Traitement des données de consommation d'électricité acquis d'un dispositif autonome de mesure préalablement conçu en vue de mieux gérer la demande

Par

Mouhammad Moustapha KONÉ

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE
LA MAÎTRISE AVEC MÉMOIRE EN GÉNIE - ÉNERGIES
RENOUVELABLES ET EFFICACITÉ ÉNERGÉTIQUE
M. Sc. A.

MONTRÉAL, LE 27 FÉVRIER 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Julio Montecinos, directeur de mémoire
Département du génie des systèmes à l'École de technologie supérieure

M. Tony Wong, codirecteur de mémoire
Département du génie des systèmes à l'École de technologie supérieure

M. Jean-Philippe Roberge, président du jury
Département de génie des systèmes à l'École de technologie supérieure

M. Georges Ghazi, membre du jury
Département de génie des systèmes à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 17 DÉCEMBRE 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à adresser mes plus sincères remerciements à toutes les personnes qui ont joué un rôle clé dans la réalisation de ce mémoire, un projet qui m'a profondément marqué et enrichi en dépit des épreuves rencontrées. À mon père, dont le soutien indéfectible et la volonté de me voir réussir sont une source d'inspiration constante, me rappelant l'importance de la persévérance face aux défis. À ma femme, dont l'amour, la patience et les encouragements quotidiens ont été un pilier essentiel, m'offrant un précieux équilibre dans les moments de stress et de doute. À ma mère, dont la foi inébranlable en mes capacités et les mots réconfortants m'ont donné la force de poursuivre, même lorsque la tâche semblait insurmontable. À mes professeurs (Julio Montecinos et Tony Wong), dont les enseignements rigoureux, les retours détaillés et l'accompagnement personnalisé ont non seulement façonné mon travail, mais aussi ma manière d'aborder la recherche avec rigueur et curiosité. Leur expertise a été un guide précieux tout au long de ce parcours académique.

Je souhaite également exprimer ma gratitude envers mes collègues et amis, qui ont partagé leurs idées et leurs expériences, enrichissant mes réflexions sur les thématiques abordées. Leur apport a été une source de motivation et d'ouverture d'esprit. À tous, je dois la réussite de ce projet, et je suis profondément reconnaissant pour votre présence et votre soutien. Merci du fond du cœur.

Traitement des données de consommation d'électricité acquis d'un dispositif autonome de mesure préalablement conçu en vue de mieux gérer la demande

Mouhammad Moustapha KONÉ

RÉSUMÉ

Ce mémoire s'attache à explorer en profondeur l'analyse et la gestion des données de consommation d'électricité d'un condominium doté de bornes de recharge pour véhicules électriques. Dans le premier chapitre, la revue offre une synthèse des fondements théoriques liés aux séries temporelles, à la classification des données et à la détection d'anomalies, en exposant une diversité d'approches méthodologiques telles que l'utilisation d'auto-encodeurs, de l'algorithme DBSCAN ou encore du LSTM, qui se révèlent pertinentes dans le contexte de la consommation énergétique. Le chapitre consacré à la méthodologie présente une approche structurée, commençant par l'enrichissement des données avec des caractéristiques, suivie de la détection d'anomalies à l'aide de techniques avancées telles que MDWS, des auto-encodeurs, DBSCAN et la forêt d'isolement, puis de leur correction via la méthode des k plus proches voisins. Cette méthodologie traite également des prévisions à l'aide de modèles tels que LSTM, SARIMA et l'empilement de modèles, ce qui permet de répondre aux enjeux énergétiques. L'étude de cas, développée dans un chapitre ultérieur, porte sur l'examen minutieux des données issues de trois unités distinctes du condominium. Les résultats, présentés et discutés dans un chapitre dédié, soulignent les performances des techniques de détection et de prédiction, en reconnaissant en particulier l'efficacité de certaines méthodes pour capturer les irrégularités et anticiper les tendances de consommation. Les discussions qui s'ensuivent explorent l'influence des véhicules électriques sur les profils de consommation, estimée comme étant une part prédominante dans les pics d'utilisation, et ouvrent la voie à des recommandations stratégiques telles que l'instauration de tarifications dynamiques adaptées aux heures de forte demande, ainsi que l'expansion réfléchie des infrastructures de recharge pour accompagner la transition énergétique de manière durable et équilibrée.

Mots-clés : Énergie électrique, véhicules électriques, analyse de données, apprentissage automatique, séries temporelles, gestion de la demande

Processing of Electricity Consumption Data Obtained from a Pre-Designed Autonomous Measurement Device for Improved Demand Management

Mouhammad Moustapha KONÉ

ABSTRACT

This document provides an in-depth analysis and management of electricity consumption data collected by an autonomous device within a condominium equipped with electric vehicle (EV) charging stations. The literature review, covered in the first chapter, provides a comprehensive synthesis of the theoretical foundations related to time series, data classification, and anomaly detection, spotlighting a wide array of methodological approaches such as the use of auto-encoders, the DBSCAN algorithm, and LSTM neural networks, which prove particularly relevant in the context of energy consumption. The methodology chapter outlines a structured and rigorous approach: beginning with data enrichment using contextual variables, followed by advanced anomaly detection techniques, including MDWS, auto-encoders, DBSCAN, and Isolation Forest, and subsequent correction using the K-nearest neighbor's method. This methodology further extends to forecasting through models such as LSTM, SARIMA, and stacking, providing a holistic framework to address energy-related challenges. The case study, detailed in a subsequent chapter, focuses on a meticulous examination of data from three distinct condominium units. The results, presented and discussed in a dedicated chapter, highlight the varied performance of different detection and prediction techniques, with recognition of the effectiveness of specific methods in capturing irregularities and anticipating consumption trends. The ensuing discussions explore the profound influence of electric vehicles on consumption profiles, which are identified as a dominant factor in usage peaks, and pave the way for strategic recommendations, such as implementing dynamic pricing tailored to peak-demand hours, alongside a thoughtful expansion of charging infrastructure to support a sustainable and balanced energy transition.

Keywords: Electrical power, electrical vehicles, data analysis, machine learning, time series, demand management

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AMI	Automated Metering Infrastructure (Infrastructure de mesure automatisée)
AMMMO	Adaptive Multi-Model Middle-Out (Multimodèle adaptatif Intermédiaire)
ARIMA	AutoRegressive Integrated Moving Average (Moyenne mobile autorégressive intégrée)
DBA	DTW Barycenter Average (Barycentre moyen du DTW)
DTW	Déformation temporelle dynamique - Dynamic Time Warping
FFD	Full Function Devices (les appareils à fonction complète)
ILM	Intrusive Load Monitoring (surveillance intrusive de la charge)
IoT	Internet of Things (internet des objets)
KNN	k-Nearest Neighbors (k plus proches voisins)
LSTM	Long Short-Term Memory (Longue mémoire à court terme)
LZ	Famille d'algorithmes de compression basée sur les travaux de Lempel et Ziv
MAE	Mean Absolute Error (Erreur moyenne absolue)
MAPE	Mean Absolute Percentage Error (Pourcentage de l'erreur moyenne absolue)
NAR	Nonlinear autoregressive neural network (réseau de neurones autorégressif non linéaire)
OS	Operating System (système d'exploitation)
RFD	Reduced Function Devices (les appareils à fonction réduite)
RMSE	Root Mean Squared Error (Racine de l'erreur quadratique moyen)
RTC	Real-time clock (Horloge temps réel)
SVM	Support Vector Machine (Machine à vecteurs de support)
SVR	Support Vector Regression

XGBoost	eXtreme Gradient Boosting
XOR	Or exclusif

LISTE DES SYMBOLES ET UNITÉS DE MESURE

h	heures
min	minutes
s	secondes
W	Watt
Wh	Watt-heure

INTRODUCTION

Le Québec, et même le Canada, assistent depuis quelques années à un boom du nombre de véhicules électriques. Selon les statistiques fournies par l'AVÉQ (2025), la variation moyenne du nombre de véhicules au cours des 5 dernières années se situe entre 30 % et 50 % par an. De même, le nombre de bornes de recharge augmente proportionnellement à celui des véhicules électriques. Cette observation suscite une attention particulière chez les acteurs du secteur énergétique. En effet, il est essentiel d'analyser l'impact des modifications apportées au réseau de distribution afin de mieux accompagner la transition énergétique en cours. Parmi les facteurs à considérer figure notamment la présence croissante de bornes de recharge dans nos villes, conséquence directe de l'augmentation rapide du nombre de véhicules électriques.

Dans un contexte de multilogement, l'installation d'un dispositif de mesure de la consommation d'électricité est envisageable lorsque la borne de recharge n'est ni partagée ni intégrée à des installations communes. Dans le cadre de ce projet, un tel dispositif a été mis en place dans un condominium doté de bornes de recharge dédiées. Les relevés effectués, tant dans les logements que dans les zones de stationnement équipées pour la recharge des véhicules électriques, génèrent un volume important de données dont l'absence d'exploitation constituerait une perte.

Cela soulève plusieurs questions :

- Comment traiter ces flux massifs de données ?
- Comment gérer les situations atypiques ou irrégulières ?
- Quels bénéfices concrets peut-on en tirer ?

Ces questions constituent la problématique centrale de ce mémoire.

De nombreuses études ont examiné séparément la consommation d'électricité résidentielle dont Brownlee (2021) ou Boudhaouia(2022) et celle liée à la recharge des véhicules électriques (Almaghrebi et al., 2024). Ce mémoire adopte une approche intégrée, analysant simultanément ces deux aspects afin de mieux comprendre le comportement énergétique global d'un ménage ou d'un individu dans ce contexte.

Les questions spécifiques de cette recherche sont :

1. Comment détecter et corriger efficacement les anomalies dans les données de consommation d'énergie issues de domiciles et de bornes de recharge pour VE ?
2. Quels modèles prédictifs sont les plus adaptés pour anticiper la demande énergétique liée à la recharge des VE en fonction des habitudes de consommation ?

Objectif principal de recherche

Cette étude consiste principalement à exploiter les données de consommation énergétique générées dans la situation telle que définie précédemment, afin de mieux gérer le réseau de distribution électrique dans un contexte de transition énergétique, où naissent de nouveaux producteurs parmi les habituels consommateurs. En particulier, le consommateur/producteur propriétaire de véhicules électriques a, lui aussi, besoin de prendre des décisions pertinentes.

Il s'agit, entre autres, de mettre en valeur les données collectées à l'aide de méthodes de traitement de données. La prédiction, à court et à moyen terme, de la demande de puissance nécessaire à la recharge des véhicules électriques aidera à planifier efficacement le délestage en période de pointe. Les modèles d'apprentissage et les techniques de traitement utilisés seront basés sur les habitudes de consommation, les données météorologiques, les caractéristiques liées aux saisons et aux jours de la semaine. Ainsi, devient-il possible de prévoir les plages horaires susceptibles de recevoir la charge déplacée (Investissement Québec, 2023)? Nous nous intéressons donc à comprendre comment mettre en valeur les données des utilisateurs afin de mieux gérer la demande.

Sous-objectifs :

- Analyse préliminaire et descriptive : Identifier les caractéristiques de consommation électrique d'une unité et extraire des informations utiles pour l'ajustement des techniques d'analyse de données.

- Identification des Irrégularités : Certaines irrégularités peuvent également intervenir du fait d'événements inhabituels ou même de la saison, dont l'impact peut être considérable sur la consommation d'un domicile où l'on a un véhicule doté d'une borne de recharge. Nous allons également proposer une méthode pour les identifier et corriger les données traitées.
- Prédiction de la consommation : Nous allons enfin présenter une technique pour estimer la consommation future. Cette démarche nous conduira à disposer des outils nécessaires pour répondre à notre problématique.

Importance de l'étude

Cette étude permettra de mettre en exergue l'intérêt de ces données pour la gestion de l'énergie, dans un contexte changeant. Face aux différents types de consommateurs énergivores et aux besoins croissants, difficiles à couvrir entièrement, les résultats de cette étude permettront de faire référence à la consommation des véhicules électriques. La conjonction des données, ici, en ne tenant compte que de la recharge effectuée à la maison, permettra de distinguer les habitudes normales de ce qui pourrait sembler anormal. On pourrait ainsi, à l'aide de cette étude, automatiser plus facilement la gestion de nos bâtiments en matière de consommation d'énergie, voire celle de nos villes.

Organisation du mémoire

Ce mémoire est structuré en 4 principaux chapitres :

- Chapitre 1 : la revue de la littérature. Elle consistera à passer en revue, de manière non exhaustive, les études déjà menées sur le sujet à l'aide d'outils pertinents pour l'acquisition et le traitement de données.
- Chapitre 2 : la méthodologie. Il s'agira de présenter la suite d'outils que l'on propose pour élucider la problématique énoncée.
- Chapitre 3 : l'étude de cas. Il y sera présenté le cas en amont du traitement des données. En effet, il s'agira de présenter les technologies utilisées pour acquérir les données, ainsi que les premières observations que l'on en tire.

- Chapitre 4 : les résultats. Ce chapitre présente les résultats obtenus après l'application de la méthodologie. Des comparaisons sont ainsi réalisées à l'aide d'indicateurs clés de performance, et ces résultats sont discutés afin d'en faire une analyse plus approfondie et de présenter des implications pour le secteur de l'énergie électrique.

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

Ce chapitre présente une synthèse des concepts fondamentaux liés à l'analyse et au traitement des séries temporelles, notamment la classification des données et la détection d'anomalies, en mettant l'accent sur leur application à la consommation d'électricité. Ces thématiques, centrales dans le cadre de ce mémoire, permettent d'explorer les outils mathématiques et les techniques d'apprentissage automatique utilisés pour modéliser, analyser et prévoir des données complexes. La revue s'appuie sur des travaux de référence pour illustrer les approches théoriques et pratiques, tout en mettant en lumière leur pertinence dans des contextes réels tels que la gestion énergétique.

Le premier axe aborde les séries temporelles, décomposées en tendance, saisonnalité et bruit, ainsi que leur modélisation pour la prévision de la consommation électrique. Le second axe traite de la classification des séries, en examinant les méthodes de partitionnement, hiérarchiques ou basées sur des distances, telles que la déformation temporelle dynamique (en anglais : Dynamic Time Warping—DTW), qui facilitent la simplification et l'analyse des données. Enfin, le dernier axe explore la détection d'anomalies, en détaillant les types d'anomalies (ponctuelles, contextuelles, collectives) ainsi que les approches statistiques ou d'apprentissage automatique pour les identifier. Ces éléments posent les bases théoriques et méthodologiques nécessaires à l'étude des données de consommation d'électricité.

Cette section explore les principales techniques de détection d'anomalies et de prévision appliquées aux séries temporelles, notamment à la consommation électrique. Elle couvre des approches variées, allant des méthodes d'apprentissage automatique non supervisées, comme la forêt d'isolement et les auto-encodeurs, aux techniques supervisées basées sur des modèles comme la moyenne mobile autorégressive intégrée (en anglais : AutoRegressive Integrated Moving Average—ARIMA) ou la machine à vecteurs de support (en anglais : Support Vector

Machine—SVM), en passant par des réseaux neuronaux avancés tels que les réseaux de neurones de type longue mémoire à court terme (en anglais : Long Short-Term Memory—LSTM). Le clustering, utilisé pour identifier des regroupements ou des anomalies, ainsi que l'empilement de modèles pour améliorer la précision des prévisions, sont également abordés. Enfin, l'imputation des données et les métriques de performance sont présentées comme des outils essentiels pour traiter les données manquantes et évaluer la qualité des modèles, offrant une vue d'ensemble des stratégies disponibles pour une analyse robuste des séries temporelles.

1.2 Séries temporelles

À l'instar de (Hyndman et al., 2018), une série temporelle est généralement présentée comme constituée de trois éléments fondamentaux :

- La tendance : c'est l'allure de la série tout au long de son évolution. Elle peut être linéaire ou non, croissante ou décroissante.
- La saisonnalité : C'est une tendance similaire observée de façon périodique. Elle peut être journalière, hebdomadaire, mensuelle etc.
- Le bruit : C'est une perturbation observée lors de l'analyse de la série et qui ne peut être prévue par le modèle.

Pour une série temporelle x_t, on l'exprime sous la forme :

$$x_t = m_t + s_t + z_t \tag{1.1}$$

où m_t, s_t et z_t représentent la tendance, la saisonnalité et le bruit aléatoire. La mise sous forme de séries temporelles des données, notamment celles de consommation d'électricité, permet l'application de plusieurs outils mathématiques, favorisant des analyses et des traitements approfondis. On parle de Popov et al. (2020), qui évoquent la prévision de la consommation d'électricité à partir de séries chronologiques issues de données incertaines. Sa méthodologie consiste à emprisonner les valeurs entre un maximum et un minimum prédéfinis afin de détecter celles qui sont incertaines, et à effectuer une fuzzification des données afin de distinguer, dans des ensembles flous, celles qui sont extrêmement petites, très petites,

moyennes, au-dessus de la moyenne et extrêmement grandes. Ceci permettrait d'obtenir une forme « logique » des données qu'il utilise ensuite pour établir une prévision de plus haute précision. On pourrait citer des dizaines d'auteurs travaillant avec cet outil mathématique, très utile à l'analyse des données et à la prévision de tout type de données.

1.3 Classification des séries

Le but ultime de la classification est de classer les séries en fonction de critères définis par les points communs entre elles. Il existe plusieurs méthodes de classification, dont les méthodes de partitionnement, les méthodes hiérarchiques descendantes et ascendantes, etc. Des distances permettent également de classer les données ; l'une d'entre elles est la distance euclidienne. Dans le cadre de la classification des données de consommation électrique, on peut par exemple classer les jours similaires, les semaines similaires, voire les mois similaires. Ça s'avère être un outil intéressant pour faire des prévisions, voire de la désaisonnalisation. Il peut souvent être nécessaire de traiter les données par lissage, filtrage, compression, etc., avant de les classer. On peut également recourir à des techniques pour atténuer l'ampleur d'un problème. En utilisant, par exemple, la transformée de Fourier, on peut obtenir une forme plus simple de notre série temporelle (Dominique, 2009). Des techniques sont disponibles dans la littérature pour classifier les données, à l'instar de Hamza, E. et Kemal, L. (2017), qui ont utilisé des matrices de covariances de caractéristiques, ou encore de Shreya, D. et al. (2020), qui proposent l'utilisation de méthodes de moyennage basées sur la distance DTW (Dynamic Time Warping), plus précisément le DBA (DTW Barycenter Average). Cette dernière est plus rapide pour établir une classification efficace. L'intérêt de ces travaux tient à l'utilisation d'outils mathématiques bien connus.

1.3.1 Détection d'anomalies

Une anomalie est un évènement perçu comme différents des autres évènements de la série temporelles. Les anomalies peuvent être dues à des phénomènes isolés ou relatifs à des comportements particuliers observés dans la série chronologique. La nature d'une anomalie peut varier selon le domaine concerné. Lors de l'observation d'une séquence de données, on

constate souvent des points différents, ce qui est normal. La détection de pics ou de baisses de consommation à des moments précis dans le graphe de la consommation électrique permet d'isoler des données qui ne seraient pas pertinentes et ne relèveraient pas de la normale.

On voit l'utilité de la détection d'anomalies dans le domaine de la finance, par exemple, afin de repérer les comportements inhabituels des marchés et d'en décider en conséquence. En utilisant l'apprentissage non supervisé, on peut détecter des fraudes ou des transactions non autorisées sur un compte bancaire donné. De plus, dans le cas de l'imagerie médicale, on peut utiliser l'apprentissage non supervisé pour analyser simultanément plusieurs radiographies et classer celles présentant une irrégularité. Ou même, dans le cas de la consommation électrique, un client qui voyage ou se rend à l'hôpital de façon urgente verrait apparaître une anomalie dans sa consommation, se traduisant par un creux.

D'après Su (2022), on peut classer les anomalies en trois groupes : l'anomalie ponctuelle, l'anomalie contextuelle et l'anomalie collective. L'anomalie ponctuelle correspond à une donnée anormale par rapport aux autres ; l'anomalie contextuelle, à une donnée anormale relativement à un contexte ; et l'anomalie collective, à un ensemble de données collectivement anormales par rapport aux autres (une donnée de cet ensemble n'est pas forcément une anomalie ponctuelle). Les relations entre ses anomalies sont mises en évidence dans la Figure 1.1.

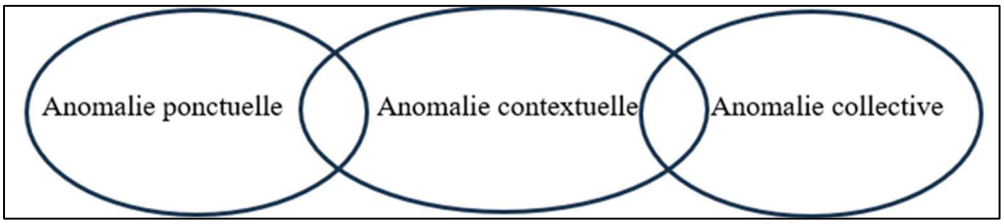


Figure 1.1 Relations entre les différents types d'anomalies
Tirée Guo et al. (2021, p.2)

Les anomalies ponctuelles et collectives sont plus faciles à détecter que les anomalies contextuelles dont la détection est un peu plus complexe. Selon Guo et al. (2021), une anomalie

contextuelle peut être soit ponctuelle, soit collective, soit aucune des deux. Ceci est illustré dans la figure suivante, qui présente les relations entre les types d'anomalies.

Afin de détecter ces différents types d'anomalies, il faut soit un algorithme spécifique choisi pour chaque type (ce qui est plus simple), soit un algorithme pour l'ensemble (ce qui est plus complexe). Il existe en général trois façons de détecter les anomalies : la méthode statistique, la méthode probabiliste et la méthode fondée sur l'apprentissage automatique. Cette dernière est plus complexe que les deux premières, mais offre l'avantage d'être plus méticuleuse selon le paramétrage choisi (Betin, 2022).

1.3.2 Détection d'anomalies ponctuelles et contextuelles

Sagoolmuang et Sinapiromsaran (2017) proposent également un algorithme pertinent dans le contexte des anomalies contextuelles. L'algorithme consiste à définir une fenêtre coulissante avec son point médian associé, puis à calculer son score à l'aide de la méthode MDWS (Mean difference window subseries). On va ainsi établir un ensemble de scores pour les différentes valeurs de la série, ainsi que des valeurs extrêmes relatives à cet ensemble, afin de détecter les anomalies. Ainsi, en deux étapes simples, on détecte les anomalies contextuelles en situant des contextes particuliers dans nos données à l'aide des fenêtres ; les données ne suivant pas « la norme » de ces contextes sont donc « hors contexte » et anormales. Toutefois, il faut noter que la détection des anomalies contextuelles va parfois mettre en évidence les anomalies ponctuelles et collectives.

1.3.3 Détection d'anomalies collectives

Outre les techniques mentionnées, diverses méthodes sont employées pour évaluer la distribution des données en cas d'anomalies collectives. Ces méthodes incluent l'intervalle interquartile (IQR), l'écart-type, l'écart moyen absolu, le coefficient de variation, entre autres. Le choix d'une méthode dépend de plusieurs facteurs, tels que sa sensibilité aux valeurs aberrantes, sa robustesse et les particularités des données à analyser. La « plage » est une mesure simple de la dispersion des données, définie comme la différence entre les valeurs maximale et minimale d'un ensemble de données. Cette mesure peut être utile pour

comprendre la dispersion globale des données de consommation électrique et pour détecter des anomalies collectives.

La « plage » peut permettre la détection d'anomalies collectives. Cependant, il est important de prendre en compte la sensibilité de la plage aux valeurs aberrantes, ainsi que les facteurs externes susceptibles d'influencer la consommation électrique. En tenant compte de ces éléments, la plage peut constituer un outil utile pour analyser les données de consommation électrique et détecter des anomalies collectives.

1.4 Techniques de détection et prédiction

Cette sous-section explore les principales techniques de détection d'anomalies et de prédiction appliquées aux séries temporelles, en mettant particulièrement l'accent sur la consommation électrique. Ces méthodes englobent des approches variées, allant des algorithmes d'apprentissage non supervisés, tels que la forêt d'isolement et les auto-encodeurs, aux techniques supervisées comme les modèles ARIMA, KNN ou SVM, en passant par des réseaux neuronaux avancés comme les LSTM. Le clustering, en tant que méthode non supervisée, permet de regrouper les données similaires ou de détecter des anomalies en fonction de l'écart par rapport aux groupes établis, tandis que l'empilement de modèles (stacking) combine plusieurs algorithmes pour améliorer la précision des prédictions. Enfin, l'imputation des données, par des méthodes telles que l'interpolation ou le KNN, et l'évaluation des performances à l'aide de métriques telles que MAE, RMSE et MAPE complètent ces approches en assurant la robustesse des analyses. Ces techniques, appuyées par des travaux de référence tels que ceux de Chicco et al. (2019) ou de Chen et al. (2018), offrent des solutions adaptées aux défis complexes de la gestion des données énergétiques.

1.4.1 Apprentissage supervisé

L'apprentissage automatique a révolutionné la science et les technologies sur plusieurs plans, et cette révolution s'est accompagnée de modes d'apprentissage présentant chacun une particularité. On évoque ici l'apprentissage supervisé, qui consiste à ce que les données soient bien structurées, par exemple sous forme de séries chronologiques la plupart du temps. Il

s'applique à bien des domaines. Il arrive que l'on entraîne, par exemple, la machine à reconnaître les figures géométriques. Cela impliquerait qu'il faut fournir à la machine le maximum de données pour qu'elle puisse apprendre suffisamment afin d'être précise lorsqu'elle est questionnée. Il faut donc fournir à la machine un jeu de données, qui peut être une liste, une matrice, etc. Relativement au type de données, il se divise en deux types : la classification (les problèmes du type « blanc ou noir », « vrai ou faux » ; on catégorise les événements) et la régression (les problèmes où l'on a des valeurs). En outre, la machine, par l'apprentissage, va définir un modèle du phénomène (le plus précis possible), à partir duquel elle sera capable de faire une prédiction (la plus réaliste possible) du phénomène appris. Afin d'estimer la précision du modèle, on définit une fonction de coût à minimiser. Cette fonction peut être de plusieurs sortes telles que les erreurs statistiques (erreur quadratique moyenne), le rapport poids-puissance en mécanique, et même en production (pour les mêmes prix de vente et d'achat, on a de bons bénéfices) : l'apprentissage va donc consister à trouver le modèle qui minimise la fonction de coût (Saint-Cirgue, 2019).

Il existe plusieurs types de modèles sur lesquels peut se baser l'apprentissage supervisé : le modèle de persistance (supposant que la valeur future sera égale à la dernière valeur observée), le modèle de moyenne mobile autorégressif intégré (ARIMA) et le modèle autorégressif (AR). Ces modèles sont utilisés dans de nombreuses applications. Tan et al. (2022) ont proposé l'utilisation de l'apprentissage supervisé et de techniques statistiques de base pour prévoir les anomalies de consommation. Nichiforov et al. (2021) ont également abordé la classification en utilisant, en plus du profil matriciel (technique d'exploration de données de séries chronologiques), KNN et SVM, tous deux des algorithmes d'apprentissage supervisé. Wen et Huang (2020) ont proposé une méthode de détection de fraude pour les cartes de crédit en combinant l'apprentissage supervisé et l'apprentissage non supervisé. Tsyganov (2021) utilise l'apprentissage supervisé dans le génie industriel pour contrôler les coûts de l'électricité. On voit à travers ces articles qu'il existe une panoplie d'usages de l'apprentissage supervisé. Les recherches dans ce domaine ont permis de progresser dans de nombreux domaines et de proposer des techniques novatrices très intéressantes.

1.4.2 Apprentissage non supervisé

Ce type d'apprentissage est complètement différent de celui précédent. Il a vu le jour avec l'avènement des réseaux de neurones artificiels, les grandes tailles de données et de puissantes machines capables de les traiter. Cet apprentissage est largement utilisé dans l'intelligence artificielle, la robotique, la santé, etc. C'est en effet l'apprentissage qui se rapproche le plus de l'apprentissage autonome et naturel, tel que l'est chez l'humain. Il est cité ici pour son utilisation avec de grandes volumes de données. Cet apprentissage consiste à analyser des données non étiquetées, c'est-à-dire des observations pour lesquelles on ne connaît pas la sortie attendue. L'algorithme cherche alors à identifier des structures, des motifs ou des regroupements semblables dans les données. Par exemple, il peut regrouper des groupes de personnes ayant des habitudes similaires ou extraire des composantes principales afin de réduire la dimensionnalité des données à analyser. On peut utiliser cet apprentissage pour classer des données ou même réduire la dimensionnalité d'un problème. Il est utilisé en imagerie médicale (Dou et al., 2019) pour améliorer la généralisation des modèles, même avec un volume de données limité.

Parmi les techniques d'apprentissage non supervisé les plus courantes, on a :

- Le clustering (le regroupement) : il consiste à identifier des points communs entre les données et à les regrouper en sous-groupes en fonction de leur similarité. Il s'agit donc de créer des sous-groupes non similaires entre eux, mais dont les éléments sont similaires. Il existe diverses méthodes de clustering, parmi lesquelles la méthode des k-moyennes, la classification hiérarchique et la classification probabiliste.
- L'association : c'est un système permettant de trouver des objets en fonction de liens selon certaines caractéristiques. C'est plus profond et plus intelligent comme apprentissage et c'est ce que l'on utilise généralement pour l'intelligence artificielle.

Ainsi, ce type d'apprentissage est relativement efficace pour des prévisions de consommation électrique (Wang, Y. et al., 2016), de trafic urbain ou encore de ventes d'épiceries du fait de l'imprévisibilité des comportements dans ces cas. Mais grâce à des séries chronologiques, l'apprentissage supervisé peut être amélioré.

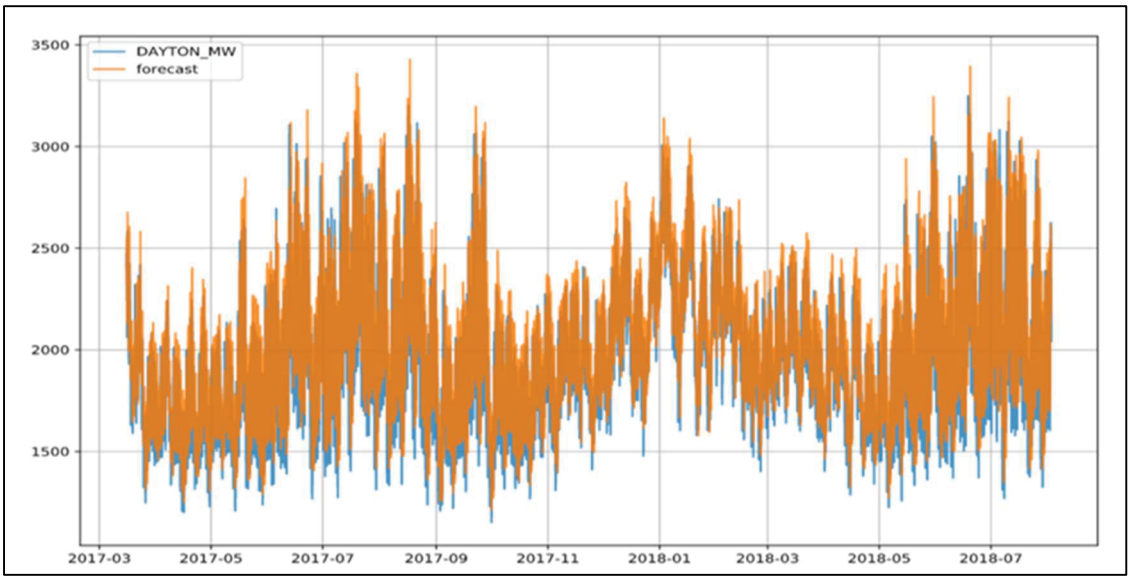


Figure 1.2 Prévision de la consommation d'énergie par l'apprentissage profond
Tirée de Eligijus (2020)

La Figure 1.2 présente la prévision obtenue par apprentissage profond concernant la consommation d'énergie. La différence entre la prévision et la réalité n'est pas si grande, de l'ordre de 15 % selon Eligijus (2020). Cette prévision est réalisée à l'aide d'un réseau de neurones à mémoire longue à court terme (LSTM). Il existe bien d'autres types de réseaux de neurones en apprentissage profond, parmi lesquels les auto-encodeurs et les CNN (réseaux de neurones convolutifs).

En particulier, l'auto-encodeur, qui a acquis une renommée dans le domaine de l'apprentissage profond pour sa capacité à apprendre des représentations efficaces à partir de données non étiquetées, est prometteur pour l'analyse des séries temporelles, y compris la consommation électrique. Comme souligné par Goodfellow et al. (2016), les auto-encodeurs peuvent extraire des caractéristiques marquantes des données, ce qui les rend potentiellement utiles pour comprendre les motifs complexes de consommation électrique. Le processus se déroule en deux étapes clés. La première est l'encodeur, qui prend les données et les comprime en un « code latent », une forme compacte qui en retient l'essence. La seconde étape est le décodeur, qui travaille à partir de ce code latent pour reconstruire les données, en tentant de les restituer le plus fidèlement possible à l'état d'entrée initial. C'est ce mécanisme d'encodage et de

décodage qui rend les auto-encodeurs si efficaces, notamment pour des tâches telles que la détection d'anomalies. Les auto-encodeurs sont ainsi particulièrement utiles pour la détection d'anomalies dans des domaines où les données normales sont très variées, tels que la consommation électrique, les images médicales ou les données de capteurs. Ils parviennent à traiter efficacement des données à grande dimensionnalité en capturant les caractéristiques essentielles et en réduisant la dimensionnalité.

En outre, leur aptitude à débruiter, décrite par Vincent et al. (2010), les rend particulièrement efficaces pour la détection d'anomalies dans les séries temporelles de consommation, en identifiant des motifs qui s'écartent de la norme. Cette capacité à modéliser et à expliquer les données est également reconnue par Chen et al. (2018), qui ont mis en avant l'importance des approches informationnelles pour l'interprétation des modèles, un aspect crucial pour les gestionnaires d'énergie souhaitant optimiser la distribution et la consommation.

1.4.3 Algorithme Forêt d'isolement

L'une des techniques les plus utilisées pour la détection d'anomalies est la forêt d'isolement. L'Isolation Forest est une méthode d'apprentissage automatique non supervisée utilisée pour détecter les anomalies dans les données de consommation électrique. Elle repose sur le principe selon lequel les anomalies sont rares et se distinguent des données normales, ce qui les rend plus faciles à isoler que les points de données normaux.

Lorsque l'Isolation Forest est appliquée à des données, elle fonctionne en construisant de nombreux arbres de décision, chacun construit de manière aléatoire. L'idée est de répéter le processus d'isolation — choisir une caractéristique et une valeur de division au hasard — jusqu'à ce qu'une observation soit isolée. Les anomalies auront tendance à être isolées plus rapidement, c'est-à-dire avec moins d'étapes, que les observations normales. Lorsque tous les arbres sont construits, le modèle est prêt à effectuer des prédictions. L'efficacité de cette technique est mentionnée par plusieurs auteurs, dont Jeremy R. (2021) et Patrick H. (2022).

En effet, pour chaque arbre, on mesure le nombre de décisions (ou tests) nécessaires pour qu'une observation *x* soit isolée dans une feuille, puis on calcule la moyenne *E* de ces longueurs de chemin sur l'ensemble des *t* arbres. Si le chemin est court, cela indique que l'observation est rapidement isolée et donc probablement une anomalie, tandis qu'un chemin long suggère que l'observation est normale. Ce concept est central pour déterminer le score d'anomalie qui permet de classer les observations comme normales ou aberrantes. Ainsi, pour *m* observations d'entraînement et *n* observations de test, le score d'anomalie qui est donné par :

$$s(x, m) = 2^{-\frac{E(h(x))}{c(m)}} \tag{1.2}$$

où $E\big(h(x)\big) = \frac{1}{t}\sum_{i=1}^{t} h_i(x)$ est la longueur moyenne du chemin pour isoler *x*, $h_i(x)$ est la longueur du chemin pour une observation *x* à travers l'arbre *i, t* est nombre d'arbres utilisés pour calculer la moyenne et $c(m)$ est une constante définie par

$$c(m) = 2H(m-1) - \frac{2(m-1)}{n}. \tag{1.3}$$

Dans (1.3), $H(m-1)$ est un nombre harmonique approximé par $\ln(m-1) + 0{,}5772156649$ qui est la constante d'Euler-Mascheroni (Bhayani, 2020).

1.4.4 Réseaux LSTM

Conçu selon le même principe que le cerveau humain, le réseau de neurones est conçu pour conserver en mémoire une information aussi longtemps que nécessaire. Selon Miled (2022), en citant Xiangyun Qing et Yugang Niu, « le LSTM est considéré comme le modèle le plus adéquat pour la prévision de données stochastiques et bruitées, telles que celles liées à la consommation, qui dépendent du temps ».

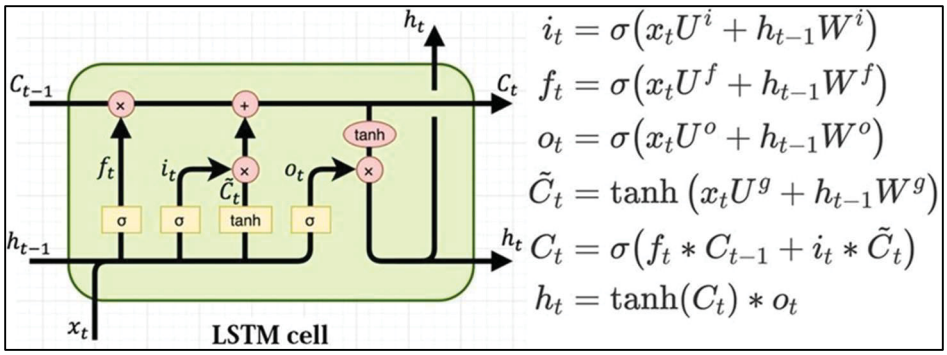


Figure 1.3 Structure de la cellule LSTM et les équations descriptives des portes de la cellule

Tiré Varsamopoulos & al. (2018, p.28)

Le LSTM est constitué de blocs de mémoire, chacun contenant une porte d'entrée, une porte de sortie et une porte d'oubli. Chaque porte est représentée par une équation telle que décrite ci-haut. En outre :

- i_t (porte d'entrée), f_t (porte d'oubli), o_t (porte de sortie basée sur l'état de la cellule) sont obtenues en appliquant une sigmoïde à l'entrée présente x_t et la sortie précédente h_{t-1}.
- C_t désigne l'état de la cellule. $\tilde{C}_t$ est obtenue en appliquant la tangente hyperbolique à l'entrée actuelle et à la sortie précédente.
- σ est une fonction sigmoïde dont la valeur est toujours comprise entre 0 et 1.
- La fonction *tanh* est la tangente hyperbolique.
- W est le poids d'une partie du réseau et U son biais.

Mostefa, S. (2021) propose un article intéressant : un guide sur la détection d'anomalies dans les séries temporelles à l'aide d'auto-encodeurs LSTM. L'article décrit comment entraîner un auto-encodeur pour reconstruire des données normales et utiliser l'erreur de reconstruction pour identifier les anomalies, en définissant un seuil d'erreur. Le modèle détecte les anomalies comme des déviations significatives par rapport aux données d'apprentissage, en s'appuyant sur le seuil de l'erreur absolue moyenne dérivée des données d'entraînement.

1.4.5 Clustering

Le clustering, en tant que méthode d'apprentissage non supervisé, constitue un outil puissant pour corriger les données de consommation électrique, en traitant les anomalies, les valeurs manquantes et les erreurs de mesure avec une précision notable. Des algorithmes comme k-means facilitent l'imputation des valeurs manquantes en exploitant les centroïdes des profils de consommation (Chicco et al., 2019), tandis que DBSCAN, fondé sur la densité, détecte efficacement les anomalies telles que les fraudes ou les pics incohérents dans les réseaux intelligents (Wang et al., 2020). Les modèles de mélange gaussiens (GMM) permettent de modéliser des distributions complexes pour corriger les incohérences des données de smart grids (Zhang et al., 2021), et les approches de deep clustering, s'appuyant sur des autoencodeurs, offrent une robustesse accrue face à des ensembles de données volumineux et hétérogènes (Yang et al., 2023). Toutefois, ces techniques présentent des défis, notamment une sensibilité aux hyperparamètres, un coût de calcul élevé pour les grands volumes de données et des difficultés à gérer les données manquantes non aléatoires. Pour répondre à ces enjeux, des recherches portant sur des approches hybrides intégrant le clustering et l'apprentissage supervisé, ainsi que des optimisations algorithmiques visant la scalabilité, s'imposent comme des orientations stratégiques.

1.4.6 Auto-encodeur

Traditionnellement, les auto-encodeurs sont composés de deux parties principales : l'encodeur et le décodeur (Vincent et al., 2010). L'encodeur compresse les données d'entrée x en une représentation de taille réduite appelée vecteur latent z tel que :

$$z = f(W_e \cdot x + b_e) \tag{1.4}$$

avec x vecteur d'entrée, W_e est la matrice de poids de l'encodeur, b_e est le vecteur de biais de l'encodeur et f est la fonction d'activation appliquée de manière élément par élément. Quant au décodeur, il reconstruit les données d'origine $\hat{x}$ à partir du vecteur latent z avec :

$$\hat{x} = g(W_d \cdot z + b_d) \tag{1.5}$$

où $\hat{x}$ est la reconstruction de l'entrée, W_d est la matrice de poids du décodeur, b_d est le vecteur de biais du décodeur et g est la fonction d'activation appliquée élément par élément.

Dans les deux cas, on observe une transformation linéaire des données d'entrée, suivie d'une non-linéarité introduite par la fonction d'activation *f*. Les fonctions d'activation couramment utilisées sont la sigmoïde, la tangente hyperbolique (*tanh*) et ReLU. En résumé, les auto-encodeurs apprennent à compresser les données d'entrée en une représentation réduite et à reconstruire les données d'origine à partir de cette représentation, en minimisant l'erreur de reconstruction.

1.4.7 Empilement de modèles (stacking)

La technique de stacking, également appelée « super learning », est une approche avancée en apprentissage automatique qui fusionne plusieurs algorithmes ou modèles de machine learning pour générer une prédiction plus fiable qu'un modèle unique. Chen et al. (2018) ont souligné que le stacking excelle particulièrement dans des situations où les modèles individuels révèlent des forces et des faiblesses distinctes. Par exemple, certains modèles peuvent être plus efficaces pour déceler des relations linéaires, alors que d'autres sont plus adaptés à modéliser des relations non linéaires. En rassemblant judicieusement ces modèles, le stacking permet de saisir un spectre plus large de relations dans les données, améliorant ainsi la précision de prédiction.

La mise en œuvre de la technique de stacking repose sur la recherche menée par Sakib Shahriar et al. (2021), qui ont démontré l'efficacité de cette approche dans divers contextes. Leur travail souligne l'importance d'explorer différentes combinaisons de modèles de machine learning afin d'optimiser la précision des prédictions. En utilisant une approche de stacking, il est possible de tirer parti des forces individuelles de chaque modèle, tout en compensant ses faiblesses, afin de créer un modèle composite offrant de meilleures performances globales.

1.4.8 Modèle SARIMA

Le modèle SARIMA est une extension du modèle ARIMA (AutoRegressive Integrated Moving Average), utilisé pour analyser et prévoir des séries temporelles (Hyndman et al.,

2018). SARIMA ajoute une composante saisonnière au modèle ARIMA, ce qui le rend utile pour les séries temporelles qui présentent des schémas saisonniers.

Le modèle SARIMA peut être dénoté comme *SARIMA* (*p*,*d*,*q*)(*P*,*D*,*Q*)*s*, où :

- *p* est l'ordre de la composante autorégressive (AR).
- *d* est l'ordre de différenciation.
- *q* est l'ordre de la composante moyenne mobile (MA).
- *P* est l'ordre saisonnier de la composante AR.
- *D* est l'ordre saisonnier de différenciation.
- *Q* est l'ordre saisonnier de la composante MA.
- *s* est le nombre de périodes par saison.

Les équations principales pour le modèle SARIMA sont :

1. **Composante AR (Autoregressive)** : Cette composante représente la relation entre une observation actuelle et ses observations précédentes. Son équation est :

$$\phi(B)\,Yt = c + \theta(B)\,Wt \tag{1.6}$$

où $\phi(B)$ est le polynôme d'ordre p et B est l'opérateur de retard.

2. **Composante I (Integrated)** : Cela implique de transformer la série temporelle afin de la rendre stationnaire en différenciant la série un nombre de fois.

$$Yt' = (1 - B)^d Yt \tag{1.7}$$

3. **Composante MA (Moving Average)** : Cette composante représente la relation entre une observation actuelle et les erreurs blanches des observations précédentes.

$$\theta(B)Wt = Yt - c \tag{1.8}$$

où $\theta(B)$ est le polynôme d'ordre q.

4. **Composantes saisonnières**: $\Phi(B^s)$ est le polynôme saisonnier d'ordre P pour la composante AR. $\Theta(B^s)$ est le polynôme saisonnier d'ordre Q pour la composante MA.

Lorsqu'on combine ces composantes, on obtient une équation globale pour le modèle SARIMA :

$$\boldsymbol{\Phi(B^s)\phi(B)(1-B)^d(1-B^s)^D Y_t = c + \Theta(B^s)\theta(B)W_t} \quad (1.9)$$

1.5 Validation croisée

Pour évaluer la robustesse de ces modèles, la validation croisée adaptée aux séries temporelles est essentielle. Selon la documentation de scikit-learn (2025), l'ordre temporel peut être préservé en divisant les données en plis où les ensembles de tests sont toujours postérieurs aux ensembles d'entraînement, évitant les biais liés à une séparation aléatoire et garantissant une évaluation cohérente avec la structure séquentielle des données énergétiques.

La validation croisée joue un rôle clé pour estimer les performances des modèles dans un contexte où les anomalies peuvent être rares et inégalement réparties, ce qui constitue un défi fréquent dans les séries temporelles énergétiques. Chandola et al. (2009) suggèrent de créer des plis respectant la temporalité, ce qui est crucial pour éviter, à l'avenir, les fuites de données vers le passé. Cette méthode est particulièrement adaptée aux données énergétiques, où les anomalies, telles que les périodes de vacances, sont souvent regroupées temporellement.

Pour gérer les plis sans anomalies, des stratégies spécifiques, telles que l'attribution de scores élevés lorsqu'aucune anomalie n'est prédite, sont recommandées afin d'éviter des métriques biaisées, conformément aux pratiques de gestion des classes déséquilibrées décrites dans la littérature. Les perspectives incluent l'intégration de caractéristiques supplémentaires (par exemple, des données météorologiques détaillées) ou l'exploration de modèles tels que les réseaux neuronaux récurrents pour améliorer la détection et l'imputation, ce qui renforce l'analyse des séries temporelles énergétiques.

1.6 Imputation des données

Les anomalies sont destinées à être retirées du jeu de données, laissant des vides à combler. Il existe plusieurs méthodes classiques pour cela, dont les différentes sortes d'interpolation, ainsi que le remplacement des données par la moyenne ou la médiane. D'autres méthodes et voies liées à l'apprentissage automatique et à l'ingénierie des caractéristiques (feature engineering).

L'imputation par la méthode des k plus proches voisins (KNN) constitue une stratégie robuste pour gérer les données manquantes, même en présence d'un grand nombre d'anomalies (Batista & Monard, 2002), s'appuyant sur un algorithme qui identifie les « k » observations les plus similaires dans l'ensemble de données pour estimer les valeurs inconnues. L'algorithme fonctionne en calculant une distance, souvent la distance euclidienne, entre la donnée incomplète et les autres données, puis en sélectionnant les voisins les plus proches en fonction de cette similarité. Cette méthode est particulièrement efficace dans les ensembles de données où les relations entre variables sont complexes, car elle n'impose pas de structure prédéfinie à ces relations, ce qui permet une estimation plus naturelle et plus précise des valeurs manquantes (Wikistat, n.d.).

Parmi les méthodes d'imputation de données : la recopie des valeurs précédentes. C'est une technique de traitement des données simple et répandue (Mathworks, 2025), appelée « imputation par la dernière observation reportée » (en anglais, « last observation carried forward » ou LOCF). Cette méthode consiste à remplacer les valeurs manquantes d'une série temporelle par la dernière valeur observée antérieure à la valeur manquante. L'imputation par la dernière observation reportée peut être utile lorsque les données sont manquantes de manière aléatoire et que la dernière observation disponible est une bonne approximation de la valeur manquante. Cependant, il est important de noter que cette méthode présente des inconvénients. Dans certaines situations, il peut être préférable d'utiliser d'autres techniques d'imputation, telles que l'interpolation linéaire, l'imputation par la moyenne ou la médiane, ou encore des modèles de prédiction pour estimer les valeurs manquantes, selon le type d'anomalie retirée. Pour les anomalies collectives, nous pourrons aisément utiliser cette technique, mais aussi pour les autres types d'anomalies.

1.7 Métriques de performance

Dans le contexte du traitement des données, les métriques essentielles pour évaluer la performance des méthodes de traitement, en particulier dans des domaines où l'IA et l'apprentissage automatique sont utilisés pour optimiser les opérations et aligner les performances sur les objectifs stratégiques d'une organisation.

Dans le contexte de l'évaluation, des méthodes de traitement des données, des KPI tels que le Mean Absolute Error (MAE), le Mean Squared Error (MSE), le Mean Absolute Percentage Error (MAPE) et le Root Mean Squared Error (RMSE) sont couramment utilisés. Ces indicateurs fournissent des mesures quantitatives de l'erreur entre les valeurs prédites par un modèle et les valeurs réelles, permettant d'évaluer l'exactitude des prédictions.

- **RMSE :**

Il s'agit de la racine carrée de la moyenne des erreurs carrées. RMSE est une métrique de performance qui mesure la magnitude moyenne de l'erreur. Une valeur de RMSE nulle indique que toutes les prédictions sont parfaites. En pratique, RMSE est utile lorsque de grandes erreurs sont particulièrement indésirables. Son unité ici est le kW.

- **MSE :**

Il s'agit de la moyenne carrée des erreurs. Comme le RMSE, le MSE mesure la qualité d'un estimateur ou d'un prédicteur. L'erreur est la quantité par laquelle une valeur prédite diffère de la valeur observée. Le MSE est toujours positif et une valeur de zéro indique une prédiction parfaite. Son unité ici est le kW au carrée (kW^2).

- **MAE :**

Il s'agit de la moyenne des valeurs absolues des erreurs. C'est une mesure linéaire qui donne une idée de la magnitude de l'erreur, sans en considérer la direction. Une MAE de zéro indique une prédiction parfaite. Son unité ici est le kW.

- **MAPE :**

Il s'agit de la moyenne des valeurs absolues des pourcentages d'erreur. Le MAPE exprime l'erreur en pourcentage, ce qui permet une interprétation intuitive en termes de « pourcentage d'erreur ». Cependant, le MAPE peut devenir infini ou indéfini si les valeurs observées contiennent des zéros.

En résumé, ces métriques d'erreur permettent d'évaluer la qualité des prédictions d'un modèle en comparant les valeurs prédites aux valeurs réelles. Elles sont essentielles au développement et à la validation des modèles prédictifs.

1.8 Conclusion

D'après ce qui précède, on remarque qu'il existe une panoplie d'articles traitant de l'IoT pour la mesure intelligente, ainsi que de l'exploitation de ces données. Outre les techniques statistiques, l'utilisation de l'apprentissage automatique pour la détection d'anomalies, la prévision de la consommation d'énergie et d'autres traitements appliqués aux séries chronologiques est pertinente au vu de la diversité des possibilités dont nous disposons. La conjugaison de ces différentes techniques nous permettra de répondre, comme il convient, à notre problématique. Dans ce contexte de transition énergétique, il est très pertinent d'aborder de telles notions qui interviennent directement dans la gestion de ladite transition. Nous sortons donc de cette revue, enrichie de nouvelles approches utiles à la prévision des données de consommation d'électricité, ainsi qu'à leur analyse.

CHAPITRE 2

MÉTHODOLOGIE

2.1 Introduction

Nous présentons dans ce chapitre notre méthodologie. Après l'acquisition des données, on commence par ajouter des caractéristiques au jeu de données. Ces caractéristiques supplémentaires, telles que la température, le jour de la semaine et le moment de la journée (éventuellement le mois si les données couvrent plusieurs années), permettent de capturer certains facteurs externes ou saisonniers qui influencent le phénomène étudié. À titre d'exemple, dans l'analyse de la consommation électrique d'un domicile, la température peut fortement influer sur l'usage du chauffage ou de la climatisation, tandis que le jour de la semaine et l'heure du jour reflètent les habitudes de présence des occupants. L'ajout de ces caractéristiques améliore la capacité des modèles d'apprentissage automatique à apprendre des motifs complexes et à faire des prédictions plus précises.

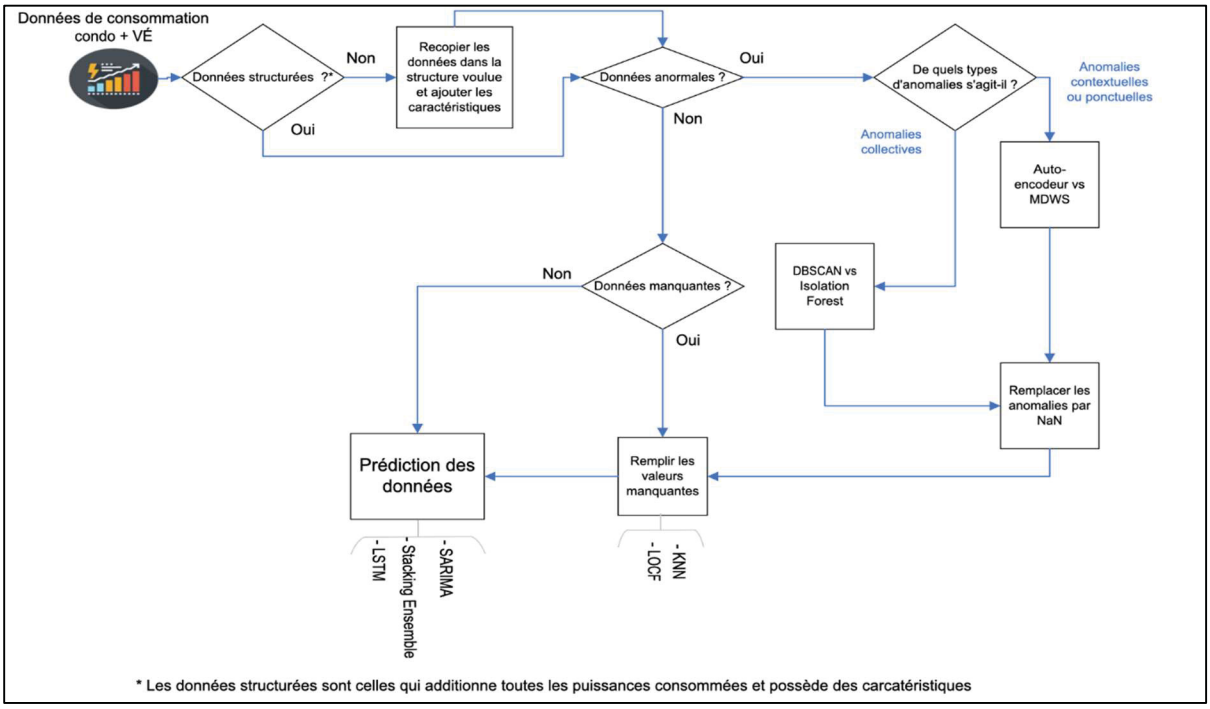


Figure 2.1 Diagramme descriptif de la méthodologie

Les données structurées correspondent à l'ensemble des puissances consommées et prennent en compte différentes caractéristiques associées. En vue d'évaluer les techniques de détection d'anomalies, les données seront étiquetées de diverses façons. On utilisera les méthodes de MDWS développées par Sagoolmuang et Sinapiromsaran (2017), l'auto-encodeur inspiré de Mostefa, S. (2021) pour détecter les anomalies ponctuelles et contextuelles, la forêt d'isolation, ainsi que DBSCAN pour détecter les anomalies collectives. En utilisant les techniques de dernière observation et de KNN relatives aux caractéristiques définies (le jour de la semaine, la température et l'heure du jour), nous imputons les valeurs manquantes et supprimons les anomalies détectées.

Dans notre méthodologie, nous inclurons la prévision ; nous comparerons trois techniques dont l'efficacité pour les prévisions est démontrée par des articles précités : le LSTM, le SARIMA et l'ensemble d'empilement regroupant plusieurs méthodes à la fois. Le diagramme de la Figure 2.1 résume la méthodologie que nous allons utiliser dans ce chapitre.

2.2 Structuration des données

Les données proviennent d'un jeu de données recensant les puissances instantanées consommées dans une résidence équipée d'une borne de recharge pour véhicules électriques. Les données provenant de plusieurs points de mesure sont collectées dans une base de données. Cette collecte est rendue possible grâce au déploiement rapide de technologies IoT (Internet of Things), qui facilitent la gestion de volumes massifs de données grâce à des solutions de communication avancées. De nombreux équipements du secteur énergétique sont conçus pour être compatibles avec l'IoT. Ce qui accroît les capacités d'utilisation, d'analyse et de valorisation de ces données, y compris en temps réel.

Nous traitons ces données à partir d'hypothèses tirées de leur observation : pour une famille classique, on distinguera des jours ouvrables, des fins de semaine, des vacances possibles, voire d'autres événements particuliers nécessitant plus ou moins de consommation qu'habituellement. Plusieurs types de traitement mentionnés dans le chapitre précédent de la révision de la littérature peuvent être appliqués à ces données : la détection d'anomalies, le

lissage, la classification pour distinguer des groupes de consommateurs, la correction des données, la compression des données, la prévision de la consommation.

Pour enrichir l'ensemble de données, de nouvelles caractéristiques ont été générées à partir des horodatages. La température simulée, spécifique au Québec, est générée mensuellement à l'aide d'une distribution gaussienne, contrainte par des plages réalistes issues de Météo Montréal.

Par ailleurs, une variable catégorielle, *Weekdays* (jour de la semaine, codée de 0 à 6), et des variables indicatrices, *Daytime* (variable binaire prenant la valeur 1 pour la plage horaire 6 h – 18 h, 0 sinon) et High_*P*_1_2 (variable binaire valant 1 lorsque la puissance excède sa moyenne), ont été construites afin de capter certaines dynamiques temporelles et contextuelles.

Ces opérations d'enrichissement ont été suivies de procédures de prétraitement de nature variée, visant à optimiser la qualité et la pertinence des données pour les analyses ultérieures.

2.3 Techniques de détection et corrections des anomalies

Dans le présent cas, on considérera que les anomalies sont de trois sortes : collectives, ponctuelles et contextuelles (les deux dernières étant détectées ensemble).

Dans l'observation et l'analyse de la consommation électrique d'un domicile, on considérera : les vacances, les fêtes, les coupures d'électricité et bien d'autres phénomènes dont la nature peut être ou non prévisible. Le but est de corriger les séries chronologiques afin que les prévisions ne soient pas corrompues par un événement non régulier (notamment dans le cas d'apprentissage supervisé) et qu'elles suivent le comportement du groupe étudié.

Peu importe la méthode utilisée, il faudra, dans le présent cas, que la valeur relevée se situe dans un intervalle défini. Relever zéro ou une valeur négative dans une séquence de données d'énergie électrique d'un foyer serait, par exemple, anormal.

Nous proposerons des techniques pour détecter et éliminer les trois types d'anomalies, à savoir : anomalie ponctuelle, anomalie contextuelle et anomalie collective.

Outre cela, nous aurons besoin d'étiqueter les anomalies ainsi que les périodes de vacances ou d'absence afin d'évaluer les performances de nos techniques. Pour cela, les données de consommation sont prétraitées en commençant par l'interpolation des températures manquantes et en générant les variables *Weekdays* et *Daytime* si absentes. Les recharges de véhicules électriques sont filtrées à l'aide d'un seuil vectorisé basé sur High_P_1_2 et sur des variations de consommation. Deux méthodes d'étiquetage sont appliquées. La première est l'IQR, qui identifie les anomalies ponctuelles à partir des quartiles de la consommation filtrée, tandis que la décomposition STL (période de 12 heures) détecte les résidus anormaux. Les périodes de vacances sont détectées par groupement temporel sur des intervalles de 12 heures, avec une faible variabilité de la consommation. Une validation croisée temporelle (3 plis) évalue la cohérence des étiquettes, et une fusion conservatrice privilégie l'IQR ou le STL en fonction de leur stabilité.

2.3.1 Détection d'anomalies ponctuelles et contextuelles

Cette section détaille deux techniques avancées de détection d'anomalies dans les séries temporelles de consommation électrique : la méthode MDWS et l'utilisation d'auto-encodeurs. La technique MDWS, inspirée de Sagoolmuang et Sinapiromsaran (2017), repose sur une analyse contextuelle des données au sein d'une fenêtre temporelle coulissante de 8 heures, permettant d'identifier des anomalies contextuelles subtiles, telles qu'une consommation inhabituelle en fonction de l'heure ou des conditions saisonnières. En complément, la technique des auto-encodeurs, décrite notamment par Mostefa (2022), exploite l'apprentissage profond pour reconstruire les données normales et détecter les écarts significatifs à partir de l'erreur de reconstruction, offrant une approche robuste pour identifier les anomalies dans des données normalisées. Ces deux méthodes, combinées à des stratégies d'imputation et d'évaluation des performances à l'aide de métriques telles que le F1-score ou la matrice de confusion, constituent des outils puissants pour une analyse précise et contextualisée des données énergétiques.

2.3.1.1 Technique MDWS

Après l'étiquetage des données, nous commençons par traiter les anomalies contextuelles, les plus subtiles à détecter. Par exemple, une puissance consommée de 4 kW lors d'une journée d'hiver à -25 °C peut être anormale, contrairement à des erreurs évidentes comme une consommation négative ou une valeur aberrante de 50 kW pour une résidence. La procédure de détection comporte l'exemple de Sagoolmuang et Sinapiromsaran (2017) : l'algorithme MDWS (Median-based Dynamic Window Scoring) et un algorithme de détection d'anomalies, adaptés pour analyser *n* données dans une fenêtre temporelle coulissante de 8 heures. Notre intérêt pour cette méthode réside dans sa capacité à situer les données dans une fenêtre temporelle choisie, ici de 8 heures, ce qui nous permet d'analyser le contexte d'une période spécifique. Cela aide à déterminer si une consommation, comme celle d'électroménagers, qui serait normale de jour, ne l'est pas la nuit, par exemple. Avant d'aller plus loin, définissons les éléments suivants :

- $Y = \{y_0, y_1, \dots, y_{n-1}\}$ est une série de n observations
- W_i est une sous-série de Y de taille $w \leq n$, associée à un y_i (point médian de la sous-série) tel que $\left\lceil \frac{w-1}{2} \right\rceil \leq i \leq n - \left\lfloor \frac{w-1}{2} \right\rfloor - 1$
- $score(y_i) = y_i - \widetilde{W}_i$ tel que $\widetilde{W}_i$ est la médiane de la sous-série

AnomalyScore= [] #On définit l'ensemble des scores d'anomalies
Pour la première fenêtre $W_{\lceil\frac{w-1}{2}\rceil}$, faire :
    Ordonner $W_{\lceil\frac{w-1}{2}\rceil}$ pour obtenir $W_{ordonné} = \{s_0, s_1, \dots, s_{w-1}\}$
    Trouver la médiane $\widetilde{W}_{\lceil\frac{w-1}{2}\rceil} = s_{\lceil\frac{w-1}{2}\rceil}$
    Calculer MDWS de $y_{\lceil\frac{w-1}{2}\rceil}$
    Ajouter $score\left(y_{\lceil\frac{w-1}{2}\rceil}\right)$ à AnomalyScore
    Garder $W_{ordonné}$
Fin Pour

Pour $i = \lceil\frac{w-1}{2}\rceil + 1, \lceil\frac{w-1}{2}\rceil + 2, \dots, n - \lceil\frac{w-1}{2}\rceil - 1$; faire :
    Retirer $y_{i-\lceil\frac{w-1}{2}\rceil-1}$ de $W_{ordonné}$
    Ajouter $y_{i+\lceil\frac{w-1}{2}\rceil}$ dans $W_{ordonné}$ et réordonner
    Trouver la médiane $\widetilde{W}_i = s_{\lceil\frac{w-1}{2}\rceil}$
    Calculer $score(y_i)$
    Ajouter $score(y_i)$ à AnomalyScore
    Garder $W_{ordonné}$
Fin Pour
Retourner AnomalyScore

Figure 2.2 Algorithmes MDWS
Tiré de Sagoolmuang (2017, p.3)

Anomalies = {}
Max = moyenne (AnomalyScore) + 3*std (AnomalyScore)
Min = moyenne (AnomalyScore) - 3*std (AnomalyScore)
Pour i = 1,…,n faire :
    Si $score_i < Min$ ou $score_i > Max$:
        Ajouter y_i à Anomalies
    Fin Si
Fin Pour
Retourner Anomalies

Figure 2.3 Détection d'anomalie
Tiré de Sagoolmuang (2017, p.3)

Nous avons besoin de définir certains paramètres. Nous disposons d'environ 148 000 observations entre janvier et août 2023. Nous avons choisi une largeur de fenêtre coulissante correspondant à un tiers d'une journée, soit 8 heures. Nous supposons que les comportements varient par tranches de 8 h pour une famille classique au Québec. Les données

sont lissées par une moyenne mobile d'une fenêtre de 9 points (seuil optimal établi par une expérimentation avec différentes valeurs afin d'éviter trop de pertes de données tout en restant efficace), afin de réduire le bruit. Les scores d'anomalie sont calculés en soustrayant la médiane glissante des consommations lissées, puis ajustés par des poids contextuels tenant compte des caractéristiques des données. Les anomalies sont identifiées lorsque les scores dépassent un seuil basé sur l'écart absolu médian, optimisé pour maximiser le F1-score (précision et rappel). Les anomalies détectées sont corrigées par interpolation linéaire et les résultats sont visualisés à l'aide d'une matrice de confusion.

2.3.1.2 Technique auto-encodeur

Pour détecter les anomalies, l'auto-encodeur est d'abord entraîné sur des données de consommation énergétique normalisées afin de reconstruire les entrées considérées comme normales. Il est ensuite utilisé pour reconstruire de nouvelles entrées. Si la reconstruction est exacte, l'entrée est jugée normale. En revanche, si la reconstruction diffère significativement de l'entrée, cette dernière est considérée comme anormale. Une mesure d'erreur, telle que l'erreur quadratique moyenne (MSE), est utilisée pour quantifier la différence entre les données d'entrée et celles reconstruites et pour déterminer si cette différence dépasse un seuil d'anomalie.

Après avoir chargé et reformaté les données de consommation afin qu'elles répondent à nos exigences, nous les normaliserons. Cette normalisation est cruciale, car elle permet de traiter efficacement les variations d'amplitude susceptibles de survenir entre différentes mesures. En nous inspirant de Mostefa, S. (2022), puis en l'ajustant à notre contexte, nous allons détecter les anomalies en utilisant l'auto-encodeur comme suit : il est composé de couches denses de 32, 16 et 8 neurones avec activation ReLU, puis de 16 et 32 neurones, avec une sortie linéaire. Il est entraîné avec l'optimiseur Adam et la perte MSE. Entraîné sur 50 époques avec un lot de 128, il prédit les données normalisées et l'erreur de reconstruction (MSE) est calculée. Les anomalies sont identifiées lorsque l'erreur dépasse le 90e percentile, seuil optimisé via une validation croisée à 4 plis évaluant la précision, le rappel et le F1-score. Les anomalies détectées sont imputées à l'aide de KNN, puis les performances sont visualisées à l'aide d'une

matrice de confusion. Ces paramètres ont été obtenus en testant plusieurs combinaisons de valeurs.

2.3.2 Détection d'anomalies collectives

Cette section explore deux approches distinctes pour la détection et la correction des anomalies collectives dans les séries temporelles de consommation électrique : la technique DBSCAN et la forêt d'isolation. La méthode DBSCAN, basée sur le clustering par densité, segmente les données en intervalles de 24 heures et identifie les périodes de faible variabilité, telles que des absences prolongées ou des perturbations, en s'appuyant sur l'écart-type comme critère clé. De son côté, la forêt d'isolation utilise des statistiques journalières, telles que la moyenne et l'écart-type, pour isoler les jours atypiques à l'aide d'arbres de décision aléatoires, afin de cibler des anomalies similaires. Les deux techniques recourent à une imputation par k plus proches voisins (KNN) intégrant des caractéristiques contextuelles pour corriger les données manquantes, et leurs performances sont évaluées à l'aide de métriques telles que la précision, le rappel, le F1-score et l'erreur quadratique moyenne (RMSE), offrant des solutions robustes face aux défis liés aux variations contextuelles imprévues.

2.3.2.1 Technique DBSCAN

La consommation électrique, caractérisée par des variations saisonnières, quotidiennes et contextuelles, nécessite un prétraitement rigoureux afin d'assurer la fiabilité de l'analyse des anomalies collectives.

La détection des anomalies collectives repose sur l'identification de périodes de faible variabilité de la consommation, indicatives d'absences prolongées ou de perturbations du réseau, à l'aide d'une approche de clustering par densité (Wang et al., 2020). Les données sont segmentées en intervalles de 24 heures et la variabilité de la consommation est évaluée à l'aide de l'écart-type. Les intervalles présentant un écart-type inférieur à un seuil prédéfini (0,15 kW) sont considérés comme anormaux, ce qui reflète des comportements inhabituels, tels que des vacances ou des pannes. Une validation croisée temporelle évalue la correspondance entre ces intervalles et les périodes de vacances identifiées, en calculant des métriques telles que la précision et le rappel. Cette méthode, bien que robuste pour détecter des anomalies

contextuelles, reste sensible aux variations induites par des facteurs externes, tels que les conditions météorologiques ou les recharges de véhicules électriques, ce qui nécessite un calibrage précis du seuil.

Les périodes identifiées comme anormales sont marquées comme valeurs manquantes et corrigées par une imputation basée sur les k plus proches voisins (KNN), en exploitant des caractéristiques contextuelles telles que le jour de la semaine, l'heure, la température et les recharges de véhicules électriques. Cette approche permet de reconstruire les données en tenant compte des motifs locaux, contrairement à une imputation simple qui ignore le contexte. La qualité de l'imputation est évaluée par la moyenne quadratique des erreurs (RMSE) sur les périodes de non-vacances, et des visualisations comparent les séries originales et corrigées, mettant en évidence les intervalles anormaux ainsi que les périodes de vacances. Les résultats, y compris les métriques de performance et les données imputées, sont archivés pour une analyse ultérieure. Toutefois, la précision de l'imputation dépend de la qualité des caractéristiques contextuelles et peut être affectée par des événements météorologiques imprévus, notamment par des événements météorologiques extrêmes.

2.3.2.2 Forêt d'isolation

La consommation électrique, marquée par des variations saisonnières, quotidiennes et contextuelles, peut être analysée à l'aide de statistiques journalières telles que la moyenne et l'écart-type, qui capturent les tendances et la variabilité typiques. L'Isolation Forest exploite la facilité d'isoler les observations atypiques dans un espace statistique pour identifier les jours anormaux. En utilisant la moyenne et l'écart-type journaliers comme caractéristiques, l'algorithme construit des arbres de décision aléatoires, où les jours nécessitant moins d'étapes pour être isolés (score d'anomalie de -1) sont classés comme anormaux, par exemple, ceux avec une moyenne inférieure à 1 kW et un écart-type inférieur à 0,15 kW, indicatifs de périodes comme des vacances ou des pannes. Une validation croisée temporelle (4 plis) évalue la correspondance entre les jours détectés et les périodes de vacances marquées, à l'aide de métriques comme la précision, le rappel et le F1-score. Cette méthode est efficace pour capturer les anomalies collectives, mais ses performances dépendent du choix du taux de contamination

et de la qualité des caractéristiques, susceptibles d'être influencées par des variations contextuelles imprévues, telles que des événements météorologiques extrêmes.

Les jours identifiés comme anormaux sont marqués comme valeurs manquantes et imputés à l'aide d'une méthode basée sur KNN, intégrant des caractéristiques contextuelles (jour de la semaine, heure, température, recharges de véhicules électriques) afin de préserver les motifs locaux de consommation. La qualité de l'imputation est mesurée par l'erreur quadratique moyenne (RMSE) sur les périodes de non-vacances, et des visualisations comparent les séries originales et corrigées, mettant en évidence les intervalles anormaux ainsi que les périodes de vacances. Une matrice de confusion illustre la performance de la détection et les données imputées sont archivées pour une analyse ultérieure. Bien que cette approche permette une correction robuste, sa précision peut être limitée par la disponibilité de données contextuelles fiables et par des variations soudaines, telles que celles induites par des pannes majeures ou des conditions climatiques extrêmes.

2.3.3 Prédiction temporelle

Cette section examine trois techniques avancées pour la prévision et l'analyse des séries temporelles de consommation électrique : le modèle LSTM, l'empilement (stacking) et le modèle autorégressif saisonnier (SARIMA). La technique LSTM, basée sur les réseaux neuronaux récurrents, exploite sa capacité à mémoriser les motifs à long terme pour prédire les consommations futures, offrant une grande précision dans des contextes où les comportements humains sont complexes. L'approche de stacking combine plusieurs modèles, tels que la forêt aléatoire, le perceptron multicouche et la machine à vecteurs de support, avec un méta-apprenant (XGBRegressor) afin d'améliorer la robustesse et la précision des prédictions. Enfin, le modèle SARIMA, adapté aux séries présentant des motifs saisonniers, intègre des composantes autorégressives, intégrées et de moyenne mobile pour capturer les tendances et les saisonnalités, ce qui nécessite un nettoyage rigoureux des données afin de garantir la stationnarité. Ces méthodes, évaluées à l'aide de métriques telles que le RMSE, le MAE et le MAPE, offrent des solutions complémentaires pour une prévision fiable et précise de la consommation énergétique.

2.3.3.1 Technique LSTM

L'approche LSTM proposée ici se déroule en deux étapes principales : l'entraînement du modèle sur les données historiques pour capturer les motifs temporels, suivi de la prédiction des valeurs futures de la série temporelle.

Dans un premier temps, l'entraînement consiste à utiliser un ensemble de données de consommation d'électricité afin d'optimiser un modèle LSTM (Brownlee, 2021). Les données, issues de la correction des données, sont prétraitées en normalisant les données de consommation sur l'intervalle [0, 1] et en vérifiant l'absence de valeurs manquantes (remplies par propagation, si nécessaire). Une transformation en problème d'apprentissage supervisé est effectuée, ce qui génère des paires entrée-sortie avec une fenêtre temporelle de 5 pas de temps (les 5 valeurs de consommation enregistrées avant une entrée à un instant donnée). Cette fenêtre permet au modèle de capturer les dépendances temporelles à court terme.

Le modèle est construit avec une architecture inspirée de Brownlee (2021), adaptée pour optimiser les performances sur un grand ensemble de données, comme dans le cas étudié ici. Il comprend :

- Deux couches LSTM : La première, avec 50 unités et une activation tanh, renvoie des séquences pour alimenter la seconde. La deuxième couche, avec 50 unités et une activation *tanh*, ne renvoie aucune séquence. Des couches de dropout (10 %) sont ajoutées après chaque couche LSTM pour réduire le risque de surajustement.
- Deux couches denses : Une couche intermédiaire de 25 unités avec activation ReLU, suivie d'une couche de sortie avec une unité générant la prédiction finale.

Le modèle est entraîné avec la fonction de perte de l'erreur quadratique moyenne (MSE), couramment utilisée pour les problèmes de régression, et l'optimiseur Adam, reconnu pour sa convergence rapide et sa robustesse (Kingma & Ba, 2014). L'entraînement est effectué sur 40 époques, avec une taille de lot de 16 et un mécanisme d'arrêt anticipé (EarlyStopping, patience = 10) pour éviter le surajustement et réduire le temps de calcul.

Dans un second temps, pour évaluer la robustesse du modèle, une validation croisée temporelle à 4 plis est utilisée, respectant l'ordre chronologique des données. Chaque pli divise les données en environ 80 % pour l'entraînement et 20 % pour le test. Cette approche, inspirée du site Scikit-learn et de l'article « Improving LSTM Performance Using Time Series Cross Validation » (LinkedIn, 2022), garantit une évaluation fiable de la généralisation du modèle sur un grand ensemble de données. Également, Bergmeir & Benítez (2012) soulignent l'intérêt de la validation croisée, car elle permet de confirmer plus fidèlement les résultats d'un modèle.

Les performances sont mesurées à l'aide de trois métriques :

- RMSE pour évaluer l'erreur globale.
- MAE pour mesurer l'erreur absolue moyenne.
- MAPE pour quantifier l'erreur relative, avec correction (pour éviter les divisions par zéro).

Le choix du LSTM repose sur sa capacité à mémoriser les dépendances à long terme dans les séries temporelles, ce qui est particulièrement adapté pour modéliser la consommation électrique d'un domicile, où les comportements humains peuvent présenter des motifs complexes et non linéaires (Ibrahim, 2021). Comparés aux modèles traditionnels comme SARIMA, les LSTM excellent dans la capture de dynamiques non linéaires, ce qui les rend adaptés à notre jeu de données volumineux (plus de 140 000 points). La validation croisée temporelle permet de garantir que le modèle généralise bien, en évaluant ses performances sur différentes périodes de la série, tout en respectant l'ordre chronologique.

2.3.3.2 Technique d'empilement (stacking)

Pour prévoir la consommation énergétique d'un domicile, nous avons structuré un ensemble de données multivariées comprenant la consommation électrique, les jours de la semaine, les températures et une variable indicatrice jour/nuit. Les données ont été prétraitées en normalisant la consommation sur l'intervalle [0, 1] pour uniformiser l'échelle, et les valeurs manquantes ont été remplies par propagation avant. La cible, définie comme la consommation

à 7 jours, est créée par un décalage temporel, ce qui transforme le problème en un apprentissage supervisé multivarié.

Le modèle d'empilement combine trois modèles de base :

- Une forêt aléatoire qui capture les relations non linéaires entre les caractéristiques, les interactions complexes entre les variables, tout en étant robuste face au bruit et aux valeurs aberrantes (Breiman, 2001, p.7).

- Un perceptron multicouche, adapté aux motifs complexes grâce à sa structure neuronale, est utile pour modéliser les influences implicites de la recharge sur la consommation (Goodfellow et al., 2016).

- SVR, une extension de SVM appliquée à la régression, est efficace pour modéliser des relations locales non linéaires et affiner les prédictions sur des plages ponctuelles, comme les pics de consommation pendant les sessions de recharge (Cortes & Vapnik, 1995).

Chaque modèle est entraîné indépendamment sur les données d'entraînement, et ses prédictions sont combinées par un méta-apprenant, XGBoost, qui optimise la précision globale en pondérant les contributions des modèles de base. Cette approche d'ensemble, inspirée des stratégies de stacking (Wolpert, 1992), exploite les forces complémentaires des modèles pour améliorer la robustesse et les performances par rapport à des modèles individuels.

La validation du modèle utilise une validation croisée temporelle à 5 plis, respectant l'ordre chronologique des données, avec environ 80 % des données pour l'entraînement et 20 % pour le test par pli. Les performances sont évaluées à l'aide de trois métriques :

- RMSE pour quantifier l'erreur globale.

- MAE pour évaluer l'erreur absolue moyenne.

- MAPE pour mesurer l'erreur relative, avec une correction ($\varepsilon = 1 \times 10^{-10}$) afin d'éviter les divisions par zéro.

Le stacking, par la combinaison de différents modèles qui se renforcent, permet de capturer les dynamiques complexes des données (par exemple, des variations liées aux températures ou aux jours de la semaine) tout en réduisant les biais et la variance des prédictions, offrant ainsi une solution robuste pour notre cas.

2.3.3.3 Modèle autorégressif saisonnier

Cette approche SARIMA présentée comprend plusieurs étapes clés : nettoyage des données, configuration et entraînement du modèle, prédiction, visualisation et évaluation des performances.

D'abord, un nettoyage rigoureux des données est essentiel pour garantir la stationnarité et l'intégrité de la série temporelle, conditions nécessaires à la précision des prédictions SARIMA. La colonne Valeur, représentant la consommation électrique, est convertie en type numérique et les valeurs manquantes sont remplacées par la moyenne de la série, suivie d'une interpolation linéaire pour assurer la continuité. Les valeurs aberrantes ne sont pas explicitement traitées, mais leur impact est limité par la robustesse du modèle SARIMA. La série est indexée temporellement à une fréquence de 5 minutes (c'est le temps maximal observé entre deux points des séries temporelles étudiées), ce qui permet de capturer les motifs réguliers. Les données sont ensuite divisées en un ensemble d'entraînement (jusqu'au 20 juillet 2023) et en un ensemble de tests (à partir du 21 juillet 2023) pour évaluer la généralisation du modèle.

Ensuite, nous déterminons les paramètres optimisés de notre modèle. Une sélection automatique des paramètres explore différents ordres (p, d, q) pour identifier la combinaison la plus adaptée. En cas d'échec ou d'incohérence, un modèle plus simple (ARIMA (0,0,1)) est utilisé en repli. Le modèle SARIMAX est ensuite ajusté aux données d'entraînement par maximisation de la vraisemblance, avec vérification de la convergence et consultation des diagnostics fournis par le résumé du modèle.

Puis, les prédictions sont générées pour une période spécifique (du 26 janvier 2023 à 2 h 20 au 21 juillet 2023 à 10:50). Les valeurs prédites sont comparées aux données réelles via un tracé

graphique montrant les séries d'entraînement et de test ainsi que les prévisions, permettant une évaluation visuelle de la capacité du modèle à capturer les motifs temporels. Pour aligner les longueurs des séries, les données réelles et les prédictions sont tronquées à la taille minimale commune, ce qui assure une comparaison équitable.

Les performances du modèle sont évaluées à l'aide de trois indicateurs : RMSE, MAE et MAPE. Ces métriques permettent de quantifier la précision du modèle SARIMA par rapport aux données observées, en mettant en évidence sa capacité à modéliser les tendances et la saisonnalité journalière.

Le modèle SARIMA est particulièrement adapté aux séries temporelles présentant des motifs saisonniers clairs, comme la consommation énergétique quotidienne. Comparé aux approches non linéaires comme les LSTM ou les modèles d'ensemble, SARIMA offre une interprétabilité élevée et une efficacité computationnelle pour les séries univariées présentant une saisonnalité marquée. La validation sur un ensemble de tests distinct et l'utilisation des indicateurs standard (RMSE, MAE, MAPE) garantissent une évaluation robuste, adaptée à un jeu de données volumineux (140 000+ points).

2.4 Conclusion

Cette section méthodologique a détaillé un ensemble d'étapes comportant des techniques avancées pour l'analyse et la prévision des séries temporelles de consommation électrique, mettant en avant des approches comme DBSCAN, la forêt d'isolation, MDWS, les auto-encodeurs, LSTM, le stacking et SARIMA. Ces méthodes, adaptées à la détection d'anomalies collectives et contextuelles ainsi qu'à la prédiction des comportements énergétiques, intègrent des prétraitements rigoureux, des imputations par KNN et des évaluations basées sur des métriques telles que MSE, RMSE, MAE et MAPE. En combinant des outils d'apprentissage supervisé, non supervisé et des modèles statistiques, ces approches permettent de capturer les motifs complexes des données tout en gérant les variations saisonnières et contextuelles. Cette méthodologie constitue une base robuste pour répondre aux enjeux de précision et de fiabilité

dans la gestion des données énergétiques, offrant des perspectives concrètes pour optimiser les prévisions et planifier l'énergie.

CHAPITRE 3

ÉTUDE DE CAS : CONDOMINIUM AVEC VÉHICULES ÉLECTRIQUES

3.1 Introduction

Ce chapitre présente une étude de cas axée sur l'analyse de la consommation électrique de trois unités d'un condominium (86230, 87112 et 89168), dorénavant désignées « A », « B » et « C », équipées de véhicules électriques, à partir de données collectées entre le 24 janvier et le 21 juillet 2023. L'objectif est d'explorer les séries temporelles de consommation afin d'identifier les tendances, les saisons et les anomalies, tout en tenant compte des spécificités propres à la recharge des véhicules électriques. Cette étude s'appuie sur des techniques de prétraitement, d'analyse descriptive et de restructuration des données afin de faciliter une compréhension approfondie des comportements énergétiques dans les condominiums.

La section commence par une présentation des séries temporelles, mettant en évidence leurs caractéristiques, telles que la variabilité liée à la fréquence élevée des mesures, les motifs saisonniers et les interruptions observées, notamment en avril. Une analyse descriptive détaille les statistiques de base, les distributions asymétriques et les pics liés aux véhicules électriques, et s'enrichit de facteurs contextuels tels que la température, l'heure du jour et le type de jour. Enfin, la restructuration des données est décrite, expliquant la transformation des fichiers bruts en un format enrichi, prêt pour des analyses avancées. Cette étude vise à poser les bases d'une gestion énergétique optimisée dans un contexte résidentiel, en présence de véhicules électriques.

3.2 Présentation des séries temporelles

Nous avons un ensemble de données concernant la consommation électrique de trois appartements spécifiques, identifiés par « A », « B » et « C ». Ces enregistrements illustrent les puissances instantanées sollicitées par ces logements, sur la période du 24 janvier 2023 au 21 juillet 2023.

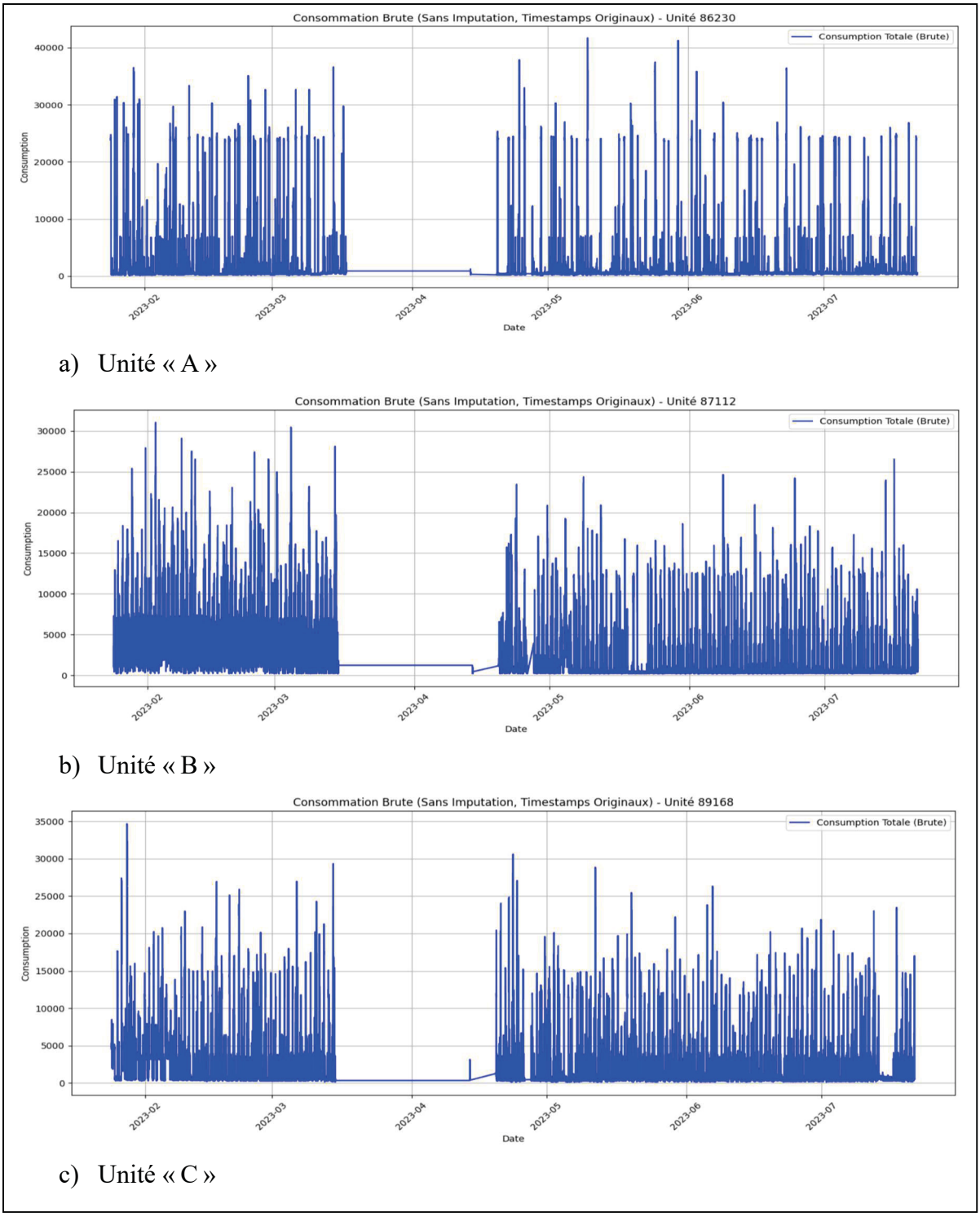


Figure 3.1 Données de consommation d'électricité brutes pour les unités « A », « B » et « C »

En prenant le temps d'analyser ces informations, plusieurs caractéristiques nous apparaissent. Premièrement, il est possible de discerner des motifs récurrents de façon journalière suggérant la présence de saisonnalités dans la consommation; la décomposition de la série nous fournira plus d'informations sur ce sujet. De plus, au fil des mois, aucune tendance ne se dessine reflétant probablement des habitudes quotidiennes ou des appareils électriques en fonctionnement constant. Une autre observation intéressante concerne le mois d'avril. Durant cette période, on constate une diminution drastique de la consommation à certains moments. Cette baisse est d'autant plus frappante qu'elle peut avoir coïncidé avec une interruption de la collecte des données, indiquant probablement une coupure d'électricité durant laquelle aucune donnée n'a pu être enregistrée, une absence prolongée ou un problème lors de l'acquisition des données.

3.2.1 Caractéristiques des données

Nous utiliserons, pour nos données, trois caractéristiques jugées pertinentes à partir des études disponibles dans la littérature, à l'instar de Brownlee (2021). Parmi elles :

- **La température extérieure** : Les températures peuvent avoir un impact significatif sur certains types de consommation, notamment sur l'énergie. Par exemple, les températures plus basses en hiver peuvent accroître la consommation d'énergie pour le chauffage. Par conséquent, l'inclusion de cette caractéristique peut améliorer la précision des prévisions.
- **Le moment de la journée : Cette caractéristique correspond au moment de la journée,** en fonction de ce qu'on fait ou du fait qu'on soit à la maison ou non. Cette caractéristique sera représentée par l'heure en raison des comportements variés des utilisateurs et du fait que certains peuvent être ou non en télétravail. C'est une caractéristique pertinente, car la consommation peut varier considérablement selon l'heure de la journée.
- **Le type de jour de la semaine** : Cette caractéristique peut influer de manière significative sur la consommation, car les habitudes de consommation varient selon le jour de la semaine. Par exemple, la consommation pourrait être différente pendant les

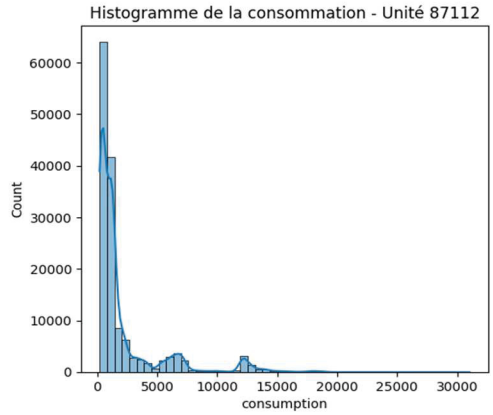
jours ouvrables et les week-ends, selon la présence ou l'absence à la maison, et à des heures différentes.

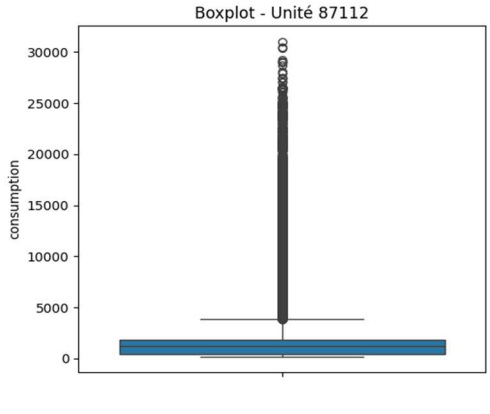
La combinaison de ces caractéristiques pourrait permettre de mieux préciser la détection d'anomalies ou la prédiction.

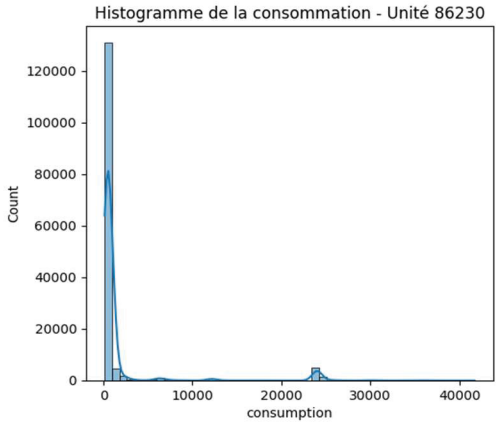
3.2.2 Analyse descriptive des séries temporelles

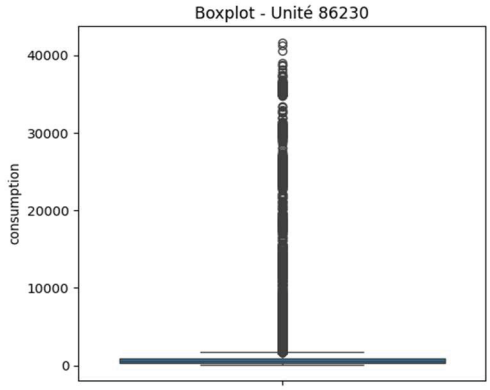
L'analyse descriptive de la consommation d'énergie électrique au sein des différentes unités commence par l'examen des statistiques de base. Nous présenterons les consommations moyennes et médianes, en indiquant le type de distribution potentielle : symétrique ou asymétrique. Nous mentionnerons également les écarts-types, qui indiquent la variabilité des valeurs autour de la moyenne. Nous examinerons les histogrammes afin d'identifier les distributions obtenues, en fonction de leur asymétrie et de leur kurtosis, ce qui suggère des courbes plus ou moins aplaties. Ensuite, les observations temporelles, représentées par les graphiques de la Figure 3.2, permettent d'identifier que la tendance des séries est stable. L'observation des boîtes à moustaches par période montrera la variation de la consommation.

a) Unité « A »

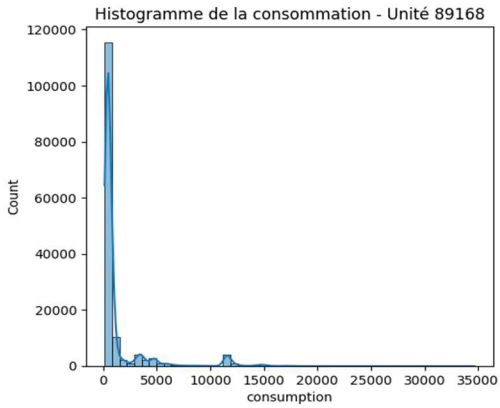


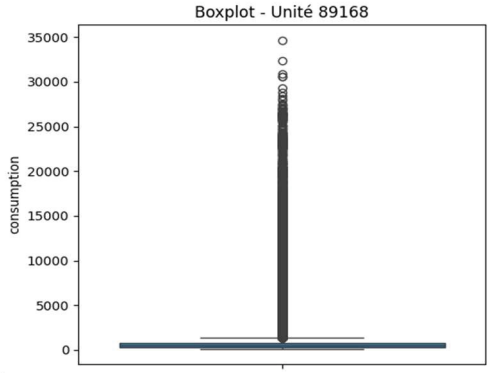






b) Unité « B »



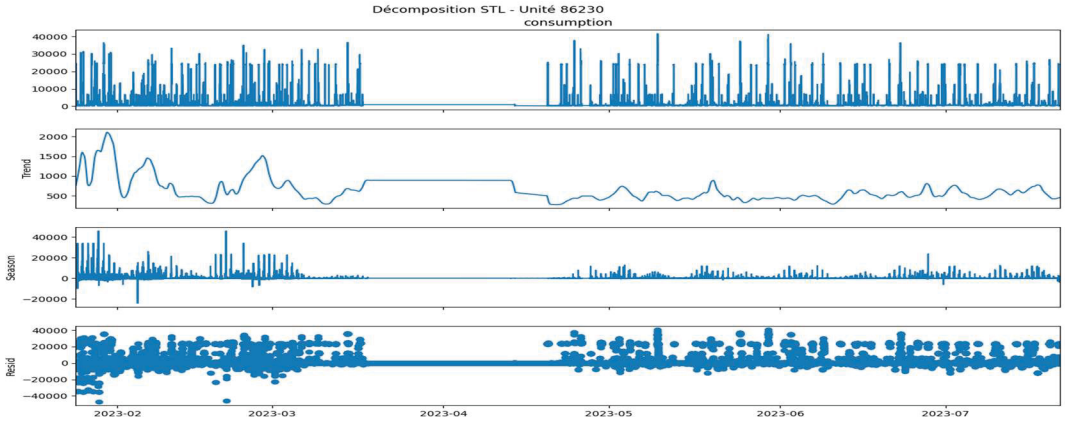


c) Unité « C »

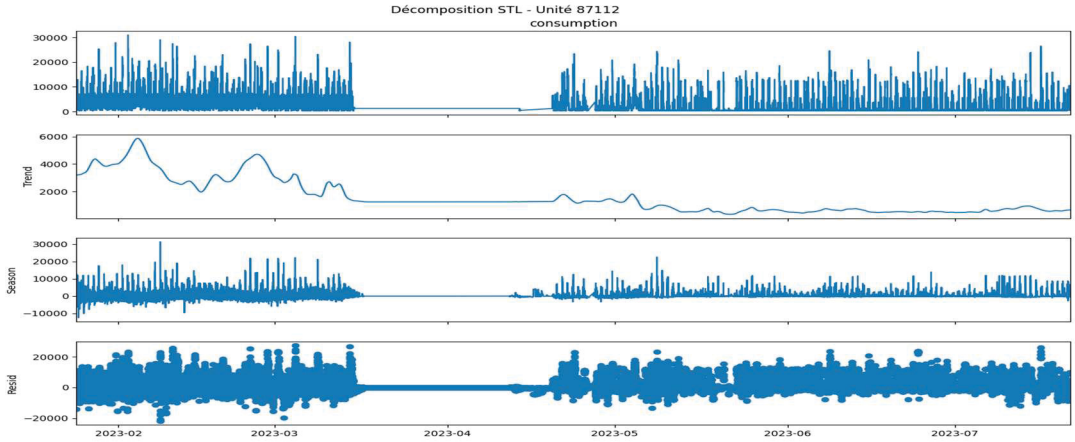
Figure 3.2 Statistiques descriptives d'un signal

L’observation de la consommation électrique des trois unités met en évidence une distribution fortement asymétrique, marquée par une prédominance de valeurs faibles et modérées malgré des pics exceptionnels dépassant régulièrement 30 000 W, comme le révèle la figure 3.2. Les histogrammes indiquent une concentration massive de la consommation en deçà de 2000 W pour les trois unités, avec une décroissance rapide au-delà de cette valeur, tandis que les boîtes à moustaches soulignent une dispersion extrême, ponctuée par un nombre significatif de valeurs aberrantes. Ces dernières se manifestent par des pointes isolées atteignant jusqu’à 30 000 W pour l’unité « A » et 35 000 W pour l’unité « C », voire 40 000 W pour l’unité « B », avec des médianes et des quartiles proches de zéro, ce qui renforce l’idée d’une consommation majoritairement basse, mais ponctuellement intense.

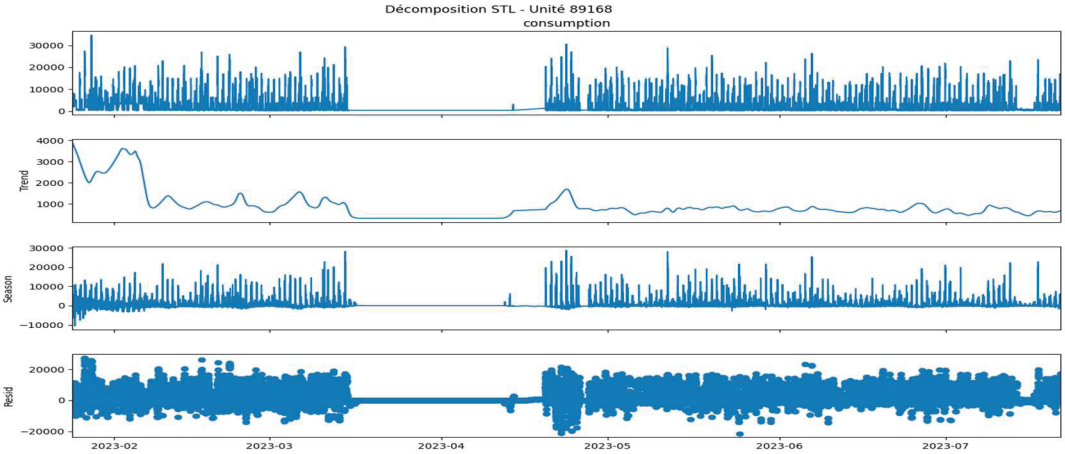
Ces pics de consommation inhabituels confirment fortement la présence d’une charge énergivore au sein du condominium, à savoir, dans le présent cas, la borne de recharge pour véhicules électriques. En effet, la recharge d’un véhicule électrique, particulièrement en mode rapide, est connue pour générer des demandes de puissance élevées sur de courtes périodes, comme l’ont documenté Almaghrebi et al. (2024) dans leurs études sur les profils de consommation liés aux infrastructures de recharge. Cette hypothèse est corroborée par l’observation de ces surconsommations sporadiques, qui coïncident probablement avec les cycles de recharge, influencés par des facteurs tels que les habitudes des utilisateurs ou les capacités des batteries. Une analyse plus fine, intégrant les horaires et les durées de ces pics, permettrait de confirmer cette tendance et d’identifier des patrons spécifiques, ouvrant la voie à une optimisation de la gestion énergétique, notamment par une planification des recharges pendant les périodes de faible demande ou par un renforcement des infrastructures électriques afin d’absorber ces variations.



a) Unité « A »



b) Unité « B »



c) Unité « C »

Figure 3.3 Décomposition des séries chronologiques

La décomposition de la série temporelle pour les trois unités, illustrée en détail dans la figure 3.3, offre une analyse approfondie des dynamiques de consommation d'énergie de ces unités. L'examen des composantes tendancielles révèle que la consommation est élevée en début d'année mais devient presque linéaire à partir du mois de mars ce qui indique une consommation d'énergie globalement constante au fil du temps. Cette constance suggère que les facteurs influençant la demande énergétique de ces unités restent relativement constants sur cette période, bien que de légères fluctuations puissent être observées.

La composante saisonnière, bien que présente, apparaît faible et caractérisée par une irrégularité notable. Cette faible saisonnalité pourrait être attribuée à la fréquence élevée des données collectées, qui capture des variations à court terme, ainsi qu'aux cycles de recharge des véhicules électriques associés à ces unités. Ces cycles irréguliers soulignent la complexité des facteurs influençant la consommation, probablement liés à l'intermittence des recharges et à l'usage variable des véhicules électriques.

Les résidus, quant à eux, dévoilent une richesse d'informations sous la forme de multiples pics soudains de consommation, particulièrement évidents, avec des valeurs allant jusqu'à 30 000 W, qui sont enregistrées. Ces anomalies indiquent la présence probable d'équipements énergivores, tels que des systèmes de recharge rapide ou d'autres appareils à forte consommation d'énergie, dont l'utilisation semble sporadique. Cette forte flottabilité de la consommation d'énergie souligne la nécessité d'une analyse plus approfondie et plus détaillée. Une étude approfondie des périodes d'activité des véhicules, notamment en identifiant les moments précis où ces pics surviennent, pourrait permettre de mieux comprendre les causes sous-jacentes, qu'elles soient liées à des comportements d'utilisation ou à des contraintes techniques. De plus, une investigation spécifique des pics anormaux, en croisant ces données avec des informations contextuelles (telles que les conditions météorologiques ou les profils d'utilisation), serait essentielle pour optimiser la gestion énergétique. Une telle approche pourrait conduire à des stratégies d'ajustement, telles que la planification des recharges pendant les périodes de faible demande ou l'amélioration des infrastructures afin de mieux

absorber ces variations, ce qui assurerait une utilisation plus efficace et durable de l'énergie. Nous proposons un tableau structuré présentant l'analyse descriptive des trois unités :

Tableau 3.1 Analyse descriptive des séries temporelles

Critères	**Unités « A »**	**Unité « B »**	**Unité « C »**
Moyenne (W)	1880.29	2156.01	1434.73
Médiane (W)	491.6	1211.24	485
Écart-type (W)	5236.4	3245.76	2838.01
Distribution	Asymétrique	Asymétrique	Asymétrique
Asymétrie	3.99	2.81	3.67
Kurtosis	14.75	8.6	14.69
Intervalle de variation (W)	126.36 – 41708.72	206.68 – 31034.48	136.6 – 34674.22
Cycle saisonnier (jours)	0.0035	0.0035	0.0035

La moyenne de consommation d'énergie pour les trois unités se situe entre 1443.51 W et 2156.01 W, tandis que les valeurs maximales atteignent plus de 40 kW, avec des écarts-types allant de 2838 W à plus de 5 kW.

L'écart-type élevé indique que la consommation d'énergie est fortement variable, alternant entre des périodes de faible consommation et des pics de forte demande. Cette variabilité peut être due à la fréquence élevée d'acquisition des données, qui capture des fluctuations fines. Si cela est pertinent pour analyser la consommation d'un véhicule électrique, ce niveau de détail est moins adapté à une analyse globale d'un appartement ou d'une maison.

Par ailleurs, un écart-type aussi élevé suggère la présence d'anomalies ou d'erreurs de mesure, ou une hausse des données normales en raison de l'existence ici d'un équipement énergivore, le VÉ. Une analyse plus approfondie sera nécessaire pour identifier et exclure ces valeurs aberrantes, afin d'éviter qu'elles ne biaisent les conclusions.

Ces observations mettent en évidence une consommation d'énergie très instable, avec des pics occasionnels extrêmement élevés. Une analyse détaillée sera menée pour mieux comprendre

les tendances sous-jacentes, les causes de ces variations et d'éventuelles anomalies dans les données.

3.3 Description du fichier généré après restructuration

Le fichier initial, nommé DatasMK_type_[Numéro de l'unité].csv, est au format CSV (Comma-Separated Values) et contient les données brutes collectées. Il inclut une colonne Date représentant les horodatages (ex. 2023-01-01 00 :00 :00), une colonne « Consumption » indiquant la consommation électrique en watts sous forme de nombres flottants (ex. 1234.56), ainsi que d'éventuelles colonnes supplémentaires non standardisées, telles que des mesures brutes de température ou des indicateurs divers caractérisant les données et permettant une analyse plus minutieuse. Une ligne typique se présente ainsi : 2023-01-01 00 :00 :00, 1234.56. Ce fichier constitue la base de données brute et nécessite un prétraitement pour être utilisable.

La transformation appliquée pour restructurer les données comprend plusieurs étapes : les valeurs manquantes sont nettoyées par interpolation linéaire, et des colonnes contextuelles sont ajoutées, notamment « Weekdays » (jour de la semaine, ex. « Monday »), « Temperatures » (température ambiante en °C, ex. 20.5), et « Daytime » (indicateur de l'heure de la journée). La fréquence des horodatages est fixée à 5 minutes afin de garantir une régularité temporelle. Le fichier résultant, DonnéesEtiquete_89168.csv, conserve le format CSV et présente une structure enrichie : Date (horodatage normalisé toutes les 5 minutes), Valeur (consommation renommée), Weekdays, Temperatures et Daytime. Un exemple de ligne est : « 2023-01-01 00:00:00, 1234.56, Monday, 20.5, 0 ». Ce fichier restructuré est prêt pour l'étape d'étiquetage et sert de base aux analyses suivantes.

3.4 Conclusion

Cette étude de cas a permis d'analyser en profondeur la consommation électrique de trois unités d'un condominium équipées de véhicules électriques, à partir de séries temporelles collectées à haute fréquence entre janvier et juillet 2023. L'examen des données a révélé des motifs saisonniers, une tendance globalement stable et des pics de consommation significatifs,

attribuables à la recharge des véhicules électriques, ainsi que des anomalies liées à des interruptions, notamment en avril. L'analyse descriptive, enrichie par des caractéristiques contextuelles telles que la température, le moment de la journée et le type de jour, a mis en évidence une distribution asymétrique et une forte variabilité, confirmant la nécessité d'un prétraitement rigoureux. La restructuration des données en un fichier enrichi, intégrant des colonnes normalisées et contextuelles, a facilité la préparation des analyses ultérieures. Ces résultats soulignent l'importance d'une approche méthodique pour comprendre et optimiser la gestion énergétique dans des contextes résidentiels intégrant des infrastructures de recharge, offrant une base solide pour des études futures portant sur la prévision et la détection d'anomalies.

CHAPITRE 4

RÉSULTATS ET DISCUSSION

4.1 Introduction

La présentation des résultats se fera dans l'ordre de la méthodologie. On va ainsi mettre en évidence des résultats chiffrés et visuels. D'une part, les résultats chiffrés, matérialisés par les indicateurs clés de performance, seront présentés sous forme de tableaux à chaque étape et pour chaque unité prémentionnée. D'autre part, les résultats visuels se présenteront sous forme de figures dont nous analyserons les indicateurs. Ces résultats seront suivis de discussions axées sur les apports de cette étude à la gestion des réseaux électriques intégrant les véhicules électriques, à travers des réglementations et des planifications bien conçues.

Les paramètres des différents modèles ont été obtenus par un processus d'optimisation systématique visant à maximiser les performances prédictives. Une recherche par grille couplée à une validation croisée temporelle a permis d'explorer différentes configurations architecturales. Dans l'exemple du LSTM on optimise les paramètres suivants : nombre de couches LSTM, nombre de neurones par couche, taux de dropout, fenêtre temporelle rétrospective, taille de batch et nombre d'époques d'entraînement. Pour chaque combinaison testée, les modèle ont été évalués sur plusieurs plis de validation. La configuration finale retenue minimise l'erreur moyenne de test tout en offrant le meilleur compromis entre performance prédictive, stabilité et efficacité computationnelle.

Nous notons que l'IQR glissant, utilisé pour étiqueter les données dans le cadre de la détection d'anomalies, identifie les valeurs hors de l'intervalle calculé sur une fenêtre temporelle mobile. Cette approche adaptative permet de mieux capturer les variations saisonnières et les changements de régime par rapport à un IQR global. La méthode reste robuste, simple et sans hypothèse distributionnelle. Cependant, elle présente certaines limitations : elle se concentre principalement sur les anomalies ponctuelles et détecte difficilement les anomalies collectives (séquences anormales). De plus, lorsque les valeurs de consommation sont très proches de zéro

ou présentent une faible variabilité, l'IQR devient très petit, rendant les seuils de détection excessivement sensibles et générant potentiellement de nombreux faux positifs. Les travaux futurs pourraient explorer des approches complémentaires comme l'IQR conditionnel (segmenté par contexte temporel) ou des méthodes alternatives telles qu'Isolation Forest et LSTM pour une détection multi-niveaux intégrant à la fois les anomalies ponctuelles et contextuelles.

4.2 Traitement des données

Afin d'évaluer nos techniques de détection d'anomalies, nous utilisons des métriques de performance. La détermination des vraies anomalies repose sur l'utilisation de la boîte à moustaches. On définit ainsi les extrêmes de la boîte comme Q_1-1.5IQR et Q_3+1.5IQR (avec Q_1 : premier quartile ou 25^e percentile, Q_3 : troisième quartile ou 75^e percentile et écart interquartile, IQR = Q_3 - Q_1) ; on fera un ratio du nombre d'anomalies détectées par les techniques proposées sur le nombre d'anomalies détectées par la boîte à moustaches. Ainsi, on étiquette les données en dehors de la boîte à moustaches comme des anomalies.

Pour l'étiquetage des anomalies collectives, nous avons porté notre attention sur le niveau minimal de consommation d'électricité des logements. En effet, la consommation d'électricité n'est jamais nulle, car bien des appareils électriques et électroniques consomment de l'énergie même éteints (RNCAN, 2014). Selon le modèle d'Hydro-Québec, un logement modeste comprend :

- Une console de jeux de type familial ;
- Un décodeur numérique pour télévision ;
- Une imprimante ;
- Un ordinateur de bureau ;
- Un ordinateur portable ;
- Un téléviseur récent.

Ceci peut représenter une consommation en mode veille de 179,6 kWh/année (Hydro-Québec, 2025A). Dans bien des foyers, cette consommation en mode veille est sans doute encore nettement plus élevée. D'ailleurs, Ressources naturelles Canada, dans une publication de 2014, soutient qu'une telle consommation représente au moins 5 % de l'électricité consommée par un ménage canadien (RNCAN, 2014). Aujourd'hui, plus d'une décennie plus tard, ce pourcentage a certainement bondi. De fait, dans une publication récente du Gouvernement du Canada, on peut lire que la consommation en mode veille peut représenter jusqu'à 10 % de la consommation totale (RNCAN, 2025).

La consommation moyenne d'électricité dans une habitation de type multilogement s'élève à environ 14 000 kWh/année (Hydro-Québec, 2025B). Cette consommation moyenne d'électricité comprend environ 22 % pour les appareils électroménagers et électroniques et 5 % pour l'éclairage. Ainsi, il y a environ 27 % ou 3 780 kWh/année d'électricité consommée par des appareils disposant de la fonction mise en veille. En utilisant les pourcentages de 5 % et 10 %, on peut estimer la consommation en mode veille. Les résultats sont présentés dans le Tableau 4.1.

Tableau 4.1 : Consommation en mode veille pour une habitation de type multilogement

Pourcentage de la consommation d'électricité	Consommation annuelle	Consommation journalière
5 %	189 kWh / année	0,52 kWh / jour
10 %	378 kWh / année	1,04 kWh / jour

En comparaison avec le modèle d'Hydro-Québec, la consommation en mode veille à 5 % correspond à un logement modeste avec très peu d'appareils électroniques. La consommation à 10 %, quant à elle, correspond à un logement comprenant :

- Une console de jeux de type familial ;
- Une console de jeux haute performance ;
- Un décodeur et enregistreur numérique ;

- Deux imprimantes ;
- Un ordinateur de bureau ;
- Deux ordinateurs portables ;
- Deux téléviseurs récents.

La consommation à 10 % offre un portrait plus réaliste du nombre d'appareils que l'on trouve dans un logement contemporain. Ainsi, la consommation de 1 kWh/jour sera utilisée comme valeur seuil. Une consommation inférieure à 1 kWh/jour sera étiquetée comme anormale.

Pour ce qui est des tracés fournis par le code, dans le cas des anomalies collectives, les lignes rouges permettent de distinguer les journées considérées comme telles, tandis que, pour les anomalies ponctuelles et contextuelles, elles se distinguent par des points rouges. Cependant, nous ne nous attarderons pas à montrer l'ensemble des résultats graphiques, mais nous montrerons seulement l'une des unités, puis l'ensemble des résultats numériques sera présenté.

4.2.1 Anomalies ponctuelles et contextuelles

Afin de détecter les anomalies, nous avons utilisé deux méthodes, à savoir l'auto-encodeur et MDWS. Nous détectons à présent les anomalies ponctuelles et contextuelles à l'aide de la méthode proposée par Sagoolmuang et Sinapiromsaran (2017) : le score de sous-séries de la fenêtre de différence médiane. En raison du fait que l'on dispose de domiciles dotés de bornes de recharge, nous constatons que la distribution des données est asymétrique et positive ; un outil aussi robuste que le réseau de neurones ou le MDWS serait requis. Les deux outils nous fournissent des observations différentes. Nous allons présenter les résultats comme mentionné précédemment.

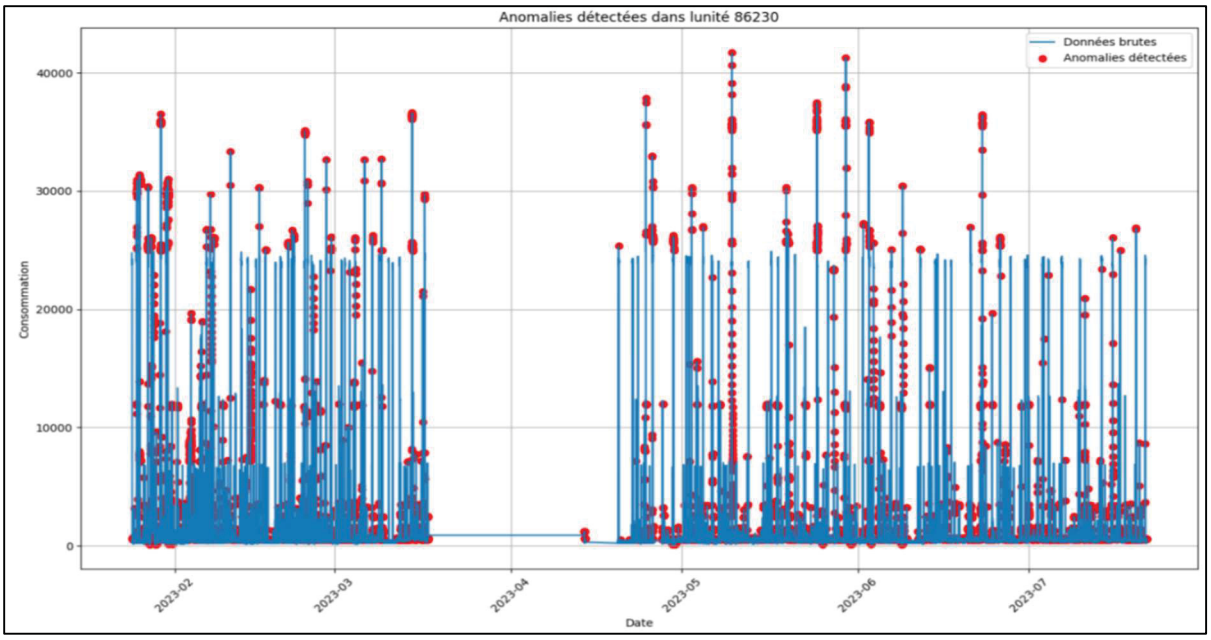


Figure 4.1 Anomalies détectées par réseau de neurones auto-encodeur : unité « A »

Le graphique de la figure 4.1 présente la détection des anomalies ponctuelles et contextuelles pour l'unité « A ». Les observations graphiques sont les mêmes pour toutes les unités : les anomalies en rouge sont particulièrement marquées coïncidant avec des variations notables dans les données brutes en bleu, ce qui reflète la présence de bornes de recharge marquant une différenciation entre ce type d'unité et une unité sans borne de recharge déjà analysée dans plusieurs articles. L'utilisation de l'auto-encodeur pour la détection d'anomalies ponctuelles et contextuelles semble efficace, car il capture des écarts significatifs au sein des trois unités. Nous discuterons de l'efficacité en abordant les résultats numériques.

4.2.2 Anomalies collectives

La détection des anomalies collectives se fait ici en vérifiant, sur une journée, les intervalles inférieurs à un certain seuil d'écart-type et/ou de moyenne. Les méthodes utilisées ici sont DBSCAN et Isolation Forest. La journée anormale détectée est représentée par une ligne rouge, tandis que les vraies anomalies précédemment imputées sont identifiées en vert. Nous dénombrons ainsi les journées anormales et les comparons aux véritables moments d'absence définis en amont.

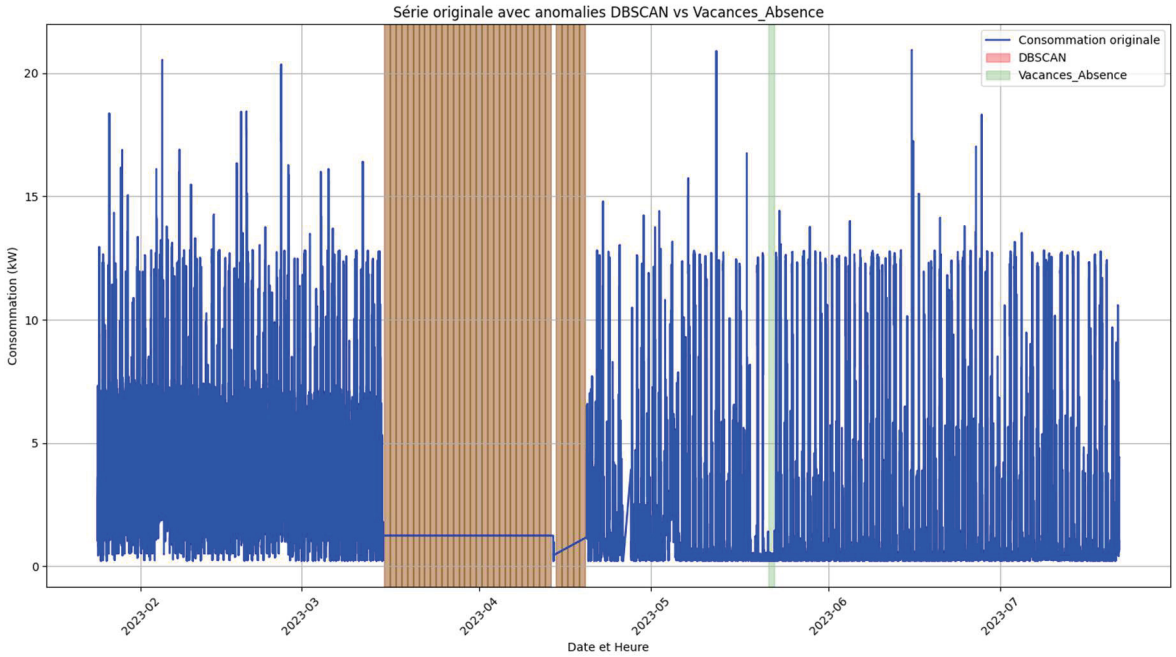


Figure 4.2 : Détection d'anomalies collectives par DBSCAN : unité « B »

La figure 4.2 présente pour l'unité « B », utilisée ici à titre d'exemple, la détection d'anomalies collectives via l'algorithme DBSCAN, comparée aux périodes de vacances et d'absences réelles. Les périodes anormales sont observées principalement pendant le mois d'avril et également en dehors et correspondent en grande partie aux moments d'absence dans cette unité, ce qui montre graphiquement l'efficacité de cet algorithme.

Le tableau suivant montre les résultats de précision, de rappel et du F1-score obtenus pour les différentes techniques en fonction de l'unité. Deux types d'anomalies ont été ciblés : les anomalies ponctuelles/contextuelles (Auto-Encodeur, MDWS) et les anomalies collectives (DBSCAN, Isolation Forest). Les résultats sont présentés dans le tableau suivant :

Tableau 4.2 Comparaison des techniques de détection

Unités	Types d'anomalies	Modèle	Précision (%)	Rappel (%)	F1-Score (%)
« A »	Anomalies ponctuelles et contextuelles	Auto-Encodeur	**55,39**	**88,76**	**66,89**
		MDWS	42,77	57,57	49,08
	Anomalies Collectives	DBSCAN	**62,5**	72,06	**65,10**
		Isolation Forest	51,7	**98,61**	60,29
« B »	Anomalies ponctuelles et contextuelles	Auto-Encodeur	**84,23**	**94,74**	**89,17**
		MDWS	44,23	66,35	52,09
	Anomalies Collectives	DBSCAN	**75,00**	**75,00**	**75,00**
		Isolation Forest	50,00	47,73	48,81
« C »	Anomalies ponctuelles et contextuelles	Auto-Encodeur	**75,64**	**83,55**	**79,40**
		MDWS	51,32	57,71	54,33
	Anomalies Collectives	DBSCAN	**100**	**100**	**100**
		Isolation Forest	75,00	69,23	71,74

Le tableau présente une évaluation des performances de différents modèles de détection d'anomalies (Auto-Encodeur, MDWS, DBSCAN, Isolation Forest) appliqués aux trois unités pour la détection des deux types d'anomalies étudiés ici.

Pour ce qui est des anomalies ponctuelles et contextuelles, l'auto-encodeur offre un bon rappel, mais une faible précision, avec un F1-Score modéré compris entre 66,89 % et 89,17 %. Ce qui fait qu'il est meilleur que MDWS dont les performances sont moins bonnes avec une précision médiane de 44,23 % et un rappel bien plus faible que celui de l'auto-encodeur (moyenne d'environ 60 %), donnant un F1-Score entre 49,08 % et 54,33 %.

Du point de vue des anomalies collectives, DBSCAN excelle avec une précision et un rappel moyens supérieurs à 80 %, alors qu'Isolation Forest, en dépit de cela, offre des résultats acceptables mais moins bons. On constate donc que le F1-Score moyen de ce dernier est inférieur de 20 % à celui de DBSCAN, ce qui montre la capacité de ce modèle à capturer les motifs anormaux tout en évitant de qualifier les normaux comme anormaux.

Ainsi, DBSCAN se distingue pour les anomalies collectives ; l'Auto-Encodeur est plus efficace pour les anomalies ponctuelles et contextuelles. Ils semblent donc s'adapter à ce contexte où la recharge d'un véhicule électrique ajoute aux données ce qui, dans d'autres contextes, constitue une anomalie évidente. MDWS, conçu pour les anomalies contextuelles, fournit des résultats moins satisfaisants que l'Auto-Encodeur ; il peut être amélioré pour ce type de contexte.

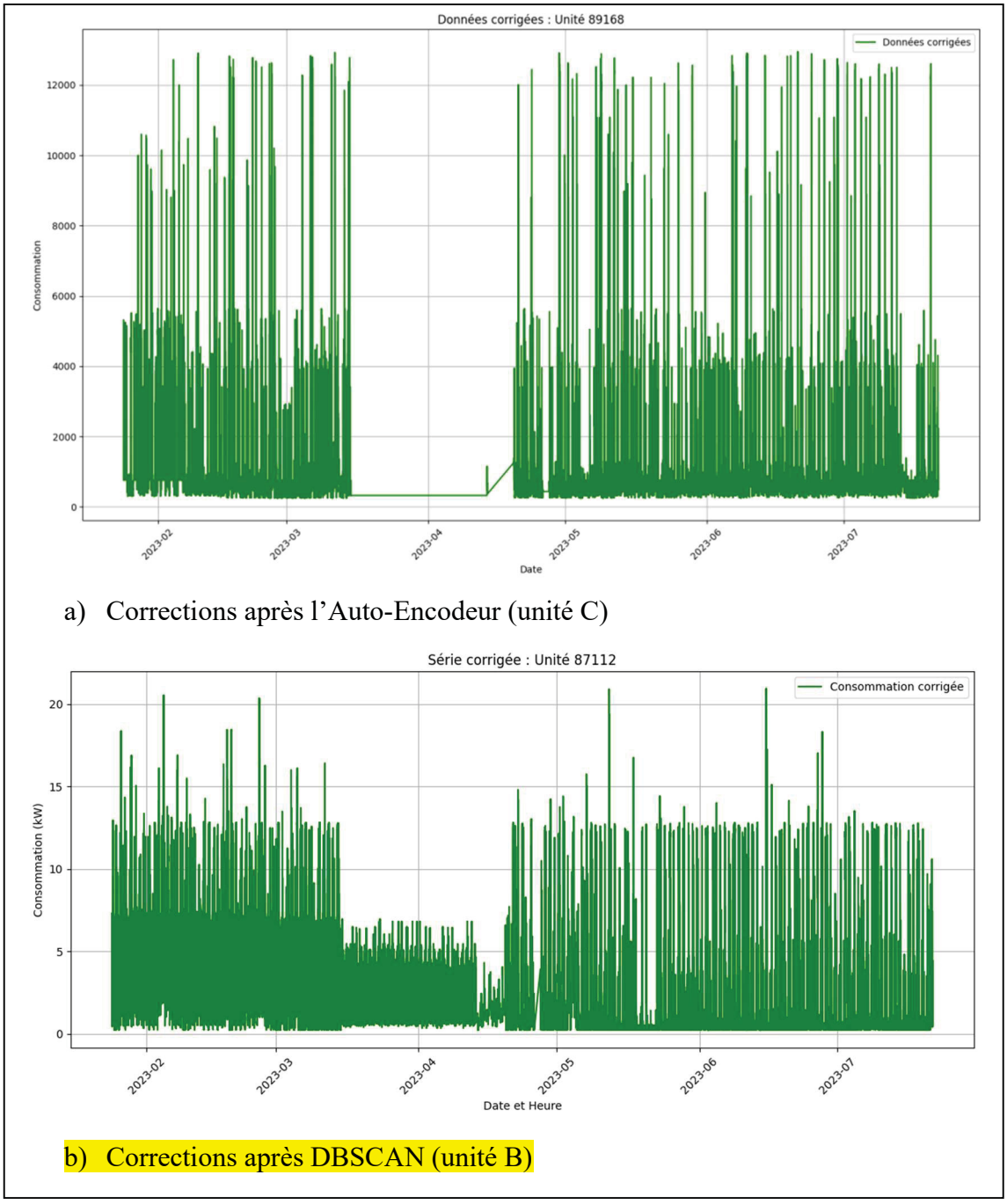
Isolation Forest présente une précision parfaite, mais un rappel faible pour les anomalies collectives, sauf pour l'unité « C ». Cela peut être dû au fait que cette dernière unité a des données recueillies sur un intervalle de temps plus régulier que celui des autres unités.

Les performances varient selon les unités, ce qui suggère que le choix du modèle dépend du type d'anomalie et des caractéristiques des données de chaque unité.

Ces résultats révèlent une complémentarité des approches : l'Auto-Encodeur est adapté aux anomalies ponctuelles, tandis que DBSCAN excelle pour les anomalies collectives, particulièrement sur des données complexes comme celles de l'unité « C ». La faible performance d'Isolation Forest suggère une limite de cette méthode pour les séries temporelles de consommation d'énergie électrique dans un domicile possédant une borne de recharge, incitant à explorer des ajustements ou des alternatives. Ces observations guideront les étapes de prédiction, en s'appuyant sur les données corrigées par les modèles les plus performants.

4.2.3 Correction des données pour « C »

Les données brutes de DonnéesEtiquete_89168.csv ont été corrigées en deux étapes pour traiter les anomalies ponctuelles, contextuelles et collectives, produisant successivement Données_1_AE_89168.csv et Données_2_DBSCAN_89168.csv.



a) Corrections après l'Auto-Encodeur (unité C)

b) Corrections après DBSCAN (unité B)

Figure 4.3 : Données corrigées à la suite de l'application de l'Auto-Encodeur et DBSCAN

Dans le premier contexte, l'imputation par KNN avec 5 voisins a été appliquée après la détection d'anomalies par un auto-encodeur, en s'appuyant sur les relations locales entre la consommation totale et la consommation de la borne de recharge. La transition abrupte vers

une consommation quasi nulle autour de mars-avril 2022, suivie d'une remontée graduelle, suggère que les valeurs manquantes ou aberrantes ont été remplacées par des estimations tirées de voisins temporels proches, ce qui reflète les motifs préexistants. Cependant, la faible variabilité observée pendant cette période pourrait indiquer un biais si les voisins sélectionnés étaient insuffisants ou peu représentatifs, ce qui pourrait sous-estimer la consommation réelle. La régularité des pics avant et après cette période montre que l'imputation a préservé les structures locales, mais l'absence de données contextuelles supplémentaires limite la précision de l'interprétation.

Dans le second contexte, l'imputation intègre une fenêtre temporelle de ±7 jours et des caractéristiques telles que les jours de la semaine, le moment de la journée, les températures et la consommation de la borne de recharge, en excluant les données de vacances. La période d'anomalie autour de mars-avril 2022 pourrait correspondre à une détection DBSCAN à faible écart-type ou à une période de vacances, imputée à partir de données « non-vacances » (non étiquetées comme des vacances) dans la fenêtre. La reprise progressive post-anomalie suggère une interpolation contextuelle réussie, ajustée aux motifs saisonniers ou horaires, ce qui renforce la robustesse grâce à l'utilisation de caractéristiques supplémentaires. Toutefois, si les données environnantes étaient rares ou biaisées (par exemple, peu de jours non-vacances), l'imputation pourrait avoir introduit des erreurs systématiques, bien que la structure générale des pics reste cohérente. Globalement, les deux approches semblent avoir corrigé les anomalies de manière satisfaisante, mais la seconde méthode, par son caractère contextuel, offre une meilleure adaptation aux variations temporelles observées.

4.3 Prédiction des séries temporelles

4.3.1 Résultats

Nous avons utilisé plusieurs techniques de prévision, parmi lesquelles Sarima, LSTM et l'empilement. Les métriques utilisées pour évaluer ces différents modèles incluent RMSE (en watts), MAE (en watts), MAPE (en %), et le temps d'exécution (en secondes), comme présenté dans le tableau 4.3. Le MAPE, malgré sa sensibilité aux valeurs proches de zéro, demeure

approprié car les consommations observées sont significativement éloignées de zéro, avec une tendance moyenne autour de 1500W. Cette métrique exprime l'erreur en pourcentage, offrant une interprétation intuitive indépendante de l'échelle des données. Elle facilite la comparaison avec la littérature scientifique et complète les perspectives fournies par le RMSE et le MAE pour une évaluation multicritères du modèle.

Tableau 4.3 Comparaison des modèles prédictifs

Unité	Modèle	RMSE (W)	MAE (W)	MAPE (%)	Temps d'exécution (s)[1]
« A »	Stacking_Ensemble	172,6	129,8	29,67	3588
	SARIMA	151,9	113,0	24,87	**10**
	LSTM	**101,5**	**59,3**	**12,61**	1805
« B »	Stacking_Ensemble	983.3	806.8	155.81	6436
	SARIMA	796.5	661.0	166.27	**10**
	LSTM	**568,0**	**345,3**	**39,34**	1537
« C »	Stacking_Ensemble	759,6	488,3	98,85	4437
	SARIMA	484,4	279,6	67,01	**10**
	LSTM	**446,1**	**182,4**	**27,27**	1938

Tout d'abord, examinons les écarts entre les métriques (RMSE, MAE, MAPE) afin d'évaluer la cohérence des prédictions. Sur « A », le RMSE du LSTM (101.5 W) dépasse largement son MAE (59.3 W), ce qui indique que les données sont correctement nettoyées des valeurs aberrantes, ce qui est cohérent avec un MAPE de 12,61 % et une distribution asymétrique (asymétrie de 3,67, écart-type de 2838,01 W). Également, sur « C », où le RMSE (446.1 W) et le MAE (182.4 W) sont plus proches, le MAPE de 29,09 % suggère une bonne capture des

[1] Ordinateur utilisé : MacBook Air 10, processeur M1, 8 cœurs et RAM 8 Go, 3,2 GHz

variations relatives, malgré une dispersion élevée (intervalle 137.8-30 609.96 W), ce qui reflète une fidélité aux données brutes (taux de reconstruction faible).

Ensuite, le temps de LSTM élevé (1500-1900 s) sur 40 époques contraste avec celui de SARIMA (10 s), ce qui suggère un entraînement intensif adapté aux motifs temporels courts (cycle de 0.0035 jours) ; le modèle ARIMA vaut la peine d'être mieux analysé en fonction des résultats que nous observons. L'ensemble d'empilement (3500-6000 s) présente un coût computationnel élevé, potentiellement dû à l'agrégation de RandomForest, MLP, SVR et XGBoost, mais son MAPE élevé pour les unités « B » et « C » indique une difficulté à généraliser à d'autres unités.

Enfin, le bon prétraitement des données favorise LSTM, dont les capacités séquentielles s'alignent avec les transformations induites par les corrections, comme l'illustrent les graphiques de consommation corrigée, et une validation croisée temporelle nous a permis de renforcer notre analyse et de la rendre plus crédible.

4.3.2 Illustration graphique

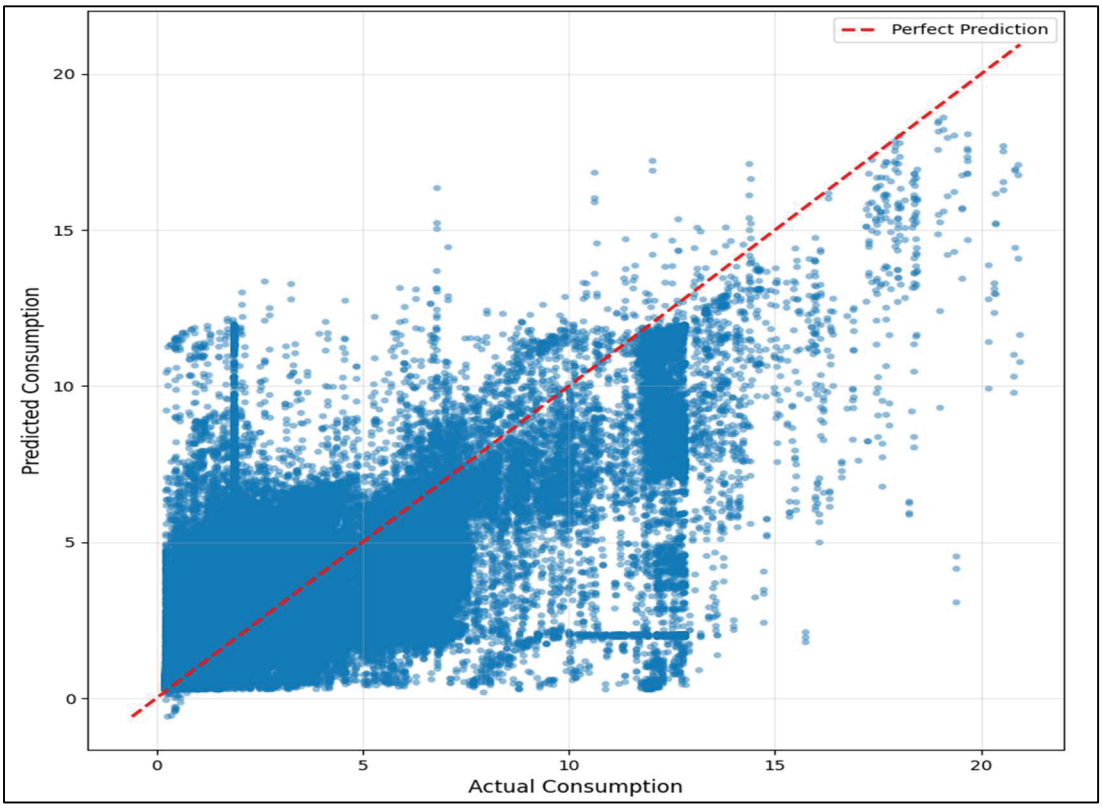


Figure 4.4 : Diagramme de dispersion des prédictions (unité B)

La Figure 4.4 présente un diagramme de dispersion confrontant les valeurs prédites aux valeurs réelles de consommation pour l'ensemble des données (entraînement et test confondus).

On observe que la majorité des points se concentre autour de la diagonale de référence, confirmant la capacité du modèle à capturer les tendances globales de consommation. Cette distribution suggère l'absence de biais systématique majeur dans les prédictions, le modèle ne surestimant ni ne sous-estimant systématiquement les valeurs de consommation.

On voit également l'augmentation de la dispersion des erreurs avec l'amplitude de la consommation. Pour les valeurs faibles à moyennes (0 à 10 kW), les prédictions sont particulièrement précises avec une dispersion réduite. En revanche, pour les consommations élevées (au-delà de 15 kW), la variance des erreurs s'accroît considérablement. Ceci indique que le modèle rencontre davantage de difficultés à prédire avec précision les pics de

consommation. Cette limitation pourrait s'expliquer par la rareté relative de ces événements dans l'ensemble d'entraînement ou par la complexité des facteurs sous-jacents à ces consommations exceptionnelles non anormales.

Les performances obtenues positionnent le modèle LSTM comme un outil viable pour la prédiction de consommation dans ce type d'application. Toutefois, plusieurs pistes d'amélioration se dégagent de l'analyse :

- Pour améliorer la prédiction des pics de consommation, l'adoption d'une fonction de perte personnalisée pénalisant davantage les erreurs sur les valeurs élevées constituent des voies prometteuses.
- Une approche hybride combinant une phase de segmentation préalable (par clustering) avec des modèles spécialisés pour chaque régime de consommation identifié pourrait permettre de mieux traiter l'hétérogénéité apparente des motifs

4.4 Interprétation des résultats

Comme l'indiquent de nombreuses études citées dans la littérature, il est important de souligner que l'intégration des véhicules électriques entraîne un changement significatif de la consommation d'énergie. En effet, ces dispositifs, particulièrement gourmands en énergie, représentent instantanément environ 80 % de la consommation totale d'une unité résidentielle. Les structures telles que les copropriétés et les immeubles à logements multiples ne sont pas initialement conçues pour gérer des augmentations de charge imprévisibles, notamment en raison de l'intégration rapide de diverses charges à l'instar des véhicules électriques. L'examen des séries temporelles, à la lumière de l'analyse descriptive menée dans le chapitre 3, révèle des tendances similaires de la consommation d'énergie, avec ou sans l'usage de véhicules électriques. Bien que l'on puisse anticiper des pics de consommation plus marqués, suggérant des anomalies contextuelles, l'identification de véhicules électriques reste possible.

L'analyse des données met également en lumière le phénomène de recharges de véhicules électriques survenant pendant et après les heures de pointe. Ce comportement de recharge peut varier selon le mode de travail des individus, qu'ils soient en télétravail ou sur leur lieu de

travail physique. Ces observations expliquent en partie pourquoi la consommation énergétique est si élevée, outre l'utilisation habituelle d'appareils énergivores.

4.4.1 Gestion des irrégularités

En analysant les anomalies détectées, nous pouvons dire que, d'une part, il n'y a pas vraiment de différences de contexte entre les saisons. Le véhicule électrique est réputé consommer plus d'énergie en hiver qu'en été. À l'exception d'une courte période au début du mois de février, la moyenne de la consommation est quasi constante sans aucune tendance.
Une remarque est que, dans les unités analysées, une période creuse ressemble à des vacances ou à une interruption de l'acquisition des données en avril. Cela porte à croire, dans le cas où il s'agit de vacances, que ce sont des logements sans enfants et que la consommation serait donc plus élevée, d'environ 20 % à 30 % par personne (Les furets, 2023). Ce genre d'aperçu offre une marge de manœuvre pour mieux établir les programmes de gestion de la demande en requérant certains utilisateurs (en toute confidentialité) d'inscrire leurs vacances afin de réduire la charge du réseau électrique durant cette période.

4.4.2 Prédiction de la consommation

L'étude a comparé les performances de prévision des modèles SARIMA, LSTM et Stacking Ensemble, en utilisant les métriques RMSE, MSE, MAE et MAPE. Le LSTM a surpassé les méthodes statistiques traditionnelles, offrant une modélisation plus précise du comportement énergétique des ménages avec des véhicules électriques. Les autres modèles utilisés ont montré des signes de surajustement, alors que l'approche LSTM a excellé, ce qui suggère que l'utilisation de réseaux de neurones pour l'analyse prédictive des comportements humains est efficace, même pour ce type de données, relativement irrégulier, par rapport aux données couramment analysées. Les ajustements de paramètres fondés sur la littérature ont été cruciaux pour affiner la précision. Ces résultats soulignent l'importance de sélectionner et d'optimiser les modèles en fonction de la spécificité des données disponibles et du comportement énergétique multifactoriel.

4.4.3 Gestion énergétique améliorée

Les données collectées et analysées jouent un rôle central dans la mise en œuvre de stratégies de réponse à la demande et de tarification dynamique, deux tactiques cruciales pour gérer efficacement la demande d'énergie, notamment dans le contexte de la recharge des véhicules électriques.

En analysant les données collectées, les fournisseurs d'énergie peuvent comprendre en détail quand la demande est à son apogée et quand elle est relativement faible. Par exemple, les périodes de forte utilisation des véhicules électriques pour la recharge peuvent être identifiées grâce à la détection d'anomalies uniquement dans les données de recharge de ces véhicules. Cette information est essentielle pour concevoir des stratégies de réponse à la demande susceptibles d'inciter les consommateurs à déplacer leur utilisation en dehors des heures de pointe, ce qui contribue à équilibrer la charge sur le réseau électrique.
Du point de vue collectif, sur la base de ces mêmes méthodes de détection d'anomalies et de prédiction, les compagnies d'électricité peuvent mettre en place des structures tarifaires variables. Par exemple, les prix peuvent être augmentés pendant les heures de pointe et réduits pendant les heures creuses pour encourager la recharge des véhicules électriques lorsque la demande sur le réseau est moindre. Cela peut aider à éviter la surcharge du réseau électrique et à réduire le besoin d'investissements coûteux en capacité supplémentaire de production d'énergie. Au niveau individuel, les données des ménages peuvent également être utilisées pour concevoir des programmes de réponse personnalisés à la demande. Par exemple, des incitatifs peuvent être offerts aux consommateurs qui acceptent de retarder la recharge de leurs véhicules électriques jusqu'à des périodes de faible demande. Ces programmes peuvent être automatisés et optimisés à l'aide de données historiques et prédictives sur la consommation d'énergie, voire de données en temps réel.

Les données analysées peuvent également aider à mieux intégrer les sources d'énergie renouvelable. Par exemple, en sachant quand l'énergie renouvelable est abondante et moins coûteuse, les consommateurs peuvent être incités à recharger leurs véhicules électriques pendant ces périodes, contribuant ainsi à une utilisation plus efficace de l'énergie propre.

En somme, les données de consommation d'énergie, lorsqu'elles sont collectées et analysées de manière approfondie, fournissent une base essentielle pour des initiatives sophistiquées de gestion de la demande d'énergie et de tarification, contribuant à une utilisation plus durable et efficace de l'énergie.

La compréhension détaillée de la demande d'énergie, en particulier pour les véhicules électriques, permet une allocation plus précise des ressources au sein de l'infrastructure énergétique. Cela aide les entreprises du secteur énergétique à investir de manière judicieuse, à éviter les dépenses excessives et à concentrer leurs efforts là où la demande est la plus forte. De plus, une analyse approfondie de la consommation d'énergie révèle des tendances cruciales dans l'utilisation des véhicules électriques, ce qui oriente stratégiquement l'expansion des capacités de recharge. L'analyse des données, réalisée ici à l'aide de la prédiction, permet à l'exploitant du réseau de déterminer où de nouvelles stations de recharge sont nécessaires, assurant une répartition optimale des charges pour répondre aux besoins des utilisateurs à des moments précis, tels que les heures de pointe, tout en n'affectant pas la charge électrique des multi-logements. Ce type de parc de recharge (avec sécurité assurée) permettrait à l'exploitant du réseau électrique d'exercer un meilleur contrôle sur la gestion de la charge.

Le renforcement proactif du réseau est également crucial. Avec l'augmentation de la demande, l'infrastructure existante doit être améliorée ; cela inclut l'installation de nouveaux transformateurs, la mise à niveau des lignes de transmission et l'intégration de solutions de stockage d'énergie pour gérer les pics de consommation (par exemple, le renforcement du concept de consommateur-producteur). Cette anticipation des besoins futurs n'est pas seulement réactive ; elle ouvre la porte à l'innovation. L'intégration de stations de recharge rapide, d'options de recharge à domicile améliorées et de technologies avancées répond plus efficacement et plus simplement à la demande.

En résumé, l'analyse détaillée de la demande d'énergie oriente les investissements stratégiques dans l'infrastructure, soutient l'expansion intelligente des capacités de recharge, favorise le renforcement préventif du réseau électrique et stimule l'adoption de technologies innovantes. Cette approche holistique assure une transition harmonieuse vers une société de plus en plus dépendante des véhicules électriques, tout en préservant la stabilité du réseau énergétique.

4.5 Conclusion

Dans ce chapitre, nous avons examiné les résultats de l'analyse des données des unités, en tenant compte des défis liés à la recharge des véhicules électriques. L'auto-encodeur s'est révélé performant pour détecter les anomalies ponctuelles, tandis que DBSCAN a excellé pour les anomalies collectives, avec des variations selon les caractéristiques des données. L'imputation par KNN, appliquée en deux étapes après les modèles de détection, a corrigé efficacement les anomalies, bien que des limites potentielles aient été observées, notamment des temps de traitement prolongés pour certains modèles. En prédiction, LSTM a surclassé les autres méthodes grâce à sa capacité à capturer les motifs temporels complexes, surpassant SARIMA qui a montré des faiblesses face à une saisonnalité irrégulière quasi inexistante.

Les discussions ont souligné l'influence significative des véhicules électriques sur la consommation, ainsi que des tendances telles que les recharges après les heures de pointe, ouvrant des perspectives sur des tarifications dynamiques et des programmes de gestion de la demande. Ces observations encouragent des investissements stratégiques dans l'infrastructure et une intégration accrue des énergies renouvelables dans la mesure où Hydro-Québec estime la consommation moyenne d'un logement à 14 MWh/an, tandis que les logements étudiés ici ont en moyenne 12 MWh/an. Ainsi, il faudra gérer le risque de surcharge en prévoyant des infrastructures s'appuyant sur des prédictions à grande échelle, en tenant compte du changement climatique et de l'intégration de charges de plus en plus énergivores dans les centres de consommation. En somme, cette étude propose une approche intelligente pour la gestion des réseaux électriques, en adaptant les méthodes de détection et de prédiction aux spécificités de chaque unité, tout en soutenant une transition énergétique durable.

CONCLUSION

Au terme de notre étude, nous pouvons dire que l'intégration des bornes de recharge pour véhicules électriques (VE) dans les domiciles peut modifier les profils de consommation et accroître ponctuellement la demande, selon les contextes observés (heures de pointe, recharges simultanées et température), sur le réseau de distribution local. Cette augmentation nécessite une surveillance attentive afin de réduire le risque de surcharge du réseau et de garantir une distribution d'énergie stable et fiable. L'impact est multidimensionnel, affectant non seulement la consommation d'énergie globale, mais aussi les dynamiques de gestion du réseau, notamment en périodes de pic où la demande peut temporairement excéder la capacité disponible. Par ailleurs, la gestion des irrégularités dans les données de consommation s'avère cruciale pour assurer une planification efficace et une réponse adaptative du réseau. Les modèles prédictifs, tels que SARIMA, LSTM et l'empilement, se sont montrés utiles dans le cadre de l'étude pour la prévision de la consommation d'énergie et de la demande de recharge, bien que leur efficacité varie. Ces systèmes utilisent des données historiques et des algorithmes pour prévoir la demande future et peuvent fournir des éléments d'aide à la décision pour réduire les risques, optimiser les opérations du réseau et planifier les investissements en infrastructures. Cependant, leur précision dépend fortement de la qualité des données et de la finesse de calibration des modèles. C'est ainsi qu'une stratégie combinant l'expansion éclairée des infrastructures de VE et l'utilisation judicieuse de la technologie prédictive est essentielle à une gestion énergétique durable et efficace.

Notre étude présente un aspect de l'analyse des données de consommation électrique de domicile avec une borne de recharge qui n'a pas encore été traité. La collecte de données s'est d'abord faite à l'aide de technologies IoT afin de faciliter l'intégration de technologies adaptées, de mieux acheminer et de stocker ces données. L'impact est d'autant plus important dans le secteur de l'énergie que l'on peut considérer que l'emploi de méthodes avancées de collecte et de traitement des données, ainsi que l'utilisation de modèles prédictifs sophistiqués, renforce la précision des prévisions de demande. Cela conduit à une gestion plus efficace du réseau, contribue à une intégration harmonieuse des VE et souligne l'importance de l'innovation continue dans l'analyse de données pour répondre aux défis énergétiques futurs.

Cette étude présente certaines limites. Premièrement, la portée des données, restreinte à un nombre de logements spécifique et une taille d'échantillon limitée, peut entraver la généralisation des résultats à des contextes plus larges. Les comportements de consommation d'énergie peuvent varier considérablement en fonction de facteurs régionaux, climatiques, culturels ou économiques non couverts par notre échantillon. Deuxièmement, la complexité des modèles prédictifs pose des défis. Bien que précis dans notre échantillonnage, ces modèles peuvent ne pas prendre en compte toutes les variables influentes, telles que des événements météorologiques extrêmes, des changements brusques dans le comportement des utilisateurs ou des évolutions technologiques des appareils de consommation d'énergie. Dans cette étude, la température est modélisée par une composante sinusoïdale contrainte par des plages réalistes, cohérente avec la saisonnalité locale. L'intégration de données météorologiques plus précises (stations locales, réanalyses, détails horaires) améliorerait la qualité des estimations. De plus, les modèles nécessitent une quantité significative de données d'entraînement, et leur efficacité pourrait diminuer en l'absence de données actualisées ou représentatives. Ces facteurs peuvent tous affecter la fiabilité des prédictions et leur applicabilité à des scénarios réels ou à plus grande échelle. Cependant, nous pouvons dire que les objectifs que l'on s'est initialement fixés ont été atteints dans la mesure où :

- Une analyse descriptive détaillée (moyenne, médiane, écart-type, asymétrie, kurtosis) a été faite et a bien mis en évidence les particularités des unités A, B et C ;
- Plusieurs techniques de détection d'anomalies (MDWS, auto-encodeur, DBSCAN, Isolation Forest) ont été appliquées, et leurs performances ont été comparées à l'aide de métriques claires (précision, rappel, F1-score). Les anomalies ont été corrigées par KNN ;
- Trois modèles (LSTM, SARIMA, Ensemble d'empilement) ont été testés et comparés sur les trois unités, avec LSTM qui s'est montré le plus performant.

Intégrer la recharge des VE à l'étude des comportements collectifs de consommation d'énergie et des scénarios de mix énergétique est crucial, car elle représente une part importante de la demande énergétique future. Au niveau communautaire, comme dans les immeubles de

copropriété ou les quartiers, la gestion coordonnée de la recharge des VE peut réduire les risques de surcharge du réseau et favoriser l'utilisation efficace des ressources énergétiques, notamment pendant les heures de pointe. En outre, l'adoption de VE au sein d'une communauté peut encourager l'installation d'une infrastructure de recharge partagée, alimentée par des sources d'énergie renouvelable, contribuant ainsi à réduire l'empreinte carbone collective. Cela soulève des questions intéressantes pour la recherche, telles que l'impact des stations de recharge communautaires sur les habitudes de consommation d'énergie et la disponibilité du réseau, ou encore la manière dont les incitatifs à la recharge à faibles émissions peuvent influencer les comportements de recharge.

Nous proposons trois perspectives de recherche que nous trouvons pertinentes pour approfondir davantage ce sujet :

- Modélisation de scénarios avancés de mix énergétique avec VE : Évaluer l'intégration des VE au sein des communautés, en tenant compte de la capacité de recharge, de la disponibilité des renouvelables et de la dynamique de la demande. L'objectif est d'identifier les conditions optimales pour une intégration harmonieuse des VE, en maximisant l'utilisation des renouvelables sans imposer un stress excessif au réseau.
- Analyse des données de microréseaux intégrant les VE : Une étude approfondie portant sur les microréseaux indépendants intégrant les VE fournirait des données concrètes sur la consommation, les habitudes de recharge, la disponibilité des renouvelables, etc. Cela permettrait d'identifier les meilleures pratiques, les défis techniques actuels et les domaines nécessitant une efficacité accrue dans le cas d'espèce.
- Intelligence artificielle pour la gestion optimisée de la recharge des VE : développer des systèmes d'IA qui apprennent les comportements des utilisateurs et les dynamiques du réseau afin de proposer des stratégies de recharge optimisées. Ces stratégies minimiseraient les coûts et les émissions, équilibreraient la demande sur le réseau et pourraient s'adapter aux évolutions des habitudes des utilisateurs ainsi qu'à la disponibilité de l'énergie.

Ces avenues de recherche, en mettant l'accent sur l'analyse technique et les données, pourraient améliorer la manière dont les VE sont intégrés dans les communautés, en mettant

l’accent sur la durabilité, l’efficacité énergétique et une transition harmonieuse vers des sources d’énergie plus vertes. Elles soulignent le rôle crucial des données et de la technologie avancée dans la navigation face aux défis liés à l’électrification des transports.

RECOMMANDATIONS

Pour tirer parti des observations sur la consommation d'énergie influencée par les véhicules électriques, plusieurs recommandations concrètes émergent. Parmi celles-ci :

- Encourager les fournisseurs d'énergie à développer la sensibilisation personnalisée (par courrier, courriel et messagerie) pour inciter les utilisateurs à programmer leurs recharges en dehors des heures de pointe, en s'appuyant sur les tendances identifiées, telles que les recharges post-travail. Cette approche permettrait de réduire les pointes de consommation et d'éviter des surcharges du réseau en incitant les ménages à décaler leurs recharges. Elle contribuerait à limiter les investissements coûteux dans les infrastructures de renforcement. Toutefois, son succès dépendrait de la mise en place d'incitatifs financiers clairs, tels que des tarifs différenciés justifiés, ainsi que de la capacité à communiquer efficacement les avantages pour l'usager. Le risque principal serait une faible adhésion si cette mesure était perçue comme une contrainte supplémentaire.
- Mettre en place des systèmes de collecte volontaire et confidentielle des périodes de vacances des ménages, afin d'ajuster la production ou la gestion de la demande pendant les creux saisonniers et de réduire les risques de surcharge du réseau. Une telle initiative améliorerait la précision des prévisions de consommation en tenant compte des absences prolongées, ce qui optimiserait la production et limiterait les pertes durant les périodes creuses. Elle soulèverait cependant des enjeux liés à la protection de la vie privée et nécessiterait des garanties solides en matière de confidentialité. Si la démarche est perçue comme transparente et bénéfique pour l'ensemble de la collectivité, elle pourrait renforcer la confiance entre les ménages et les fournisseurs d'énergie.
- Investir dans des infrastructures de recharge collective rapide accessibles dans les copropriétés, avec des points de contrôle sécurisés, pour mieux répartir la charge et répondre aux besoins croissants. Ces infrastructures seraient gérées par l'exploitant du réseau (Hydro-Québec) en collaboration avec le fournisseur d'énergie. Ce type d'investissement favoriserait une gestion centralisée de la recharge et réduirait la multiplication des bornes individuelles. Bien que coûteuses au départ, ces installations

généreraient des économies d'échelle à long terme et permettraient de démocratiser l'accès à la recharge, notamment pour les ménages n'ayant pas de garage privé. La réussite de cette mesure reposerait sur la définition d'un cadre réglementaire clair précisant les responsabilités entre les copropriétés, les exploitants et les fournisseurs d'énergie.

- Promouvoir l'utilisation de modèles prédictifs, tels que les LSTM, pour anticiper les pics de consommation, en intégrant des données historiques afin d'affiner les prévisions et de guider les décisions tarifaires dynamiques. L'intégration de modèles prédictifs améliorerait la capacité d'anticiper les variations de la demande et d'optimiser la planification énergétique. Cela permettrait de réduire les coûts liés à une mauvaise anticipation et de renforcer l'efficacité du réseau. Toutefois, la dépendance à des algorithmes d'intelligence artificielle imposerait un encadrement rigoureux afin d'éviter des inégalités entre les consommateurs. L'acceptabilité sociale serait conditionnée par la transparence des modèles utilisés et l'équité des tarifs qui en découleraient.
- Anticiper, à l'échelle des bâtiments, la nécessité d'augmenter leurs capacités grâce à la prévision statistique. En mettant en place une telle approche, il serait possible d'identifier plus tôt les besoins de renforcement des infrastructures locales, comme les transformateurs ou les câbles, ce qui réduirait les risques de panne et améliorerait la planification des investissements. Cela impliquerait un rôle accru pour les gestionnaires immobiliers, qui deviendraient des acteurs clés de la gestion énergétique. La mesure pourrait être consolidée par l'instauration d'audits énergétiques réguliers, garantissant la fiabilité et la continuité des prévisions.
- Explorer des solutions de stockage d'énergie domestiques, telles que des batteries (y compris celles de véhicules électriques) couplées à des panneaux solaires, pour stabiliser le réseau face aux variations saisonnières et soutenir une transition vers les énergies renouvelables. L'adoption de solutions de stockage renforcerait l'autonomie énergétique des ménages et contribuerait à stabiliser le réseau en période de forte demande. Bien que leur coût initial demeure élevé, ces dispositifs pourraient s'avérer rentables grâce à la revente d'énergie, notamment via des mécanismes de type

« vehicle-to-grid ». Néanmoins, cette option risque de creuser un écart entre les ménages selon leur capacité d'investissement. Elle supposerait également un cadre normatif strict garantissant la sécurité des installations et le recyclage des batteries en fin de vie.

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Almaghrebi, A., James, K., Al Juheshi, F., & Alahmad, M. (2024). Insights into Household Electric Vehicle Charging Behavior: Analysis and Predictive Modeling. *Energies*, *17*(4), 925. Repéré à https://doi.org/10.3390/en17040925

AVÉQ. (2025, 17 juin). *Statistiques SAAQ-AVÉQ sur l'électromobilité au Québec en date du 31 mars 2025 [Infographie]*. AVÉQ. Repéré à https://www.aveq.ca/actualiteacutes/statistiques-saaq-aveq-sur-lelectromobilite-au-quebec-en-date-du-31-mars-2025-infographie

Batista, G. E. A. P. A., & Monard, M. C. (2002). *A study of K-nearest neighbour as an imputation method*. Repéré à https://www.researchgate.net/publication/220981745_A_Study_of_K-Nearest_Neighbour_as_an_Imputation_Method

Bhayani, A. (2020). *Isolation Forest algorithm for anomaly detection*. *Medium*. Repéré à https://medium.com/@arpitbhayani/isolation-forest-algorithm-for-anomaly-detection-f88af2d5518d

Boudhaouia, A. (2022). Analyse, classification et prédiction de la consommation d'eau et d'électricité par des techniques de machine learning. Repéré à https://theses.hal.science/tel-03562074/

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. Repéré à https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Brownlee, J. (2021). *How to develop LSTM models for multi-step time series forecasting of household power consumption*. Machine Learning Mastery. Repéré à https://www.machinelearningmastery.com/how-to-develop-lstm-models-for-multi-step-time-series-forecasting-of-household-power-consumption

Chandola, V., Banerjee, A., & Kumar, V. (2009). Outlier Detection: A Survey. Repéré à https://www.researchgate.net/publication/242403027_Outlier_Detection_A_Survey

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8837–8846.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Repéré à https://arxiv.org/abs/1603.02754

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273–297.

Dominique. (2009). Classification de séries temporelles. Applications à la prévision et à la désaisonnalisation. Dominique Ladiray. Insee, Département Statistiques de Court Terme. Repéré à https://journees-methodologie-statistique.insee.net/wp-content/uploads/2009/S01_0_PRESENTATION_LADIRAY_JMS2009.PDF

Dou, Q., Chen, H., Yu, L., Qin, J., & Heng, P.-A. (2019). *Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss*. arXiv. Repéré à https://arxiv.org/abs/1811.06042

Bujokas, E. (2020). *Energy consumption time series forecasting with python and LSTM deep learning model*. Medium. Repéré à https://medium.com/data-science/energy-consumption-time-series-forecasting-with-python-and-lstm-deep-learning-model-7952e2f9a796

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Guo, S., Liu, Y., & Su, Y. (2021). Comparison of Classification-based Methods for Network Traffic Anomaly Detection. *IMCEC 2021—IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference*, 360–364. Repéré à https://doi.org/10.1109/IMCEC51613.2021.9482274

Hydro-Québec. (2025A). Calculer la consommation en mode veille (charges fantômes). Repéré à https://www.hydroquebec.com/residentiel/espace-clients/consommation/outils/calculette-charges-fantomes.html

Hydro-Québec. (2025B). Consommation selon les caractéristiques de l'habitation. Repéré à https://www.hydroquebec.com/residentiel/espace-clients/consommation/outils/utilisation-electricite.html

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Repéré à https://otexts.com/fpp2/seasonal-arima.html?utm

Ibrahim, K. (2021). Forecasting the Future Power Consumption of Germany using LSTM(RNN) and DNN. Repéré à https://towardsdatascience.com/forecasting-the-future-power-consumption-of-germany-using-lstm-rnn-and-dnn-d8e05e7fdc0a

Jeremy R. (2021). Isolation Forest : Comment détecter les anomalies dans une dataset ? Repéré à https://datascientest.com/isolation-forest

Makers, R. T. I. (2022). *L'IoT et les énergies renouvelables*. RT One Blog. Repéré à https://blog.rtone.fr/iot-et-énergies-renouvelables

MathWorks. (2025). Analyse des séries temporelles : étapes, types et exemples. Repéré à https://fr.mathworks.com/discovery/time-series-analysis.html

Mohammed, A. S., Asteris, P. G., Koopialipoor, M., Alexakis, D. E., Lemonis, M. E., & Armaghani, D. J. (2021). Stacking Ensemble Tree Models to Predict Energy Performance in Residential Buildings. *Sustainability, 13*, 8298. Repéré à https://doi.org/10.3390/su13158298

Patrick, H. (2022). *La détection d'anomalies en Machine Learning non supervisé*. CTIF Metal Blog. Repéré à https://metalblog.ctif.com/2022/10/03/la-detection-danomalies-en-machine-learning-non-supervise/

Rajasekaran, R. G., Manikandaraj, S., & Kamaleshwar, R. (2017). Implementation of Machine Learning Algorithm for predicting user behavior and smart energy management. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 24–30. https://doi.org/10.1109/ICDMAI.2017.8073480

RNCAN. (2014). *Consommation en mode veille - même éteints, vos appareils veillent encore*. Ressources naturelles Canada. Repéré à https://publications.gc.ca/site/fra/375195/publication.html

RNCAN. (2025). Faites fuir l'alimentation fantôme hors de votre maison. Repéré à https://publications.gc.ca/site/fra/375195/publication.html

Investissement Québec. (2023). *Plus de 7 millions de dollars dans RVE pour démocratiser l'accès à la recharge électrique en Amérique du Nord* [Communiqué]. https://www.investquebec.com/quebec/fr/salle-de-presse/communiques/Plus-de-7-

Millions-de-dollars-dans-RVE-pour-democratiser-l-acces-a-la-recharge-electrique-en-Amerique-du-Nord.html

Sagoolmuang, A., & Sinapiromsaran, K. (2017, June 23). Median-difference window subseries score for contextual anomaly on time series. *2017 8th International Conference on Information and Communication Technology for Embedded Systems, IC-ICTES 2017—Proceedings*. https://doi.org/10.1109/ICTEmSys.2017.7958772

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water, 11*(5), 910. Repéré à https://doi.org/10.3390/w11050910

Varsamopoulos, S. (2019). Neural Network based Decoders for the Surface Code. [Dissertation, TU Delft]. Repéré à https://doi.org/10.4233/uuid:dc73e1ff-0496-459a-986f-de37f7f250c9

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research, 11*(Dec), 3371–3408.

Wang, Y., Chen, Q., Kang, C., & Xia, Q. (2016). Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications. *IEEE Transactions on Smart Grid, 7*, 2437–2447. https://doi.org/10.1109/TSG.2016.2548565

Yu, X., Peng, Y., Li, F., Wang, S., Shen, X., Mai, H., & Xie, Y. (2020). Two-level data compression using machine learning in time series database. *Proceedings—International Conference on Data Engineering, 2020-April* 1333–1344. https://doi.org/10.1109/ICDE48307.2020.00119

Zhang, A. (2019). *A gentle introduction to XGBoost for applied machine learning*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/