The smart grid | SmartGrid.gov, 2019. URL https://www.smartgrid.gov/the_smart_grid/smart_grid.html. (consulté le 2021-01-30).

Sokal Robert R, Sneath Peter HA et others : Principles of numerical taxonomy. *Principles of numerical taxonomy.*, 1963.

Solargis : Solar resource maps of france, 2021. URL https://solargis.com/maps-and-gis-data/download/france. (Consulté le 2021-02-11).

SRD : Historique | SRD energies, 2020. URL https://www.srd-energies.fr/mieux-nous-connaitre/historique. (Consulté le 2020-11-23).

SRD OpenData : SRD OpenData, 2021. URL https://opendata.srd-energies.fr/pages/page-accueil/. (Consulté le 2021-01-27).

Susanto Ferry, De Souza Paulo et He Jing : Spatiotemporal interpolation for environmental modelling. *Sensors*, 16(8):1245, 2016.

Thurner Leon, Scheidler Alexander, Schäfer Florian, Menke Jan-Hendrik, Dollichon Julian, Meier Friederike, Meinecke Steffen et Braun Martin : pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, 33(6):6510–6521, 2018.

Tureczek Alexander, Nielsen Per Sieverts et Madsen Henrik : Electricity consumption clustering using smart meter data. *Energies*, 11(4):859, 2018.

UE : 2030 climate and energy framework, 2021a. URL `https://ec.europa.eu/clima/policies/strategies/2030_en`. (Consulté le 2020-05-10).

UE : Accord de paris, 2021b. URL `https://ec.europa.eu/clima/policies/international/negotiations/paris_fr`. (Consulté le 2020-05-10).

van der Meer D.W., Widén J. et Munkhammar J. : Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, 81:1484–1512, 2018. ISSN 13640321. URL `https://linkinghub.elsevier.com/retrieve/pii/S1364032117308523`.

Vaz AGR, Elsinga B, Van Sark WGJHM et Brito MC : An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in utrecht, the netherlands. *Renewable Energy*, 85:631–641, 2016.

Vendramin Lucas, Campello Ricardo JGB et Hruschka Eduardo R : Relative clustering validity criteria : A comparative overview. *Statistical analysis and data mining : the ASA data science journal*, 3(4): 209–235, 2010.

Vogt Mike, Marten Frank et Braun Martin : A survey and statistical analysis of smart grid co-simulations. *Applied energy*, 222:67–78, 2018.

Wan C., Zhao J., Song Y., Xu Z., Lin J. et Hu Z. : Photovoltaic and solar power forecasting for smart grid energy management. *CSEE Journal of Power and Energy Systems*, 1(4):38–46, 2015. ISSN 2096-0042. Conference Name : CSEE Journal of Power and Energy Systems.

Wang Xiaoyue, Mueen Abdullah, Ding Hui, Trajcevski Goce, Scheuermann Peter et Keogh Eamonn : Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.

Wen Ruofeng, Torkkola Kari, Narayanaswamy Balakrishnan et Madeka Dhruv : A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv :1711.11053*, 2017.

Wildi Théodore et Sybille Gilbert : *électrotechnique*. De Boeck Supérieur, 2000.

Yang Dazhi, Kleissl Jan, Gueymard Christian A., Pedro Hugo T.C. et Coimbra Carlos F.M. : History and trends in solar irradiance and PV power forecasting : A preliminary assessment and review using text mining. *Solar Energy*, 168:60–101, 2018. ISSN 0038092X. URL `https://linkinghub.elsevier.com/retrieve/pii/S0038092X17310022`.

Yang Shan-lin, Shen Chao et others : A review of electric load classification in smart grid environment. *Renewable and Sustainable Energy Reviews*, 24:103–110, 2013.

Zamo M., Mestre O., Arbogast P. et Pannekoucke O. : A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part i : Deterministic forecast of hourly production. *Solar Energy*, 105:792–803, 2014a. ISSN 0038092X. URL `https://linkinghub.elsevier.com/retrieve/pii/S0038092X13005239`.

Zamo M., Mestre O., Arbogast P. et Pannekoucke O. : A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. part II : Probabilistic forecast of daily production. *Solar Energy*, 105:804–816, 2014b. ISSN 0038092X. URL `https://linkinghub.elsevier.com/retrieve/pii/S0038092X14001601`.

Zhang Jie, Florita Anthony, Hodge Bri-Mathias, Lu Siyuan, Hamann Hendrik F, Banunarayanan Venkat et Brockway Anna M : A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175, 2015.

Zimmerman Ray Daniel, Murillo-Sánchez Carlos Edmundo et Thomas Robert John : Matpower : Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems*, 26(1):12–19, 2010.

Annexe A

Présentation des modèles

A.1 Modèles statistiques

A.1.1 Modèle ARMA

Le modèle à moyenne mobile auto-régressif (Auto-regressive moving average ARMA) est l'un des modèles statistiques les plus connus et populaires dans le domaine d'analyse de séries temporelles. Le modèle ARMA se décompose en deux parties, une partie moyenne mobile (MA) avec un ordre p et une partie auto-régressive (AR) avec un ordre q.

Pour une série Y_t stationnaire, un modèle $ARMA(p,q)$ est exprimé pour chaque t dans T de la manière suivante :

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{A.1}$$

où ε_t est un bruit blanc et où les deux polynômes $(1 - \phi_1 x - ... - \phi_p x^p)$ et $(1 + \theta_1 x + ... - + \theta_q x^q)$ n'ont aucun facteur en commun {Brockwell et Davis, 2016}.

Lorsque $p = 0$ le modèle ARMA se transforme en un modèle $AR(q)$,

$$Y_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

et inversement lorsque $q = 0$ il se transforme en un modèle $MA(p)$.

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i}$$

Le modèle ARMA est un modèle puissant lorsque nous avons une auto-corrélation forte dans la série temporelle. Toutefois, il nécessite la stationnarité de la série temporelle pour être appliqué et la détermination des paramètres p et q est empirique. Pour les séries non stationnaires, un autre modèle a été développé.

A.1.2 Modèle ARIMA

Le modèle à moyenne mobile auto-régressif intégré (autoregressive integrated moving-average) $ARIMA(p,q,d)$ pour une série non stationnaire Y_t s'exprime par

$$Y_t = (1-B)^d X_t \tag{A.2}$$

où X_t est un processus $ARMA(p,q)$, d est un entier non négatif et B est un opérateur de décalage vers l'arrière tel que $B^j Y_t = Y_{t-j}$ et $B^j \varepsilon_t = \varepsilon_{t-j}$ pour tout j dans $\mathbb{N}$.

Le modèle ARIMA est largement utilisé dans les travaux sur la prévision de la production solaire. Il est un outil puissant et très populaire dans la prévision des séries temporelles, il capture bien les cycles périodiques dans les séries.

A.2 Modèles machine learning

Les algorithmes Random forest (RF) et Gradient boosting machine (GBM) utilisés dans la prévision se fondent sur un algorithme appelé arbre de décision.

A.2.1 Arbre de décision

Les modèles fondés sur des arbres sont beaucoup utilisés dans le domaine du Machine Learning en général. Dans ces méthodes nous partitionnons l'espace des variables explicatives sous forme de plusieurs sous-ensembles disjoints. La méthode la plus connue dans cette famille de modèles est la méthode CART (Classification And Regression Tree).

Supposons que nous avons un problème de régression avec une variable cible Y_{t+1} et deux variables explicatives Y_t et I_t. Dans notre cas de prévision solaire par exemple, Y_{t+1} est la production à un instant $t+1$, Y_t est la production à un instant t et I_t est l'irradiation solaire à l'instant t. Notre équation est donc de la forme $Y_{t+1} = f(Y_t, I_t)$. Nous cherchons avec le modèle CART à trouver une fonction $\hat{f}$ approximant la fonction f. Tout d'abord, nous commençons par diviser notre espace en deux régions puis par modéliser la variable cible par la moyenne de Y_{t+1} dans ces deux régions. Dans cette opération, nous choisissons la variable et le seuil de séparation entre les régions en minimisant une certaine erreur d'apprentissage pour avoir un meilleur ajustement. Le choix de variable signifie l'ordre de partitionnement selon l'utilisation des deux variables Y_t et I_t. C'est-à-dire, nous pouvons commencer par diviser Y_t et ensuite diviser I_t ou le contraire nous commençons par I_t et après nous divisons par Y_t. Le seuil de séparation signifie la valeur de la variable explicative à utiliser pour diviser l'espace. Dans la deuxième phase d'algorithme, nous divisons les deux régions en autres deux sous-régions, ainsi de suite jusqu'à ce qu'une condition d'arrêt d'algorithme soit appliquée. La figure A.1 illustre le fonctionnement de cet algorithme pour deux variables explicatives {Hastie *et al.*, 2009}.

D'une manière générale, supposons que nous avons m variables explicatives et une variable cible avec N observations, c'est-à-dire (y_i, x_i) pour tout i dans $[1,N]$ avec $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,m})$. L'algorithme de l'arbre de régression doit décider l'ordre des variables et les seuils de séparations dans le partitionnement de l'espace. Pour K régions différentes, nous obtenons une modélisation générale de la variable cible Y.

$$y = f(x) = \sum_{k=1}^{K} c_k \mathbb{1}\{x \in R_k\} \tag{A.3}$$

Si nous fixons par exemple l'erreur quadratique $\sum_{i=1}^{N} e_i^2$ comme un critère de minimisation, le c_k optimal est tout simplement la moyenne de y_k dans la région R_k.

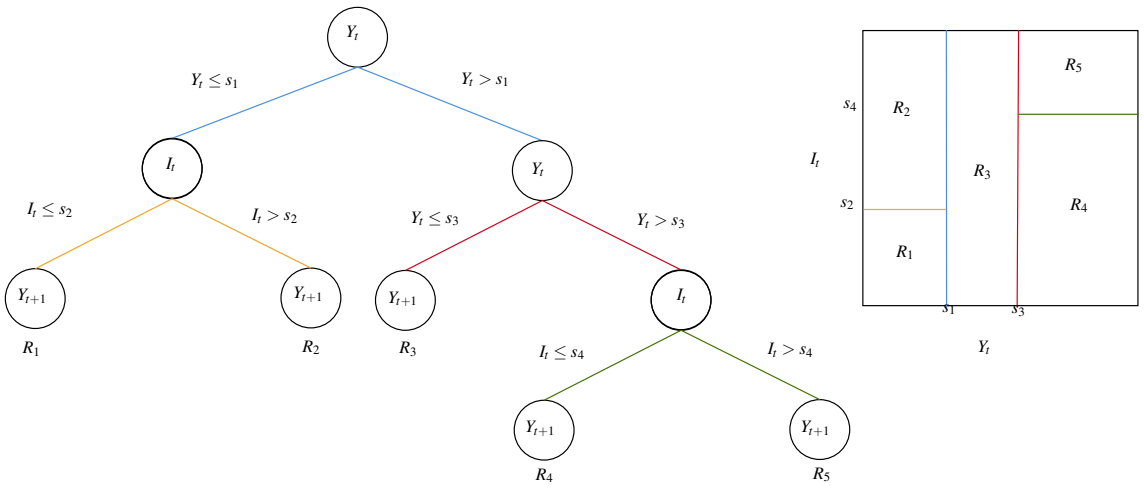


FIGURE A.1 – Exemple de l'arbre de décision (Adapté de {Hastie *et al.*, 2009})

$$\hat{c_k} = \frac{1}{D_k} \sum_{l=1}^{D_k} (y_l | x_l \in R_k) \tag{A.4}$$

Avec D_k égale au cardinal de l'ensemble $\{x \in R_k\}$

Cependant la détermination d'une partition optimale de l'espace des variables explicatives en minimisant l'erreur quadratique est généralement un problème NP-difficile, c'est-à-dire qu'il est irréalisable en calcul vu le nombre de solutions faisables possibles. Nous trouvons dans la littérature plusieurs algorithmes heuristiques calculant une partition selon plusieurs paramètres comme la taille de l'arbre, le nombre de racines, un seuil minimum, un temps de calcul maximum, etc. Pour plus d'informations théoriques, se référer au livre {Hastie *et al.*, 2009}.

A.2.2 Gradient boosting machine (GBM)

Le Boosting est une technique d'apprentissage très populaire dans le domaine d'apprentissage machine. Cette technique a été conçue au début sur des problèmes de classification, mais elle peut être utilisée ainsi dans des problèmes de régression. La motivation derrière son développement est de construire une procédure combinant les prédictions de plusieurs modèles « faibles » pour produire un modèle plus précis. Nous avons vu précédemment que les arbres de décision divisent l'espace de toutes les variables explicatives en multiples régions R_j, j dans $\{1,2,...,J\}$. Ces régions sont présentées par les nœuds terminaux des arbres. La prédiction est donc une constante λ_j affectée à chacune de ces régions R_j {Hastie *et al.*, 2009}. C'est-à-dire que si un x appartient à une région R_j, cela implique que $f(x) = \lambda_j$. Un arbre peut être formulé de la manière suivante

$$A(x,\Theta) = \sum_{j=1}^{J} \lambda_j \mathbb{1}(x \in R_j) \tag{A.5}$$

Où Θ présente les différents paramètres de l'arbre, $\{(R_1,\lambda_1),...,(R_J,\lambda_J)\}$. J est un méta-paramètre de l'algorithme. Les paramètres optimaux sont trouvés en minimisant le risque empirique

$$\hat{\Theta} = arg \min_{\Theta} \sum_{j=1}^{J} \sum_{x_i \in R_j} L(y_i, \lambda_j) \tag{A.6}$$

avec L la fonction objectif à minimiser.

Nous formulons notre problème ici par un problème d'optimisation combinatoire qui est généralement NP-difficile. Nous nous contentons donc de trouver des solutions sous-optimales approximatives via des algorithmes heuristiques.

Parmi ces algorithmes heuristiques, nous trouvons l'algorithme GBM qui est beaucoup utilisé dans la communauté d'apprentissage machine.

Nous formulons notre problème décrit en équation (A.6) par une équation récursive

$$\hat{\Theta} = arg \min_{\Theta_m in} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i, \Theta_m)) \tag{A.7}$$

Pour résoudre cette équation, nous pouvons l'approximer par les méthodes issues de l'optimisation numérique. L'erreur de la prédiction de $f(x)$ sur les données de l'apprentissage est la suivante

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{A.8}$$

Notre objectif est de minimiser la fonction L pour une fonction f donnée. Comme nous avons défini dans (A.3), la fonction f est contrainte d'être une somme de plusieurs arbres. Si nous ignorons cette contrainte, la minimisation de l'équation (A.8) peut être formulée comme une optimisation numérique d'une fonction

$$\hat{f} = arg \min_{f} L(f) \tag{A.9}$$

avec $f = \{f(x_1), f(x_2), ..., f(x_N)\}$ un vecteur de valeurs de la fonction f pour chaque point de données x_i, $i \in \{1, 2, ..., N\}$.

La solution donnée par l'optimisation numérique de l'équation (A.9) est de la forme

$$f_M = \sum_{m=0}^{M} h_m \tag{A.10}$$

Où h_m un vecteur dans $\mathbb{R}^N$.

$f_0 = h_0$ est une estimation initiale et chaque fonction f_m successive se fonde sur la base du vecteur de paramètres f_{m-1} obtenu à partir des vecteurs précédents. Il existe plusieurs méthodes d'optimisation numérique pour résoudre cette équation. Parmi ces méthodes, nous trouvons l'algorithme du gradient. Dans cet algorithme, nous choisissons $h_m = \rho_m g_m$ avec ρ_m un scalaire et g_m un vecteur dans $\mathbb{R}^N$ le gradient de la fonction $L(f)$ évalué à l'étape $f = f_{m-1}$. Les composants i du vecteur g_m se définissent par

$$g_{i,m} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x_i)=f_{m-1}(x_i)} \tag{A.11}$$

avec i dans $\{1, 2, ..., N\}$.

Le scalaire ρ_m s'appelle la longueur du pas. Il est donné par

$$\rho_m = arg \min_{\rho} L(f_{m-1} - \rho g_m) \tag{A.12}$$

La solution actuelle de notre problème d'optimisation est mise à jour par

$$f_m = f_{m-1} - \rho_m g_m \tag{A.13}$$

Le processus se répète jusqu'à ce qu'une condition d'arrêt soit atteinte.

L'algorithme 1 présente le fonctionnement général de l'algorithme GBM. La première étape de l'algorithme est d'initialiser le modèle par une constante, cette constante est un arbre composé d'un seul nœud terminal. Les composantes négatives $-r_{i,m}$ du gradient sont appelées des résidus généralisés ou des pseudo-résidus {Hastie *et al.*, 2009}.

Algorithm 1: Algorithme gradient boosting machine pour régression {Hastie *et al.*, 2009}

1 : Initialiser $f_0(x) = arg\ min_\lambda \sum_{i=1}^{N} L(y_i, \lambda)$

2 : Pour $m = 1$ jusqu'à M :

(a) Pour i dans $\{1, 2, ..., N\}$ calculer :

$$r_{i,m} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

(b) Entraîner un arbre de régression pour les cibles $r_{i,m}$ donnant les régions $R_{j,m}$, $j \in \{1, 2, ..., J_m\}$.

(c) Pour j dans $\{1, 2, ..., J_m\}$ calculer

$$\lambda_{j,m} = arg \min_{\lambda} \sum_{x_i \in R_{j,m}} L(y_i, f_{m-1}(x_i) + \lambda)$$

(d) Mettre à jour $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \lambda_{j,m} \mathbb{1}(x \in R_{j,m})$

3 : Retourner $\hat{f}(x) = f_M(x)$

A.2.3 Random forest (RF)

Dans le domaine de la statistique, le Boostrap est une méthode d'estimation des paramètres d'une population en utilisant un échantillonnage répété avec remplacement dans les données observées de cette population. Dans le domaine de machine learning, il est utilisé comme une méthode d'évaluation des incertitudes et de la précision de l'estimation d'un paramètre ou d'une prédiction. Le Boostrap peut être utilisé avec d'autres méthodes comme la méthode de Bagging pour améliorer la qualité de prédiction d'un modèle. Pour expliquer la méthode de Bagging, considérons par exemple un problème de régression. Nous souhaitons faire apprendre notre modèle sur un ensemble de données $Z = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, pour obtenir une fonction $\hat{f}$ à partir des données explicatives X. L'agrégation du Bootstrap ou le Bagging calcule une moyenne de la prédiction fournie par $\hat{f}$ sur un ensemble d'échantillons Bootsrap, ce qui diminue la variance du modèle. Par exemple, pour chaque échantillon Bootsrap Z^{*j} , $j = 1, 2, ..., B$, nous entraînons notre modèle et nous obtenu une prédiction $\hat{h}^{*j}$. L'estimation du Bagging se définit par

$$\hat{f}_{bag} = \frac{1}{B} \sum_{j=1}^{B} \hat{h}^{*j} \tag{A.14}$$

Un bon exemple d'utilisation de cette méthode est l'arbre de régression. Supposons que nous avons un modèle CART avec une fonction $\hat{f}$ et des données d'entrées X. Nous réalisons sur les données d'apprentissage B échantillons boostrap. Nous entraînons ensuite un arbre pour chaque échantillon. Cet arbre a des caractéristiques différentes des autres arbres et de l'arbre original. Par exemple, il peut avoir un nombre différent de nœuds, un ordre différent des variables, etc. L'estimation Bagging est donc la prédiction moyenne des données d'entrée X à partir de ces B arbres.

Les forêts aléatoires (Breiman {2001}) constituent une modification considérable de Bagging en construisant une grande collection d'arbres non corrélés, afin de calculer la moyenne de prédiction fournie par ces arbres. Dans le Bagging nous calculons la moyenne de nombreux modèles bruités et approximativement non biaisés. L'idée de Random forests (voir l'algorithme 2) est d'améliorer la réduction de la variance de Bagging en réduisant la corrélation entre les différents arbres, sans trop augmenter la variance. Ceci est réalisé dans un processus de croissance d'arbres par une sélection aléatoire de variables explicatives d'entrée. C'est-à-dire, lors de la création d'un arbre sur un échantillon Boostrap, avant chaque partitionnement de l'arbre, nous sélectionnons un nombre K de variables inférieur au nombre total M de variables explicatives disponibles. Après un nombre B d'arbres créés, la prédiction de l'arbre de décision égale

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{j=1}^{B} A(x,\Theta_j) \tag{A.15}$$

Avec $A(x,\theta_j)$ décrit le $j$ème arbre dans la forêt aléatoire en termes de variables partitionnées, les seuils de séparation de chaque nœud et les valeurs des nœuds terminaux dans l'arbre {Hastie *et al.*, 2009}. L'algorithme 2 présente le fonctionnement global de l'algorithme RF.

Algorithm 2: Algorithme Random forest pour régression {Hastie *et al.*, 2009}

B : Nombre d'arbres
n_{min} : Taille minimale de nœuds
K : Nombre de variables dans un arbre
M : Nombre total de variables explicatives
1 : Pour $j = 1$ jusqu'à B
- (a) Tirer un échantillon bootstrap Z^* de taille N à partir de données d'apprentissage.
- (b) Cultiver un arbre de forêt aléatoire A_j aux données Z^* bootstrapées, en répétant récursivement les étapes suivantes pour chaque nœud terminal de l'arbre, jusqu'à ce que n_{min} soit atteint.
  - i. Sélectionner aléatoirement K de variables à partir des M variables explicatives.
  - ii. Choisir la meilleure variable/le meilleur seuil de séparation parmi les m variables.
  - iii. Partitionner le nœud en deux nœuds fils.

2 : Retourner l'ensemble d'arbres $\{A_1, A_2, ..., A_B\}$
Prédiction : $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{j=1}^{B} A_j(x)$

A.2.4 L'importance des variables explicatives dans le processus de décision

Dans les deux algorithmes GBM et RF, l'importance des variables explicatives est déterminée en calculant l'influence de chaque variable dans le processus de décision. Cette influence pour une variable explicative donnée est caractérisée par deux points : si cette variable a été sélectionnée pour être divisée dans la construction des arbres et de combien l'erreur de prévision sur tous les arbres s'est améliorée. Lorsqu'un nœud dans un arbre est divisé sur la base d'une spécificité numérique ou catégorielle, l'erreur réduite est la différence d'erreur entre ce nœud et ses nœuds enfants.

Résumé

Les gestionnaires de réseaux de distribution (GRD) d'électricité ont connu ces dernières années une intégration importante des moyens de production d'énergie renouvelable (EnR). De plus, avec l'apparition de nouveaux usages de l'énergie, notamment la mobilité électrique, les microgrids et les technologies de stockage, la gestion du réseau est devenue de plus en plus complexe, complexité qui ira croissante dans les années à venir. Dans ce contexte, SRD, GRD dans le département de la Vienne, a financé des travaux de recherche dans le domaine des smartgrids, notamment sur l'optimisation dynamique du schéma d'exploitation de son réseau de distribution d'électricité haute et moyenne tension, et cette thèse sur la prévision, la prédiction et l'estimation des valeurs des flux énergétiques circulant sur ce réseau. Dans une première phase, la thèse propose une approche de sélection des données de la consommation d'énergie les plus pertinentes et étudie leur influence sur l'efficacité de l'optimisation. Pour cela une méthodologie de réduction de dimensionnalité de données est proposée. Elle utilise des techniques issues du domaine de l'apprentissage automatique non supervisé et d'analyse de données temporelles. Cette méthodologie permet de détecter des similitudes dans les données afin de les regrouper dans des groupes homogènes. La seconde phase élabore une méthodologie d'estimation de la production d'énergie des installations photovoltaïques (PV) distribuées dans le réseau de distribution de SRD en utilisant les méthodes d'interpolation spatiale. En effet, la plupart des producteurs avec des capacités de production basse tension ne sont pas instrumentés pour une mesure en temps réel. Au contraire, les producteurs moyenne et haute tension sont instrumentés et permettent des mesures à grain fin sous forme de séries temporelles. Le but de cette étude est d'estimer les productions de milliers de petits producteurs en exploitant les données des moyens et gros producteurs de référence équipés de compteurs communicants. Finalement, le problème de la prévision de la production photovoltaïque est abordé. Le but de cette étude est d'élaborer une prévision ponctuelle court terme d'un horizon d'une heure pour gérer l'intermittence de la production solaire et une prévision probabiliste long terme pour planifier et optimiser le réseau sur un horizon d'un à trois mois. Nous montrons que les algorithmes d'apprentissage automatique avec une approche globale améliorent les prévisions fournies par des méthodes naïves. La pertinence des prévisions obtenues par rapport au cadre applicatif a été validée à l'aide d'un estimateur d'état du réseau pour quantifier les différences de pertes, de chutes de tension et d'élévation de tension entre un état prévu et un état réel.

Mots-clés : Apprentissage automatique ; Électricité–Production–Prévision ; Réseaux électriques intelligents ; Séries chronologiques ; Intelligence artificielle appliquée à l'énergie

---

Abstract

In recent years, electricity distribution system operators (DSOs) have seen a significant integration of renewable energy production. In addition, with the development of new energy uses, such as electric mobility, microgrids and storage technologies, grid management has become more and more complex, a complexity that will increase in the years to come. In this context, SRD, DSO in the department of Vienne in France, has funded research in the field of smartgrids, notably on the dynamic optimization of the operating scheme of its high and medium voltage electricity distribution network, and this thesis on the forecasting, prediction and estimation of the values of the energy flows circulating on this network. In the first phase, the thesis proposes an approach to select the most relevant energy consumption data and study their influence on the efficiency of the optimization. For this purpose, a data dimensionality reduction approach is proposed. It uses techniques from the unsupervised machine learning and temporal data analysis fields. This methodology allows to detect similarities in the data in order to group them in homogeneous groups. The second phase develops a methodology for estimating the energy production of photovoltaic installations distributed over the SRD distribution network using spatial interpolation methods. In fact, most of the generators with low voltage production capacities are not instrumented for real time measurement. On the other hand, medium and high voltage generators are instrumented and allow fine-grained time series measurements. The goal of this study is to estimate the production of thousands of small generators by exploiting the data of medium and large generators of reference equipped with communicating meters. Finally, the problem of forecasting photovoltaic production is considered. The goal of this study is to develop a short-term point forecast with a horizon of one hour to manage the intermittency of the solar production and a long-term probabilistic forecast to schedule and optimize the network over a horizon of one to three months. We show that machine learning algorithms with a global approach improve the forecasts provided by naive methods. The pertinence of the obtained forecasts to the application framework has been validated using a loadflow estimator to quantify the differences in losses, voltage drops and voltage rises between a forecasted state and an actual state.

Keywords : Machine learning ; Electric power production–Forecasting ; Smart power grids ; Time-series analysis ; Artificial intelligence applied in energy

---

Secteur de recherche : Informatique et applications

LABORATOIRE D'INFORMATIQUE ET D'AUTOMATIQUE POUR LES SYSTÈMES
Ecole Nationale Supérieure de Mécanique et d'Aérotechnique
Téléport 2 – 1 avenue Clément Ader – BP 40109 – 86961 FUTUROSCOPE CHASSENEUIL CEDEX
Tél : 05.49.49.80.63 – Fax : 05.49.49.80.64