Modélisation et exploration de données liées à la consommation

Anissa Ticherahine

Université de Haute Alsace

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur
(MSII, ED 269)
Institut de Recherche en Informatique, Mathématiques, Automatique et Signal
(IRIMAS, EA 7499)

Modélisation et exploration de données liées à la consommation

Thèse

préparée par

Anissa TICHERAHINE

présentée pour obtenir le grade de
Docteur de l'Université de Haute Alsace
Discipline : Mathématiques appliquées et applications mathématiques

soutenue publiquement le 30/11/2021 devant le jury composé de :

Pr. Nikolaos Limnios, Université de Technologie de Compiègne, Président
Pr. Hocine Fellag, Université Mouloud Mammeri de Tizi-Ouzou, Rapporteur
Pr. Stéphane Galland, Université de Technologie de Belfort-Montbéliard, Rapporteur
Pr. Alain Dieterlen, Université de Haute Alsace, Examinateur
Pr. Abdenacer Makhlouf, Université de Haute Alsace, Directeur de thèse
Pr. Patrice Wira, Université de Haute Alsace, Directeur de thèse

« The music is not in the notes, but in the silence between. »

— Wolfgang Amadeus Mozart —

(1756-1791)

Remerciements

Ce travail de thèse a été effectué au laboratoire IRIMAS de l'Université de Haute-Alsace, financé par le gouvernement algérien grâce à une Bourse d'Excellence.
Je remercie chaleureusement tous ceux qui ont contribué à la réussite de ce travail.
Tous d'abord, j'aimerais exprimer ma sincère gratitude et ma profonde reconnaissance à mon directeur de thèse, Mr Abdenacer Makhlouf et à mon co-directeur de thèse, Mr Patrice Wira pour m'avoir encadré. Merci de votre orientation, de votre savoir, de vos conseils, de votre disponibilité ainsi que votre encouragement durant ces années de recherche.
Je remercie également les membres du jury : Mr Hocine Fellag, Professeur à l'Université de Tizi-Ouzou ; Mr Nikolaos Limnios, Professeur à l'Université de Technologie de Compiègne ; Mr Stéphane Galland, Professeur à l'Université de Technologie de Belfort-Montbéliard et Mr Alain Dieterlen ; Professeur à l'Université de Haute Alsace, pour leurs intérêt et pour avoir accepté de rapporter sur mon travail, sans oublier les membres du comité de suivi de thèse Mr Camille Laurent-Gengoux, Professeur à l'Université de Lorraine et Mr Jonathan Weber, Maitre de Conférences à l'Université de Haute Alsace.
J'adresse également mes remerciements au Ministère de l'Enseignement Supérieur et de la Recherche Scientifique d'Algérie qui a financé ma thèse. Un grand remerciement au personnel du Consulat Général d'Algérie à Strasbourg pour leur gentillesse, leur accueil chaleureux et leurs bons services.
J'en profite également pour remercier ma chère Soumia, avec qui j'ai eu l'occasion de travailler. Merci pour son aide et sa sympathie et les bons moments que nous avons passés ensemble.
Merci à tous les doctorants avec qui j'ai partagé le bureau du Département de Mathématiques et l'équipe IMTI du département ASI de IRIMAS situé à l'IUT de Mulhouse en particulier à Aida qui a patiemment répondu à toutes mes questions au début de ma thèse et avec qui j'ai eu l'opportunité de collaborer sur des travaux de recherche.
Enfin, j'adresse mes remerciements les plus chaleureux et toute mon affection à mon père et ma mère décédée, à qui je dédie le fruit de cette réussite, et j'espère qu'ils seront fiers de ce que j'ai accompli, à mon chère frère Ismail et à mes sœurs pour m'avoir soutenue tout au long de mes études. Je tiens tout particulièrement à remercier mon mari, Ali, qui m'a encouragé et m'a motivé à préparer cette thèse. Merci à toute ma famille, mes amies et à tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce travail.

Abstract

The thesis focuses on the theoretical and applied aspects of the evaluation, modeling and exploration of consumption data. To establish and verify the models and methods, representative data on water and electricity consumption have been collected automatically using smart meters installed in several buildings. These data represent consumption very precisely and constitute a large database.

The first objective of the thesis is purely mathematical and it consists in developing statistical methods and tools, in particular on chronological series. The analysis of time series is dealt primarily by the identification method, in particular by the Box-Jenkins method. It is a question of developing these aspects within the framework of algebraic statistics. After a theoretical study, the work consisted in considering real data of water consumption and electricity consumption. A comparative study of methods was carried out based on the large amounts of data collected. We have set up hybrid models combining deterministic and stochastic time series models, exponential smoothing models and neural network models with the parameters already defined to create a better prediction model of the consumption processes and to obtain a smaller error in the forecast consumption. These models have been tested and validated on the basis of real data. In practice, they make it possible to predict consumption. The second objective is to detect anomalies and we have found a new WLICTD (Water Leakage Indicator based on the Consumption Temporary Density) approach for the detection of water leaks. The third objective is to solve the problem of data storage ; we have evaluated data reduction approaches such as the self-organizing map.

The proposed contributions have been validated on real consumption data. Hybrid models have shown better performance for improving forecasting. As well as the WLICTD approach makes it possible to use raw data and detect water leaks in just a few hours.

Key words : Smart meter ; load curve ; time series ; the Box-Jenkins method ; hybrid model ; deterministic and stochastic models ; exponential smoothing model ; neural network models ; WLICTD ; self-organizing map.

Résumé

La thèse porte sur les aspects théoriques et appliqués de l'évaluation, modélisation et exploration des données de consommation. Pour établir et vérifier les modèles et les méthodes, des données représentatives des consommations d'eau et d'électricité ont été collectées automatiquement à l'aide de compteurs intelligents installés dans plusieurs bâtiments. Ces données représentent de manière très précise les consommations et constituent une grande base de données.
Le premier objectif de la thèse est purement mathématique et il consiste à développer des méthodes et des outils statistiques, en particulier les séries chronologiques. L'analyse des séries chronologiques est traitée essentiellement par la méthode d'identification, en particulier par la méthode de Box-Jenkins. Il s'agit de développer ces aspects dans le cadre de statistiques algébriques. Après une étude théorique, le travail a consisté à considérer des données réelles de consommation d'eau et de consommation d'électricité. Une étude comparative de méthodes a été effectuée en s'appuyant sur de grandes quantités de données collectées. Nous avons mis en place des modèles hybrides combinant des modèles déterministes et stochastiques des séries temporelles, des modèles de lissage exponentiel et les modèles de réseaux de neurones avec les paramètres déjà définis pour créer un meilleur modèle de prédiction des processus de consommation et pour obtenir une plus petite erreur dans la consommation prévue. Ces modèles ont été testés et validés sur la base de données réelles. Ils permettent en pratique de faire la prédiction de la consommation. Le deuxième objectif est de détecter les anomalies et nous avons trouvé une nouvelle approche WLICTD (Water Leakage Indicator Based on the Consumption Temporary Density) pour la détection des fuites d'eau. Le troisième objectif consiste à résoudre le problème de stockage des données ; nous avons évalué des approches de réduction de données comme la carte auto-organisatrice.
Les contributions proposées ont été validées sur des données réelles de la consommation. Les modèles hybrides ont montré de meilleure performance pour l'amélioration de la prévision. Ainsi que l'approche WLICTD a permis d'utiliser les données brutes et détecter les fuites d'eau en quelques heures seulement.

Mots clefs : Compteur intelligent ; courbe de charge ; série chronologique ; la méthode de Box-Jenkins ; modèle hybride ; modèle déterministe et stochastique ; modèle de lissage exponentiel ; modèle de réseaux de neurones ; WLICTD ; carte auto-organisatrice.

Table des matières

Liste des acronymes

ACF	Autocorrelation function Fonction d'autocorrélation
ACFP	Partial autocorrelation function Fonction d'autocorrélation partielle
AIC	Akaike Information Criterion Critère d'information Aikaike
ANN	Artificial Neural Networks Réseaux de neurones artificiels
AR	Autoregressive model Modèle autorégressif
ARIMA	Auto Regressive Integrated Moving Average
ARMA	Autoregressive Moving Average Modèle mixte
BIC	Bayesian Information Criterion critère d'information Bayésien
CdC	Courbe de charge
EM	Expectation-Maximization Espérance-Maximisation
ES	Exponential Smoothing Modèle de lissage exponentiel
ETS	Error, Trend, Season
IoT	Internet of things Internet des objets
GMM	Gaussian Mixture Model Modèle de mélange gaussien
KDE	Kernel Density Estimation L'estimation par noyau
LSTM	Long Short-Term Memory Réseau récurrent à mémoire court et long terme
MA	Moving-Average model Modèle à moyenne mobile

MLP	Multilayer Perceptron
	Perceptron multicouche
MNF	Minimum Night Flow
	Débit minimum nocturne
RMSE	Root Mean Square Error
	Erreur quadratique moyenne
SARIMA	Modèle "Seasonal Autoregressive Integrated Moving Average"
SES	Simple Exponential Smoothing
	Lissage exponentiel unique
SOM	Self-organizing card
	Carte auto-organisatrice
WDN	Water Distribution Networks
	Les réseaux de distribution d'eau
WLICTD	Water Leakage Indicator based on the Consumption Temporary Density
	Indicateur de fuite d'eau basé sur la densité temporaire de consommation

Table des figures

Liste des tableaux

1 Introduction générale

La demande en eau et en électricité est essentielle et elle ne cesse d'augmenter de manière significative. En raison des récents progrès technologiques et de l'émergence du paradigme de l'Internet des objets, un nouveau type des compteurs intelligents d'eau et d'électricité est apparu. Ces compteurs permettent de surveiller la consommation d'eau et d'électricité. Notre objectif est d'analyser et d'exploiter ces données de consommation puis trouver des modèles et des outils statistiques qui permettent de prédire la consommation et de détecter les anomalies. C'est dans ce cadre général que se situent les travaux de recherche présentés dans cette thèse intitulée " **Modélisation et exploration de données liées à la consommation** ".
La collecte des données de la consommation d'eau et d'électricité représente l'enregistrement des écarts de temps que suit à une intégration d'un litre d'eau ou un watt-heure de la puissance électrique. La stratégie de collecte de données à l'aide de compteurs intelligents transmet les données en temps réel. Pour comprendre ces données, des études sur l'aspect statistique sont nécessaires. Cela permet d'estimer les quantités de consommation moyenne, minimum et maximum pour les deux ressources (eau et électricité). Il permet également de distinguer les consommations anormales des normales grâce à des représentations graphiques. Afin d'effectuer une analyse approfondie des données, nous utilisons des courbes de charge (CdC). Une courbe de charge définit la grandeur qui décrit l'évolution de la consommation dans le temps. Elle permet de voir le comportement de l'utilisateur et d'extraire un ensemble de descripteurs d'un client donné. De plus, la CdC représente l'évolution de consommation sur une période donnée. Ainsi, le profil de charge est une dérivée de la CdC sur la même période. Comme la consommation d'eau et d'électricité ne reste jamais constante, elle varie de temps en temps et ces changements de charge peuvent être tracés sur une base d'une demi-heure, d'une heure ou même chaque minute tout au long de la journée. La courbe ainsi obtenue est connue sous le nom de CdC journalière, mais elle peut également être prolongée pour n'importe quelle période de temps, c'est-à-dire qu'elle peut être tracée sur un mois ou sur une année. Le profil de charge ou la CdC sont des outils simples et efficaces pour évaluer l'utilisation et la demande d'eau et d'électricité, mais aussi l'efficacité et la fiabilité du transport d'eau et d'électricité.
Une série temporelle est une suite finie de valeurs numériques représentant l'évolution de la quantité de consommation au fil du temps. La prédiction de séries temporelles est un sujet

majeur en aide à la décision, car elle impacte tous les sujets en fonction du temps. Récemment, les réseaux de neurones ont été beaucoup utilisés pour prédire des séries chronologiques. Dans [1], trois types de réseaux de neurones artificiels sont utilisés : les réseaux de neurones à corrélation en cascade (CCNN), les réseaux de neurones à régression généralisée (GRNN) et les réseaux de neurones à anticipation (FFNN) pour la prévision mensuelle de la consommation d'eau. De plus [2], a utilisé un filtre de Kalman pour améliorer l'ANN pour prédire le niveau d'eau quotidien, tandis que l'article [3] a présenté deux modèles de prédiction quotidienne étendus, des modèles autorégressifs linéaires et non linéaires pour prédire la consommation d'eau urbaine. Dans [4], la méthode de régression multiple corrigée par la formule de régression linéaire est proposée pour la prévision des données sur l'eau. Plus recemment, les modèles hybrides qui sont des combinaisons de modèles individuels avec des paramètres spécifiques deviennent de plus en plus populaires et très utilisables. Un modèle hybride de l'ANN et du modèle de moyenne mobile intégrée autorégressive (ARIMA) est appliqué à trois bases de données annuelles : les données sur les taches solaires, les données sur le lynx canadien et le taux de change de la livre sterling pour la prévision des séries chronologiques [5]. Les auteurs de l'article [6] ont proposé une combinaison de l'approche Box-Jenkins et ANN. Dans le même contexte [7] a utilisé une approche hybride qui combine ARIMA et ANN pour la prévision mensuelle de l'eau urbaine et le modèle additif Hault-Winters pour la prévision trimestrielle. De plus [8] a proposé un nouveau modèle hybride d'ARIMA et de réseau neuronal à fonction radiale (RBF-NN) sur des données hebdomadaires. Un ANN combiné à un algorithme génétique (AG) est présenté dans l'article [9] pour la prédiction hydrologique de séries chronologiques de données mensuelles. De plus, une combinaison de SARIMA et de modèles de régression linéaire multiple pour prédire la demande de chaleur urbaine est détaillée dans [10]. L'article [11] présente une comparaison entre le modèle ARIMA et ANN appliqué sur les données de précipitations et de ruissèlement mensuels du bassin du lac Urmia. Des concours ont été mis en place pour la prédiction de séries temporelles avec différentes bases de données comme le concours de prévision M4 [12]. Une méthode hybride qui mélange le modèle de lissage exponentiel (ES) avec des réseaux de neurones avancés à mémoire à long terme (LSTM) est proposée dans [13]. Toutes les approches mentionnées fonctionnent pour une base de données spécifiques. Après avoir appliqué de nombreux modèles de séries chronologiques aux données de consommation d'eau et d'électricité, nous avons constaté que les modèles hybrides donnent les meilleurs résultats.
Dans le cadre de la mesure de l'eau, de nombreux principes et algorithmes différents ont été proposés dans divers travaux de recherche à des fins telles que l'estimation de la quantité d'eau retenue, la détection et le contrôle des fuites, le suivi de la consommation des utilisateurs, la réduction du temps nécessaire à la réparation de la fuite, etc. Les auteurs de [14] ont défini un débit minimum nocturne (MNF) qui consiste en un seuil défini pour une zone isolée où la demande en eau est généralement faible. Dans [15], une méthode basée sur un décisionnaire à logique floue (FLDM) a été proposée pour détecter les fuites. Elle est basée sur une technique qui s'appuie sur des ensembles flous pour fournir une sorte de modèle grossier du WDN qui prend en compte sa typologie (matériau, longueur, diamètre et âge des canalisations), son environnement (demande, topographie et pressions de fonctionnement) et même ses conditions d'exploitation (taille de la population, logement et caractéristiques socio-économiques, niveau

de vie). Des techniques basées sur la fusion de capteurs de données ont été utilisées dans [16] et [17]. Cependant, ils ne sont pas efficaces pour détecter toutes les fuites. L'approche proposée dans [18] détecte les fuites en se basant sur l'apprentissage de réseaux de neurones artificiels, mais cette approche nécessite un grand ensemble de données. Le travail de [19] associe un seuil MNF à un autre seuil qui est une période sans consommation nulle (PWNC). Il est capable de détecter la plupart des petites fuites d'eau au cours de la journée en utilisant le débit d'eau circulant dans un point de mesure. De plus, la détection des fuites importantes repose sur la courbe de charge maximale. Ces approches sont basées sur des données échantillonnées toutes les minutes. Nous proposons un algorithme de détection de fuite basé sur la densité temporaire de consommation d'eau. Cette nouvelle approche prend en compte la consommation d'eau de manière temporelle en utilisant des données de consommation d'eau en temps réel. Les données sont collectées et analysées de manière itérative à tous les instants, cela permet de détecter les fuites en quelques heures seulement. L'approche a été comparée à d'autres méthodes qui ont été appliquées à nos données telles que la MNF, la courbe de charge maximale et la détection à l'aide de débit.
Ayant recueilli beaucoup d'observations, il s'agit souvent d'un gros corpus de données qui peut poser des problèmes de stockage. Nous visons donc à les réduire à un petit nombre de paramètres qui définissent aussi clairement que possible la position et la dispersion des observations et qui ainsi facilitent les interprétations. Pour résoudre le problème de stockage et augmenter l'exploration de données, nous avons utilisé des modèles qui estiment les courbes de charge en remplaçant les données par des paramètres de modèle, des données d'échantillonnage, et les classes de regroupement selon le type de consommation. Ensuite, nous avons suggéré l'apprentissage de la carte auto-organisatrice pour réduire les données de la courbe de charge d'eau et d'électricité.
L'objectif de notre travail est d'analyser les données de consommation pour comprendre le comportement des utilisateurs. Parmi les raisons de développement des modèles est la nécessité de disposer d'une méthode permettant de prévoir et de contrôler la consommation. Le but de la détection des fuites d'eau est de conserver des ressources limitées, et minimiser les dommages.

Contributions

Nous avons obtenu dans cette thèse les résultats suivants.

- Nous avons suggéré d'utiliser la carte d'auto-organisation pour réduire les données afin de les rendre plus facilement exploitable.
- Nous avons proposé de nouvelles combinaisons des modèles de séries temporelles pour prédire la consommation. Le premier modèle hybride combine le modèle déterministe, le modèle de Holt Winter et la combinison de deux modèles de réseaux de neurones (LSTM et MLP) pour prédire la consommation d'eau. La deuxième combinaison des modèles SARIMA et LSTM est utilisée pour prédire la consommation de la puissance électrique.
- Nous avons introduit une nouvelle approche nommée WLICTD qui détecte les fuites d'eau dans un délai raisonnable. Elle est basée sur la densité temporelle et utilise des

données brutes inchangées.

Organisation du manuscrit

Ce rapport de thèse se décompose en quatre chapitres, en plus de l'introduction et de la conclusion :

1. Analyse des données
2. Réduction des données
3. Modélisation des données
4. Détection des fuites d'eau.

— Le **chapitre 2** décrit les compteurs intelligents qui fournissent des données de consommation d'eau et d'électricité en temps réel. La collecte des données est présentée dans la deuxième section. Ensuite, une analyse statistique des données est effectuée à l'aide de statistiques descriptives et de représentations graphiques. Pour bien comprendre le comportement des utilisateurs, une analyse des CdC est effectuée. Cette analyse nous a permis d'avoir une idée générale sur les données et de comprendre les modes de vie des utilisateurs. Nous avons également pu identifier des anomalies.

— Le **chapitre 3** présente des méthodes de réduction des données telle que la réduction de la numérosité par des méthodes paramétriques comme la régression linéaire et des méthodes non paramétriques comme l'échantillonnages et le regroupement. Les points de données sont remplacés par deux paramètres de la droite de régression dans la réduction paramétrique, et par des points d'échantillonnages ou des points de regroupement dans la réduction non paramétrique. Nous avons utilisé une carte auto-organisée pour réduire les données en remplaçant les points de données par des neurones de la meilleure carte et cela nous a donné de bons résultats.

— Le **chapitre 4** développe le contexte de modélisation et de prévision à l'aide de modèles hybrides. Nous avons fourni des méthodologies pour la modélisation des courbes de charge eau-électricité. Nous avons commencé par des approches numériques avec des méthodes d'interpolation et d'approximation (interpolation de Lagrange, interpolation au sens de Tchebychev, interpolation de spline cubique, approximation par les moindres carrés, les courbes de Bézier) et des méthodes probabilistes par l'estimation de densité par noyau et le modèle de mélange gaussien. Ensuite, nous avons appliqué des modèles de séries temporelles (modèles déterministes paramétriques et non paramétriques, modèles stochastiques, modèles de lissage exponentiel) et le modèle hybride qui combine un modèle déterministe paramétrique et un modèle stochastique. Pour prédire la consommation d'eau et d'électricité, nous avons utilisé des modèles de séries temporelles, des modèles de réseaux de neurones (LSTM, MLP) et des modèles hybrides qui combinent ces modèles.

— Le **chapitre 5** propose une nouvelle approche pour la détection des fuites d'eau nommée Water Leakage Indicator based on the Consumption Temporary Density (WLICTD). L'ap-

proche proposée permet de détecter toutes les fuites d'eau dans une période raisonnable (environ de 3 heures) en utilisant les données brutes. D'autres méthodes de détection de fuites d'eau telles que le Minimum Night Flow (MNF), la détection en utilisant la courbe de charge maximale et la détection par la fonction de débit ont été utilisées.

2 Analyse des données

Dans cette thèse nous allons nous intéresser à la consommation d'eau et d'électricité dans plusieurs bâtiments. Ce chapitre fournit d'abord une description détaillée de la plateforme utilisée pour générer les données. Ensuite, les données acquises sont expliquées et analysées. L'analyse des données est faite par les courbes de charge et des études statistiques.

2.1 La plateforme IoT

L'internet des objets (IoT) [20] est un développement de l'architecture internet en tant que réseau d'appareils connectés. Ces dispositifs peuvent être des objets qui communiquent entre eux et avec leurs utilisateurs [21]. Le concept de l'internet des objets vise à rendre l'architecture du système d'information importante pour la mise en œuvre de bâtiments intelligents plus immersive [22]. Les compteurs IoT peuvent enregistrer et transmettre aussitôt des impulsions qui correspondent à la petite quantité d'eau ou d'électricité consommée en quelque minutes ou secondes. Grâce à cette technologies nous pouvons obtenir rapidement et en permanence des informations sur la consommation.
La figure 2.1 représente un modèle de bâtiment intelligent où la consommation d'eau et d'électricité est surveillée par des compteurs intelligents [19], [23]. Un compteur intelligent est un objet connecté avec des capacités de communication qui permettent de collecter et de transmettre des informations sur la consommation. Ce compteur permet de détecter une impulsion. Cette impulsion représente une rotation de la petite turbine à l'intérieur du compteur d'eau. Chaque fois qu'un litre d'eau est consommé, la turbine réalise une rotation complète et génère ainsi un événement. L'impulsion dans les compteurs électriques est une intégration d'un Watt-heure de la puissance et envoie une impulsion au raspberry. Chaque impulsion correspond à un événement. L'événement produit consiste à horodater les données et à les envoyer à un serveur de stockage via internet. Cela permet de surveiller régulièrement les ressources en eau et en électricité et de suivre la consommation quotidienne.

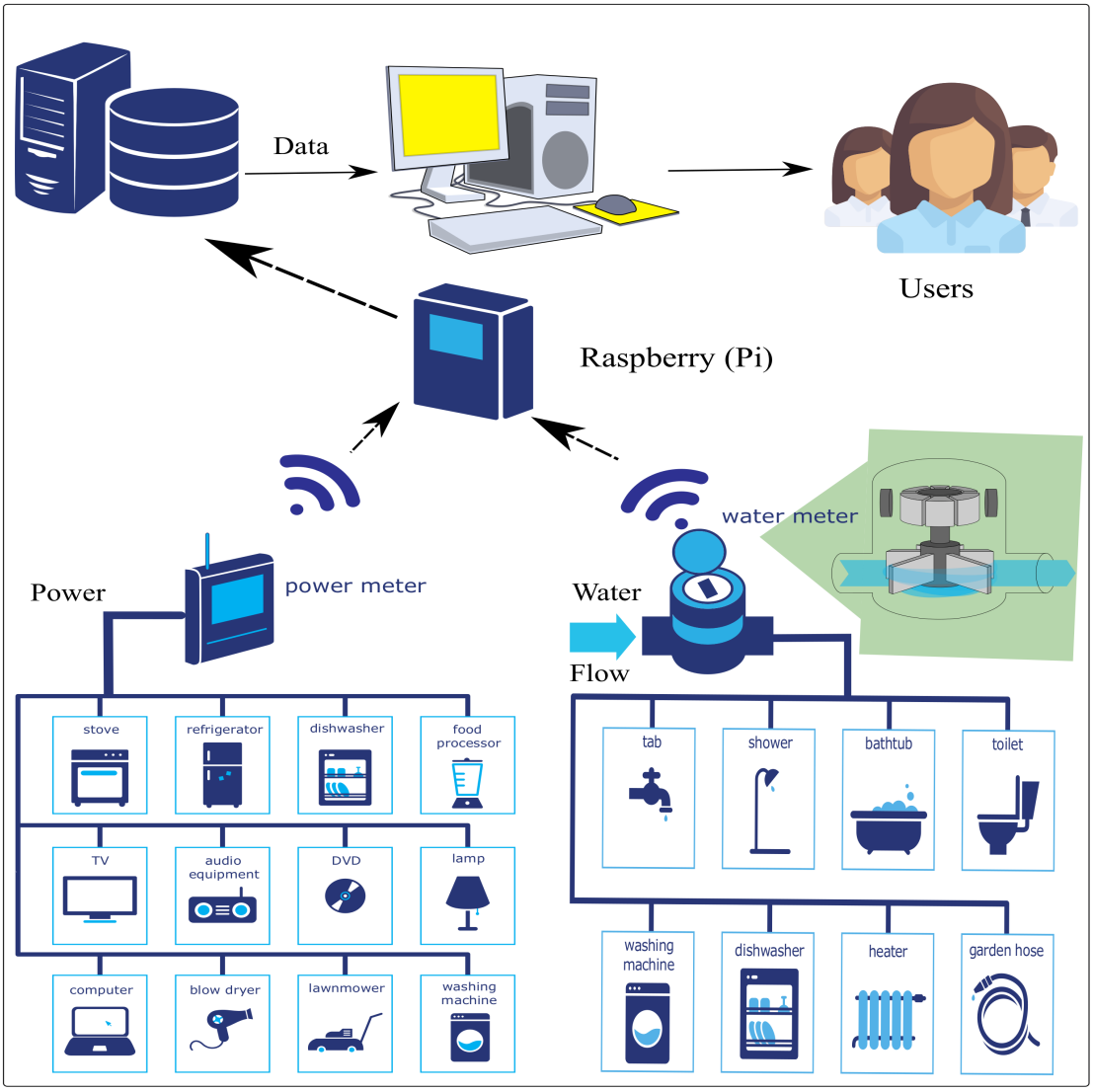

FIGURE 2.1 – La plateforme IoT intégrée dans un bâtiment

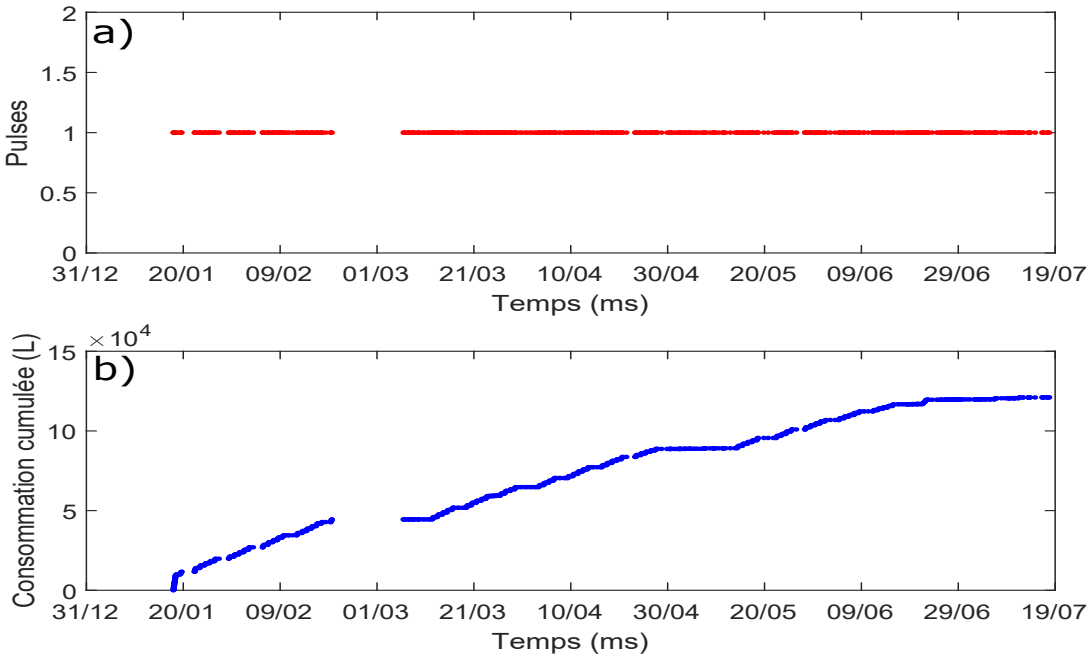

FIGURE 2.2 – Les données brutes de la consommation d'eau au restaurant universitaire de l'IUT de Mulhouse du 17/01/2018 au 17/07/2018.

2.2 Description des données

Les données brutes sont construites à partir d'un compteur intelligent et ces données représentent les écarts de temps en millisecondes $\triangle t_i = t_{i+1} - t_i$ tel que pour chaque $\triangle t_i, \forall i \in I$ (I le nombre des données) nous consommons un litre d'eau si les compteurs sont petits (un Wh pour l'électricité) sinon nous consommons 10 litres d'eau et dans tous les cas chaque impulsion est un tour de la petite turbine des compteurs d'eau ou une intégration d'un Wh de la puissance électrique.
L'évolution de la consommation d'eau au restaurant universitaire de l'IUT de Mulhouse est représentée par la figure 2.2 pour la période du 17 janvier 2018 au 17 juillet 2018. La figure 2.3 représente les données de la consommation d'électricité à l'IUT de Mulhouse du 22 octobre 2019 au 16 septembre 2020.
Les points rouges sur les figures .a) correspondent aux impulsions qui indiquent la consommation de 1 L d'eau ou 0.1 Wh d'électricité dans chaque intervalle de temps. L'espace entre les points rouges correspond aux périodes de week-end et des vacances. La courbe colorée en bleu dans les figures .b) représente la consommation cumulée au fil du temps.

2.3 Analyse statistique

L'analyse statistique implique la collecte et l'interprétation de données, nous devons d'abord savoir comprendre, afficher et résumer de grandes quantités d'informations quantitatives, avant

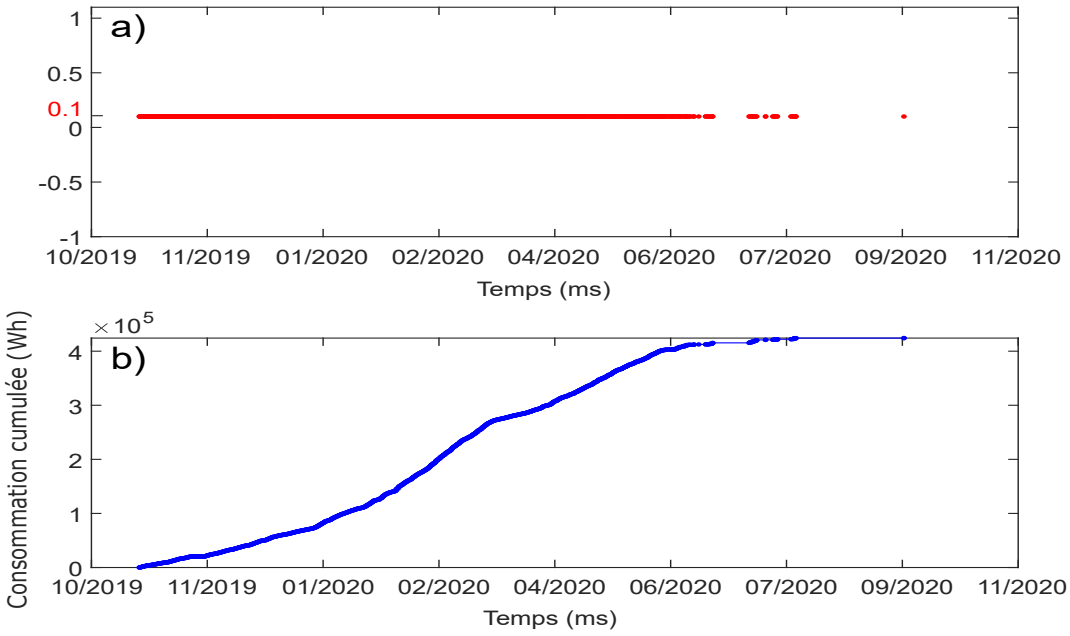

FIGURE 2.3 – Les données brutes de la puissance électrique à l'IUT de Mulhouse du 22/10/2019 au 16/09/2020

d'entreprendre une analyse plus sophistiquée.
L'analyse statistique des données quantitatives est importante dans toutes les sciences exactes et sociales. Il est très difficile d'interpréter un seul ensemble de données. Il y a beaucoup d'informations contenues dans les données, mais elles sont difficiles à cerner. Nous avons besoin de moyens de comprendre les caractéristiques importantes des données et de les résumer de manière significative.
L'utilisation de graphiques et de statistiques récapitulatives pour comprendre les données est une première étape importante dans la réalisation de toute l'analyse statistique.

2.3.1 Représentations graphiques

Il est important d'étudier la forme de la distribution d'une variable aléatoire. Ceci est plus facilement examiné à l'aide de diagrammes. Un graphique à barres est presque toujours préférable. Les principaux avantages de l'utilisation d'un diagramme à barres sont qu'il montre la forme générale des données et de acquérir une compréhension empirique des caractéristiques importantes de la distribution des données. En mettant les graphiques côte à côte avec la même échelle, nous pouvons comparer les distributions de différents groupes. Le diagramme permet de voir les anomalies dans la consommation, les jours ouvrable et les autres jours de week-end même les jours de vacances.
Comme les données sont discrètes, nous les affichons graphiquement à l'aide d'un diagramme

avec des hauteurs de barres représentant des consommations journalières. La figure 2.4 montre la variation de la consommation d'eau chaque jour dans un diagramme. A travers les barres, nous remarquons que la plus grande consommation d'eau représentée par la deuxième barre orange est due à la présence de fuite d'eau, alors que les jours de fuites représentés dans les autres barres orange n'ont pas été remarqués car ils ne sont pas différents des autres barres de jours représentées en bleues. Il est également possible de distinguer les deux jours du week-end, représentés en rouge, du reste des jours ouvrables, avec peu ou pas de consommation, et ce sont les deux barres qui suivent les cinq barres bleues pour les jours ouvrables. Cependant que parfois, nous ne trouvons que quatre barres bleues, suivies ou précédées d'une barre noire avec une très petite valeur de consommation (indiquée en zoom et nous l'avons représentée avec des étoiles noires) et cela revient à s'adapter aux jours fériées. Alors qu'il a été observé qu'il y avait une consommation inhabituellement élevée un jour de week-end, représentée par la barre rose, car c'est un jour de portes ouvertes. De plus, les barres brunes représentent également les jours de week-end, avec une augmentation de la consommation due à une fuite. Les jours de vacances sont définis par des barres vertes, dans lesquelles une légère consommation d'eau est observée, et en revanche, dans d'autres barres des vacances, une augmentation significative de la consommation par rapport aux jours qui leur correspondent, et ceci est dû à le fait que le groupe des doctorants et les enseignants ne sont pas concernés par des mêmes période de vacances pour les étudiants d'autres niveaux d'enseignement.
La figure 2.5 représente un diagramme de la consommation électrique par jours à l'IUT de Mulhouse. La consommation d'électricité est observée de manière large, car les jours ouvrables normaux sont représentés par des barres bleues, tandis que la consommation diminue après chaque cinq barres bleus qui sont représentées par des barres rouges. Cela est dû au fait qu'ils coïncident avec les jours de week-end. De plus, nous pouvons voir que la consommation d'électricité varie en quelque sorte certains autres jours, représentée par des barres verts, qui représentent les vacances. Aussi, la consommation d'électricité diminue très significativement le jour représenté par la barre noire et c'est parce que c'est un jour férié. Les barres blanches représentent les jours au cours desquels l'information a été perdue et la consommation réelle d'électricité n'a pas été précisée.

2.3.2 Statistiques descriptives

En plus des techniques graphiques utilisées pour analyser les données, il est souvent utile d'avoir des résumés quantitatifs d'aspects spécifiques des données. Les mesures brèves les plus simples peuvent être divisées en deux types. Les premiers sont connus comme des mesures de localisation qui sont typiques des données et les seconds comme des mesures de propagation qui résument la variabilité des données.
Mesures de localisation :
La moyenne de l'échantillon est la mesure de localisation la plus importante, principalement en raison de ses propriétés mathématiques. En particulier, la médiane de l'échantillon est un estimateur de localisation beaucoup plus robuste, et beaucoup moins sensible que la moyenne de l'échantillon aux asymétries et aux valeurs inhabituelles des données. Le mode est la valeur

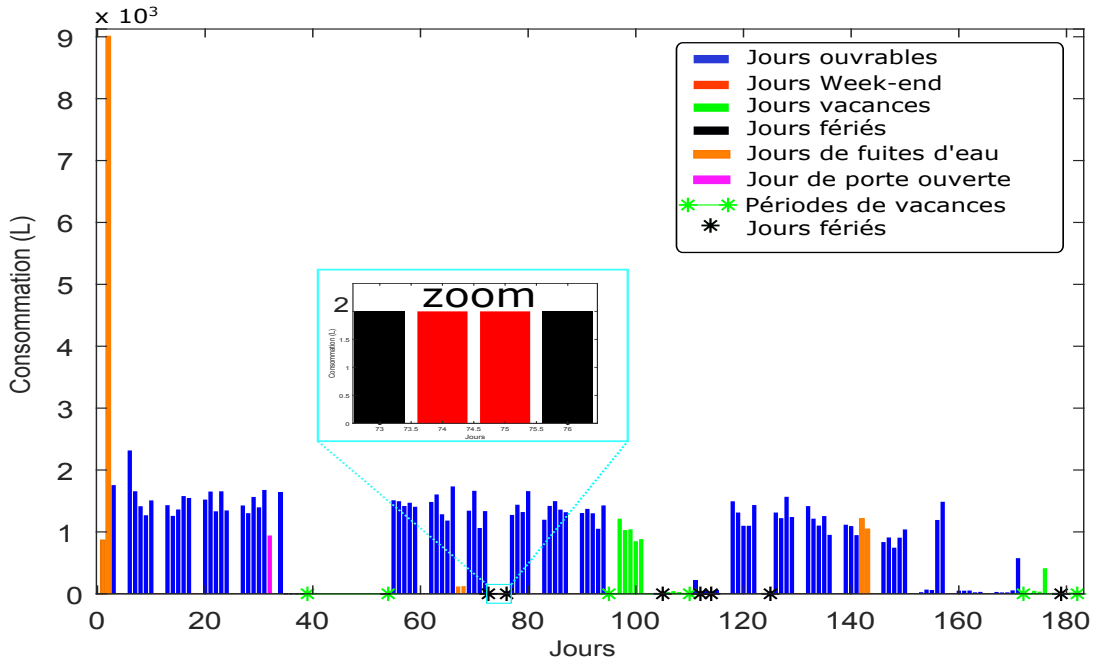

FIGURE 2.4 – Diagramme de la consommation d'eau au restaurant universitaire

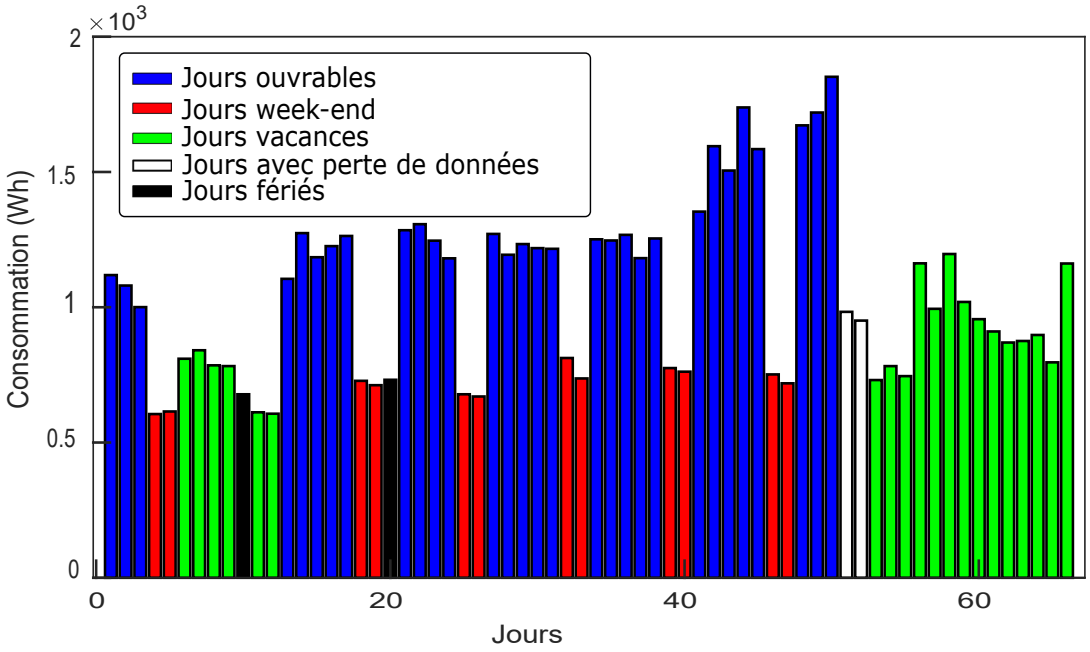

FIGURE 2.5 – Diagramme de la puissance électrique à l'IUT de Mulhouse

qui apparaît le plus fréquemment dans les observations.
Lorsque la distribution des données est à peu près symétrique, les trois mesures seront de toute façon très proches les unes des autres. Cependant, si la distribution est très asymétrique, il peut y avoir une différence considérable, et les trois mesures pourraient être utiles pour comprendre les données.
Supposons que nous ayons un échantillon de taille n de données quantitatives. Nous désignerons les observations par $x_1, x_2, ...x_n$.

La moyenne :

Il s'agit de la mesure de localisation la plus importante et la plus largement utilisée. L'échantillon moyen d'un ensemble de données est :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La médiane :

La médiane de l'échantillon est l'observation du milieu lorsque les données sont classées par ordre croissant. On notera les observations classées $x(1), x(2), ..., x(n)$. Si le nombre d'observations est pair alors il n'y a pas de nombre du milieu, et donc la médiane est définie comme étant la moyenne de l'échantillon des deux observations du milieu.

$$Med = \begin{cases} x(\frac{n+1}{2}) \quad si \;\; n \; impair, \\ \dfrac{x(\frac{n}{2}) + x(\frac{n}{2}+1)}{2} \quad si \; n \; pair. \end{cases}$$

Parfois, l'utilisation de la médiane est préférée à la moyenne, en particulier lorsque les données sont asymétriques ou contiennent des valeurs aberrantes. Cependant, ses propriétés mathématiques sont moins faciles à déterminer que celles de la moyenne de l'échantillon, ce qui rend la moyenne de l'échantillon préférable pour l'analyse statistique formelle.

Mode :

Le mode est la valeur qui se produit avec la plus grande fréquence.

Mesures de propagation :

la connaissances des mesures de localisation des données seulement n'est pas suffisante. Nous devons également savoir à quel point il est concentré ou étalé. Autrement dit, nous devons savoir quelque chose sur la diversité des données. Les mesures de propagation sont un moyen de quantifier numériquement cette idée.

Variance et écart type :

La variance est la distance quadratique moyenne des observations par rapport à leur valeur moyenne. La variance de l'échantillon est donnée par :

$$Var = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{X})^2$$

L'écart type de l'échantillon σ est simplement la racine carrée de la variance de l'échantillon. Elle est préférée comme mesure récapitulative car elle est exprimée dans les unités des données

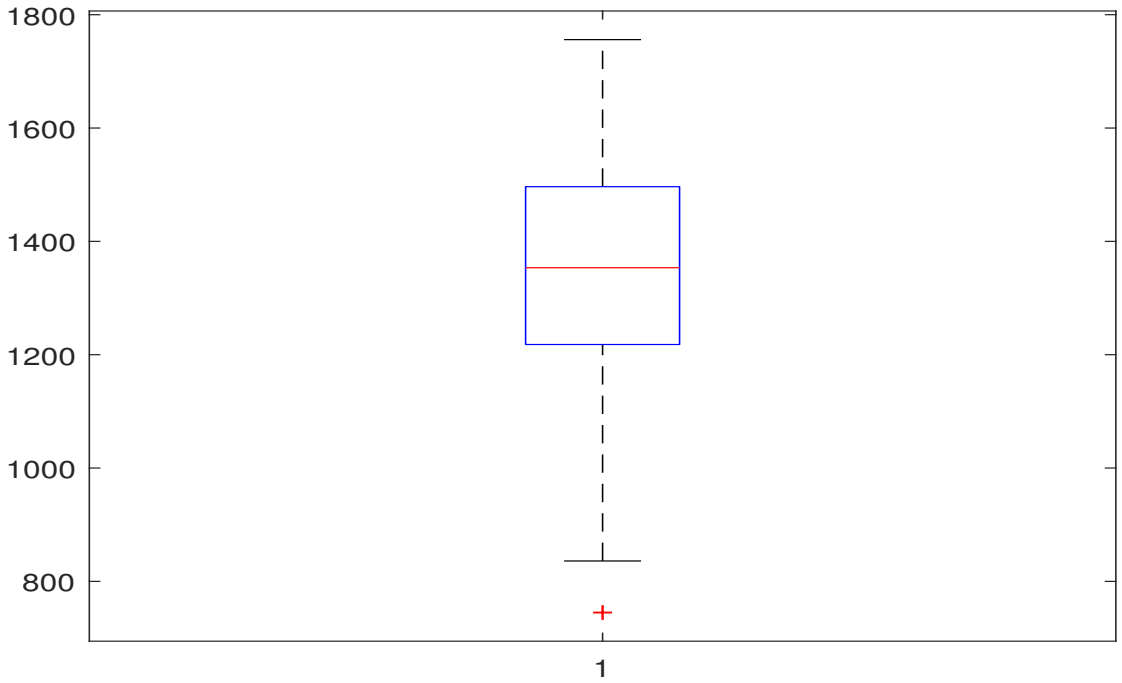

FIGURE 2.6 – Boite à moustache de la consommation journalière d'eau au restaurant universitaire

d'origine. Cependant, il est souvent plus facile d'un point de vue théorique de travailler avec des variances. Ainsi, les deux mesures sont complémentaires.

Quartiles :

Étant donné que la médiane contient la moitié des données de moins qu'elle, le quartile inférieur (Q1) a un quart des données en dessous, et le quartile supérieur (Q3) a un quart des données au-dessus.

Diagrammes en boîtes de moustaches : est une description graphique utile des principales caractéristiques d'un ensemble d'observations. Il existe de nombreuses variantes de la boîte à moustaches. La forme la plus simple est construite en dessinant une boîte rectangulaire qui s'étend du quartile inférieur au quartile supérieur et est divisée en deux à la médiane. De chaque extrémité de la boîte, une ligne est tracée vers les observations maximales et minimales.

Nous considérons les données de la consommation d'eau dans les jours ouvrable normales au restaurant universitaire. La consommation moyenne d'eau au restaurant universitaire est estimé par 1344 litres par jours et la consommation maximum égale à 1756 litres. La faible consommation journalières peut être estimée à 745 litre. La valeur de médiane n'est pas loi de la moyenne $Med = 1353.5$. La consommation d'eau qui se produit avec la plus part des jours est $Mode = 1099$. Donc, la distribution des données est à peut prés symétrique.

La boîte à moustache pour ces données est donnée ci-dessous avec la figure 2.6. Remarquez comment la symétrie de la consommation d'eau apparaît très clairement sur le graphique.

La consommation moyenne d'électricité à l'IUT de Mulhouse égale à 1.32×10^3 qui est défirent de la valeur de médiane 1.25×10^3 et le mode qui estime par 10^3. Alors, les données de la consommation électrique dans les jours ouvrable sont asymétrique.

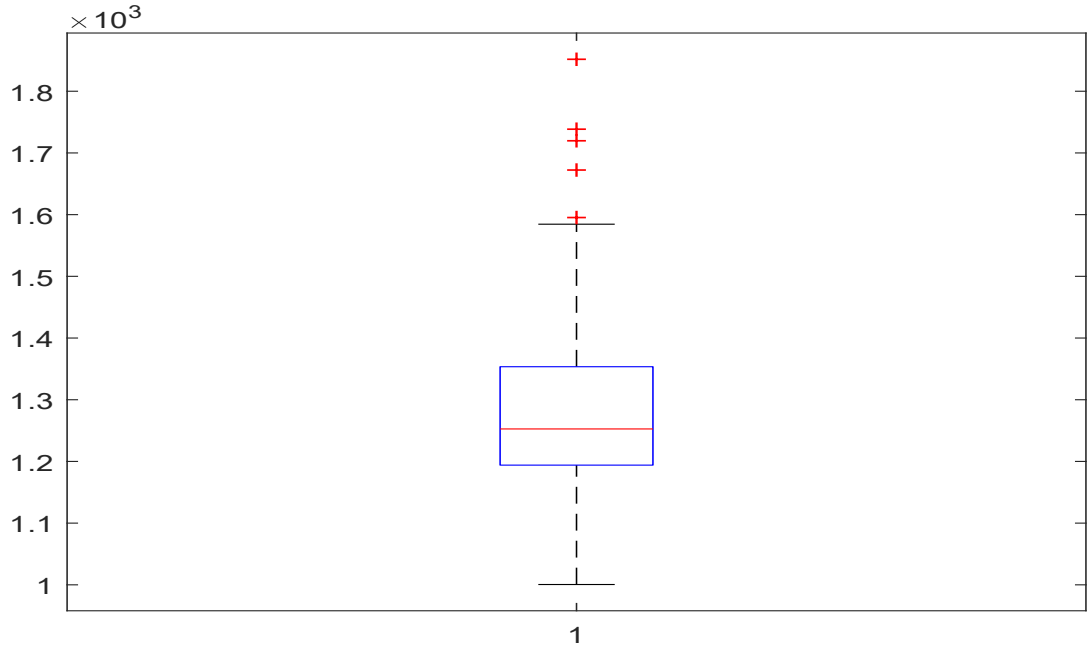

FIGURE 2.7 – Boite à moustache de la consommation journalière d'électricité à l'UT de Mulhouse

La figure 2.7 montre les principales caractéristique de la puissance dans les jours ouvrables par la boîte à moustache. Nous notons que la distribution des données est asymétrique.

2.4 Les courbes de charge

Dans cette partie, nous allons analyser les données par des courbes de charge journalières, hebdomadaires et mensuelles dans plusieurs bâtiments.
La courbe de charge est une représentation de la consommation cumulée d'eau ou d'électricité atteinte par un compteur sur une période donnée (jours, semaines, mois,...). Cette courbe définit le mode de vie des utilisateurs [24]. L'analyse de la courbe de charge pourrait permettre de mettre à jour les anomalies d'une consommation. La modélisation des courbes de charge est une tâche difficile en raison de la diversité des courbes de charge pour des jours sélectionnés qui représente à la fois la non-coïncidence de la consommation et la variété illimitée des caractéristiques des utilisateurs.
Parmi les utilisations de la courbe de charge, il y a la détermination des activités et des dispositifs utilisés dans la consommation. Bien qu'une bonne analyse de la courbe de charge permette d'envisager les usages électriques et hydropiques dans un équipement donné.

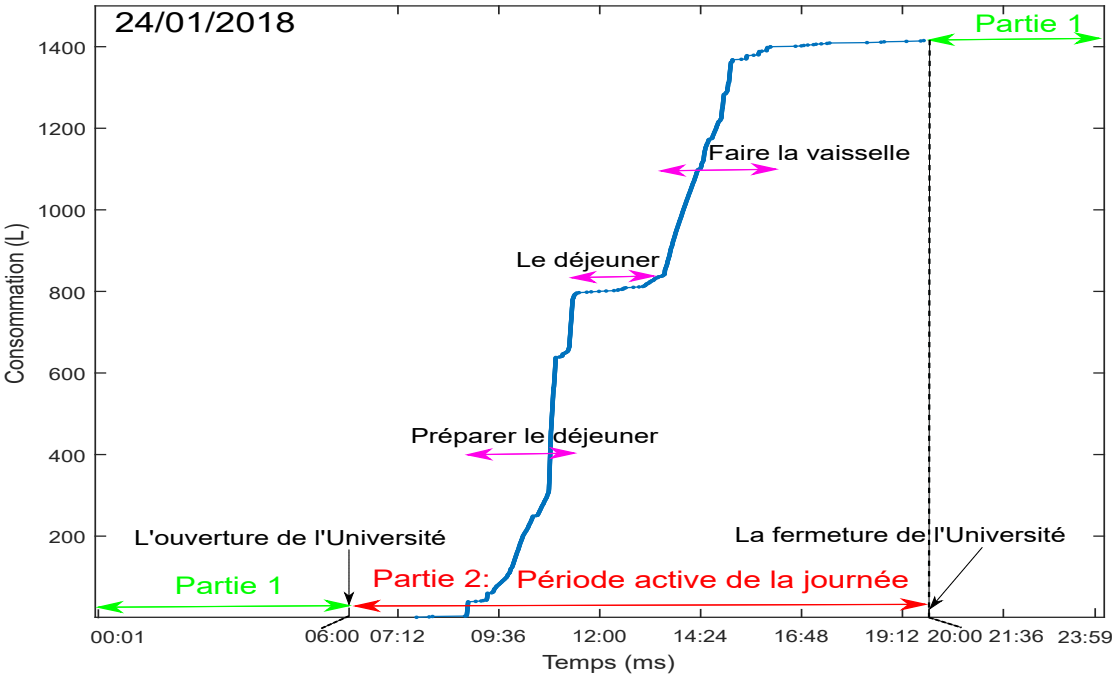

FIGURE 2.8 – La courbe de charge journalière de la consommation d'eau (restaurant universitaire).

2.4.1 Courbes de charge journalières

La courbe de charge quotidienne est une fonction croissante $y(t)$ qui représente les cumulés de la consommation pendant la journée où t désigne les instants de mesure [25]. Elle reflète des habitudes des différents consommateurs. Ainsi que la construction de la courbe de charge ultime peut détecter la charge de pointe quotidienne. Nous pouvons voir les changements de consommation lors du traçage de la courbe de charge quotidienne sur une base heure ou minute ou pour n'importe quelle période de temps.
La figure 2.8 donne un exemple de la courbe de charge journalière de la consommation d'eau dans un bâtiment tertiaire. La courbe de charge quotidienne de l'eau peut être divisée en deux parties, la partie 1 est représentée par les flèches vertes sur la figure 2.8 qui commence à minuit jusqu'à 6 heures du matin (ouverture de l'université) et de 20 heures (fermeture de l'université) jusqu'à la fin du jour où il n'y a pas de consommation d'eau dans les restaurants universitaires. La partie 2 est la période active de la journée où il y a de consommations. Nous remarquons dans cette dernière partie, trois changements dans la courbe. Tout d'abord, la première croissance qui s'explique par les consommations successives d'eau pour la préparation des repas et le lavage des légumes. Ensuite, nous avons la planéité de la courbe pendant les heures de déjeuner des étudiants lorsque l'eau est moins consommée. Puis il y a une deuxième croissance suite à la forte consommation de faire la vaisselle et de nettoyage de restaurant. Le début et la fin de chaque changement varient d'un jour à l'autre et dépendent des femmes de ménage et des étudiants.
Les courbes de charge journalières nous permettent de distinguer les jours normaux des jours anormaux. La figure 2.9 représente quelque courbes de charge journalières de la consom-

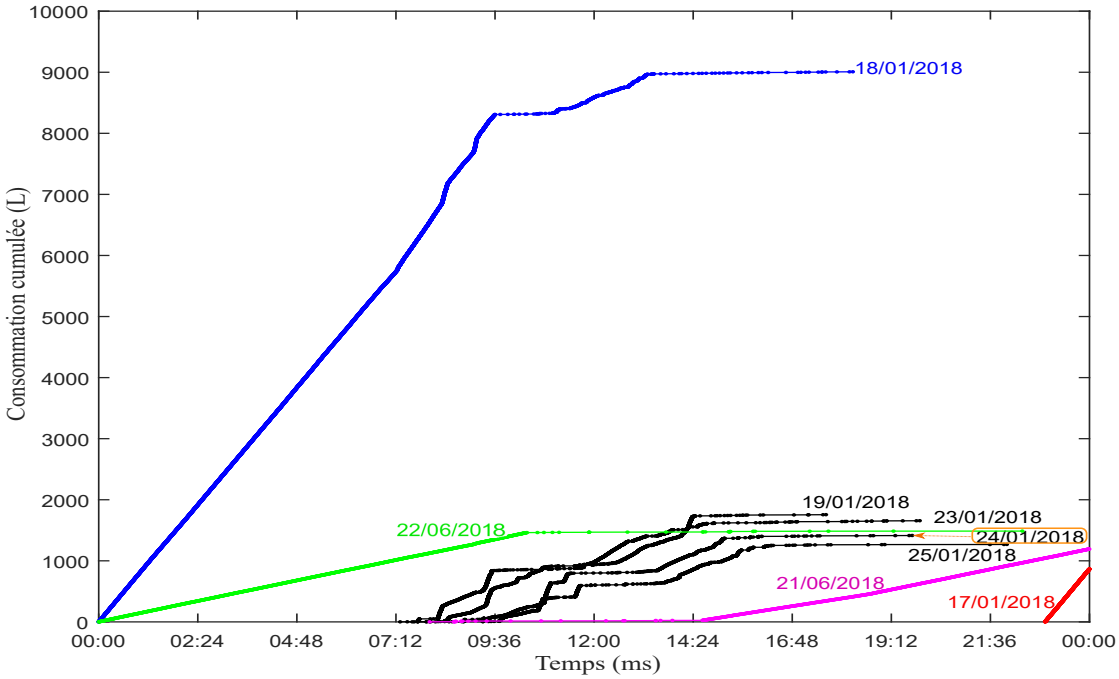

FIGURE 2.9 – Exemple de quelques courbes de charge journalière de la consommation d'eau.

mation d'eau dans un restaurant universitaire. Une observation préliminaire montre que le comportement des courbes de charge change d'un jour à l'autre. Chaque courbe présente une distribution différente, cela signifie que la consommation d'eau n'est pas régulière mais varie dans le temps. Une observation plus approfondie révèle certaines similitudes entre certaines courbes (c'est-à-dire les courbes de charge en noire), qui peuvent s'expliquer par les journées d'activité habituelles au restaurant universitaire. Malgré les points communs présentés par ces courbes, il reste très difficile de trouver un modèle précis capable de les représenter toutes. Les courbes colorées restantes sont des courbes spéciales qui illustrent les jours anormaux (situations de fuites). La courbe rouge marque le début d'une fuite à la fin de la courbe du 17 janvier 2018. Cette fuite n'a pas été détectée et se poursuit jusqu'au lendemain, soit le 18 janvier 2018 (représentée par la courbe bleue) jusqu'à ce qu'elle soit détectée au milieu de la journée. La deuxième fuite qui est représenté par les courbes respectivement en rose et vert a commencé le 21 juin 2018 et elle a été détectée le 22 juin 2018.

La figure 2.10 montre quelques courbes de charge journalières de la consommation électrique à l'IUT de Mulhouse. Nous observons que les courbes en noir ont la même forme qui signifient la consommation normale dans les jours ouvrables. Au temps que les autres courbes sont similaires, ressemblent à une droite et avec moins de consommation. Les deux courbes en rouges sont des courbes de jours de week-end et les courbes en vertes sont des jours de vacances où la consommation est inférieure aux jours normales. Les courbes en bleus sont représentés deux jours fériés où l'université est fermée.

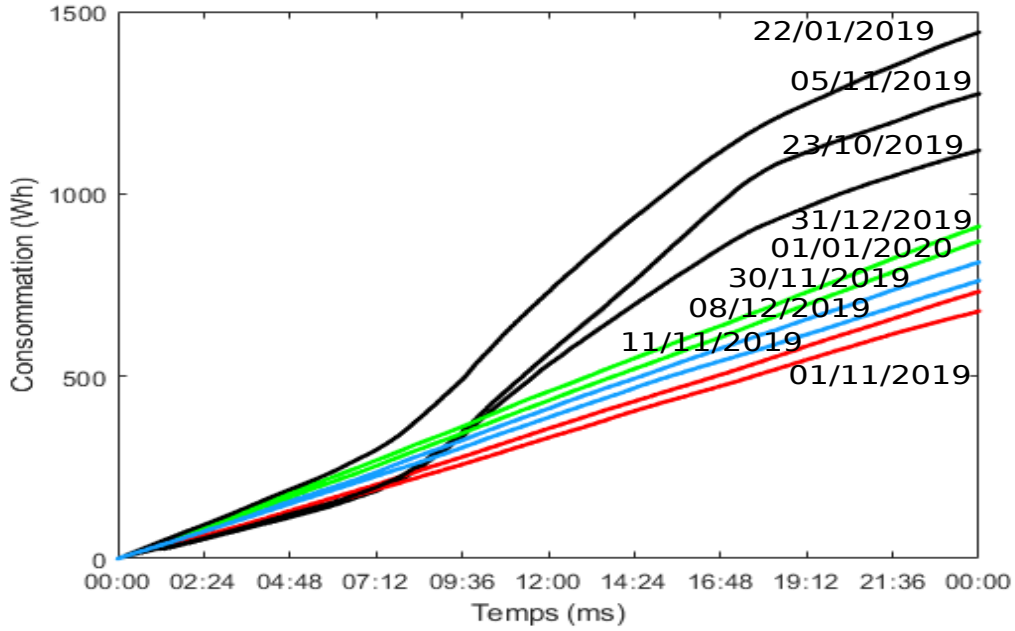

FIGURE 2.10 – Quelques courbes de charge journalière de consommation électrique

2.4.2 Courbes de charge hebdomadaires

Grâce à la courbe de charge hebdomadaire, nous différencions les jours ouvrables de la semaine, les week-ends et même les jours fériés.
La figure 2.11 représente certaines des courbes de charge hebdomadaires de la consommation d'eau dans un restaurant universitaire. Nous remarquons que les courbes de charge en noir sont assez similaires. Notez que chaque courbe noire contient cinq courbes de charge journalières pour les jours ouvrables (de lundi à vendredi) et quelques points les week-end. Dans la courbe de charge en vert, il y a six courbes de charge journalières, en fait il y a une grand consommation le samedi qui justifie par la porte ouverte de l'université. Ainsi que la courbe en orange a que quatre courbes de charge journalières où il n'y a pas la CdC journalière de lundi qui correspond au jour férié. Une semaine de vacances a marqué par la courbe en rose. La courbe en rouge est différente des autres car cette semaine elle était cohérente avec la semaine des examens où il n'y a pas beaucoup de consommation, mais il y a une grande consommation jeudi et vendredi (21 et 22 juin 2018) qui sont deux jours de fuites au restaurant universitaire.
La figure 2.12 représente quelques courbes de charge hebdomadaires de la consommation électrique à l'IUT de Mulhouse. Les courbes en rouge sont représentées deux semaine de vacances de Noël où la consommation pas grande et elle se consomme de la même manière. Alors que les autres courbes diffèrent d'elles et elles sont similaires entre elles qui représentent des semaines normales. Nous remarquons que la courbe de semaine normale comme celle en bleu a 5 courbes de charge journalières des jours ouvrables suit de deux courbes de charge journalières de week-end.

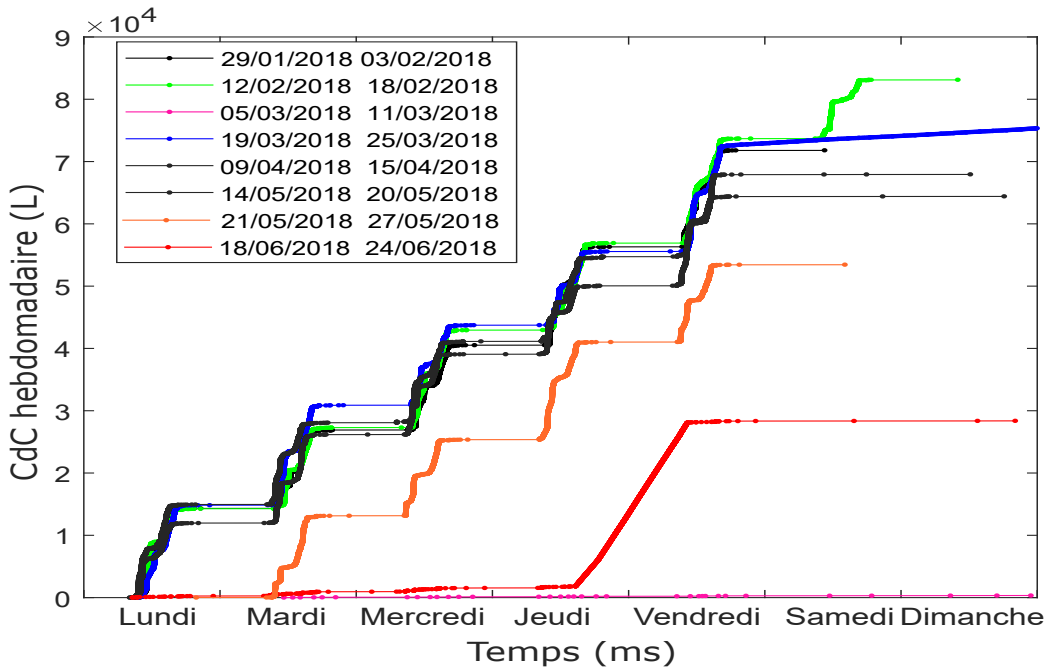

FIGURE 2.11 – Quelques courbes de charge hebdomadaires dans le restaurant de l'IUT

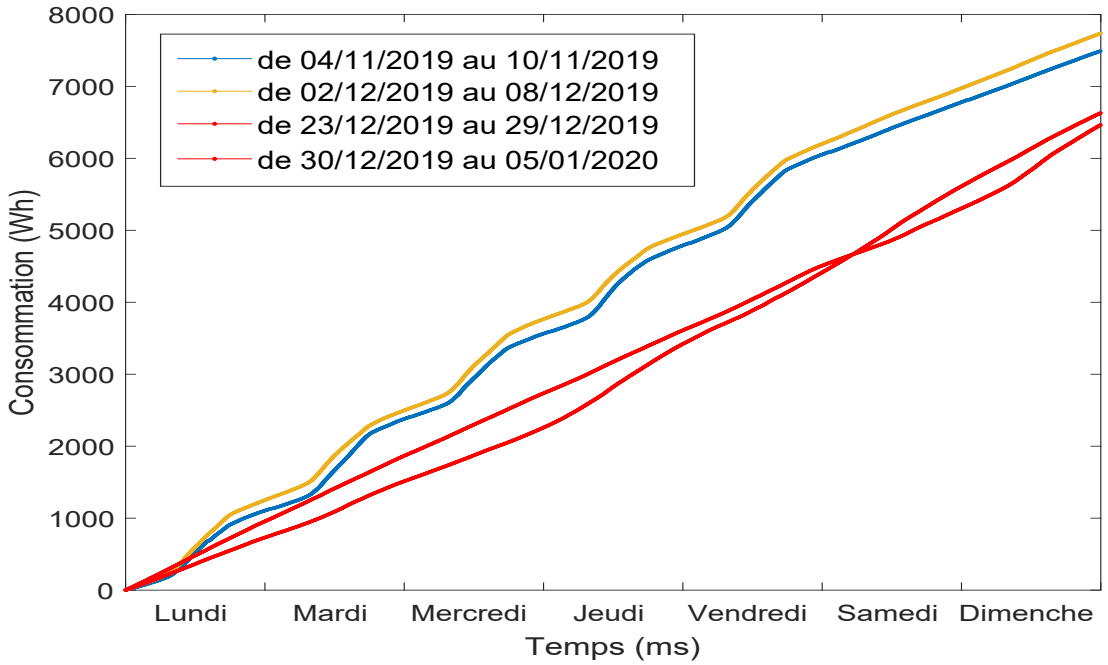

FIGURE 2.12 – Exemple de quelques courbes de charge hebdomadaires à l'IUT de Mulhouse

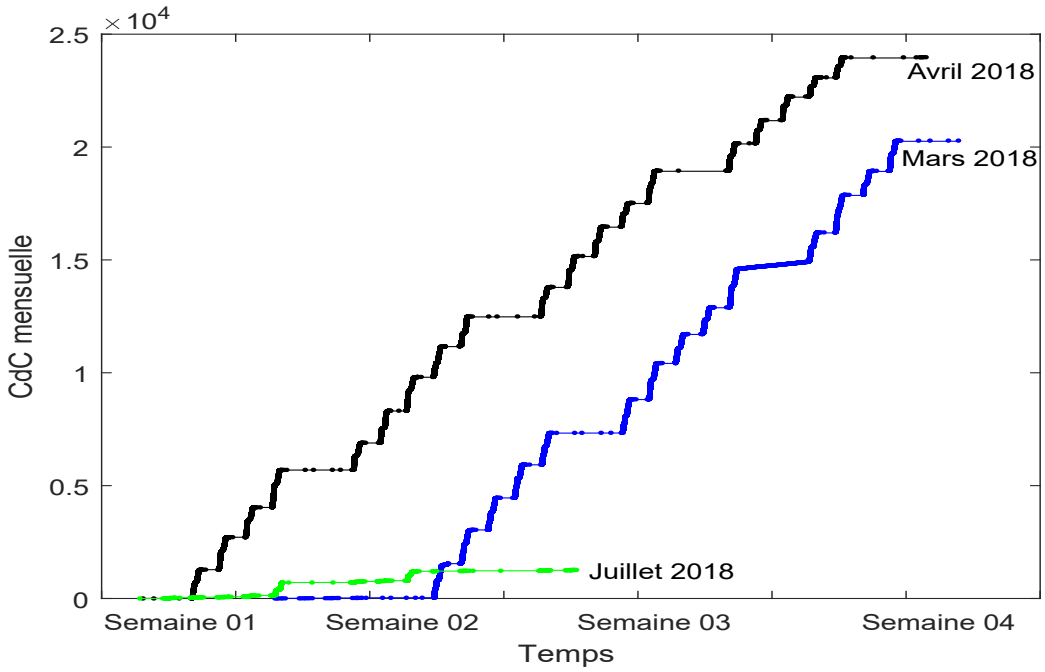

FIGURE 2.13 – Quelques courbes de charge mensuelles dans le restaurant de l'IUT

2.4.3 Courbes de charge mensuelles

À partir des courbes de charge mensuelles nous pouvons voir les jours des vacances au restaurant de l'IUT.
La figure 2.13 représente les courbes de charge de la consommation d'eau dans le restaurant universitaire au mois de mars, avril et juillet. La courbe en noir du mois d'avril contient quatre CdC hebdomadaire qui est signifié une situation normale. Bien que la courbe en bleu a que trois CdC hebdomadaire normale et la première semaine a mois de consommation car c'est une semaine de vacance. Le mois de juillet est le début des vacances d'été qui est représenté par la courbe de moins consommation en verte. La petite consommation en juillet est celle des enseignants et des doctorants qui prennent les vacances après les étudiants.

3 Réduction des données

La réduction des données est devenue une tache très importante dans les applications d'exploration de données. L'exploration de données est l'extraction de connaissances à partir d'une grande base de données. Construire des modèles à partir de données en utilisant un ensemble d'approches statistiques et l'intelligence artificielle permet d'extraire un maximum de connaissance.
La réduction des données est une transformation de données numériques en une forme corrigée, structurée et simplifiée. Nous avons utilisé des techniques de réduction des données pour résoudre le problème de stockage (libérer de la capacité sur un périphérique de stockage) et augmenter la durée de vie des appareils. De plus, cela permet d'économiser de l'énergie et augmenter l'efficacité de l'exploration de données.
Dans ce chapitre, nous avons appliqué les études connues dans la littérature telle que la réduction de la numérosité pour la réduction des données de la CdC journalière de la consommation d'eau et d'électricité. Ainsi, nous allons expliquer le principe des cartes auto-organisatrices en détaillant leurs algorithmes d'apprentissage. Ensuite, nous montrons son application pour réduire les données de consommation.

3.1 La réduction de la numérosité

La réduction numérosité réduit le nombre de variables en choisissant des représentations alternatives de formes plus petites. Elle utilise des techniques qui peuvent être paramétriques ou non paramétriques. Dans les techniques paramétriques, des modèles ou fonctions sont utilisés pour estimer les données. Parmi les méthodes non paramétriques, nous mentionnons l'échantillonnage et le regroupement.

3.1.1 La régression linéaire

La régression linéaire est une méthode paramétrique de la réduction de la numérosité. La réduction à l'aide de la régression linéaire est réalisée en remplaçant les données par deux

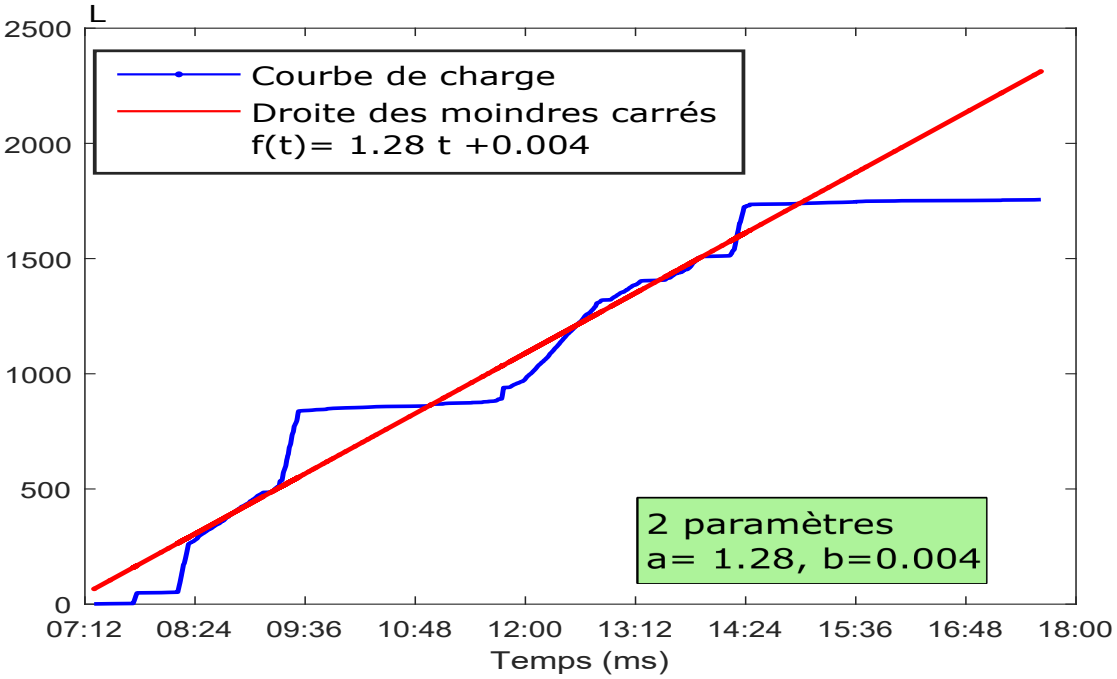

FIGURE 3.1 – La réduction des données de la consommation d'eau au restaurant universitaire le 19/01/2018 par la régression linéaire

paramètres de la droite de régression $f(t) = at + b$. Pour estimer les paramètres de la droite de régression, la méthode des moindres carrés est utilisée.

La droite de régression des données de consommation d'eau du 19/01/2018 est donnée par $1.28t + 0.004$. La droite de la regression permet de réduire les données de la CdC de la consommation d'eau de 1756 à 2 avec un coefficient de corrélation de 2.3×10^{-10}.

La figure 3.1 représente la droite de régression linéaire en rouge et la CdC de la consommation d'eau du 19/01/2018 en bleu.

La réduction des données avec la droite de régression linéaire définie par $10^9 t - 2.56 \times 10^9$ réduit les données de la consommation électrique de 26409 à 2 avec un coefficient de corrélation de 0.6×10^{-3}. La figure 3.2 montre les résultats de la réduction des données de la consommation électrique le 07/01/2021 par la régression linéaire. La courbe en bleu représente les données et la droite en rouge est la droite de régression linéaire.

La régression linéaire permet de réduire les données de consommation à deux paramètres seulement mais on perd beaucoup d'informations sur la consommation et la droite est loin de la courbe aux bords et les coefficients de corrélation sont loin de 1 et -1.

3.1.2 L'échantillonnage des données

L'échantillonnage des données consiste à sélectionner un sous-ensemble de données afin d'identifier des informations significatives qui concernent globalement l'ensemble des données. L'échantillonnage est l'une des techniques fondamentales de statistique. Cette technique pré-

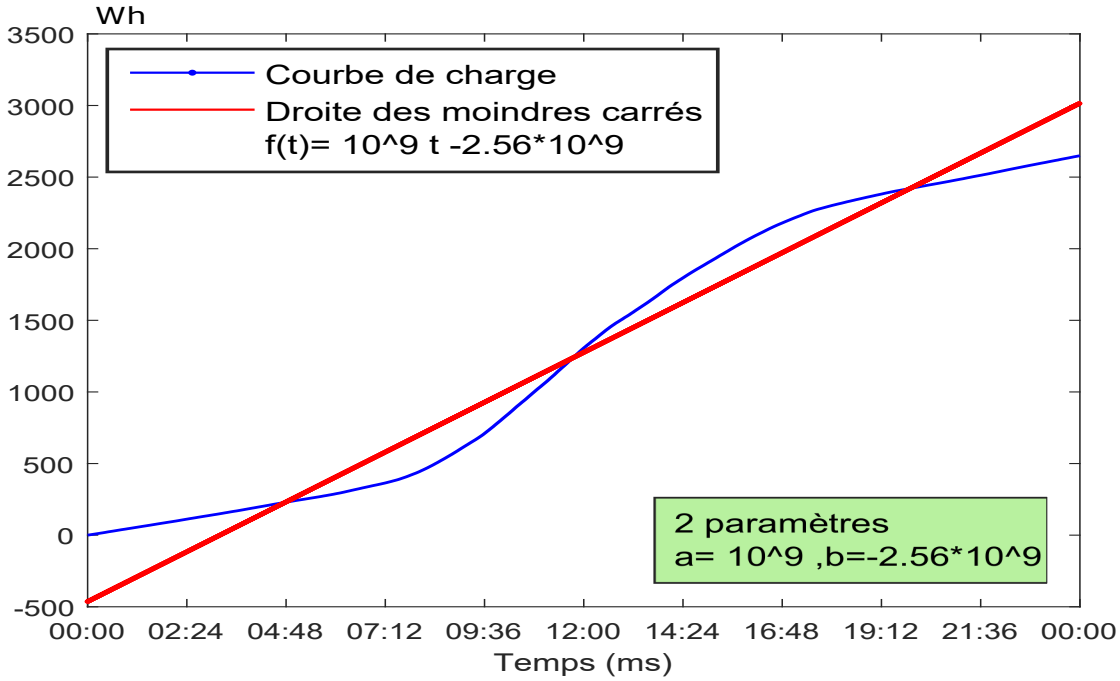

FIGURE 3.2 – La droite de régression de la consommation électrique à l'IUT de Mulhouse le 07/01/2020

sente plusieurs avantages : une étude restreinte sur une partie de la population, un moindre coût, une collecte des données plus rapide que si l'étude avait été réalisée sur l'ensemble de la population, et la réalisation de contrôles descriptifs. La génération d'échantillons permet notamment de tester une hypothèse sur un échantillon, puis de la valider sur un autre et d'obtenir des ensembles de données avec une taille plus petite tout en préservant des propriétés des données d'origine. On peut également déduire des propriétés de la population à partir de celles de l'échantillon par inférence statistique.

L'échantillonnage peut être utilisé comme une technique de réduction des données, car il permet à un grand ensemble de données d'être représentées par un échantillon (ou sous-ensemble) beaucoup plus petit. Il existe plusieurs méthodes d'échantillonnage des données pour réduire la taille d'un grand ensemble de données, et comme nos données sont horodatées et en millisecondes, nous pouvons échantillonner en heures, en minutes et même en seconde les données de la courbe de charge journalière.

L'échantillonnage en heure (en minute) consiste à calculer le nombre de litres d'eau ou Watt-heure de puissances électrique consommé chaque heure (chaque minute). La courbe de charge journalière de la consommation d'eau au restaurant universitaire le 19/01/2018 contient 1756 points qui sont représentés par la courbe en bleu dans la figure 3.3. La réduction des données par l'échantillonnage en heure permet de n'avoir que 24 points de données colorés en vert dans la même figure. L'échantillonnage en minute diminue des données de la consommation d'eau pendant la journée à 1440 points de données qui sont illustrés par les points rouges de la figure 3.3.

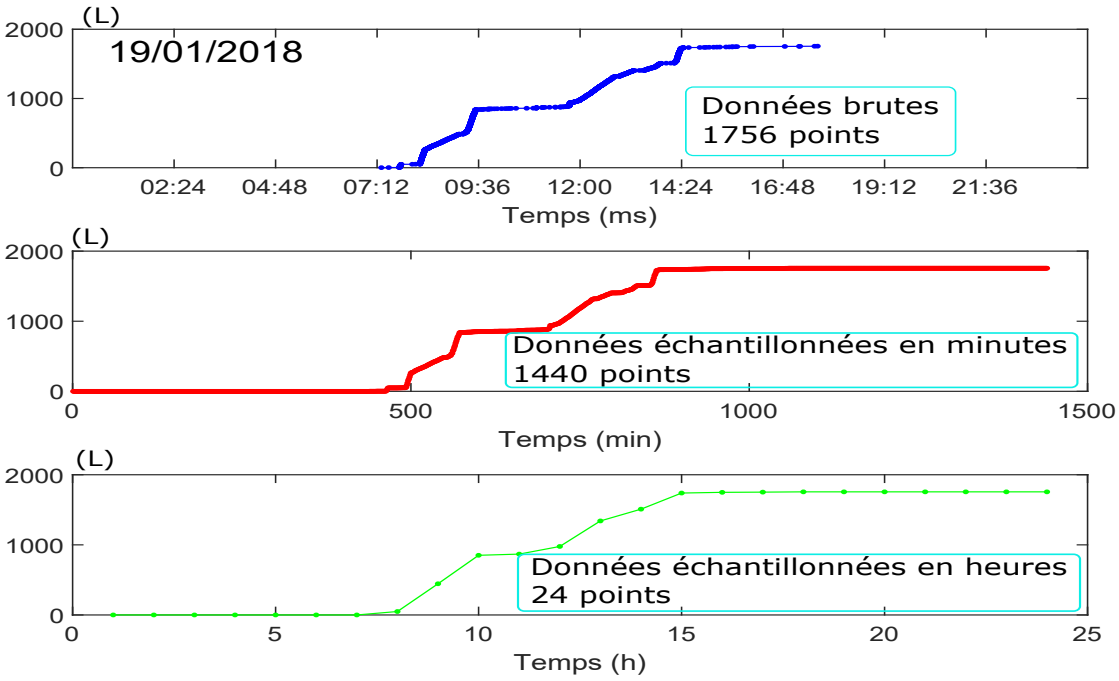

FIGURE 3.3 – Échantillonnage de la consommation d'eau au restaurant universitaire le 19/01/2018

La figure 3.4 montre la diminution des données de la puissance électrique à l'IUT de Mulhouse le 07/01/2020 de 26409 points à 24 points par l'échantillonnage en heures et à 1440 points par l'échantillonnage en minute.

3.1.3 Regroupement des données selon 3 types de consommation

Le regroupement des données crée un aperçu permettant d'identifier des modèles, des tendances, des irrégularités ou des valeurs aberrantes. Le regroupement permet de déterminer le nombre d'enregistrements et la valeur ou la quantité concentrée par l'une des mesures ou l'un des identificateurs de votre choix. Les données de regroupement sont plus simples. Le regroupement de données peut aider à afficher un sous-ensemble de données réduit.
Il existe plusieurs façons de créer un groupe. Nous avons créé des groupes des données à partir de la manière de consommation selon 3 types de consommation (petite, moyenne, et grande consommation). Puis, on a réduit les données en sélectionnant le début et la fin de chaque groupe.
Pour chaque jour on va regrouper les données par 3 classes selon 3 types de la consommation :

* **Grande consommation** ; si les $\triangle t_i = t_i - t_{i-1}$ sont dans le premier tiers entre la consommation minimale et la consommation maximale, c'est à dire :

$$\min_j(\triangle t_j) \leq \triangle t_i \leq \frac{\max_j(\triangle t_j) - \min_j(\triangle t_j)}{3}$$

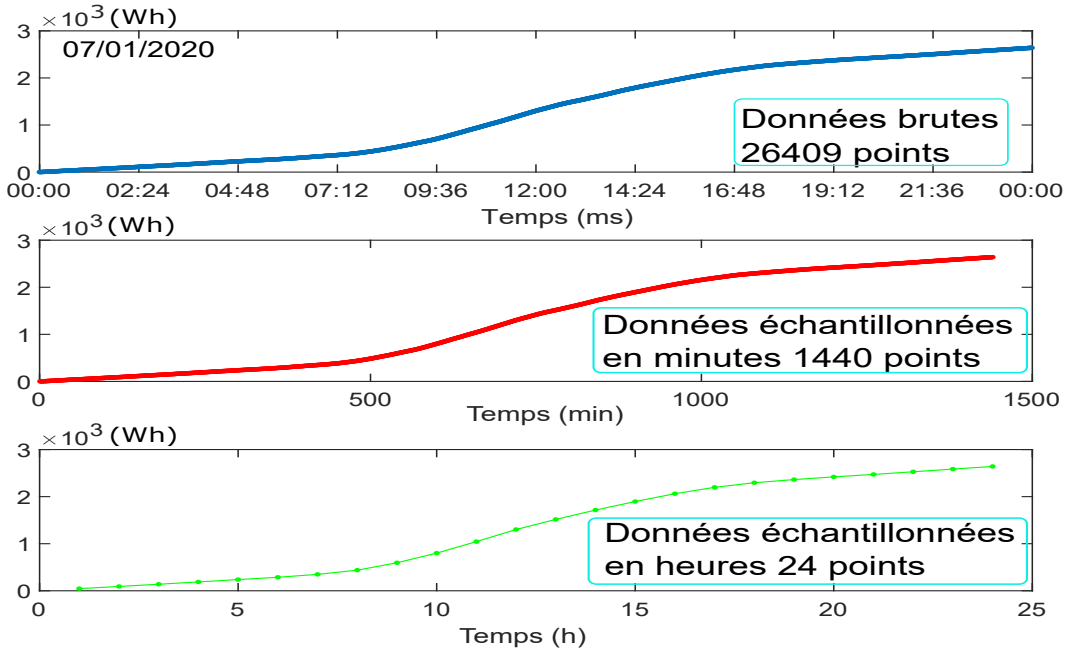

FIGURE 3.4 – La réduction des données de la puissance électrique à l'IUT de Mulhouse le 07/01/2020 par l'échantillonnage des données

* **Petite consommation** ; si les $\triangle t_i$ sont dans la troisième tiers c'est à dire :

$$2\frac{\max_j(\triangle t_j) - \min_j(\triangle t_j)}{3} \leq \triangle t_i \leq \max_j(\triangle t_j)$$

* **Moyenne consommation** ; si

$$\frac{\max_j(\triangle t_j) - \min_j(\triangle t_j)}{3} < \triangle t_i < 2\frac{\max_j(\triangle t_j) - \min_j(\triangle t_j)}{3}$$

Les figures 3.5 et 3.6 représentent le regroupement des données de la consommation d'eau et de l'électricité selon 3 types de consommation.
Dans le premier temps, on classe chaque point des données brutes de la courbe de charge journalière selon son écart de temps au type de consommation précédent. Puis, on les représente par des couleurs indiquant le type de consommation.
Dans la deuxième partie, on relie le premier et le dernier point dans chaque classe si la classe contient au moins 2 points, sinon (si on a 2 points au plus qui n'appartient pas à la même classe) on calcule la moyenne de ces $\triangle t_i$ et on vérifie à quelle classe la moyenne appartient pour dessiner la droite qui passe par le premier et la dernier point avec la couleur correspondante.
Dans la troisième partie on représente les classes par un histogramme. C'est-à-dire que chaque groupe est représenté par une barre colorée par type de consommation.
La réduction des données avec le regroupement des données de la consommation d'eau au restaurant universitaire selon 3 types de consommation permet de garder 11 points au lieu de 1756 points montrés dans la figure 3.5. Dans la CdC journalière de la consommation d'électricité

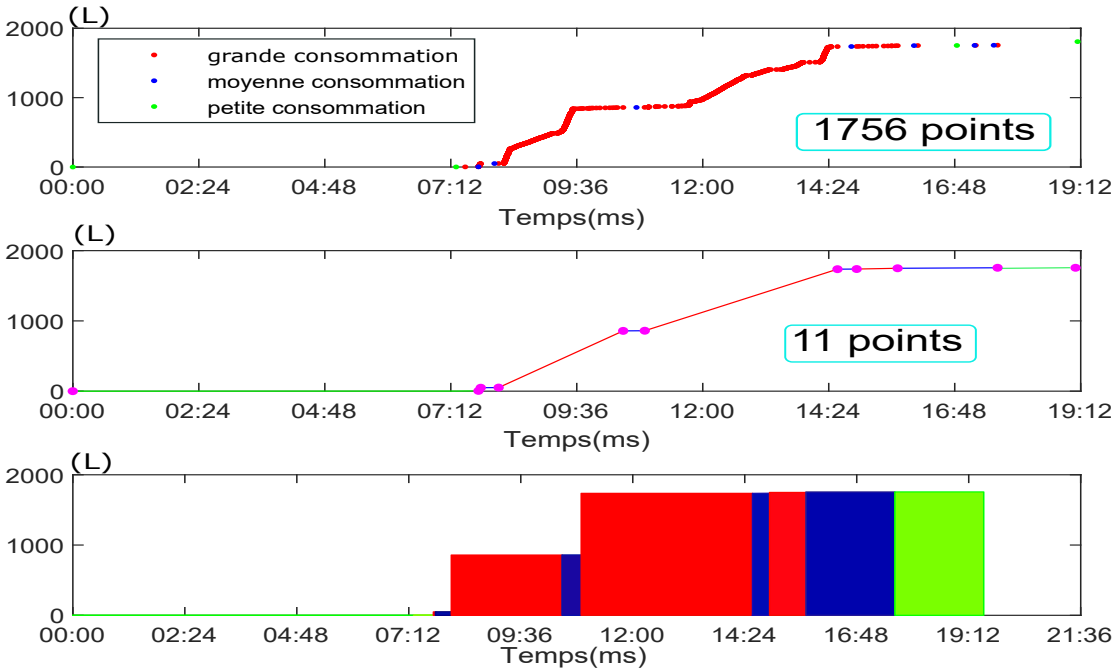

FIGURE 3.5 – Regroupement en trois types de consommation d'eau dans le restaurant de l'IUT le 19/01/2018

à l'IUT de Mulhouse, on a 26409 points et après le regroupent on ne garde que 2266 points.

3.2 Carte auto-organisatrice

La carte auto-organisatrice (SOM) ou carte topologie a été inspirée par le principe neuronal du cerveau des mammifères et inventée par Kohonen (1984) [26]. Il s'agit d'un type de réseau de neurones artificiels dont l'apprentissage se déroule de manière non supervisée. Une carte auto-organisatrice est composée d'une grille de neurones organisés dans une seule couche. Chaque neurone possède des poids qui définissent sa position dans l'espace des données. Quand la grille est unidimensionnelle, chaque neurone a deux voisins. Une carte auto-organisatrice est évaluée pour ses capacités de quantification et ses capacités de préservation de la topologie. Nous visons à utiliser une carte auto-organisée unidimensionnelle sur les données de courbe de charge pour réduire les données. Le choix d'une carte auto-organisatrice unidimensionnelle a fourni la dimension minimale de la carte qui a permis d'obtenir une courbe lisse grâce à l'avantage d'avoir au maximum deux voisins par neurone.

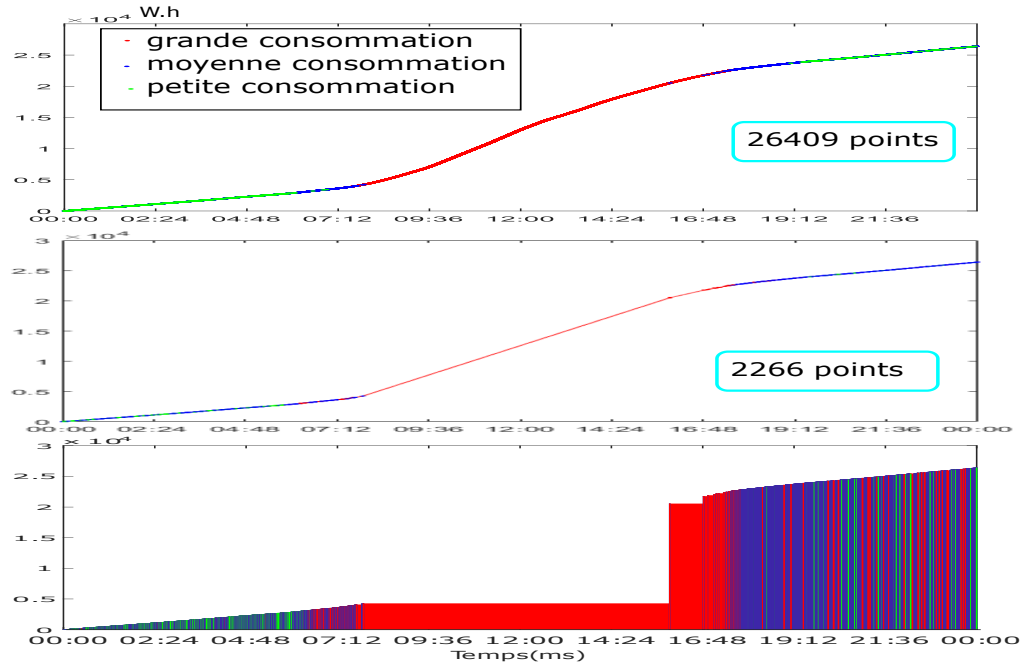

FIGURE 3.6 – Regroupement en trois types de consommation électrique à l'IUT de Mulhouse le 07/01/2020

3.2.1 Algorithme d'apprentissage

Principe

On considère une carte auto-organisatrice de dimension P avec K neurones, le vecteur référent du neurone n_k est reconnu par w_k avec $k \in 1,...,K$ et $w_k \in R^P$.
Après une initialisation aléatoire des valeurs de chaque neurone, on soumet une à une les données à la carte auto adaptative pour l'apprentissage qui s'effectue en deux étapes :

1- **La compétition entre les neurones.**
Lorsqu'on présente un vecteur d'entrée, on cherche le neurone qui répondra le mieux au stimulus. Celui dont la valeur sera la plus proche de la donnée présentée.

2- **L'adaptation des poids des neurones.**
Le neurone gagnant sera gratifié d'un changement de valeur pour qu'il réponde encore mieux à un autre stimulus de même nature que le précédent. Par là même, les neurones voisins du gagnant seront également gratifiés un peu avec un facteur multiplicatif du gain inférieur à un.

C'est donc toute la région de la carte autour du neurone gagnant qui se spécialise. À la fin de l'algorithme, lorsque les neurones arrêtent de bouger, ou très peu, à chaque itération, la carte auto-organisatrice couvre toute la structure des données.

Formalisation mathématique

La carte de l'espace d'entrée est réalisée en ajustant les vecteurs de référence w_k. L'adaptation se fait par un algorithme d'apprentissage dont la force réside dans la compétition entre neurones et dans l'importance accordée à la notion de voisinage. Un nouveau cycle d'adaptation est fait pour chaque vecteur d'entrée. Pour chaque vecteur de données v dans la séquence, on détermine le neurone vainqueur, c'est-à-dire le neurone dont le vecteur référent approche v le plus.

$$s = \arg \min_{k \in Grille} \|v - w_k\|$$

Le neurone vainqueur n_s et ses voisins n_r déplacent leurs vecteurs référents vers le vecteur d'entrée v.

$$w_s(new) = w_s - \mu(v - w_s)$$

$$w_r(new) = w_r - \varepsilon\mu(v - w_r)$$

Où μ représente le coefficient d'apprentissage et ε la fonction qui définit l'appartenance au voisinage.

Algorithme

L'algorithme 1 résume les étapes de l'apprentissage des cartes auto-organisatrices.

Algorithm 1 Apprentissage des cartes auto-organisatrices

Initialisation : Initialiser les vecteurs de référence $w_k^{(0)}$ avec des valeurs aléatoires.
Apprentissage :
for Chaque vecteurs de données d'entrée v_i **do**
  for Chaque vecteurs de référence de la carte w_k; $k = 1...K$ **do**
    $d_k = norm(v_i,\ w_k)$
  end for
  $s = \arg\min_{1<k<K} d_k$
  $w_s^{(i)} = w_s - \mu(v_i - w_s)$
  for Chaque vecteurs de référence (w_r) au voisinage du vecteur vainqueur (w_s) **do**
    $w_r^{(i)} = w_r - \varepsilon\mu(v - w_r)$
  end for
end for

3.2.2 Applications

Dans la CdC journalière de la consommation d'eau de 19/01/2018, on a 1756 points. On prend une carte unidimensionnelle initiale de 190 neurones colorés en rose dans la figure 3.7. On applique l'algorithme d'apprentissage 1 cinq fois. Chaque fois, on obtient une nouvelle représentation de la carte et on arrête à la carte rouge qui couvre la structure des données. Donc, avec la

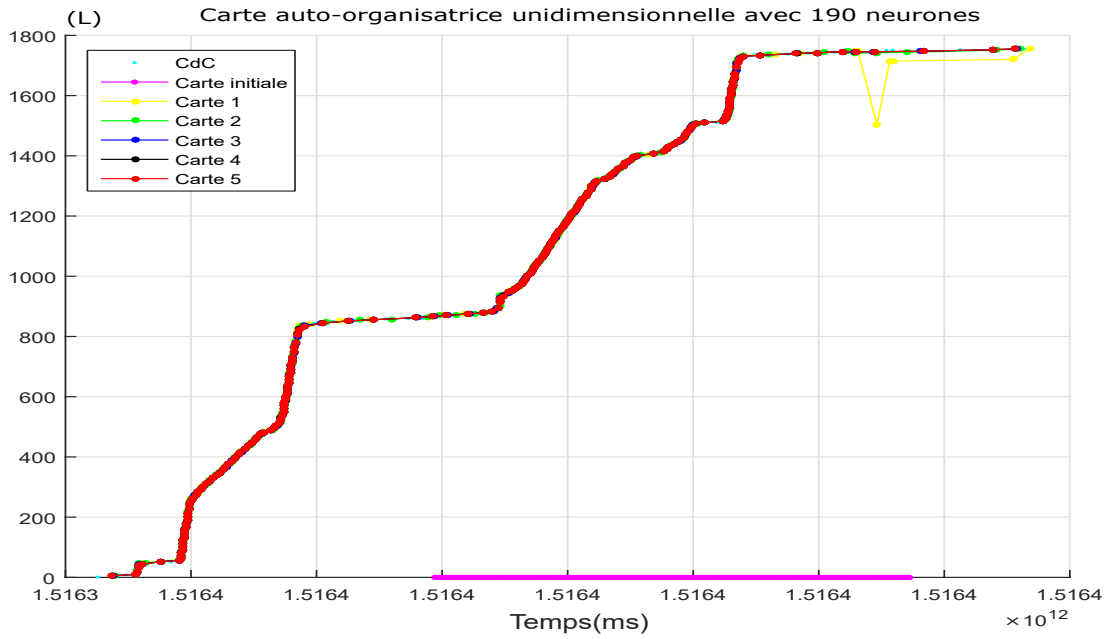

FIGURE 3.7 – La carte auto-organisatrice des données de la CdC journalière de la consommation d'eau le 19/01/2018

carte auto-organisatrice on a réduit le nombre de points de données de 1756 à 190 points. Les paramètres de la carte sont sélectionnés par des tests empiriques.
L'application d'une carte auto-organisatrice sur la CdC électrique de 07/01/2020 permet de réduire les données de 26409 points à 270 points.
Les figures 3.7 et 3.8 représentent les résultats de la réduction des données d'eau et d'électricité avec la carte auto-organisatrice.

3.3 Comparaison

Pour comparer les méthodes que nous avons utilisées pour réduire les données de consommation, nous avons adopté deux critères :
1- Nombre de paramètres à stocker.
2- Erreur ; on a utilisé l'erreur moyenne quadratique. Puisque le nombre de points de réduction x^* est inférieur au nombre de points des données x, nous remplaçons le nombre de points de réduction par l'évaluation du temps de données dans la droite qui relie chacun des deux points de réduction. On définit l'erreur par :

$$Erreur = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t - a_i t + b_i)^2} \tag{3.1}$$

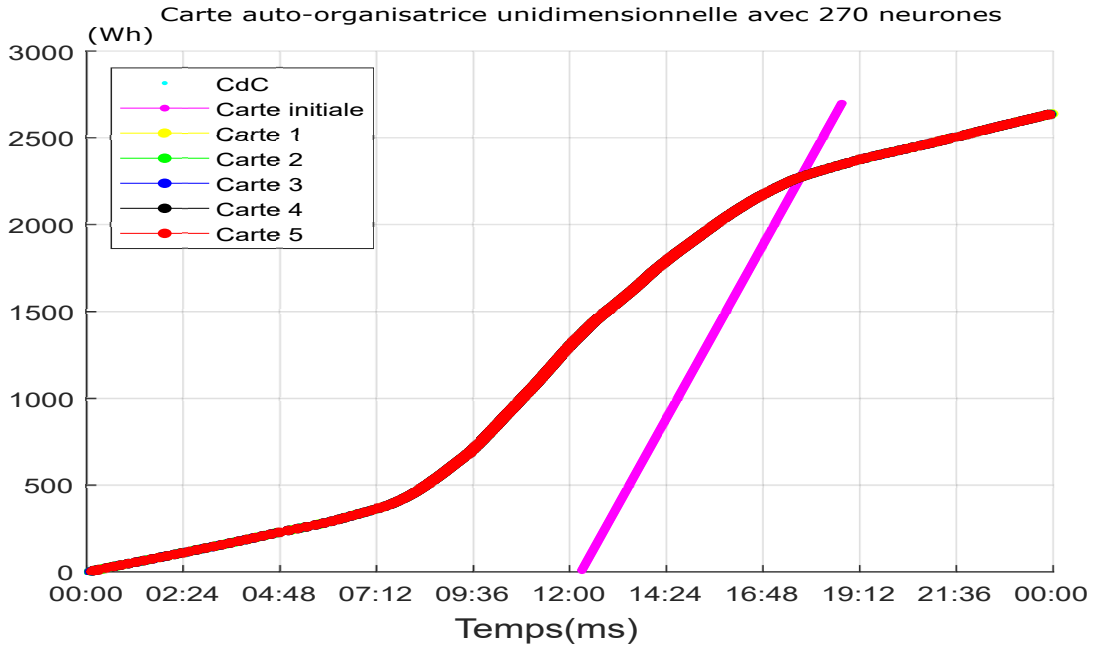

FIGURE 3.8 – La carte auto-organisatrice des données de la CdC journalière de la puissance électrique le 07/01/2020

TABLE 3.1 – Comparaison des méthodes de réduction des données

	CdC d'eau "19/01/2018" à 1756 points		CdC d'electricité "07/01/2021" à 26409 points	
	Nombre de paramètres	Erreur	Nombre de paramètres	Erreur
Régression linéaire	2	105.82	2	179.06
Échantillonnage en heure	24	54.29	24	5.12
Échantillonnage en minute	1440	1.88	1440	0.08
Regroupement	11	114.49	2266	70.51
Carte auto-organisatrice	190	2.82	270	8.98

Tels que $a_i = \dfrac{x_k^* - x_{k+1}^*}{t_k^* - t_{k+1}^*}$; $b_i = x_k^* - a_i t_k^*$ avec $t \in [t_k^*,\ t_{k+1}^*]$

Le tableau 3.1 montre les résultats de la comparaison selon les deux critères. Le nombre minimum de paramètres est donné par régression linéaire mais l'erreur est grande. Pour les données de consommation d'eau du 19/01/2018, l'erreur minimale est donnée par l'échantillonnage en minutes mais le nombre de points est de 1440, ce qui n'est pas loin du nombre des données 1756. La deuxième erreur minimale est donnée par la carte d'erreur qui réduit les données à 190 points, ce qui est un nombre acceptable. L'echantillonnage en heure permet de réduire les données de la consommation électrique du 07/01/2021 à 24 points avec une erreur faible.
Après avoir appliqué plusieurs méthodes de réduction des données aux CdC journalières de la consommation d'eau et d'électricité, nous concluons que la carte auto-organisatrice extrait un maximum de connaissances sur les données de consommation d'eau et que l'échantillonnage en heure est le meilleur moyen de réduire les données de la consommation électrique.

4 Modélisation des données

4.1 Introduction

La demande d'eau et d'électricité est nécessaire et augmente constamment. Par conséquent, nous nous efforçons de développer des modèles afin de pouvoir comprendre la consommation d'eau et d'électricité. La modélisation adaptative de la courbe de charge permet de générer un profil de consommation. Dans l'objectif de trouver un modèle quotidien de la consommation d'eau et d'électricité d'un bâtiment tertiaire à partir de sa courbe de charge, nous mentionnons les méthodes d'interpolation et d'approximation ainsi que des modèles avec résidus.
La recherche d'un modèle permettant de prévoir avec précision la consommation est l'un des défis majeurs des systèmes d'approvisionnement en eau et en électricité. La prévision de la consommation d'eau et d'électricité est essentielle pour la gestion de cette ressource, qui n'existe qu'en quantité limitée. Ainsi, différents modèles de prévision ont été proposés. Ces modèles sont généralement basés sur des outils de traitement et d'analyse de séries chronologiques, et sur différentes techniques. En fait, ces modèles ont récemment démontré leur succès dans les prévisions de consommation.
Dans ce chapitre, nous présentons des modèles de courbe de charge en utilisant les méthodes d'interpolation et d'approximation. Nous mentionnons la méthode d'interpolation de Lagrange, la méthode d'interpolation au sens de Tchebychev, le modèle d'approximation par la méthode des moindres carrés, l'interpolation de spline cubique et les courbes de Bézier. Ainsi que des modèles probabilistes tels que le modèle de mélange gaussien et l'estimation de la densité par noyau. Nous mentionnerons également les modèles avec résidus : Modèle déterministe, modèle de moyenne mobile intégrée autorégressive saisonnière SARIMA, modèle de Holt-Winters et le modèle hybride. En plus, nous suggérons plusieurs méthodes de prévision des séries chronologiques, telles que des méthodes déterministes, méthodes stochastiques, des réseaux de neurones, des méthodes de lissage et des méthodes hybrides.

4.2 Interpolation et approximation

L'objectif est de déterminer une fonction la plus proche possible des données pour les modéliser.

4.2.1 Interpolation polynomiale de Lagrange

Le problème d'interpolation consiste à chercher une fonction qui passe par tous les points d'interpolation.
Soient $(x_i,\ y_i);\ pour\ i = 0,\ 1,\ ...,\ n$, un nuage de points, nous appelons x_i un nœud d'interpolation et les couples $(x_i,\ y_i);\ i = 0,\ 1,\ ...,\ n$, les points de collocation.
Nous définissons le polynôme d'interpolation de Lagrange [27] par (4.1) :

$$p(x) = \sum_{j=0}^{n} y_j L_j(x);\ x \in R \tag{4.1}$$

tel que $(L_j)_{j=0}^{n}$ forme une base de Lagrange.

$$\begin{aligned} L_j(x) &= \frac{(x-x_0)(x-x_1)...(x-x_{j-1})(x-x_{j+1})...(x-x_n)}{(x_j-x_0)(x_j-x_1)...(x_j-x_{j-1})(x_j-x_{j+1})...(x_j-x_n)} \\ &= \prod_{i \neq j} \frac{x-x_i}{x_j-x_i} \end{aligned}$$

avec :

$$L_j(x_i) = \begin{cases} 0 & si\ \ j \neq i, \\ 1 & si\ \ j = i. \end{cases}$$

Cela signifier que le polynôme $L_j(x)$ de degré n prend la valeur 1 en x_j et s'annule à tous les autres points de collocation.

Example 4.1 *Nous considérons un nuage de points $(x_i,\ y_i);\ i = 1,...,24$, tels que x_i représente le temps en heures et y_i est le nombre de litre d'eau consommer de minuit à l'instant x_i le 25/05/2018 au restaurant universitaire. Le nuage de points représente des données échantillonnées en heures de la consommation d'eau.*
La figure 4.1 représente le polynôme d'interpolation de Lagrange d'ordre 23 par la courbe en rouge. La courbe en bleu est la courbe de charge données par les points d'interpolation.

Example 4.2 *Nous appliquons la méthode d'interpolation de Lagrange sur les données échantillonnées en heures de la consommation cumulée de l'électricité le 28/01/2020 dans un bureau des doctorants à l'IUT de Mulhouse. Dans la figure 4.2, la courbe en rouge représente un polynôme d'interpolation de Lagrange d'ordre 23 applique sur les 24 points d'interpolation. La courbe de charge de la puissance est donnée par la courbe en bleu.*

Bien que l'erreur entre le polynôme d'interpolation de Lagrange et le nuage de points soit de 0, la courbe du polynôme diverge aux bords de l'intervale de temps.

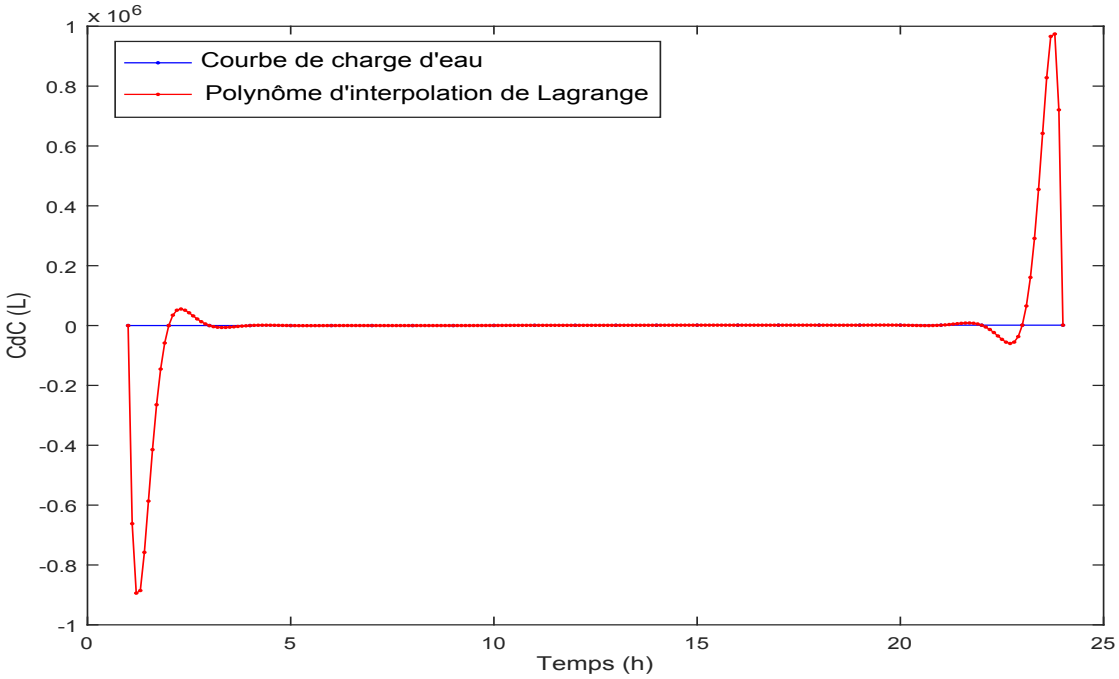

FIGURE 4.1 – La courbe de charge d'eau du 25/05/2018 estimée par le polynôme d'interpolation de Lagrange

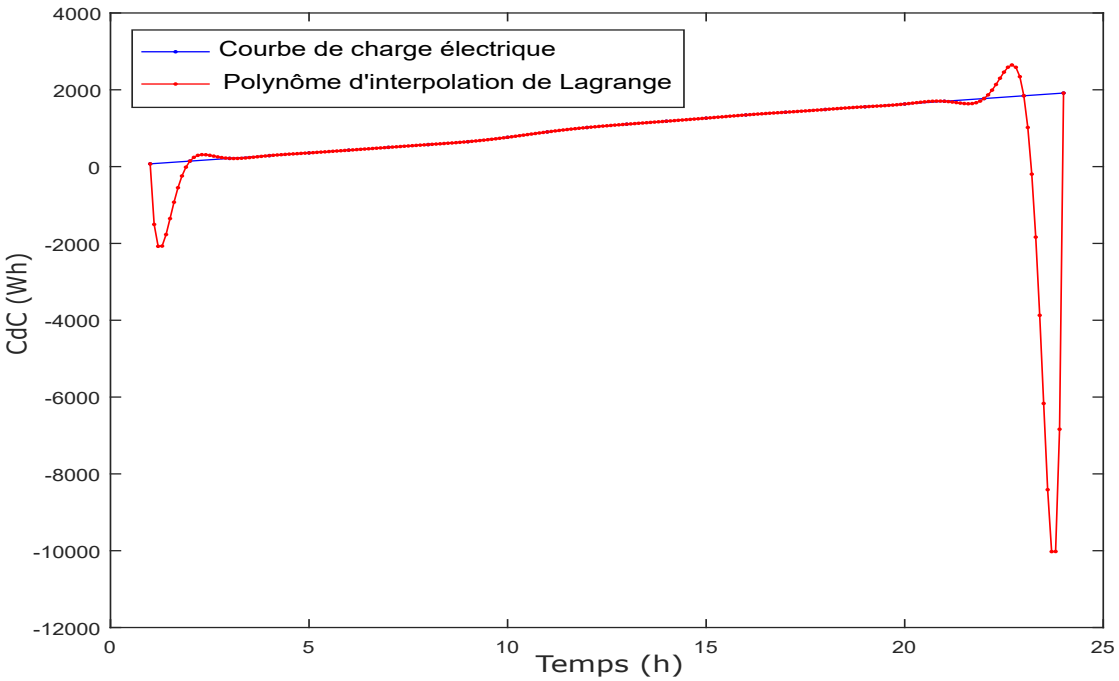

FIGURE 4.2 – L'estimation de la courbe de charge électrique du 28/01/2020 par la méthode d'interpolation de Lagrange

4.2.2 Interpolation de Tchebychev

Les polynômes de Tchebychev [28] constituent un outil important dans le domaine de l'interpolation. En effet, les racines des polynômes de Tchebychev minimise l'erreur d'interpolation de Lagrange.

Définition 4.1 *Les polynômes de Tchebychev sont des fonctions définies sur l'intervalle* $[-1,\ 1]$ *par :*

$$T_n(x) = \cos(n \arccos x)$$

Et nous pouvons les définir par la relation de récurrence suivante :

$$T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x)$$

Polynômes de Tchebychev réduits :
Nous appelons polynôme réduit (ou normalisé) de Tchebychev le polynôme définit par :

$$T_n^*(x) = \frac{1}{2^{n-1}} T_n$$

Les racines des polynômes de Tchebychev :
T_n admet $n+1$ racines simples définissent par :

$$t_k = cos(\frac{2k+1}{2n+2})\pi);\ k = 0,\ 1,\ ...,\ n.$$

Si les nœuds appartiennent à l'intervalle $[a,\ b]$ alors les racines de polynôme de Tchebychev sont généralisées par :

$$t_k = \frac{a+b}{2} + \frac{b-a}{2} \cos(\frac{2k+1}{2n+2}\pi) \tag{4.2}$$

Définition 4.2 *Le polynôme d'interpolation de Tchebychev est le polynôme d'interpolation de Lagrange qui est équivalent au polynôme d'approximation des moindres carrés avec $m = n$ appliqué sur les points de Tchebychev.*

Example 4.3 *Nous utilisons le même nuage de points de l'exemple 4.1. Puis, nous construisons les abscisses de Tchebychev en utilisant l'équation (4.2). Ensuite, nous définissons les points de Tchebychev* $(t_k,\ C_k);\ k = 1,\ ...,\ 24$ *où* C_k *est la consommation à l'instant* t_k*, ce que nous trouvons en se base sur des données horodatées. Finalement, nous interpolons les points de Tchebychev par la méthode d'interpolation de Lagrange. L'erreur quadratique moyenne entre le nuage de points et le polynôme d'interpolation de Tchebychev égale à* 449.28.
La figure 4.3 représente la méthode d'interpolation de Tchebychev appliquée sur les données de la

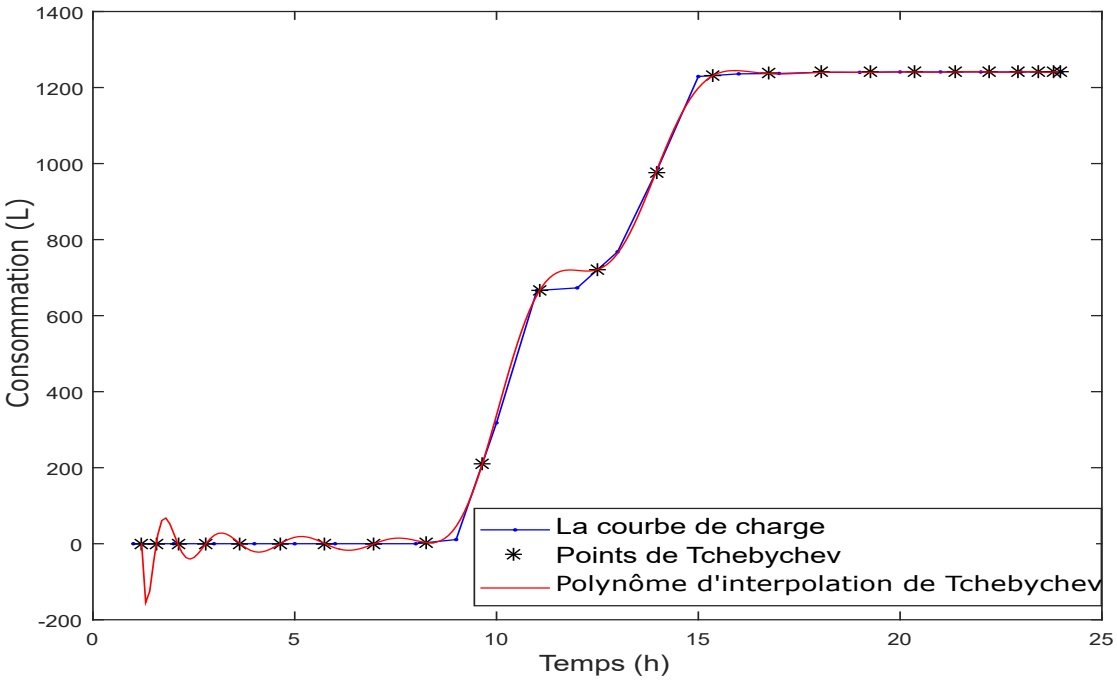

FIGURE 4.3 – Le polynôme d'interpolation de Tchebychev qui estime la courbe de charge d'eau du 25/05/2018

courbe de charge de la consommation d'eau du 25/05/2018. Les points d'interpolation sont donnés par les points en bleu et les points de Tchebychev sont définis par les étoiles en noir. Ainsi que le polynôme d'interpolation de Tchebychev d'ordre 23 *est représenté par la courbe en rouge.*

Example 4.4 *Nous représentons le polynôme d'interpolation de Tchebychev par la courbe en rouge dans la figure 4.4 qui approche les points d'interpolation définir dans l'exemple 4.2. Tous d'abord, nous calculons les points de Tchebychev qui sont représentés par les étoiles en noir. Après, nous estimons la courbe de charge électrique donnée par le polynôme d'interpolation au sens de Tchebychev d'ordre* 23 *(en bleu) avec un erreur quadratique moyenne égale à* 6.79.

La méthode d'interpolation au sens de Tchebychev corrige la divergence du polynôme d'interpolation de Lagrange aux bordx dans la courbe de charge électrique mais pas exactement au niveau de la courbe de charge de consommation d'eau.

4.2.3 Interpolation par spline

Une spline est une fonction définie par morceaux constituée d'un polynôme $s_k(x)$ de degré k, sur chaque intervalle entre nœuds $[x_{i-1},\ x_i]$. Le type de fonction spline fréquemment utilisé est la spline cubique [29].

Spline cubique

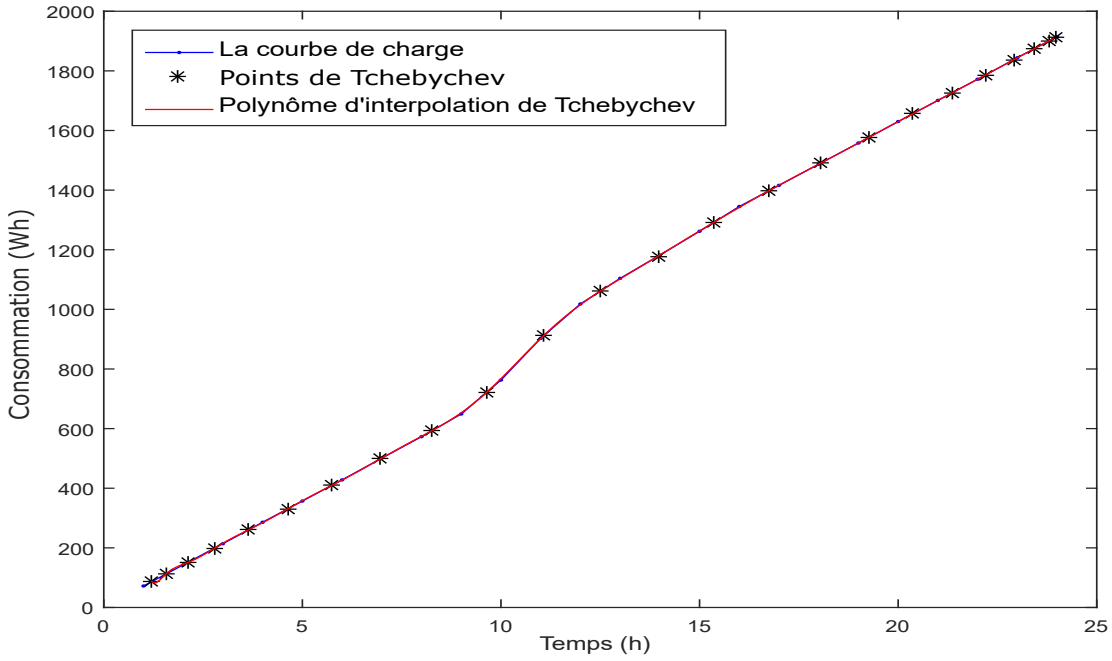

FIGURE 4.4 – L'estimation de la courbe de charge électrique du 28/01/2020 par le polynôme d'interpolation de Tchebychev

La fonction de spline cubique $s_3(x)$ est continue, ainsi que ses dérivées premières et secondes $s_3'(x),\ s_3''(x)$.
Sur chaque intervalle $[x_{i-1};\ x_i]$, la fonction de spline cubique $s_3(x)$ est un polynôme $p_i(x)$ de degré 3, tels que : $\forall i = 1,\ ...,\ n$

$$\begin{cases} p_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i;\ \forall i = 1,\ ...,\ n. \\ p_i(x_i) = y_i;\ \forall i = 1,\ ...,\ n. \\ p_i(x_{i-1}) = y_{i-1};\ \forall i = 1,\ ...,\ n. \\ p_i'(x_i) = p_{i+1}'(x_i);\ \forall i = 1,\ ...,\ n-1. \\ p_i''(x_i) = p_{i+1}''(x_i);\ \forall i = 1,\ ...,\ n-1. \end{cases} \tag{4.3}$$

Le polynôme d'interpolation locale en splines cubiques [29] est défini par l'équation (4.4).

$$\begin{aligned} p_i(x) = {} & \frac{(x_i - x)^3 m_{i-1} + (x - x_{i-1})^3 m_i}{6h_i} + \frac{(x_i - x) y_{i-1} + (x - x_{i-1}) y_i}{h_i} - \\ & \frac{h_i}{6}[(x_i - x) m_{i-1} + (x - x_{i-1}) m_i];\ \forall x \in [x_{i-1},\ x_i],\ \forall i = 1,\ ...,\ n \end{aligned} \tag{4.4}$$

tels que $h_i = x_i - x_{i-1}$ et m_i sont des solutions du système suivant :
$\frac{h_{i-1}}{6} m_{i-1} + \frac{h_{i-1} + h_i}{3} m_i + \frac{h_{i+1}}{6} m_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}, \forall i = 1, ..., n-1.$
Pour résoudre le système, nous ajoutons deux conditions supplémentaires :
$p_1''(x_0) = 0,\ p_n''(x_n) = 0$; correspondant aux splines cubiques naturelles [30].

Pour trouver l'équation (4.4), nous posons $p_i''(x_i) = m_i$ et $p_i''(x_{i-1}) = m_{i-1}$.
Nous avons le polynôme local de splines cubiques de degré 3 alors la dérivé seconde est de degré 1, d'ou $p"_i(x) = m_{i-1}\frac{x_i - x}{h_i} + m_i\frac{x - x_{i-1}}{h_i}$.
En intégrant p_i'' deux fois il vient :
$p_i(x) = \frac{(x_i - x)^3 m_{i-1} + (x - x_{i-1})^3 m_i}{6h_i} + r_i x + l_i$.
r_i, et l_i sont des constant d'intégration, nous les déterminons à l'aide des conditions
$p_i(x_i) = y_i,\ p_i(x_{i-1}) = y_{i-1}$, nous trouvons
$r_i = \frac{y_i - y_{i-1}}{h_i} - \frac{h_i}{6}[m_i - m_{i-1}],\ l_i = \frac{y_{i-1}x_i - y_i x_{i-1}}{h_i} + \frac{h_i}{6}[m_i x_{i-1} - m_{i-1}x_i]$.
Après quelques manipulation algébrique nous obtenons l'équation (4.4). Pour obtenir les valeurs $m_i;\ \forall i = 0,\ ...,\ n$, nous utilisons les conditions de la continuité des dérivées premières $p_i'(x_i) = p_{i+1}'(x_i);\ \forall i = 1,\ ...,\ n-1$, avec $m_0 = m_n = 0$.
Nous avons $p_i'(x) = \frac{h_i^2 - 3(x_i - x)^2}{6h_i} m_{i-1} + \frac{3(x - x_{i-1})^2 - h_i^2}{6h_i} m_i + \frac{y_i - y_{i-1}}{h_i}$.
Alors $\frac{h_{i-1}}{6} m_{i-1} + \frac{h_{i-1} + h_i}{3} m_i + \frac{h_{i+1}}{6} m_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}},\ \forall i = 1,\ ...,\ n-1$.
Il ne reste d'impose les conditions aux limites de splines cubiques naturelles pour résoudre le système. L'écriture matricielle de ce système est donnée par $Am = b$ tels que :
$m = [m_1,\ ..., m_{n-1}],\ b = [b_1, ..., b_{n-1}];\ b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}$ et

$$A = \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & & \dots & 0 \\ \frac{h_2}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & & & : \\ 0 & & \ddots & & & \\ : & & & \frac{h_{n-2}}{6} & \frac{h_{n-2}+h_{n-1}}{3} & \frac{h_{n-1}}{6} \\ 0 & \dots & & 0 & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} \end{pmatrix}.$$

Example 4.5 *L'application de la méthode d'interpolation spline sur les données échantillonnées en heures de la consommation d'eau du 25/05/2018 consiste à définir pour chaque intervalle de temps qui duré une heure un polynôme de degré* 3. *Nous trouvons* 23 *polynômes de degré* 3. *Ils sont représentés par la courbe rouge de la figure 4.5. Les points en bleu sont les points d'interpolation.*

Example 4.6 *La figure 4.6 montre les résultats de l'interpolation par les splines cubiques appliquée sur les données échantillonnées en heures de la consommation d'électricité du 28/01/2020. La courbe en rouge représente la fonction définie par morceaux avec l'interpolation de spline cubique appliquer sur les points d'interpolation en bleu.*

L'interpolation avec les splines cubiques fournit une erreur nulle mais la courbe de la fonction spline qui estime la courbe de charge de la consommation d'eau diminue de temps en temps tandis que la courbe de charge est cumulative donc elle est croissante.

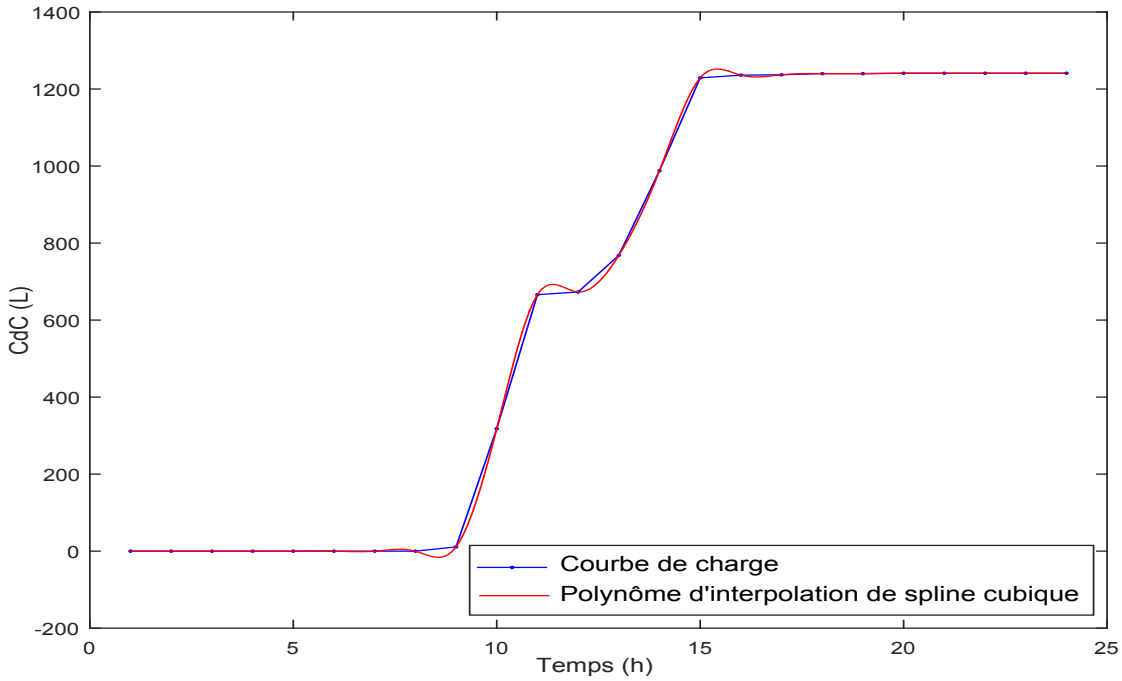

FIGURE 4.5 – L'interpolation par les splines cubiques appliquer sur les consommations d'eau du 25/05/2018

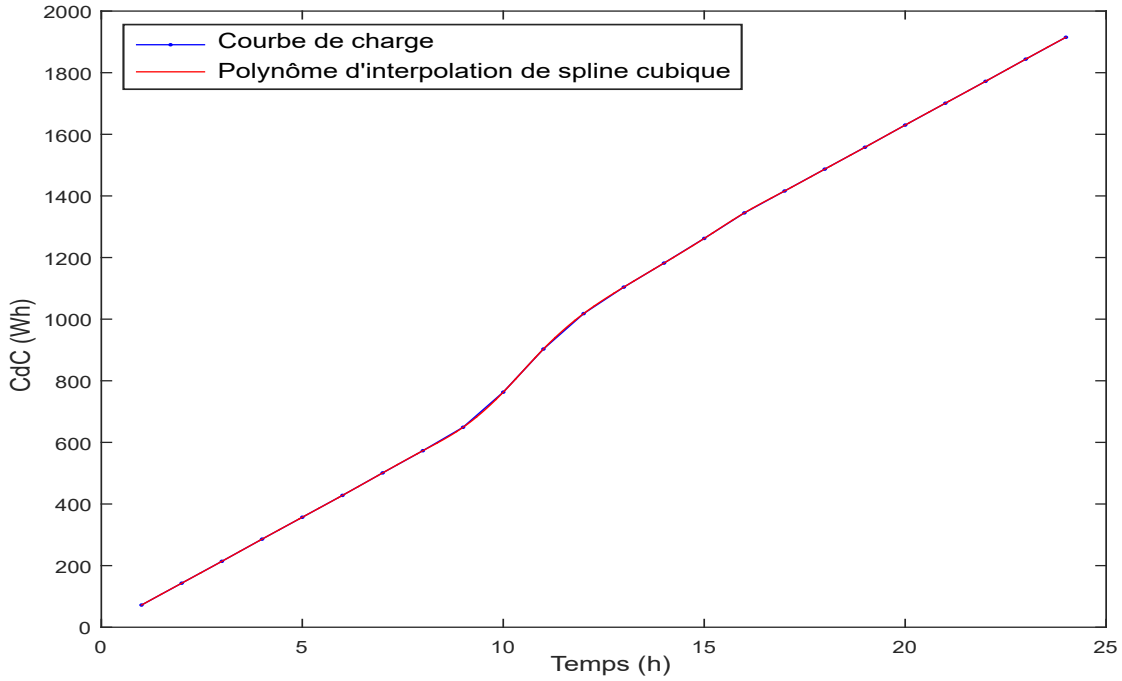

FIGURE 4.6 – L'estimation de la courbe de charge de la puissance électrique du 28/01/2020 par l'interpolation de spline cubique

4.2.4 Méthode des moindres carrés

Nous considérons un nuage de points $M_i = (x_i,\ y_i),\ i = 1,\ ...,\ n$. Soient $B = \{\phi_1,\ ...,\ \phi_m\}$ une base des fonction ϕ_k donnée et $a = (a_1,\ ...,\ a_m)$; m paramètres inconnus tel que $m \leq n$.
La méthode des moindres carrés [31] consiste à donner une approche par la fonction (4.5).

$$f(x, a) = \sum_{k=1}^{m} a_k \phi_k(x) \tag{4.5}$$

C'est à dire, elle consiste à minimiser la somme des carrés des erreurs $\varepsilon_i = (y_i - f(x_i,\ a)),\ \forall i = 1,\ ...,\ n$;

$$\min_a \sum_{i=1}^{n} \varepsilon_i^2 = \min_{a_k} \sum_{i=1}^{n} (y_i - \sum_{k=1}^{m} a_k \phi_k(x_i))^2$$

Pour trouver le minimum, nous dérivons par rapport aux paramètres et nous résolvons ce système :

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial a_k} = 0;\ \forall k = 1,\ ...,\ m$$

Nous avons : $\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial a_k} = -2 \sum_{i=1}^{n} [y_i - f(x_i,\ a)] \frac{\partial f(x_i,\ a)}{\partial a_k}$
Nous posons : $\frac{\partial f(x_i,\ a)}{\partial a_k} = X_{ik}$
Alors : $\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial a_k} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} [y_i - f(x_i, a)] X_{ik} = 0$
D'où : $\sum_{i=1}^{n} f(x_i,\ a) X_{ik} = \sum_{i=1}^{n} y_i X_{ik};\ \forall k = 1,\ ...,\ m.$
Donc :

$$\begin{cases} f(x_1, a)X_{11} + ... + f(x_n, a)X_{n1} & = & y_1 X_{11} + ... + y_n X_{n1} \\ . \\ . \\ . \\ f(x_1, a)X_{1m} + ... + f(x_n, a)X_{nm} & = & y_1 X_{1m} + ... + y_n X_{nm} \end{cases}$$

Alors : $\begin{pmatrix} X_{11} & . & . & . & X_{n1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{1m} & . & . & . & X_{nm} \end{pmatrix} \times \begin{pmatrix} f(x_1, a) \\ . \\ . \\ . \\ f(x_n, a) \end{pmatrix} = \begin{pmatrix} X_{11} & . & . & . & X_{n1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{1m} & . & . & . & X_{nm} \end{pmatrix} \times \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}$

Nous posons : $M = \begin{pmatrix} X_{11} & . & . & . & X_{1m} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{n1} & . & . & . & X_{nm} \end{pmatrix} \quad et \quad Y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}$

Alors : $M^T \times \begin{pmatrix} f(x_1,\ a) \\ . \\ . \\ . \\ f(x_n,\ a) \end{pmatrix} = M^T \times Y$

Donc : $M^T \times \begin{pmatrix} \phi_1(x_1) & . & . & . & \phi_m(x_1) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \phi_1(x_n) & . & . & . & \phi_m(x_n) \end{pmatrix} \times \begin{pmatrix} a_1 \\ . \\ . \\ . \\ a_m \end{pmatrix} = M^T \times Y$

Et nous avons $X_{ik} = \dfrac{\partial f(x_i,\ a)}{\partial a_k} = \phi_k(x_i)$

Alors : $M^T \times \begin{pmatrix} X_{11} & . & . & . & X_{1m} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{n1} & . & . & . & X_{nm} \end{pmatrix} \times \begin{pmatrix} a_1 \\ . \\ . \\ . \\ a_m \end{pmatrix} = M^T \times Y$

Donc : $M^T \times M \times a = M^T \times Y$
D'où : $a = inv(M^T \times M) \times M^T \times Y$
On commence par le modèle le plus simple : la droite, aprés on continue avec des polynômes.

La droite des moindres carrés

L'approximation par la méthode des moindres carrés linéaires consiste à trouver les deux coefficients a et b de la droite qui minimise la somme des carrés des erreurs entre les points d'approximation et les points estimés par la méthode. La droite des moindres carrés définit par : $f(x) = ax + b$, tels que :
$a = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ et $b = \overline{y} - a\overline{x}$, avec $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ et $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$.

Example 4.7 *La droite de régression représentée par la droite en rouge dans la figure 4.7 est définie par :* $f(x) = 75.21\ x - 281.31$, *avec une erreur au moyenne quadratique égale à* 202.77.
La figure 4.7 représente la régression linéaire par la méthode des moindres carrés appliquer sur les données échantillonnées en heures de la consommation d'eau du 25/05/2020.

Example 4.8 *L'approximation par la droite de régression linéaire est données par la fonction* $f(x) = 83.6\ x - 39.85$ *et elle est représenté par la ligne rouge dans la figure 4.8. Cette figure représente les résultats de la méthode des moindres carrés linéaire sur les données échantillonnées en heures de la consommation d'électricité du 28/01/2020. L'erreur au moyenne quadratique de la régression linéaire égale à* 37.54

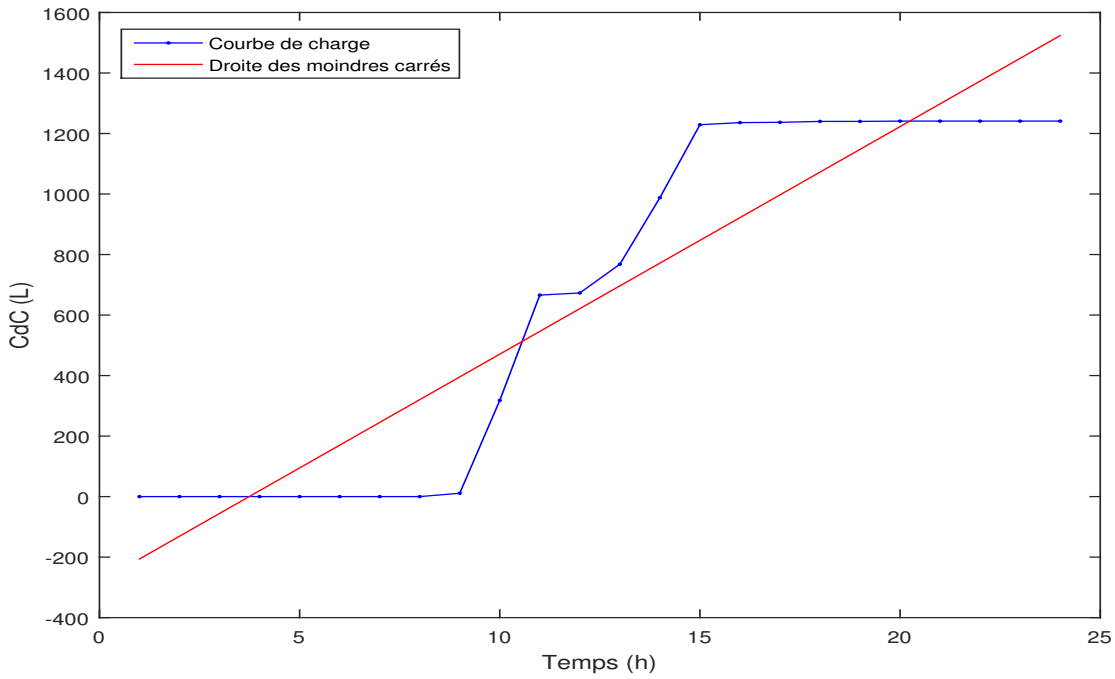

FIGURE 4.7 – La régression linéaire de la consommation d'eau par la méthode des moindres carrés

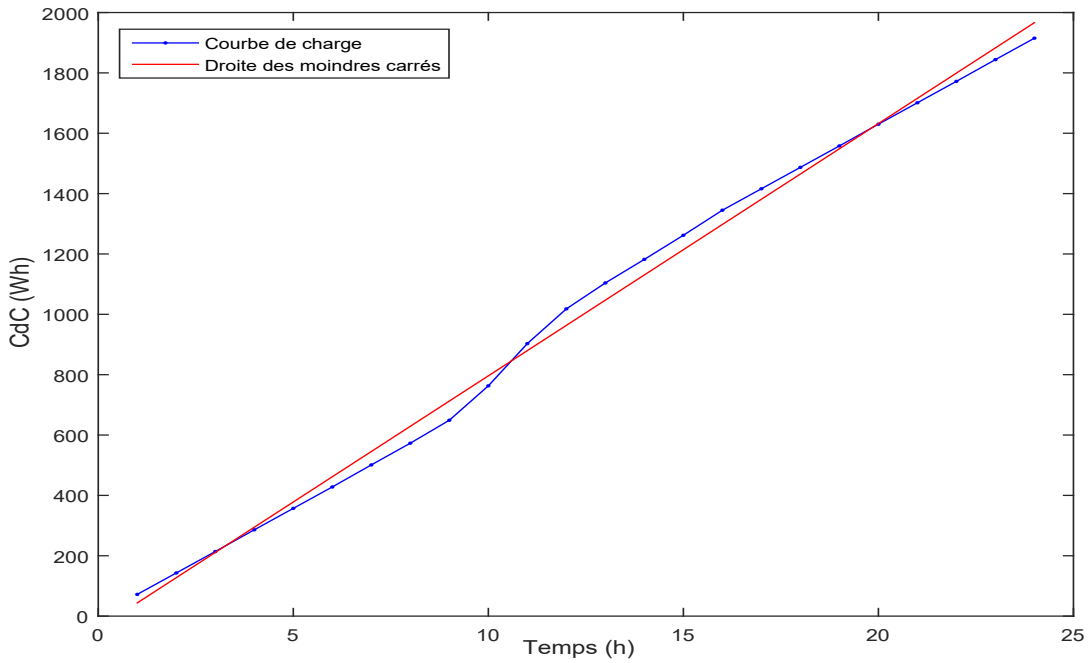

FIGURE 4.8 – L'approximation de la puissance électrique le 28/01/2020 par la droite des moindres carrés

Moindres carrés pour polynômes

Nous considérons $B = \{1,\ x,\ x^2,\ ...,\ x^m\}$ une base polynomiale.
Soient $(x_i,\ y_i);\ i = 1,\ ...,\ n$ un nuage de points tel que $m < n - 1$.
$f(x,\ a) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m = y;\ a_i;\ i = 0,\ ...,\ m$ sont des paramètres inconnus.
Multiplions successivement la relation précédente par $x, x^2, ..., x^m$, nous obtenons le système d'équations suivant :

$$\begin{cases} a_0 + a_1 x + a_2 x^2 + & ... & + a_m x^m = y \\ a_0 x + a_1 x^2 + a_2 x^3 + & ... & + a_m x^{m+1} = yx \\ . & & \\ . & & \\ . & & \\ a_0 x^m + a_1 x^{m+2} + a_2 x^{m+3} + & ... & + a_m x^{2m} = y x^m \end{cases}$$

Nous écrivons les relations du système évaluées sur tous les points $(x_i,\ y_i);\ i = 1,\ ...,\ n$. Puis, nous additionnons toutes les équations du même ordre.
Nous obtenons :

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & . & . & . & \sum_{i=1}^n x_i^m \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & . & . & . & \sum_{i=1}^n x_i^{m+1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sum_{i=1}^n x_i^m & \sum_{i=1}^n x_i^{m+1} & . & . & . & \sum_{i=1}^n x_i^{2m} \end{pmatrix} \times \begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ a_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \\ . \\ . \\ \sum_{i=1}^n y_i x_i^m \end{pmatrix}$$

Enfin, les valeurs $a_i;\ i = 1,\ ...,\ n$ s'obtiennent par la résolution de l'équation matricielle suivante :

$$\begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ a_m \end{pmatrix} = inv \begin{pmatrix} n & \sum_{i=1}^n x_i & . & . & . & \sum_{i=1}^n x_i^m \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & . & . & . & \sum_{i=1}^n x_i^{m+1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sum_{i=1}^n x_i^m & \sum_{i=1}^n x_i^{m+1} & . & . & . & \sum_{i=1}^n x_i^{2m} \end{pmatrix} \times \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \\ . \\ . \\ \sum_{i=1}^n y_i x_i^m \end{pmatrix}$$

Example 4.9 *Les points de l'approximation sont les données échantillonnées en heures de la consommation d'eau durant un jour. Les coefficients polynomiaux calculés selon la méthode des moindres carrés sont donnés par le vecteur a dans (4.6).*

$$a = (-1.1 \times 10^7,\ 3.93 \times 10^7,\ -6.1 \times 10^7,\ 5.59 \times 10^7, -3.43 \times 10^7,\ 1.51 \times 10^7,\ -4.99 \times 10^6,\ 1.27 \times 10^6,\ -2.57 \times 10^5,\ 4.14 \times 10^4,\ -5.4 \times 10^3,\ 5.74 \times 10^2,\ -49.89,\ 3.55,\ -0.21,\ 9.69 \times 10^{-3},\ -3.67 \times 10^{-4},\ 1.1 \times 10^{-5},\ -2.55 \times 10^{-7},\ 4.40 \times 10^{-9},\ -5.33 \times 10^{-11},\ 4.04 \times 10^{-13},\ -1.44 \times 10^{-15}). \tag{4.6}$$

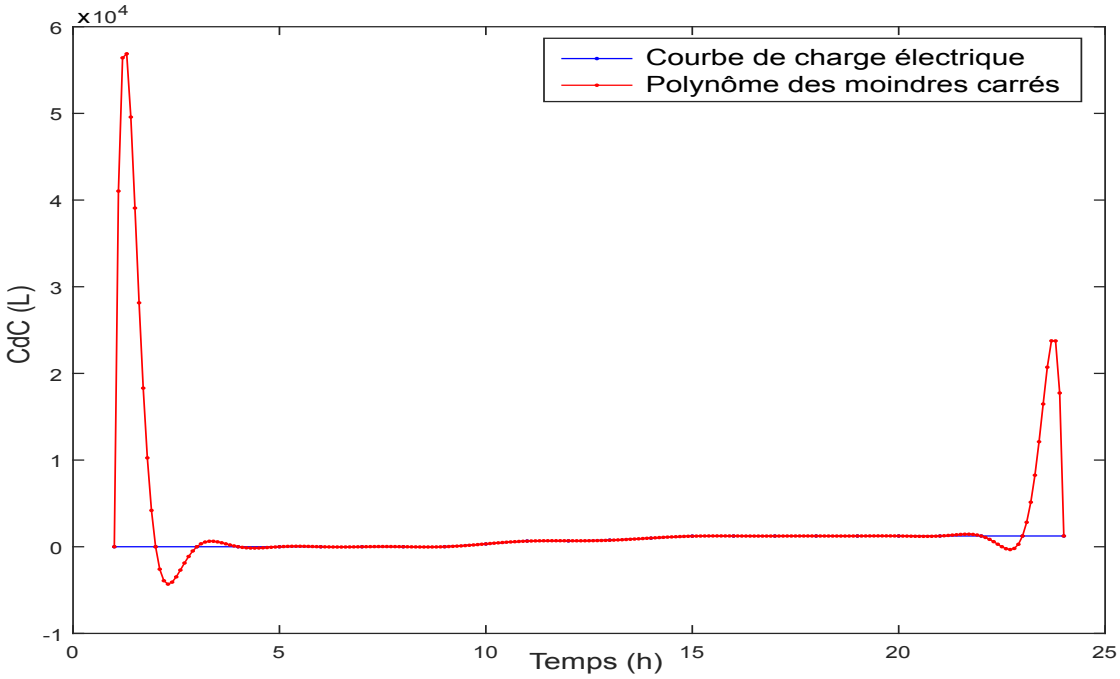

FIGURE 4.9 – L'approximation des données de la consommation d'eau le 25/05/2018 par la méthode des moindres carrés

La figure 4.9 reflète le polynôme d'ordre 22 *estimé par la méthode des moindres carrés sur les points de la courbe de charge du 25 mai 2018 avec un erreur au moyenne quadratique égale à* 8.15.

Example 4.10 *L'équation polynomiale d'ordre 10 de la régression des moindres carrés est donnée par l'équation (4.7) :*

$$f(t) = -149.25 + 413.37t - 288.48t^2 + 121.31t^3 - 28.8t^4 + 4.11t^5 - 0.37t^6 + 0.02t^7 - 6.84 \times 10^{-4}t^8 + 1,28 \times 10^{-5}t^9 - 1.03 \times 10^{-7}t^{10}. \quad (4.7)$$

La figure 4.10 représente le polynôme d'ordre 10 *estimé par la méthode des moindres carrés sur les points d'approximation de la puissance le 28 janvier 2020. L'erreur au moyenne quadratique est* 6.68.

Le polynome des moindres carrés estime la CdC mieux que la droite des moindres carrés. L'estimation de la CdC de la consommation d'eau par la droite des moindres carrés donne une grande erreur due aux pics et les variations de la CdC. La courbe du polynôme des moindres carrés qui estime la CdC de la consommation d'eau diverge aux extrémités.

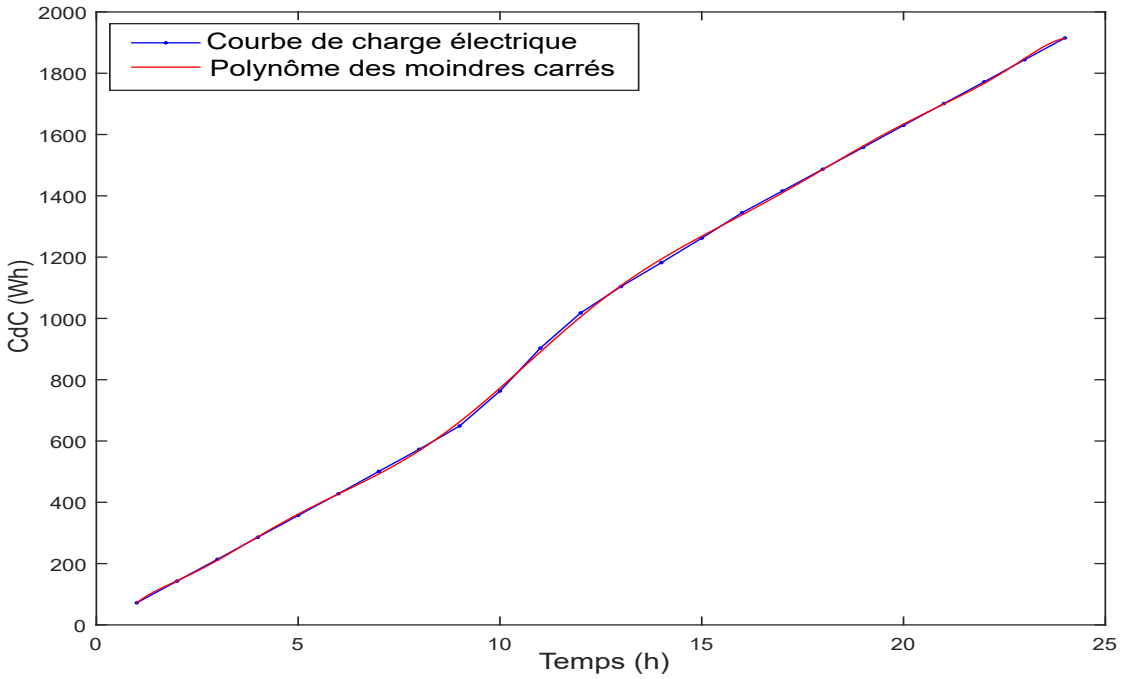

FIGURE 4.10 – L'approximation par les moindres carrés des données de la consommation d'électricité le 28/01/2020.

4.2.5 Courbes de Bézier

À une allure donnée, nous associons une courbe paramétrique obtenue par des interpolations linéaires successives.
Les courbes de Bézier [32] sont des courbes paramétriques de n points $p_i = (x_i,\ y_i),\ i = 0,\ 1,\ ...,\ n-1$, qui passent par le premier et le dernier points. Nous définissons la courbe de Bézier associée à ces points par :

$$p(t) = \sum_{i=0}^{n-1} B_i^{n-1}(t) \times p_i = \begin{cases} \sum_{i=0}^{n-1} B_i^{n-1}(t) \times x_i. \\ \sum_{i=0}^{n-1} B_i^{n-1}(t) \times y_i. \end{cases} \tag{4.8}$$

Nous demandons que les coefficients $B_i^{n-1}(t)$ soient positifs, de somme 1 et qu'ils dépendent de t de manière la plus régulière possible. La solution adoptée par Bézier consiste à prendre les polynômes de Bernstein qui sont donnés par :

$$B_i^n(t) = C_i^n t^i (1-t)^{n-i} = \frac{n!}{i!(n-i)!} t^i (1-t)^{n-i};\ t \in [0,1].$$

Propriétés des polynômes de Bernstein :
Les polynômes de Bernstein vérifient ces propriétés :

* $B_i^n(0) = 0$ pour $i > 0$.
* $B_0^n(0) = 1$.
* $B_i^n(1) = 0$ pour $i < n$.

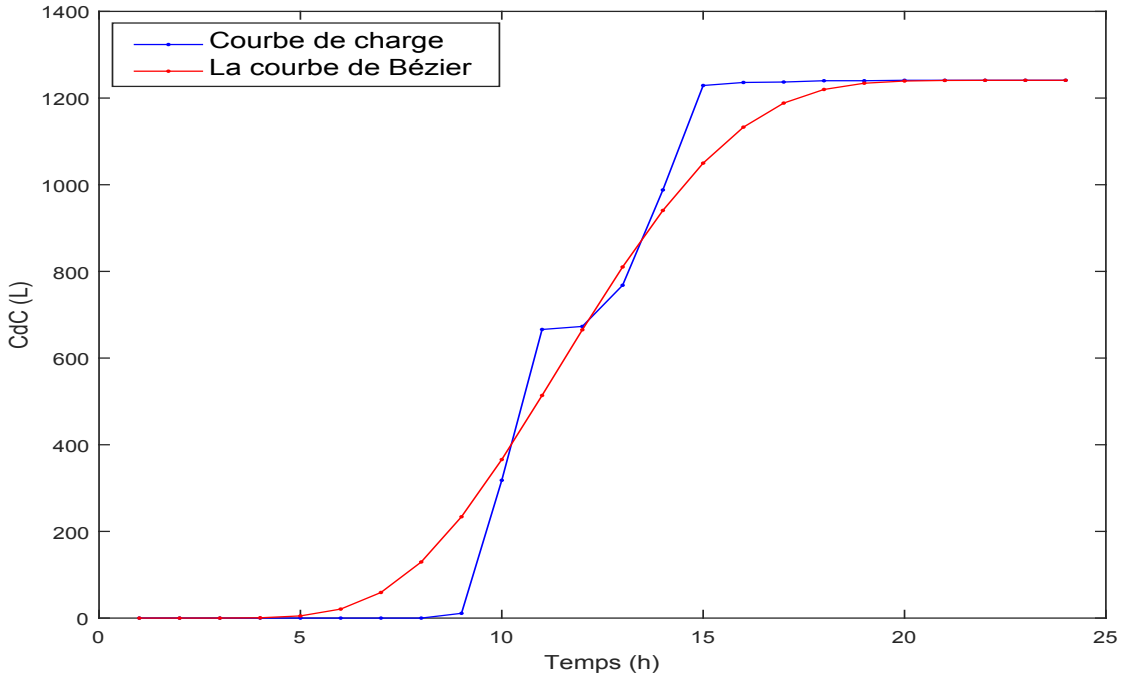

FIGURE 4.11 – L'estimation de la courbe de charge d'eau par la courbe de Bézier

* $B_n^n(1) = 1$.

Example 4.11 *La courbe de Bézier qui estime la CdC de la consommation d'eau avec des données échantillonnées en heures est définie par :*

$$p(t) = \begin{cases} g(t) = & 23t + 1. \\ h(t) = & t^8 10^7(-1.4t^{15} + 128t^{14} - 657t^{13} + 2019t^{12} - 3996t^{11} + 4905t^{10} - 2508t^9 - \\ & 3169t^8 + 8656t^7 - 10184t^6 + 7577t^5 - 3767t^4 + 1219t^3 - 230t^2 + 17t + 0.5). \end{cases}$$

Cette courbe est représentée en rouge dans la figure 4.11 et elle estime la courbe de charge de la consommation d'eau en bleu.

Example 4.12 *La courbe de Bézier appliquée aux données de la consommation électrique est*

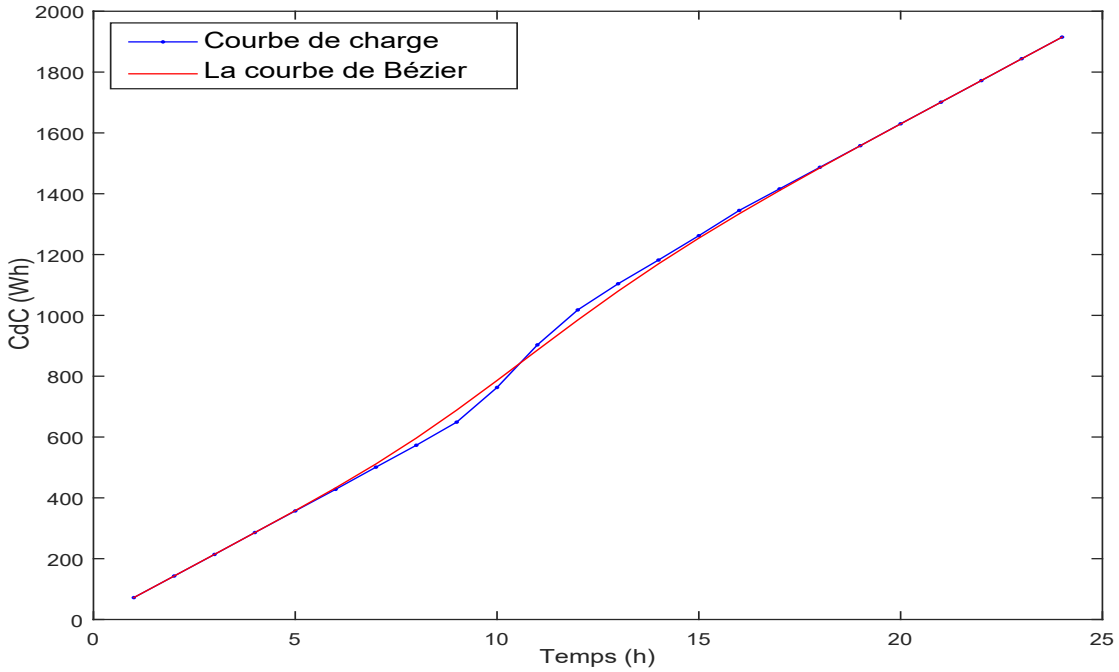

FIGURE 4.12 – L'approximation de la courbe de charge électrique par la courbe de Bézier

définie par :

$$p(t) = \begin{cases} g(t) = & 7084t^3(t-1)^{20} - 759t^2(t-1)^{21} - (t-1)^{23} - 44275t^4(t-1)^{19} + 201894t^5 \\ & (t-1)^{18} - 706629t^6(t-1)^{17} + 1961256t^7(t-1)^{16} - 4412826t^8(t-1)^{15} + \\ & 8171900t^9(t-1)^{14} - 12584726t^{10}(t-1)^{13} + 16224936t^{11}(t-1)^{12} - \\ & 17577014t^{12}(t-1)^{11} + 16016924t^{13}(t-1)^{10} - 12257850t^{14}(t-1)^9 + \\ & 7845024t^{15}(t-1)^8 - 4167669t^{16}(t-1)^7 + 1817046t^{17}(t-1)^6 - 639331t^{18} \\ & (t-1)^5 + 177100t^{19}(t-1)^4 - 37191t^{20}(t-1)^3 + 5566t^{21}(t-1)^2 + 24t^{23} + \\ & 46t(t-1)^{22} - 529t^{22}(t-1). \\ h(t) = & 506506t^3(t-1)^{20} - 54142t^2(t-1)^{21} - 72(t-1)^{23} - 3161235t^4(t-1)^{19} + \\ & 14401772t^5(t-1)^{18} - 50574447t^6(t-1)^{17} + 140474961t^7(t-1)^{16} - \\ & 318213786t^8(t-1)^{15} + 623515970t^9(t-1)^{14} - 1033091598t^{10}(t-1)^{13} + \\ & 1376415404t^{11}(t-1)^{12} - 1492694112t^{12}(t-1)^{11} + 1352286012t^{13}(t-1)^{10} \\ & -1031293780t^{14}(t-1)^9 + 659472330t^{15}(t-1)^8 - 347142312t^{16}(t-1)^7 + \\ & 150108189t^{17}(t-1)^6 - 52425142t^{18}(t-1)^5 + 14433650t^{19}(t-1)^4 - 3012471 \\ & t^{20}(t-1)^3 + 448316t^{21}(t-1)^2 + 1915t^{23} + 3289t(t-1)^{22} - 42412t^{22}(t-1). \end{cases}$$

La figure 4.12 représente la courbe de Bézier qui estime la courbe de charge électrique du 28/01/2020 dans un bureau des doctorants.

4.2.6 Estimation de la densité de probabilité

La distribution des données peut être décrite par une fonction de densité de probabilité. Cette densité est la meilleure façon de comprendre les données et elle aide à construire un modèle qui définit le processus qui génère les données. La densité peut être estimée par des méthodes paramétriques comme le modèle de mélange gaussien ou des méthodes non paramétriques telles que l'estimation par noyau [33].

Modèle de mélange gaussien

Les modèles de mélange gaussien (en anglais Gaussian Mixture Models, GMM) sont apparus dans les travaux de Pearson. Ils sont utilisés avec succès dans de nombreuses disciplines et ils constituent un outil commun de classification des données. Ils permettent de modéliser des ensembles de données numériques.
Le modèle de mélange gaussien est un modèle statistique qui s'exprime en fonction de la densité du mélange gaussien. Il sert à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme somme de plusieurs gaussiennes, appelées noyaux [34].

Définition 4.3 *Le modèle de mélange gaussien est défini par :*

$$g(t) = \sum_{i=1}^{k} p_i f_i(t);\ f_i(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(t-m_i)^2}{2\sigma_i^2})$$

avec : p_i *la proportion de la ième gaussienne telle que* $0 < p_i < 1$, $\sum_{i=1}^{k} p_i = 1$,
f_i *la densité du ième gaussienne de moyenne* m_i *et d'écart type* σ_i,
k *le nombre de gaussiennes tel que* k *soit inférieur au nombre de points de l'échantillon d'une variable aléatoire.*
Donc, il s'agit de déterminer la variance, la moyenne et la proportion de chaque gaussienne. Ces paramètres sont optimisés selon le critère de maximum de vraisemblance. Cette optimisation est effectuée en utilisant la procédure itérative appelée Espérance-Maximisation (EM).

Définition 4.4 *La fonction de vraisemblance est donnée par :*

$$L(X,\theta) = \prod_{j=1}^{n} \sum_{i=1}^{k} p_i f_i(x_j)$$

$\theta = \{p_i,\ m_i,\ \sigma_i;\ i = 1,...,k\}$ *sont des paramètres inconnus du modèle.*
Il est plus facile de maximiser la fonction log-vraisemblance au lieu de la fonction de vraisemblance et la valeur qui maximise $\ln(L(X,\theta))$ *est la même que celle qui maximise* $L(X,\theta)$.
La fonction log-vraisemblance s'écrit :

$$l(X,\theta) = \ln L(X,\theta) = \sum_{j=1}^{n} \ln \sum_{i=1}^{k} p_i f_i(x_j)$$

Estimation des paramètres du modèle de mélange gaussien
L'algorithme EM est une technique itérative, permettant de maximiser la vraisemblance des paramètres de modèles probabilistes. Il fonctionne en deux étapes :

- Étape E, la phase expectation : elle consiste à calculer l'espérance conditionnelle de la fonction de vraisemblance ; $Q(\theta, \theta^{(0)}) = E[L(X, \theta) \mid \theta^{(0)}]$ où $\theta^{(0)}$ est l'initialisation des paramètres du modèle.
- Étape M, la phase maximisation : procède à la maximisation de la fonction Q.

Définition 4.5 *L'estimation de la CdC par le modèle de mélange gaussien est donnée par sa fonction de répartition :*

$$G(t) = \int_{-\infty}^{t} g(y)dy$$

Example 4.13 *La distribution des données de la consommation d'eau du 25/05/2018 normalisés et échantillonnées en minute définie par un mélange de deux gaussiennes* $N(594, 32.5)$ *et* $N(827, 46.32)$ *qui sont représentées dans la figure 4.13(1).*
La fonction de répartition de la densité de mélange gaussien représentée par la courbe en rouge dans la figure 4.13(2) est définie par :

$$G(t) = \int_{-\infty}^{t} g(x)dx = \int_{-\infty}^{t} (0.54 f_1(x) + 0.46 f_2(x))dx;$$

telles que $f_1 \hookrightarrow N(594, 32.5)$, $f_2 \hookrightarrow N(827, 46.32)$.
Cette densité estime la courbe de charge cumulée (en bleu) de la consommation d'eau pondant la journée au restaurant universitaire.
Pendant la nuit, il n'y a pas de consommation d'eau dans le restaurant universitaire, donc l'estimation de la CdC normalisée représentée dans la figure 4.14 est donnée par :

$$\begin{cases} 0 & Si\ 0 < t < 515 \\ G(t) & Si\ 516 < t < 1171 \\ 1 & Si\ 1172 < t < 1440 \end{cases}$$

La figure 4.15 représente le modèle de mélange gaussien (la courbe en rouge) qui estime la CdC de la consommation d'eau au restaurant universitaire le 25/05/2018 (la courbe en bleu).

Estimation par noyau

L'estimation par noyau (kernel density estimation en anglais ou KDE) est une méthode non-paramétrique pour estimer la densité de probabilité d'une variable aléatoire. Elle est aussi appelée méthode de Parzen-Rosenblatt [35].

Définition 4.6 *Soit* $x_1, x_2, ..., x_n$ *un échantillon d'une variable aléatoire ; l'estimation par noyau*

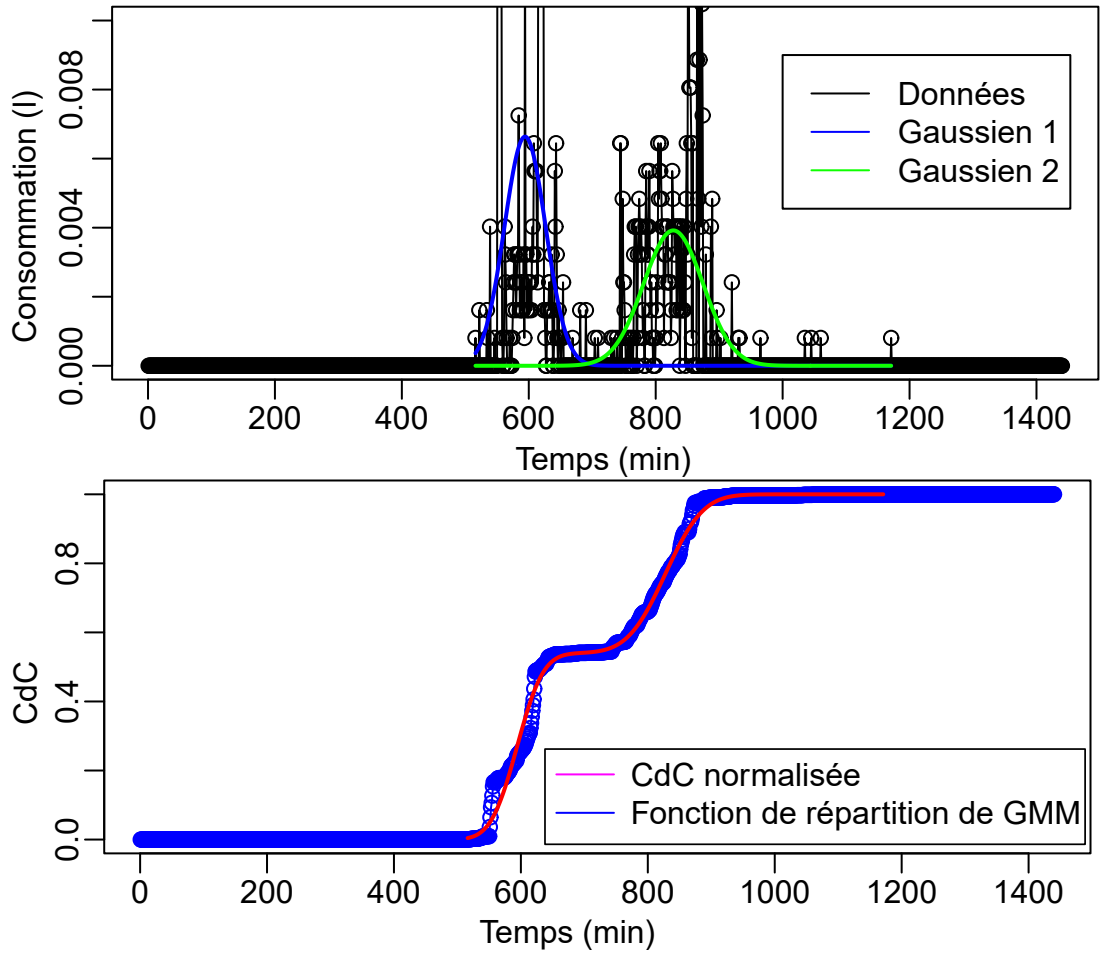

FIGURE 4.13 – L'estimation paramétrique de la densité des données de la consommation d'eau le 25/05/2018

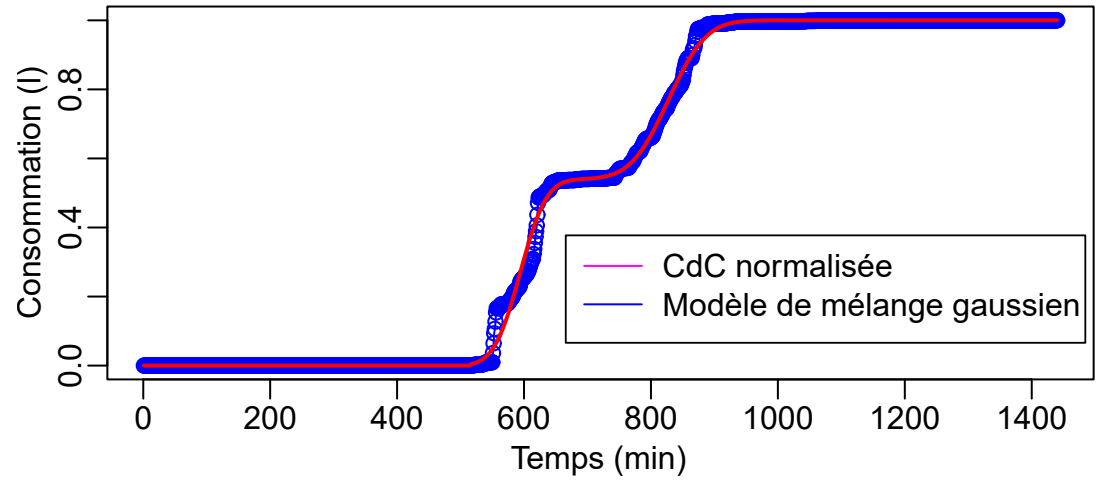

FIGURE 4.14 – Le modèle de mélange gaussien d'une CdC d'eau normalisé

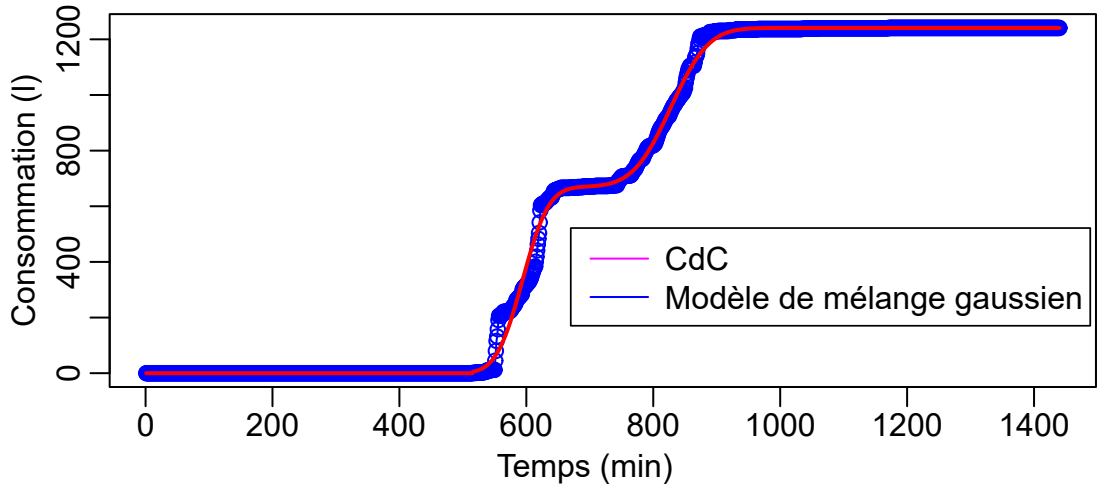

FIGURE 4.15 – Le modèle de mélange gaussien de la CdC d'eau du 25/05/2018

consiste en une somme de n fonctions, une pour chaque observation. Elle est définie par :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{4.9}$$

K est un noyau souvent choisi comme la densité de la fonction gaussienne standard,
$K(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$
h paramètre de lissage (fenêtre).

Pour estimer la densité de probabilité d'une courbe de charge, on utilise l'estimation par noyau cumulative qui est définie par :

$$F_h(x) = \int_{-\infty}^{x} f_h(y)dy = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-x_i}{h}} K(y)dy \tag{4.10}$$

Dans le restaurant universitaire, il n'y a pas de consommation pendant la nuit, donc cette période est séparée lors de l'utilisation des données échantillonnées. Donc le modèle qui estime la CdC normalisée des données d'eau échantillonnée en minute est défini par :

$$G(t) = \begin{cases} 0 & Si\ 0 < t < t_1 \\ F_h(t) & Si\ t_1 < t < t_2 \\ 1 & Si\ t_2 < t < 1440 \end{cases} \tag{4.11}$$

Example 4.14 *L'estimation de la CdC normalisée représentée dans la figure 4.16 est donnée par :*

$$\begin{cases} 0 & Si\ 0 < t < 515 \\ F_{0.0003}(t) & Si\ 516 < t < 1171 \\ 1 & Si\ 1172 < t < 1440 \end{cases}$$

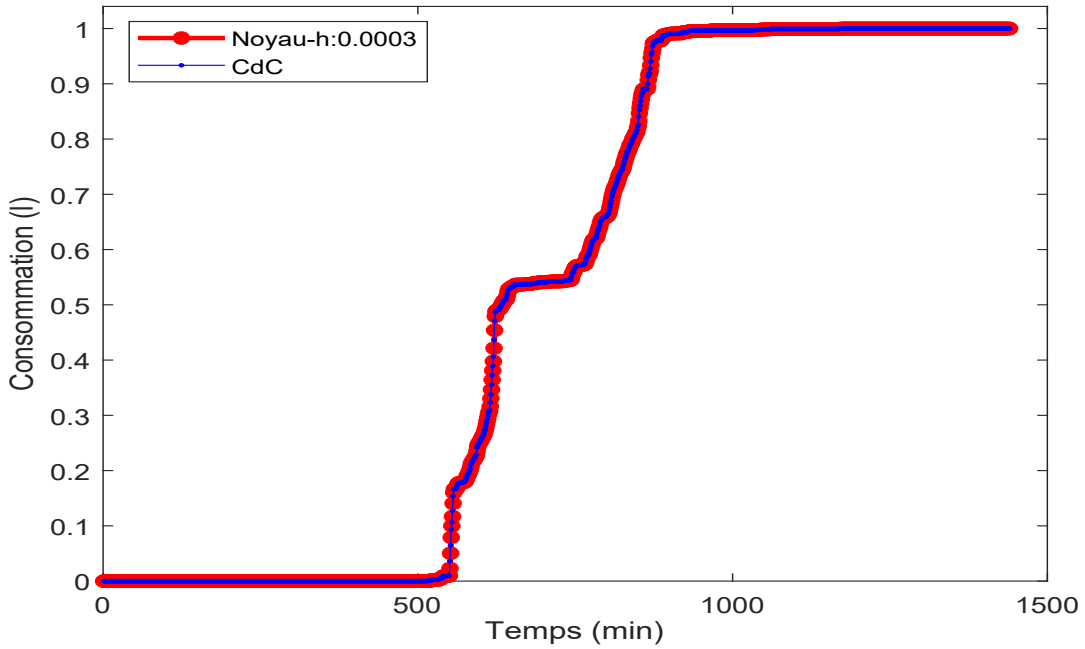

FIGURE 4.16 – La distribution de la CdC d'eau normalisé estimée par noyau

TABLE 4.1 – Comparaison entre la méthode d'estimation par noyau et le modèle de mélange gaussien

	RMSE (CdC) 19/01/2018	RMSE (CdC) normalisé	Nombre de paramètres	Temps de calcule
KDE	1.73	0.0014	1	16.85 s
GMM	13.62	0.01	6	1.64 s

La figure 4.17 représente l'estimation de la CdC de la consommation d'eau le 25/05/2018 en bleu par la courbe en rouge en utilisant l'estimation par noyau.

Comparaison

Le tableau 4.1 montre les résultats selon trois critères de comparaison entre les méthodes d'estimation de densité KDE et GMM. Nous avons obtenu l'erreur quadratique moyenne minimale entre la CdC et la densité estimée en utilisant la méthode d'estimation par noyau KDE qui a le moins nombre de paramètres par rapport à la méthode d'estimation paramétrique GMM. Mais la méthode prend beaucoup de temps pour choisir son paramètre contrairement à la méthode d'estimation par mélange gaussien.

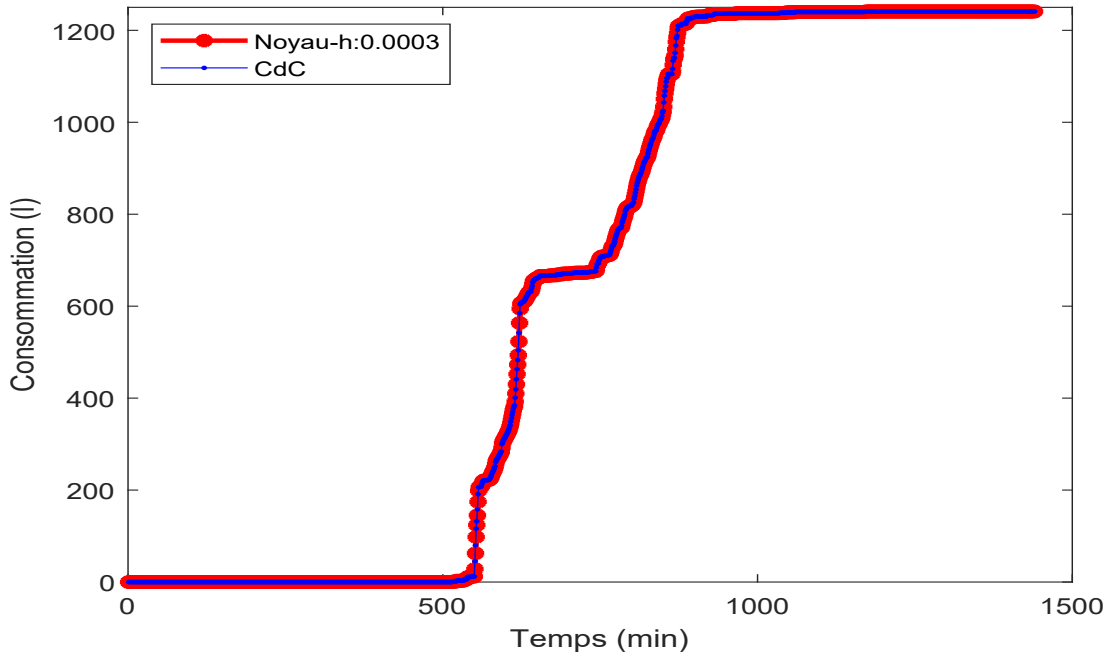

FIGURE 4.17 – L'estimation par noyau de la CdC de la consommation d'eau du 25/05/2018

4.3 Séries temporelles

L'étude des séries temporelles permet d'analyser les données. Nous nous intéressons à l'évolution au cours du temps d'un phénomène, dans le but de décrire, expliquer puis prévoir ce phénomène dans le futur.

Définition 4.7 *Une série temporelle (ou série chronologique) est une suite réelle finie* $(X_t)_{1\leq t\leq n}$ *d'observations correspondant à la même variable où* t *représente le temps.*

Les composantes d'une série chronologique :
En général, une série chronologique contient trois éléments essentiels [24] tels que :

1. La tendance Z_t qui correspond à l'évolution au long terme et le mouvement fondamental de la série.
2. La saisonnalité S_t qui correspond à un phénomène périodique d'une période identifiée p.
3. Le bruit ε_t (erreur ou une composante résiduelle) est la partie aléatoire de la série.

Example 4.15 *La figure 4.18 représente la série temporelle avec des données de la consommation d'eau échantillonnées par heures pendant 7 semaines dans un restaurant universitaire. Elle décrit le profil de consommation d'un bâtiment tertiaire dont la consommation est clairement présente pendant les cinq jours de travail (les Week-days) est faible les deux jours du week-end [19].*

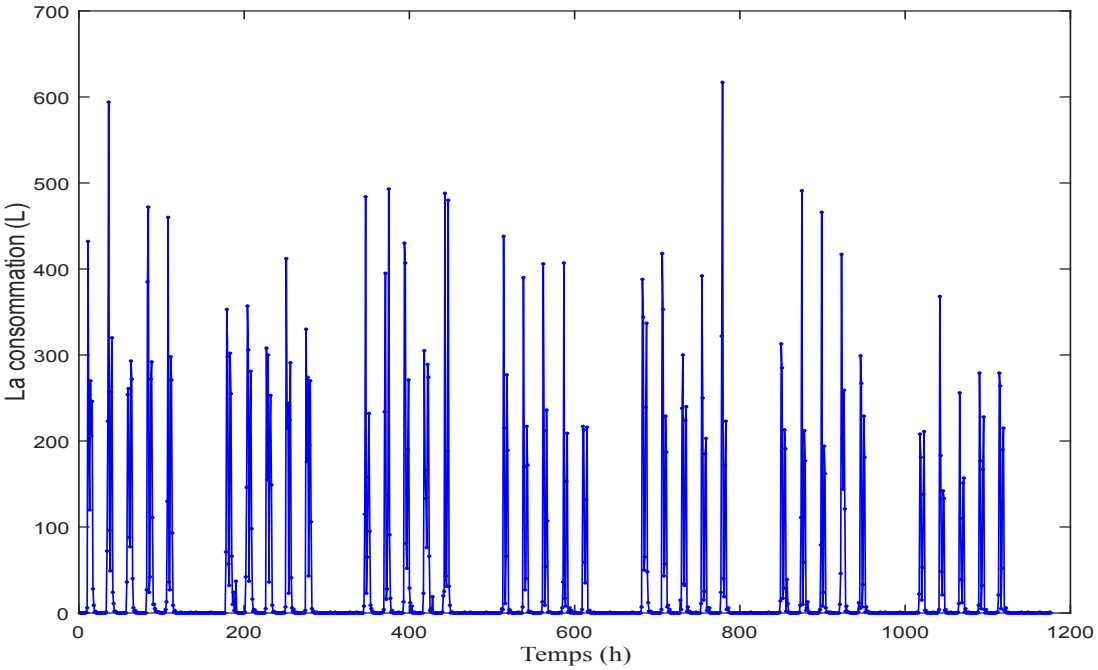

FIGURE 4.18 – La série temporelle de la consommation d'eau

Pour appliquer des modèles de prédiction de séries chronologiques, nous divisons la série chronologique en deux sous-séries : l'ensemble d'entraînement se compose de 6 semaines (80 % de données) et l'ensemble de test contient la dernière semaine (20 % de données).

Modèles de décomposition additifs et multiplicatifs

Nous considérons deux modèles liant les composantes de la série temporelle, ces modèles portent le non de schéma de décomposition [24].

1. **Schéma additif :** est une somme représentée selon l'équation (4.12) telle que :

$$X_t = Z_t + S_t + \varepsilon_t \tag{4.12}$$

2. **Schéma multiplicatif :** la série X_t représente par le produit de trois composantes telle que :

$$X_t = Z_t \times S_t \times (1 + \varepsilon_t) \tag{4.13}$$

Choix du schéma de décomposition de la série temporelle

Trois critères sont disponibles permettant de choisir un schéma de décomposition du modèle déterministe [36].

1. **Méthode de Bande :** Deux droites sont utilisées, l'une passe par les points minimum et l'autre passe par les points maximum pendant chaque période.
   * Si les deux droites sont à peu près parallèles, le modèle est additif.

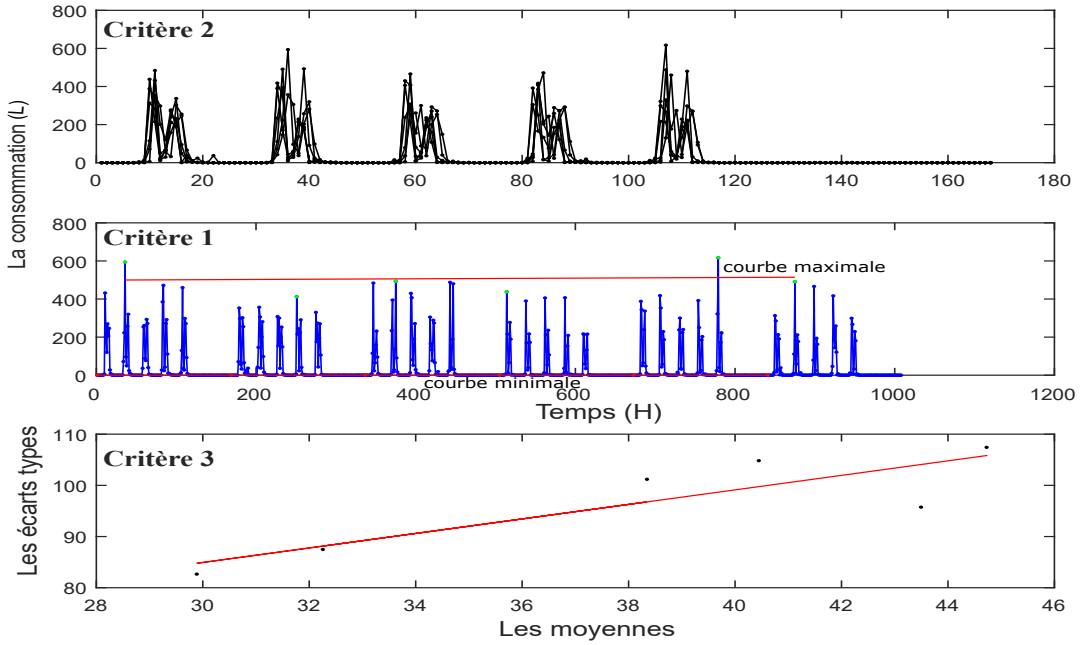

FIGURE 4.19 – Critères de choix du schéma de décomposition de la série temporelle de l'exemple 4.15

   * Sinon, le modèle est multiplicatif.

2. **Méthode du profil :** Elle utilise les courbes en superpositions (les courbes des séries chronologiques dans chaque période).
   * Si les différentes courbes sont à peu près parallèles alors le modèle est additif.
   * Sinon, le modèle est multiplicatif.
3. **Méthode du tableau Buys et Ballot :** Elle se base sur la droite des moindres carrés appliquée aux points de coordonnées moyens et l'écart type des courbes de superposition.
   * Si la pente des moindres carrés est très proche de 0 alors le modèle est additif.
   * Si la pente des moindres carrés n'est pas nulle alors le modèle est multiplicatif.

La figure 4.19 représente les trois critères de choix de modèle. La droite représentée dans le troisième graphe a une pente égale à 1.4182 qui est loin de zéro donc le modèle est multiplicatif.

Modèles déterministes et non déterministes des séries temporelles

* **Le modèle déterministe :** Dans ce modèle nous supposons que la série est une fonction du temps et d'une variable ε_t qui représente l'erreur (la différence entre la réalisation et le modèle proposé).
* **Le modèle stochastique :** Le modèle stochastique est une fonction qui dépend du temps sachant que la variable de bruit ε_t est un processus stochastique [36] et [37].

4.3.1 Modèles déterministes

Après le choix du modèle, le modèle déterministe est donnée par : $X_t = Z_t \times c_i^* \times (1+\varepsilon_t)$ si le modèle est multiplicative, sinon $X_t = Z_t + c_i^* + \varepsilon_t$.

Estimation paramétrique de la tendance et de la saisonnalité :

L'estimation paramétrique de la tendance et de la saisonnalité est détaillée ci-après :

* Nous estimons la tendance par la droite de régression des moindre carré. Elle est donnée par : $Z_t = at + b$.
* Nous calculons les rapports (la série corrigée de la tendance) : $R_t = X_t / Z_t$ si le modèle est multiplicatif, sinon $R_t = X_t - Z_t$.
* Nous déterminons les coefficients de saisonnalité : $c_i = \frac{1}{p}\sum_{j=0}^{n} R_{i+j\times p};\ i = 1,\ ...,\ p$.
* Nous calculons les coefficients de saisonnalité normalisés : $c_i^* = \dfrac{c_i}{\frac{1}{p}\sum_{i=1}^{p} c_i}$ si le modèle est multiplicative, sinon $c_i^* = c_i - \frac{1}{p}\sum_{i=1}^{p} c_i,\ i = 1,\ ...,\ p$.

Application
La figure 4.20 représente l'estimation paramétrique de la série temporelle décrite dans l'exemple 4.15. La série temporelle est donnée par la courbe bleue et la série en vert est la série estimée par paramètre par le modèle déterministe multiplicatif. La droite en rose est la tendance calculée par la méthode des moindres carrés.

Estimation non paramétrique de la tendance et de la saisonnalité :

Le filtrage permet d'écarter les pics et les trous autrement dit il permet de lisser la courbe afin de dégager une tendance. Le filtre le plus employé est la moyenne mobile.
Appliquer la moyenne mobile sur la série temporelle, c'est donc filtrer la série pour éliminer certaines composantes.
La méthode de filtrage de moyenne mobile (Moving Average) :
La méthode de moyenne mobile centrée d'ordre p est donnée par :
1- Si p impair ; $p = 2m+1$ alors,

$$M.M_{p,t} = \frac{1}{p}\sum_{i=-m}^{m} X_{t+i};\ \forall t = m+1, ..., n-m \tag{4.14}$$

2- Si p est pair ; $p = 2m$ alors,

$$M.M_{p,t} = \frac{1}{p}(\frac{1}{2}X_{t-m} + \sum_{i=-m+1}^{m-1} X_{t+i} + \frac{1}{2}X_{t+m}). \tag{4.15}$$

La théorie non paramétrique de l'estimation de la tendance et l'effet saisonnier [36] sont données par les étapes suivantes :

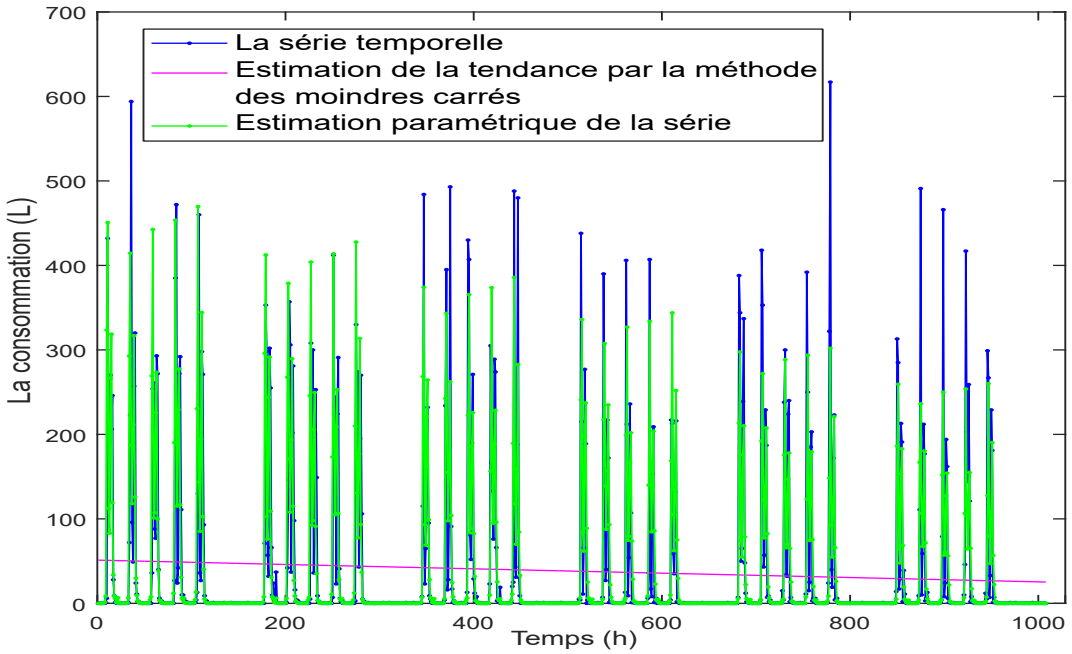

FIGURE 4.20 – Estimation paramétrique du modèle déterministe de la série temporelle de l'exemple 4.15.

* Nous estimons la tendance par la méthode de moyenne mobile centrée d'ordre p telle que $Z_t = M.M_{p,t}$.
* Nous calculons les rapports (la série corrigée de la tendance) : $R_t = X_t / Z_t$ si le modèle est multiplicatif, sinon $R_t = X_t - Z_t$
* Nous déterminons les coefficients de saisonnalité : $c_i = \frac{1}{p}\sum_{j=0}^{n} R_{i+j\times p};\ i = 1,...,p.$
* Nous calculons les coefficients de saisonnalité normalisés : $c_i^* = \dfrac{c_i}{\frac{1}{p}\sum_{i=1}^{p} c_i}$ si le modèle est multiplicatif, sinon $c_i^* = c_i - \frac{1}{p}\sum_{i=1}^{p} c_i, i = 1,...,p.$

Application

La figure 4.21 représente l'estimation non paramétrique de la même série temporelle. La tendance estimée par la méthode des moyennes mobile est dessiné en rose. L'estimation non paramétrique de la série temporelle colorée en bleu est donnée par la série verte.

Prévision des valeurs futures :

Nous prenons 20% des données de la série temporelle dans l'exemple 4.15 pour faire le test avec l'extrapolation de la série estimée paramétriquement :

Pour prédire une valeur de la série à l'instant $n+h$ où $h \geq 1$, nous utilisons les estimations de la tendance et de la saisonnalité. La valeur de prédiction X_{n+h} est donnée par :

$X_{n+h} = a \times (n+h) + b + c_i^*$; si le modèle additif, sinon $X_{n+h} = (a \times (n+h) + b) \times c_i^*$; avec $i = h$ si $h \leq p$ sinon $i =$ le reste de la division de h sur p [36].

Pour l'estimation non paramétrique, nous faisons pareillement mais après l'estimation de la tendance de la série du modèle déterministe non paramétrique par la droite de régression.

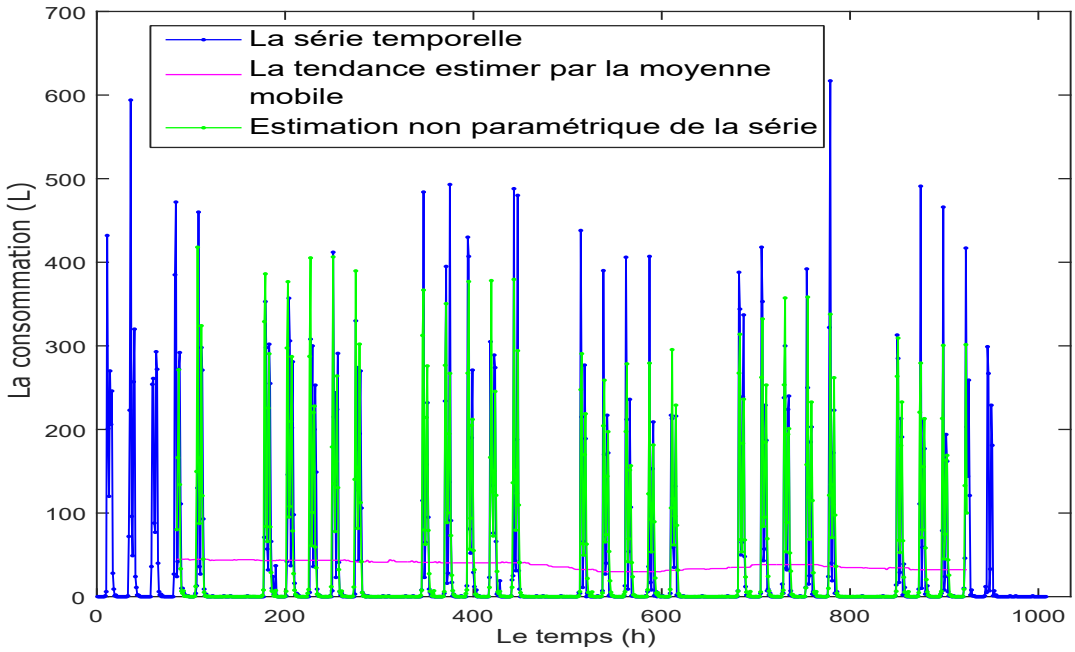

FIGURE 4.21 – Estimation non paramétrique du modèle déterministe de la série temporelle de l'exemple 4.15.

Applications

La figure 4.22 représente la prédiction du modèle déterministe estimé par paramètre pendant une semaine par la courbe en rouge.

La prédiction du modèle déterministe estimé par la moyenne mobile pendant une semaine est représentée par la courbe en rouge dans la figure 4.23.

4.3.2 Modèles stochastiques

Soit (Ω, F, p) un espace probabilisé, Ω l'espace fondamental des événements, F la tribu engendrée par Ω et p une probabilité.

(E, ξ) est un espace mesurable, ξ est une tribu définie sur l'ensemble (E) [38].

Définition 4.8 ***Une variable aléatoire*** *est une fonction mesurable X définie sur l'espace probabilisé (Ω, F, p) à valeur dans (E, ξ), c'est a dire :*

$$\forall B \in \xi,\ X^{-1}(B) \in F.$$

Définition 4.9 ***Un processus stochastique*** *est une suite de variable aléatoire $((X_t)_{t \in T})$ indéxée par le temps $t \in T$.*

$$\forall (w, t) \in (\Omega, F, p) \times T \longrightarrow X_t(w) \in E.$$

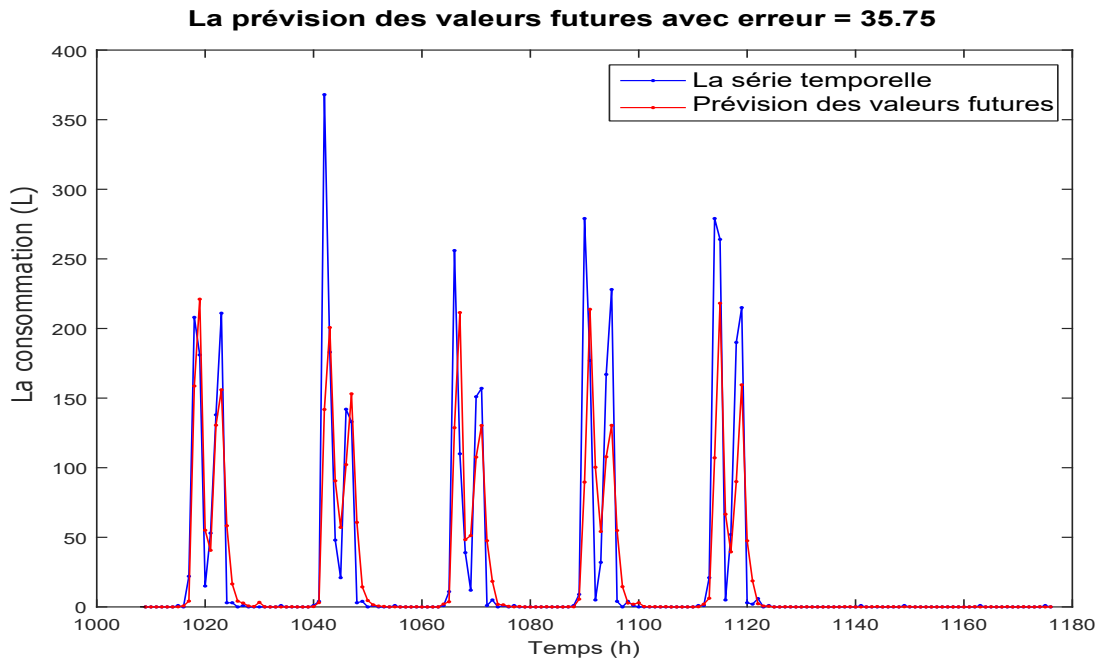

FIGURE 4.22 – Prévision avec le modèle déterministe estimée paramétriquement

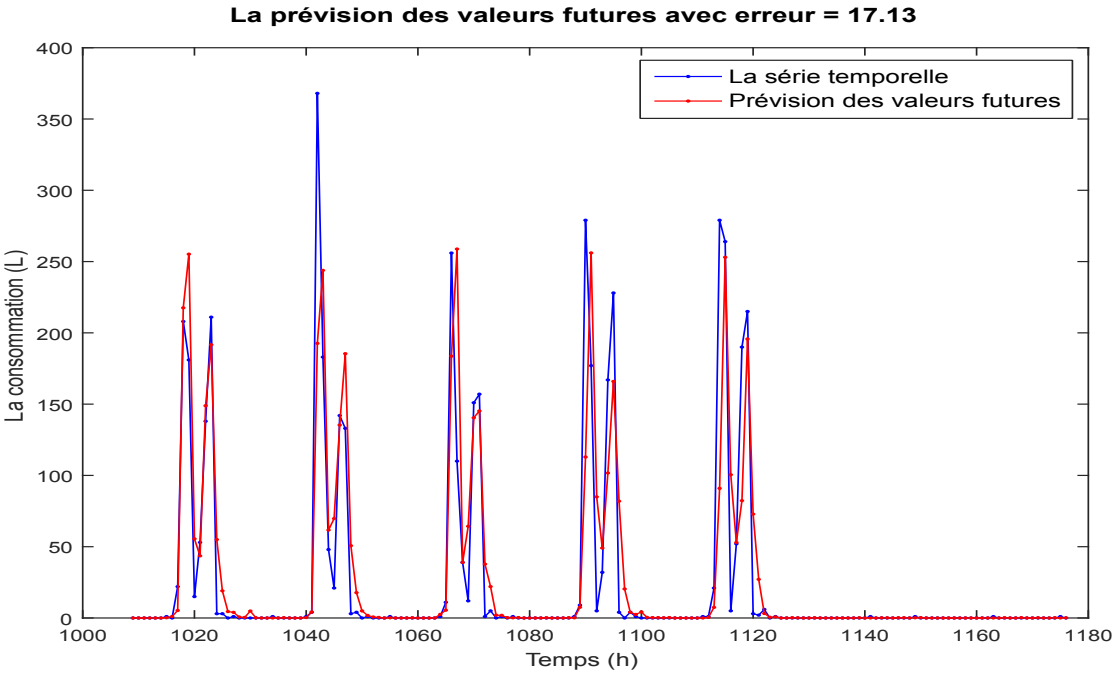

FIGURE 4.23 – Prévision avec le modèle déterministe estimée par les moyennes mobiles

Remarque : La série temporelle est une réalisation d'un processus stochastique.

Stationnarité d'un processus stochastique

Soit (X_t) un processus stochastique, et pour tout entier h et tout $t \in R$:

- (X_t) est **stationnaire au sens fort** si (X_t) et (X_{t+h}) suivent la même loi.
- (X_t) est **stationnaire au sens faible** si :

1. $E(X_t) = \mu$
2. $var(X_t) = \delta^2$
3. $cov(X_t, X_{t+h})$ ne dépendent que de h.

Fonction d'autocorrélation (ACF) :

Soient X_t un processus stochastique d'espérance $E(X_t)$, variance $var(X) = \delta^2$ et de la fonction d'autocovariance $\gamma(h) = cov(X_t, X_{t+h});\ \forall h \in N$.

La fonction d'autocorrélation est une fonction de N à $]-1, 1[$, noté $\rho(h)$ et elle est définie par :

$$\rho(h) = corr(X_t, X_{t+h}) = \frac{\gamma(h)}{\delta^2}$$

Fonction d'autocorrélation partielle (ACFP) :

Pour tout $h > 2$ la fonction d'auocorrélation partielle notée $\pi(h)$ est donnée par :

$$\pi(h) = corr(X_t - E_L(X_t|X_{t+1}, ..., X_{t+h-1}),\ X_{t+h} - E_L(X_{t+h}|X_{t+1}, ..., X_{t+h-1}))$$

telle que E_L soit l'espérance conditionnelle.

Opérateur de retard :

L'opérateur de retard d'ordre 1 est défini par :

$$BX_t = X_t - X_{t-1}$$

Et nous définissons par récurrence l'opérateur de retard d'ordre k par : $B^k X_t = X_t - X_{t-k}$ [39].

Opérateur de différenciation :

L'opérateur de différenciation [39] d'ordre d est donné par :

$$\nabla^d X_t = \underbrace{\nabla o ... o \nabla}_{d\ \ fois}(X_t)$$

pour $d = 1 : \nabla X_t = X_t - X_{t-1}$

pour $d = 2 : \nabla^2 X_t = X_t - 2X_{t-1} + X_{t-2}$.

Critère d'information Aikaike et Bayésien :

Le critère d'information Aikaike (AIC) ou le critère d'information Bayésien (BIC) sont basées sur la notion de vraisemblance pour choisir le modèle le mieux adapté par minimisation de l'un des critères.

$$AIC = 2 \times k - 2\ln(L)$$

$$BIC = -2\ln(L) + \ln(n)k$$

Modèle autorégressif AR(p)

Le modèle d'auto-régression d'ordre p d'une série temporelle stationnaire X_t est défini par (4.16) :

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t \tag{4.16}$$

tels que :
$\varepsilon_t = X_t - \sum_{i=1}^{p} \alpha_i X_{t-i}$ sont des résidus.
$\alpha_1, \alpha_2, ..., \alpha_p \in R$ sont des paramètres du modèle AR d'ordre p que nous pouvons estimer par les moindres carrés.
Soit α^* l'estimation de $\alpha = \{\alpha_1, ..., \alpha_p\}$, alors

$$\alpha^* = \arg\min_{\alpha} \sum_{t=p+1}^{n} (X_t - \sum_{i=1}^{p} (\alpha_i X_{t-i}))^2$$

Nous posons : $M = \begin{bmatrix} X_p & X_{p-1} & ... & X_1 \\ X_{p+1} & X_p & ... & X_2 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{n-1} & X_{n-2} & ... & X_{n-p} \end{bmatrix}$, $Y = \begin{bmatrix} X_{p+1} \\ X_{P+2} \\ . \\ . \\ . \\ X_n \end{bmatrix}$

Donc : $\alpha^* = (M^T M)^{-1} M^T Y$
Le modèle AR(p) peut s'écrire sous la forme (4.17) :

$$\Phi(B) X_t = \varepsilon_t \tag{4.17}$$

avec : $\phi(B) = 1 - \alpha_1 B + \alpha_2 B^2 + ... + \alpha_p B^P$ où B l'opérateur de retard [40].

Applications

La figure 4.24 représente le modèle d'auto-régression d'ordre 8 appliqué aux données de la consommation d'eau échantillonnées en heures pendent 6 semaines. La courbe en noir représente notre série temporelle et la courbe en bleu est la prédiction en heures pendant une semaine par le modèle d'auto-régression.

Modèle à moyenne mobile MA(q)

Le modèle à moyenne mobile d'ordre q (MA(q)) [40] est donné par (4.18) :

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \tag{4.18}$$

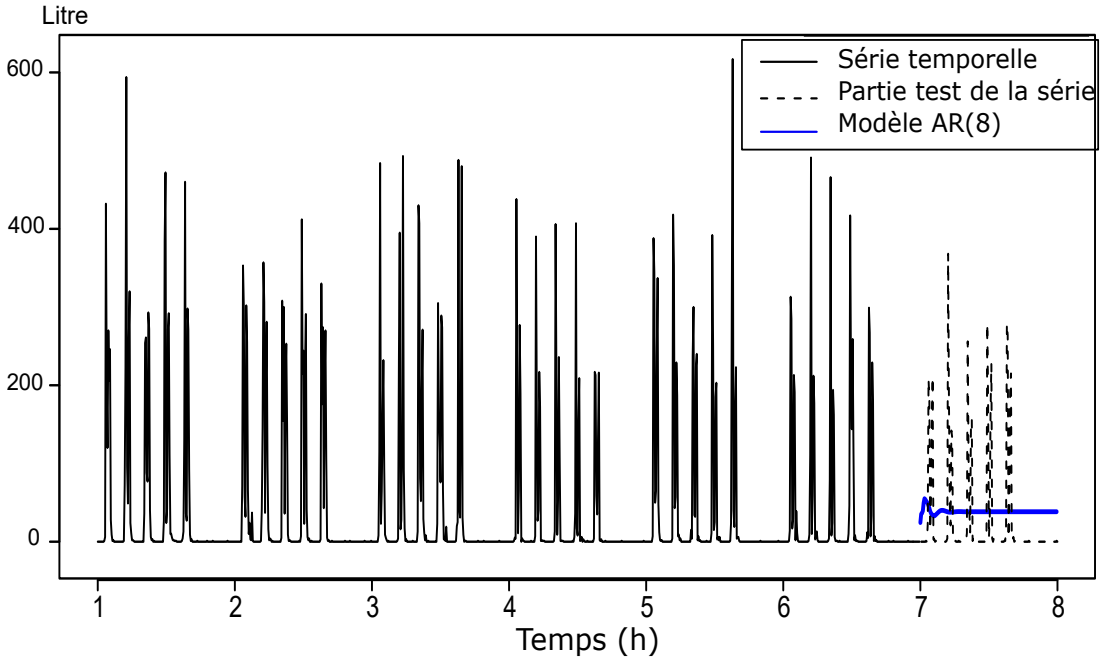

FIGURE 4.24 – Le modèle AR(8) pour la prédiction de la consommation d'eau

où ε_t est un bruit blanc et $\beta_i;\ i = 1,\ ...,\ q$ sont des réels.
La forme de MA(q) avec l'opérateur de retard B est définie par (4.19) :

$$X_t = \Psi(B)\varepsilon_t \tag{4.19}$$

avec : $\Psi(B) = 1 - \beta_1 B - \beta_2 B^2 - ... - \beta_q B^q$.

Application

Nous appliquons le modèle de moyenne mobile à la série temporelle spécifiée dans l'exemple 4.15.
D'après la méthode de sélection du modèle par le critère AIC, nous trouvons que le modèle MA(7) est le modèle choisi pour prédire la consommation d'eau mais comme les données sont périodiques, nous remarquons que les valeurs prédites représentées par la courbe en bleu dans la figure 4.25 ne sont pas bonnes.

Modèle mixte ARMA(p, q)

Le modèle mixte noté ARMA(p, q) est le modèle qui consiste à assembler le modèle d'auto-régression d'ordre p et le modèle de moyenne mobile d'ordre q pour des séries temporelles stationnaires [40]. Il est défini par (4.20) :

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \tag{4.20}$$

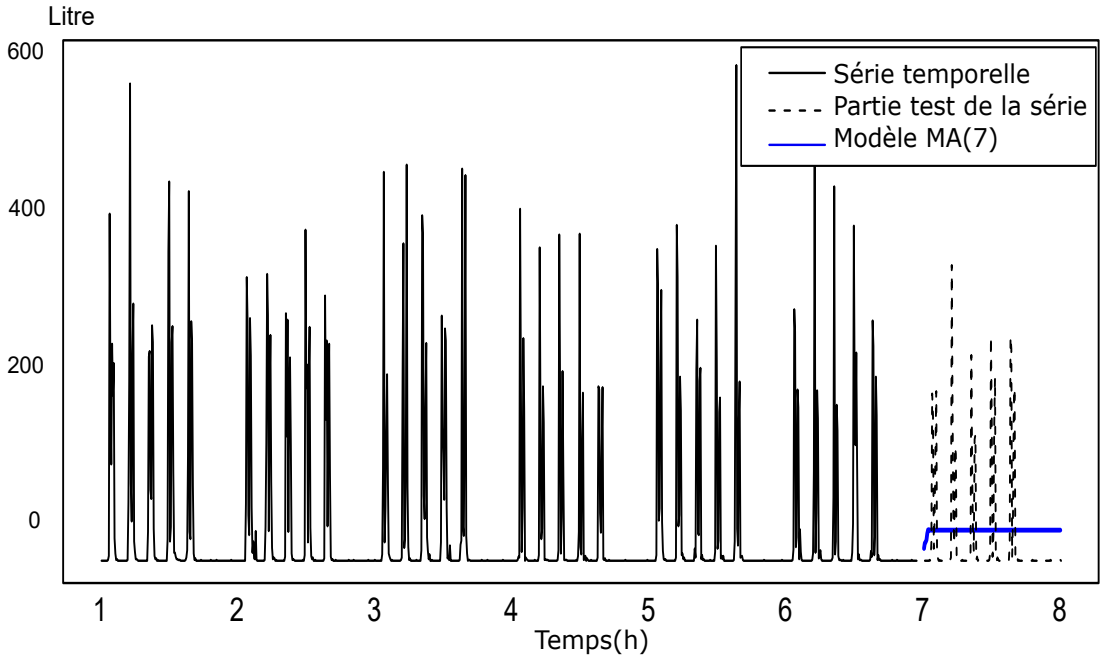

FIGURE 4.25 – La prévision de consommation d'eau avec le modèle MA(7).

ou par la formule (4.21) :

$$\Phi(B)X_t = \Psi(B)\varepsilon_t \tag{4.21}$$

où $\phi(B) = 1 - \alpha_1 B + \alpha_2 B^2 + ... + \alpha_p B^P$ et $\Psi(B) = 1 - \beta_1 B - \beta_2 B^2 - ... - \beta_q B^q$.

Application

La figure 4.26 représente les résultats de la prévision de la consommation d'eau avec le modèle mixte. La série temporelle utilisée est la même que pour l'exemple 4.15 et les paramètres des modèles ARMA 6,2 sont sélectionnés à l'aide du critère AIC.

En raison de la saisonnalité de la série nous rejetons le modèle ARMA(6,2) représenté dans la figure 4.26.

Modèle ARIMA(p, d, q)

Le modèle Autoregressive integrated moving average (ARIMA) [40] est adapté à une série stationnaire par différentiation et non périodique et il est défini par (4.22) :

$$X_t = \sum_{i=1}^{p} \alpha_i \nabla^d X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \tag{4.22}$$

ou par la formule (4.23) :

$$\Phi(B)(I - B)^d X_t = \Psi(B)\varepsilon_t \tag{4.23}$$

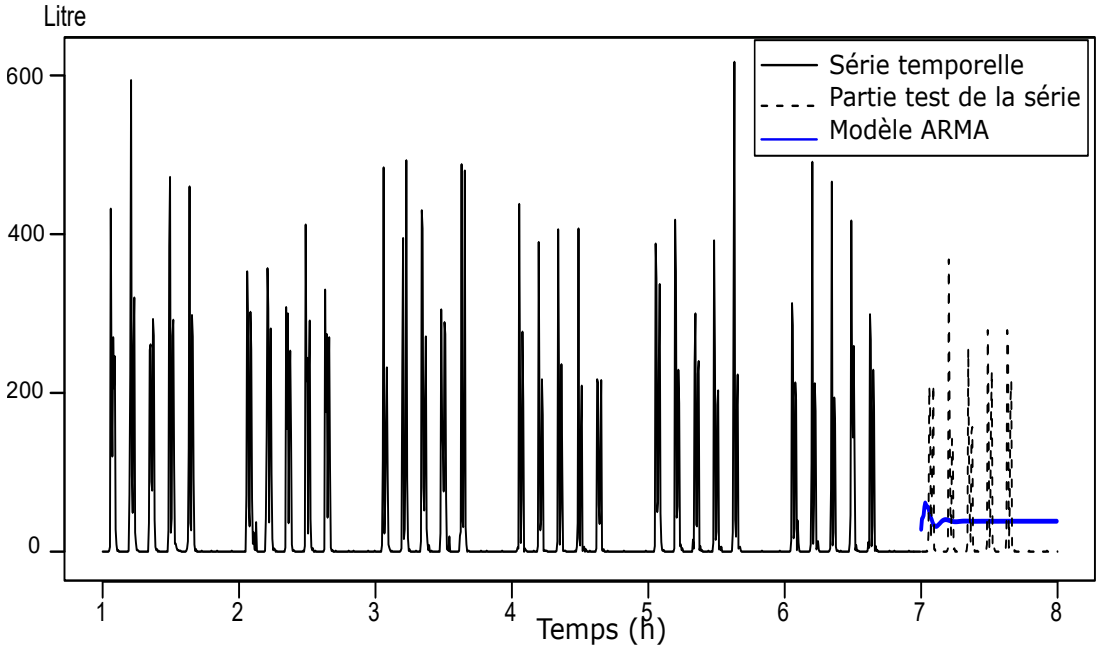

FIGURE 4.26 – La prévision de la consommation d'eau par le modèle ARMA(6,2).

où p est le nombre de termes d'autorégression, q est le nombre de termes de moyenne mobile et d est le nombre de différentiations.
Remarque : Notre série temporelle de l'exemple 4.15 est stationnaire donc le modèle ARIMA est équivalent au modèle ARMA.

Modèle SARIMA

Le modèle seasonal autoregressive integrated moving average (SARIMA)$(p,d,q)(P,D,Q)_s$ est le complété du modèle ARIMA sur les séries avec saisonnalité [41], [42]. Il est défini par (4.24) :

$$\Phi(B)F(B)(I-B)^d(I-B^S)^D X_t = \Psi(B)G(B)\varepsilon_t \tag{4.24}$$

telles que $\Phi(B)$, $F(B)$, $\Psi(B)$, $G(B)$ sont des polynômes de degré respectivement p, P, q, Q et ε_t le bruit.
p : l'ordre de l'autorégressive.
d : l'ordre de la différence de la série.
q : l'ordre de la moyenne mobile.
P : l'ordre de la partie autorégression saisonnière.
D : l'ordre de la différence saisonnière.
Q : l'ordre de la moyenne mobile saisonnière.

Méthodologie de Box et Jenkins

La méthodologie de Box et Jenkins consiste à la modélisation non déterministe des séries temporelles. Elle est donnée par 5 étape [43] :

1. **La stationnarité de la série :**
   * Si la série est saisonnière, nous commençons par l'élimination de la périodicité (désaisonalisation) avec l'opérateur $\nabla_s^D = (1 - B^s)^D$ ensuite nous vérifions la stationnarité de la série.
   - La présentation graphique permet de voir la stationnarité de la série.
   - Avec ACF nous pouvons décider si la série est stationnaire ou pas. La fonction de l'autocorrélation décroit rapidement vers zéro si la série est stationnaire.
   - Test d'hypothèse de non stationnarité de Dickey-Fuller augmenté (ADF) est donné par la stationnarité de la série en rejetant l'hypothèse nulle du test au seuil de 5%.
   * Si la série est non saisonnière nous faisons la différenciation simple des données d fois et nous nous arrêtons dans l'ordre de différenciation d qui rend la série stationnaire.
   La série représentée dans la figure 4.18 est saisonnière de période 7×24 heures donc, $s = 168$ et $D = 1$.
   Le test de stationnarité ADF indique que la série résiduelle est stationnaire avec $p - value = 0.01 < 0.05$ d'où $d = 0$.
2. **Identification :**
   Nous sélectionnons le modèle par l'estimation des paramètres p, q, P, Q pour les séries stationnaires par :
   - Le graphe de la fonction d'autocorrélation et la fonction d'autocorrélation partielle :
   * Pour trouver le paramètre p du modèle AR(p) : Les autocorrélations sont dans une enveloppe à décroissance géométrique et $\pi(k)$ sont identiquement nulles pour $k > p$.
   * Pour trouver le paramètre q de modèle MA : Les autocorrélations partielles sont dans un enveloppe à décroissance géométrique et $\rho(k)$ sont identiquement nulles pour $k > q$.
   * Pour trouver les paramètres p et q pour les modèles ARMA et ARIMA : Les autocorrélations partielles sont identiquement nulles après l'ordre p et Les autocorrélations partielles sont identiquement nulles après l'ordre q.
   * Pour trouver les paramètres p, q, P et q pour les modèles SARIMA : Les auto-corrélations partielles sont identiquement nulles après l'ordre p et les auto-corrélations partielles sont identiquement nulles après l'ordre q, P égale à l'ordre multiple de la saisonnalité de l'autocorrélograme partielle et Q est donnée par l'ordre multiple de la saisonnalité de l'autocorrélograme.
   - Souvent, le choix des ordres p, q, P et Q en regardant les autocorrélations et autocorrélations partielles n'apparaisse pas de manière évidente. Nous pouvons dans ce cas sélectionner un modèle en minimisant un critère pénalisé de type AIC ou BIC.
   Le meilleur modèle au sens de l'AIC est ici le modèle : SARIMA$(4,0,1)(0,1,0)[168]$.
3. **Estimation des modèles :** C'est à dire trouver les coefficients des polynômes du modèle par la méthode des moindres carrés ordinaires, la méthode de Yule-Walker ou la méthode du maximum de vraisemblance.

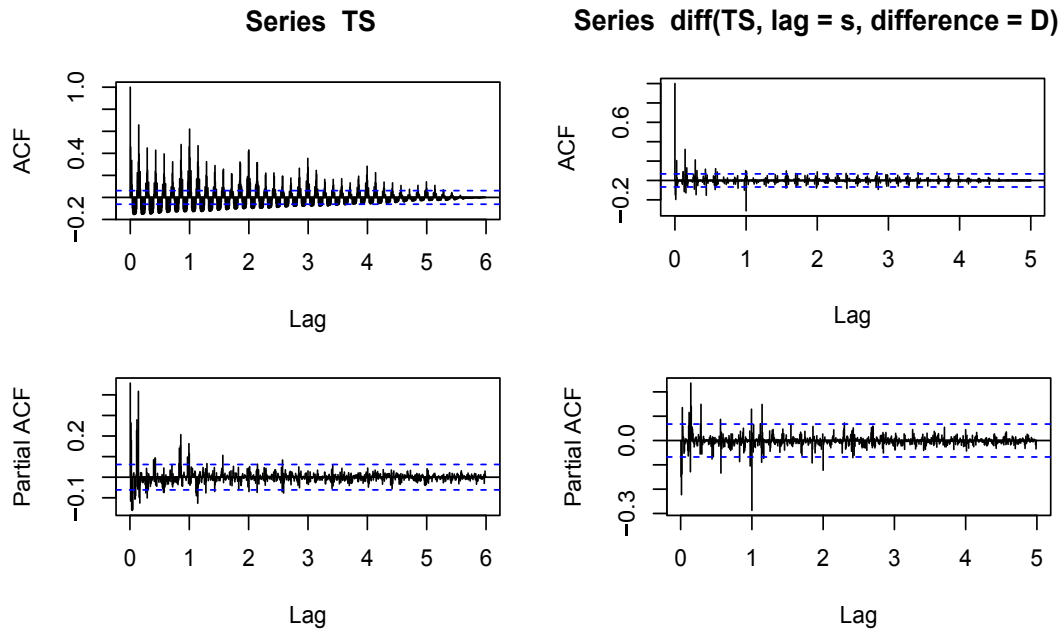

FIGURE 4.27 – L'autocorrélograme et l'autocorrélograme partielle de la série temporelle et la série corrigée par les effets saisonniers

La méthode la plus utilisée est celle du maximum de vraisemblance.
Nous posons : $\theta = (\alpha_1, ..., \alpha_p,\ f_1, ..., f_P,\ \beta_1, ..., \beta_q,\ g_1, ..., g_Q)$, avec $\alpha_i,\ \beta_i,\ f_i,\ g_i$ sont les coefficients des polynômes $\Phi(B),\ F(B),\ \Psi(B),\ G(B)$, l'estimation de θ est donnée par l'équation :

$$\theta^* = \arg\max L(X_1,\ ...,\ X_n,\ \theta) \tag{4.25}$$

Les coefficients estimés sont données par : $\theta = (0.16,\ -0.14,\ -0.01,\ 0.19,\ -0.35)$.
Remarque : Les trois premières étapes de la modélisation non déterministe sont faites automatiquement sur le logiciel R avec la commende (auto.arima) [44].

4. **Validation des modèles :** Pour faire la prédiction avec le modèle d'estimation, les résidus doivent vérifier certaines propriétés statistiques [39]. Nous utilisons le test Ljung-Box Q pour valider si les résidus estimés satisfont à l'exigence d'une séquence de bruit blanc avec $p - valeur > 0,05$, [42], [44].
   Le test Ljung-Box Q a suggéré que la série des résidus comprenait du bruit blanc ($X - squared = 0.014,\ df = 1,\ p - value = 0.91$).
5. **Prévision :** La confirmation de notre choix du modèle est fait par simuler la prévision selon un échantillon test.
   La formule de prévision pour le modèle ARIMA(p,d,q) est donnée par :

$$X_{t+k} = \sum_{i=1}^{p} \alpha_i \nabla^d X_{t+k-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t+k-i}$$

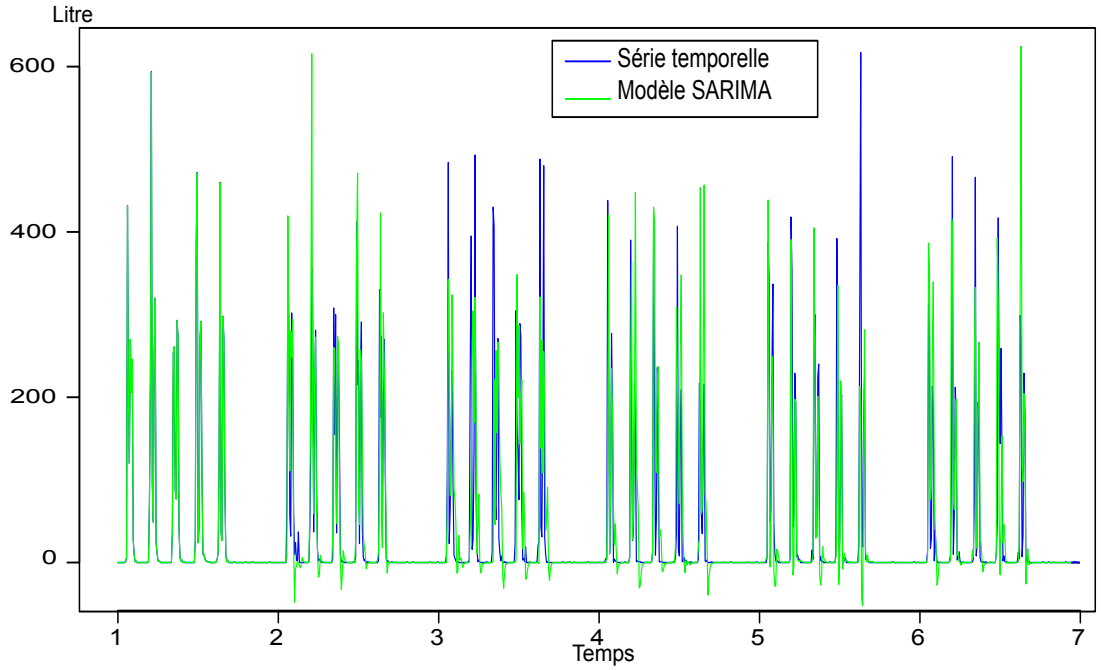

FIGURE 4.28 – Résultat du modèle SARIMA(4,0,1)(0,1,0)[168]

Si $d = 0$ alors nous avons le modèle ARMA(p, q).
Si $d = 0$ et $q = 0$ alors nous avons le modèle AR(p).
Si $d = 0$ et $p = 0$ alors nous avons le modèle MA(q).

Applications

Le modèle SARIMA$(4,0,1)(0,1,0)_{168}$ est le meilleur modèle qui minimise les critères d'information (AIC et BIC). Les résultats d'entrainement du modèle SARIMA sont représentés dans la figure 4.28. Le modèle SARIMA$(4,0,1)(0,1,0)_{168}$ est définie par l'équation (4.26) :

$$(0.16\ B - 0.14\ B^2 - 0.01\ B^3 + 0.19\ B^4)(I - B^{168})X_t = (-0.35\ B)\epsilon_t. \tag{4.26}$$

La figure 4.29 donne la prédiction de 168 heures estimé par le modèle SARIMA$(4,0,1)(0,1,0)_{168}$ avec l'erreur dans la partie entrainement égale à 57.71 et dans la partie test égale à 35.75.

4.3.3 Lissage exponentiel

Les méthodes de lissages exponentiels sont des outils permettant de réaliser des prévisions à partir des observations d'une série temporelle [45], [46] et [47].
Nous présentons trois types de lissage exponentiel :

* Le lissage exponentiel simple.
* Le lissage exponentiel double.
* Le lissage exponentiel triple de Holt-Winters.

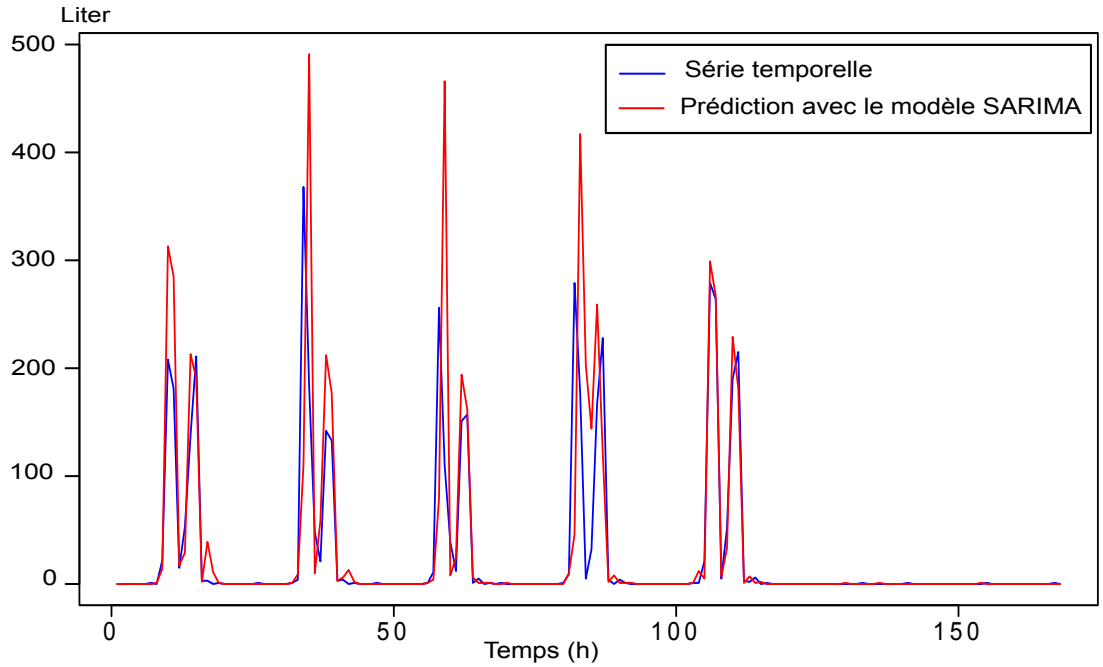

FIGURE 4.29 – La prédiction avec le modèle SARIMA(4,0,1)(0,1,0)[168]

Lissage exponentiel simple

Le lissage exponentiel simple appelé aussi lissage exponentiel unique est une méthode de prévision de série temporelle pour des données sans tendance et sans saisonnalité ETS(A, N, N) [46]. Il est défini par :

$$\begin{cases} Y_t = \alpha X_t + (1-\alpha) Y_{t-1};\ t > 1. \\ Y_1 = X_1 \end{cases} \tag{4.27}$$

Avec $0 < \alpha < 1$ paramètre de lissage.
Pour faire la prédiction avec le modèle de lissage exponentiel simple nous utilisons l'équation (4.28).

$$X_{t+k} = Y_t;\ \forall k = 1, 2, ...;\ t > n. \tag{4.28}$$

La prévision avec la méthode de lissage simple appliqué à la série temporelle de l'exemple 4.15 est représenté dans la figure 4.30. Le paramètre de lissage simple $\alpha = 0.39$ est fixé en minimisant l'erreur au moyenne quadratique de la partie d'entrainement et les erreurs sont données par : $RMSE - Train = 92.69,\ RMSE - Test = 73.31$.
RMSE-Train est l'erreur au moyenne quadratique de la partie entrainement.
RMSE-Test est l'erreur au moyenne quadratique de la partie test.

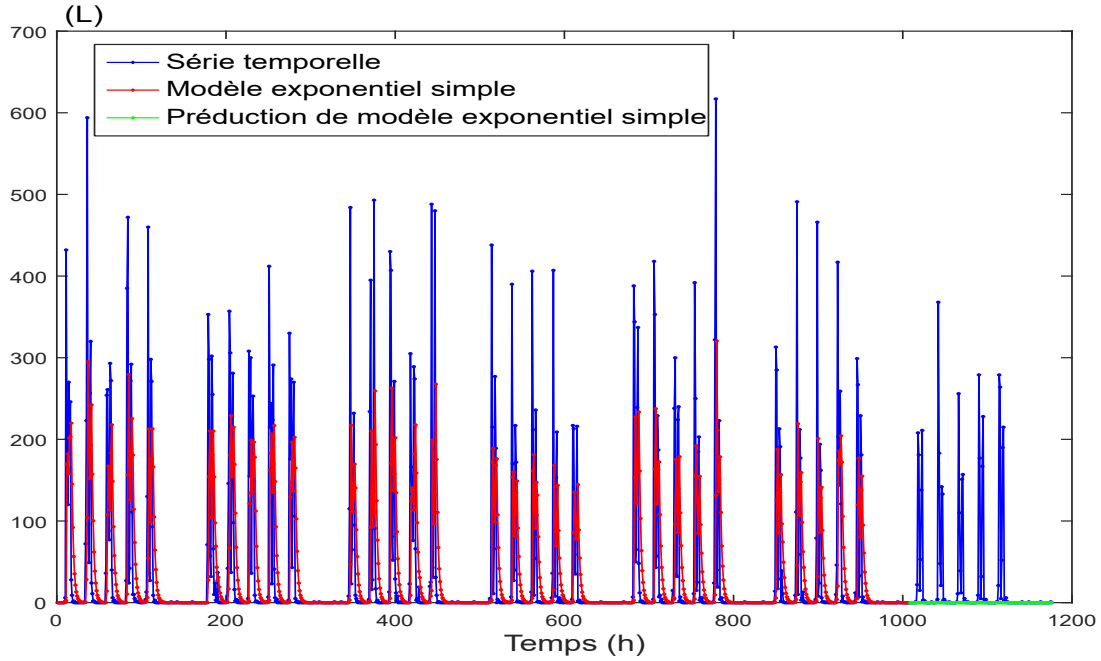

FIGURE 4.30 – Résultat obtenu avec le modèle exponentiel simple

Le lissage exponentiel double

Le lissage exponentiel double est une extension du lissage exponentiel simple qui prend en charge la tendance des données de la série temporelle.

1. **Le lissage exponentiel de Holt :**

   * **Tendance additive : ETS(A, A, N)** est un lissage exponentiel double avec une tendance linéaire.

$$\begin{cases} Y_t = \alpha X_t + (1-\alpha)(Y_{t-1} + T_t);\ t > 2. \\ T_t = \beta(Y_t - Y_{t-1}) + (1-\beta)T_{t-1};\ t > 2. \\ Y_1 = X_1. \\ Y_2 = X_2. \\ T_2 = X_2 - X_1. \end{cases} \tag{4.29}$$

   tels que $0 < \alpha,\ \beta < 1$ les paramètres de lissage. La prédiction par la méthode de lissage exponentiel double est donnée par (4.30) :

$$X_{t+k} = Y_t + kT_t;\ \forall k = 1,2,...;\ t > n. \tag{4.30}$$

   La figure 4.31 représente les résultats de la méthode de lissage exponentiel double de Holt avec les paramètres $\alpha = 0.39,\ \beta = 10^{-4},\ RMSE - Train = 92.70,\ RMSE - Test = 73.67$.

   * **Tendance multiplicative : ETS(A, M, N)** est un lissage exponentiel double avec

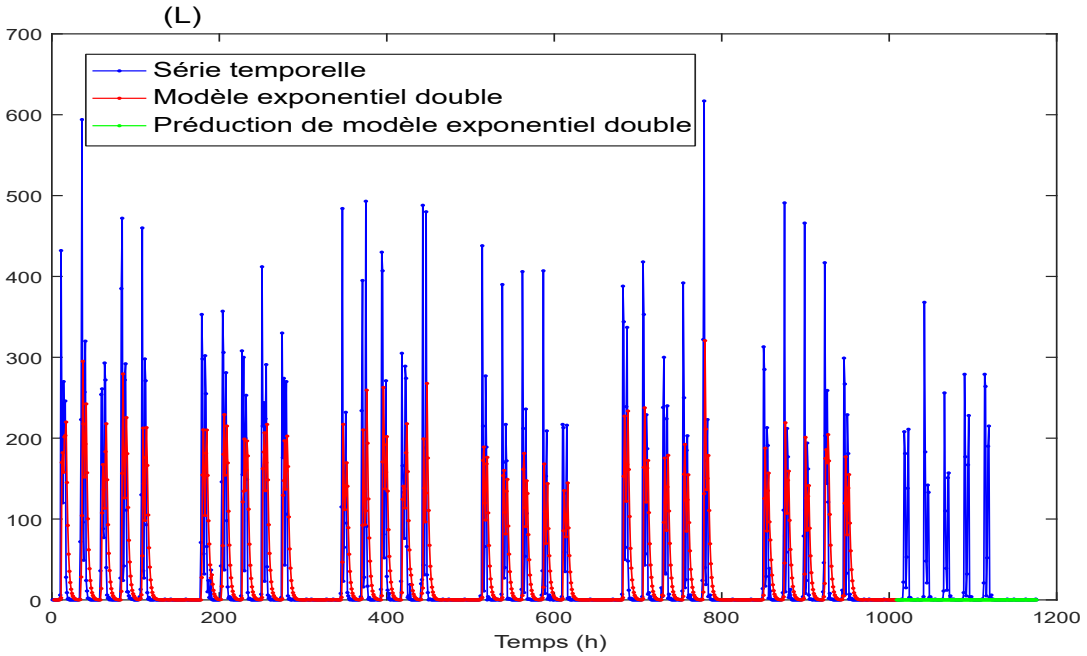

FIGURE 4.31 – Résultat du modèle exponentiel double ETS(A, A, N)

une tendance exponentielle.

$$\begin{cases} Y_t = \alpha X_t + (1-\alpha)(Y_{t-1} \times T_t);\ t > 2. \\ T_t = \beta(Y_t / Y_{t-1}) + (1-\beta)T_{t-1};\ t > 2. \\ Y_1 = X_1. \\ Y_2 = X_2. \\ T_2 = X_2 / X_1. \end{cases} \tag{4.31}$$

tels que $0 < \alpha,\ \beta < 1$ les paramètres de lissage. La prédiction par la méthode de lissage exponentiel double est donnée par (4.32) :

$$X_{t+k} = Y_t \times T_t^k;\ \forall k = 1, 2, ...;\ t > n. \tag{4.32}$$

2. **Le lissage exponentiel amorti**

   * **Amortissement additive : ETS(A, Ad, N)** amortit la tendance de manière linéaire.

$$\begin{cases} Y_t = \alpha X_t + (1-\alpha)(Y_{t-1} + \phi T_t);\ t > 2. \\ T_t = \beta(Y_t - Y_{t-1}) + (1-\beta)\phi T_{t-1};\ t > 2. \\ Y_1 = X_1. \\ Y_2 = X_2. \\ T_2 = X_2 - X_1. \end{cases} \tag{4.33}$$

tels que $0 < \alpha,\ \beta < 1$ sont des paramètres de lissage et ϕ est le facteur d'amorti. La

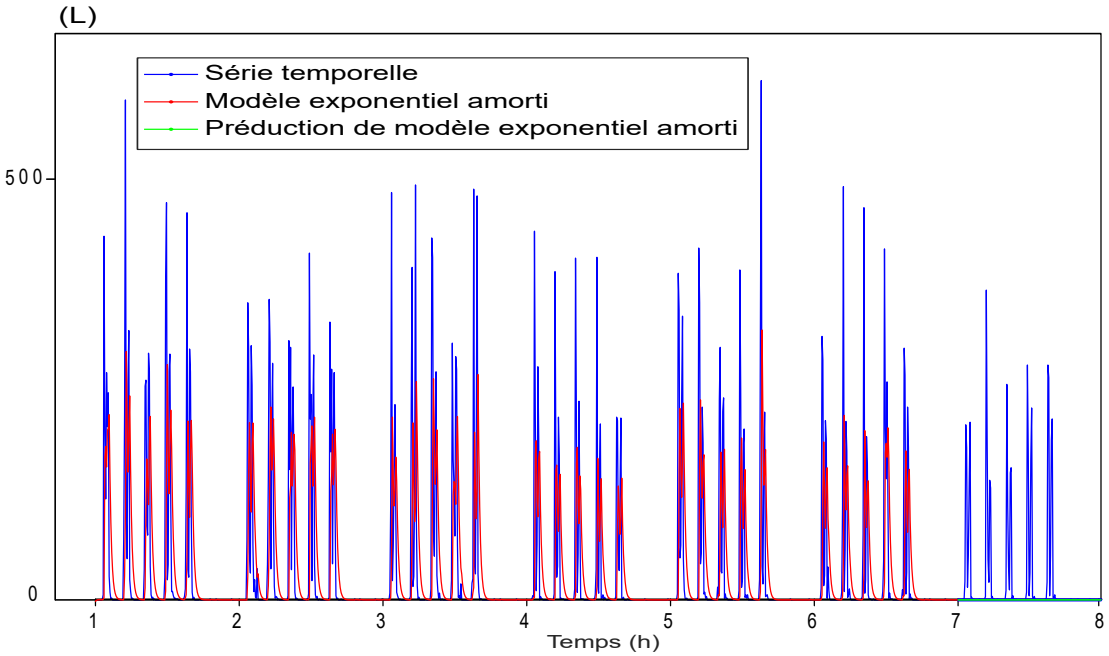

FIGURE 4.32 – Résultat du modèle exponentiel double ETS(A, Ad, N)

prédiction par la méthode de lissage exponentiel double est donnée par (4.34) :

$$X_{t+k} = Y_t + \sum_{i=1}^{k} \phi^i T_t;\ \forall k = 1,2,...;\ t > n. \tag{4.34}$$

La méthode exponentielle double amortie avec les paramètres $\alpha = 0.39,\ \beta = 10^{-4}$ et $\phi = 0.8$ est représentée dans la figure 4.32 avec $RMSE - Train = 92.69,\ RMSE - Test = 73.31$.

* **Amortissement multiplicative : ETS(A, Md, N)** amortit la tendance de manière exponentielle.

$$\begin{cases} Y_t = \alpha X_t + (1-\alpha)(Y_{t-1} \times \phi T_t);\ t > 2. \\ T_t = \beta(Y_t / Y_{t-1}) + (1-\beta)\phi T_{t-1};\ t > 2. \\ Y_1 = X_1. \\ Y_2 = X_2. \\ T_2 = X_2 - X_1. \end{cases} \tag{4.35}$$

tels que $0 < \alpha,\ \beta < 1$ sont des paramètres de lissage et ϕ est le facteur d'amorti. La prédiction par la méthode de lissage exponentiel double est donnée par (4.36) :

$$X_{t+k} = Y_t + T_t^{\sum_{i=1}^{k} \phi^i};\ \forall k = 1,2,...; t > n. \tag{4.36}$$

Lissage avec Holt-Winters

La méthode de Holt-Winters a été introduite par Holt et Winters dans [48].
La méthode de Holt-Winters compte sur l'estimation du niveau de la série désaisonnalisée a_t, de la pente de la tendance b_t et de la saisonnalité S_t. Elle permet de lisser la courbe de la série de l'observation de période p qui contient le terme de la tendance et la saisonnalité et de prédire le valeur future de la consommation [45], [48] et [46].

1. **Méthode de Holt-Winter additive :**
   Le modèle de prédiction est donné par :

   $$X_{n+k} = a_n + kb_n + S_{n+k-p}$$

   où :

   $$\begin{cases} a_n = \alpha(X_n + S_{n-p}) + (1-\alpha)(a_{n-1} + b_{n-1}). \\ b_n = \beta(a_n - a_{n-1}) + (1-\beta)b_{n-1}. \\ S_n = \gamma(X_n - a_n) + (1-\gamma)S_{n-p}. \end{cases}$$

   tels que α, β et γ sont des paramètres de lissage qui appartiennent à l'intervalle $]0,1[$, on les choisit en minimisant la somme des carrés des erreur de prédiction.
   Initialisation : pour $j = 1, ..., p$

   $$\begin{cases} a_j = \frac{\sum_{i=1}^{p} X_i}{p}. \\ b_j = \frac{1}{p}\sum_{i=1}^{p} \frac{X_{i+p} - X_i}{p}. \\ S_j = X_j - a_j. \end{cases}$$

2. **Méthode de Holt-Winter multiplicative :**
   Le modèle de prédiction est donné par :

   $$X_{n+k} = (a_n + k \times b_n) \times S_{n+k-p}$$

   Où :

   $$\begin{cases} a_n = \alpha(\frac{X_n}{S_{n-p}}) + (1-\alpha)(a_{n-1} + b_{n-1}). \\ b_n = \beta(a_n - a_{n-1}) + (1-\beta)b_{n-1}. \\ S_n = \gamma(\frac{X_n}{a_n}) + (1-\gamma)S_{n-p}. \end{cases}$$

   tels que α, β et γ sont des paramètres de lissage qui appartiennent à l'intervalle $]0,1[$ nous les choisissons en minimisant de la somme des carrés des erreur de prédiction.
   Initialisation : pour $j = 1, \ ..., \ p$

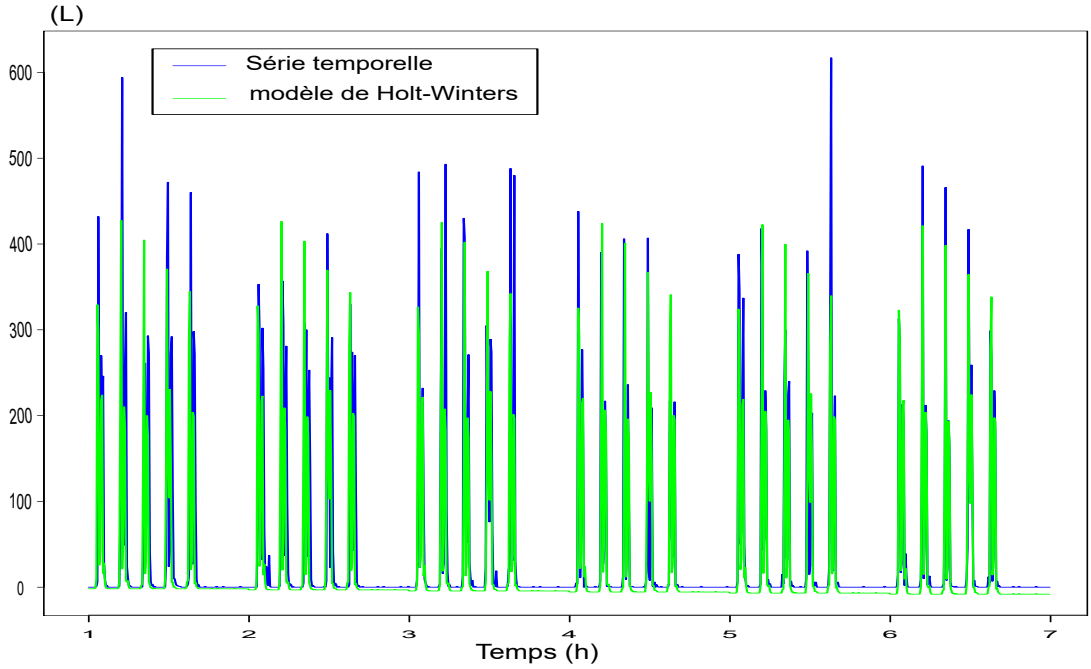

FIGURE 4.33 – Résultat de modèle de Holt-Winters additif

$$\begin{cases} a_j = \frac{\sum_{i=1}^{p} X_i}{p}. \\ b_j = \frac{1}{p}\sum_{i=1}^{p} \frac{X_{i+p} - X_i}{p}. \\ S_j = \frac{X_j}{a_j}. \end{cases}$$

Le choix des paramètres de modèle de Holt-Winters est fais automatiquement à l'aide du logiciel R avec la fonction HoltWinters. Les paramètres inconnus sont déterminés en minimisant l'erreur de prédiction au carré

Application

La courbe lissée avec la méthode de Holt-Winters est représentée sur la figure 4.33 en vert. Les meilleurs paramètres du modèle de Holt-Winters pour des données échantillonnées en heures de la série de l'exemple 4.15 sont : $\alpha = 1,\ \beta = 0.4479909, \gamma = 4.380685e-12$.

La prédiction avec la méthode de Holt-Winters représenté dans la figure 4.34 est validée par une erreur égale à 62.15 dans la partie d'entrainement et de 44.99 dans la partie test.

4.3.4 Réseaux de neurones artificiels

Un réseau de neurones artificiels est un système d'apprentissage automatique. Il s'est inspiré du fonctionnement des cerveaux humains. La théorie mathématique et informatique des réseaux de neurones a été développée par McCulloch et Pitts déjà en 1943 [43], avant qu'elle connaisse un engouement très important à l'heure actuelle.

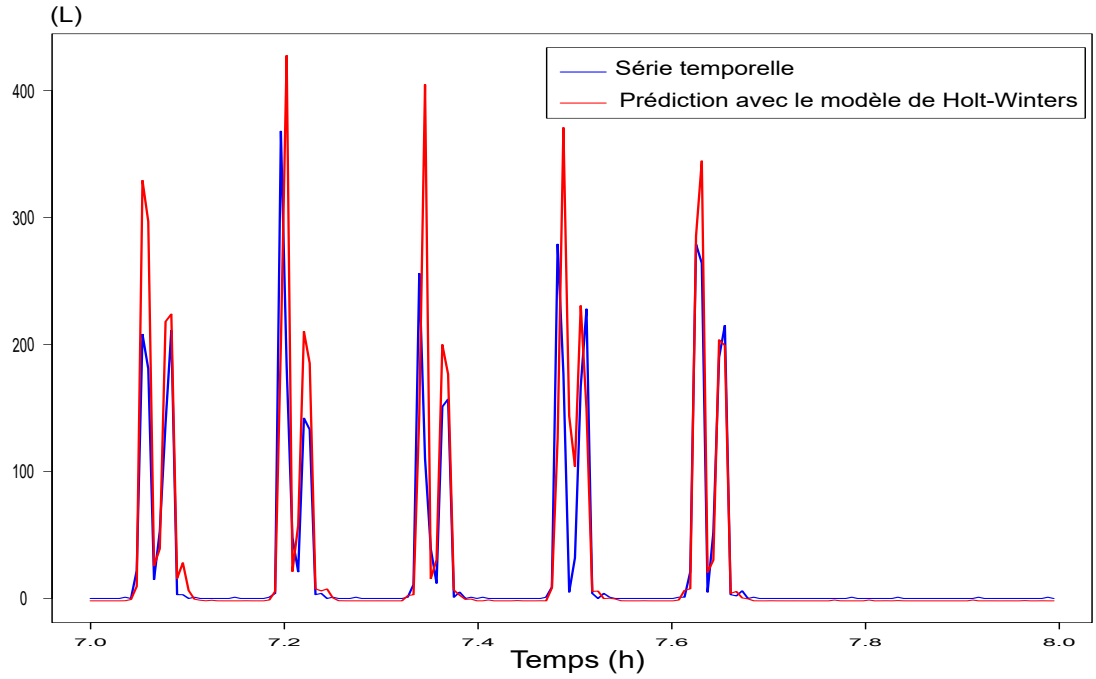

FIGURE 4.34 – La prédiction avec le modèle de Holt-Winters additif

Un réseau de neurones est un ensemble de neurones avec une propension naturelle à stocker des connaissances expérimentales et à les rendre utilisables [49]. Il présente les connaissances acquises par le réseau via un processus d'apprentissage qui peut être réalisé à travers un séquençage d'information ou un cycle récurrent. De plus, il est caractérisé par ces forces de connexion inter-neurones appelées poids synaptiques qui sont utilisées pour stocker les connaissances. En général, un réseau de neurones est constitué d'un ensemble de neurones interconnectés interagissant de manière non linéaire. La sortie de chaque neurone est une combinaison non linéaire de ses entrées et qui est définie en fonction de la nature et de la structure du réseau. Les réseaux de neurones artificiels ont été largement utilisés pour modéliser et prédire les séries chronologiques. Le plus grand avantage de ces réseaux est leur capacités à modéliser une relation non linéaire complexe sans avoir aucune hypothèse a priori sur la nature de la relation [43] et ceci uniquement à partir de données.

Les composants de réseau de neurones

Les composants principaux d'un réseau de neurones sont :

1. Les neurones : ensemble de fonctions.
2. Les couches (layer) : des groupements de neurones.
3. Les poids et les biais : des valeurs numériques.
4. La fonction d'activation : qui est une formule mathématique appliquée aux valeurs numériques en sortie d'un neurone artificiel. Il existe un grand nombre de fonctions d'activation, comme :
   * Sigmoïde (σ) :
   $$f(v_i) = \sigma(v_i) = (1 + \exp(-v_i))^{-1}.$$

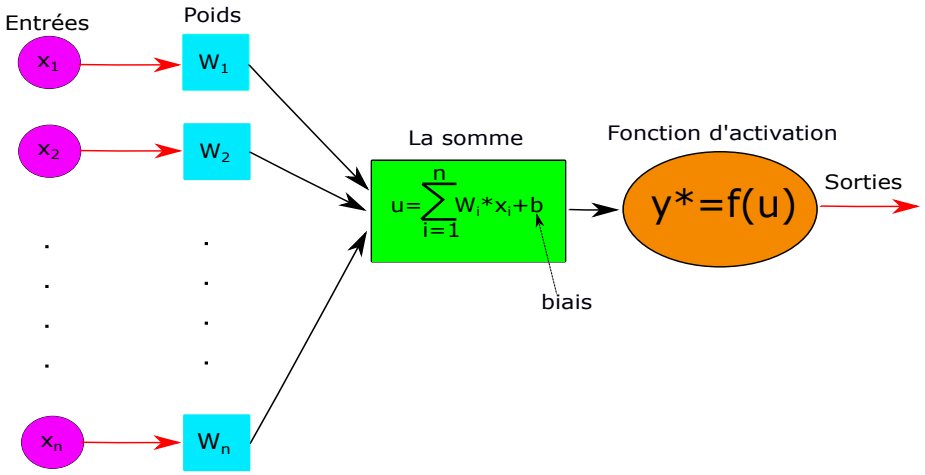

FIGURE 4.35 – Architecture d'un perceptron

* Fonction de tangente hyperbolique (tanh) :

$$f(v_i) = \tanh(v_i) = \frac{\exp(2v_i) - 1}{\exp(2v_i) + 1}.$$

* Unité linéaire rectifiée (ReLU) :

$$f(v_i) = \max(0,\ v_i);\ \forall v_i \in R.$$

Perceptron :
Le perceptron est un neurone qui reçoit les variables d'entrée $x_1, ..., x_n$, et qui calcule la sortie $y = f(\sum_1^n w_i \times x_i + b)$, avec w_i et b qui sont des poids et des biais. Le perceptron est le plus simple des types de réseau de neurone (figure 4.35). Il est considéré comme un algorithme d'apprentissage supervisé dans le domaine de l'apprentissage automatique.
L'algorithme du perceptron consiste à :

→ Initialiser aléatoirement les poids et les biais.

→ Calculer la sortie y^*.

→ Recalculer les poids et le biais par :

$$w_i = w_i + (y - y^*)x_i,\ b = b + (y - y^*).$$

Types de réseau de neurones : Nous distinguons différents types de réseaux de neurones.

1. **Réseau de neurones à propagation avant,** en anglais feedforward neural network est un réseau de neurone qui ne contient pas de cycles dans le réseau. Il se caractérise par le fait que l'information ne se déplace que dans une seule direction à partir des nœuds d'entrée vers les nœuds de sortie. Le plus connu est le perceptron multicouche.
   Perceptron multicouche (MLP) : est un réseau de neurones à propagation avant. C'est

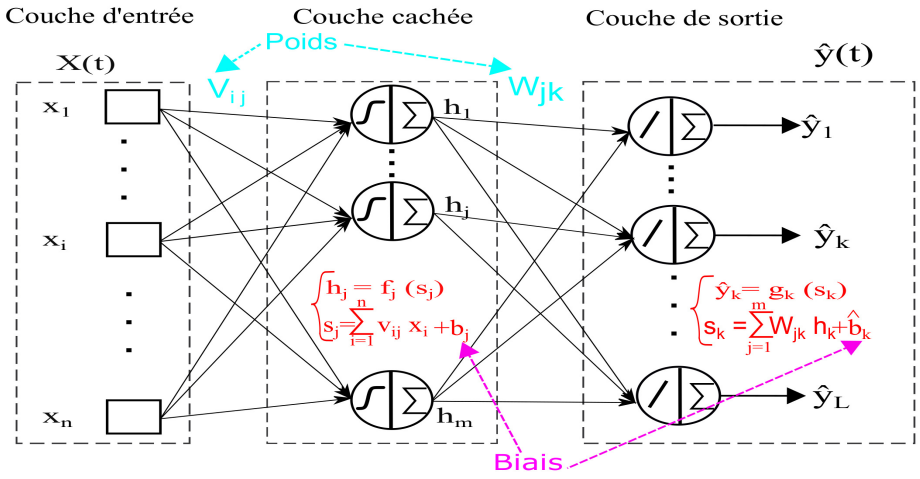

FIGURE 4.36 – Perceptron multicouche (MLP)

une extension du premier réseau de neurones artificiels, le perceptron. MLP est un réseau à propagation direct où l'information circule de la couche d'entrée vers la couche de sortie uniquement en passant par les couches cachées. Un MLP contient au moins trois couches. Il utilise une technique d'apprentissage supervisée appelée rétropropagation pour la mise à jour des poids [50].
Le vecteur d'entrée du MLP est constitué des échantillons d'une série chronologique précédente tel que : $X(t) = [x_1, x_2, ..., x_i, .., x_n]$. Le vecteur de couche cachée est $h = [h_1, h_2, ..., h_j, .., h_m]$ dont le calcul des sorties des mesures est décrit par :

$$\begin{cases} h_j = f_j(s_j);\ \forall j = 1, ...,\ m \\ s_j = \sum_{i=1}^{n} v_{ij} x_i + b_j, \end{cases} \tag{4.37}$$

avec j le numéro du neurone dans la couche cachée.
Le vecteur de sortie est $\hat{y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_k, .., \hat{y}_L]$ qui est décrit par :

$$\begin{cases} \hat{y}_k = g_k(s_k);\ \forall k = 1, ..., L \\ s_k = \sum_{j=1}^{m} w_{jk} h_k + \hat{b}_k, \end{cases} \tag{4.38}$$

tels que k soit le numéro du neurone de sortie.
$V = [v_1, v_2, ..., v_j, .., v_m]$ et $W = [w_1, w_2, ..., w_k, .., w_L]$ représentent respectivement la matrice de poids de la couche d'entrée à la couche cachée et la matrice de poids de la couche cachée à la couche de sortie (figure 4.36).
La rétropropagation est un généralisation de l'algorithme des moindres carrés moyens. C'est une méthode classique de correction des erreurs qui se base sur le calcul du gradient afin de converger vers les meilleurs poids de manière itérative (figure 4.37).
L'algorithme de rétropropagation :
1 - Initialiser les poids $w_{jk}^{(n)}$ et le biais $b_j^{(n)}$; $\forall n = 1, ...,$ nombre de couche.

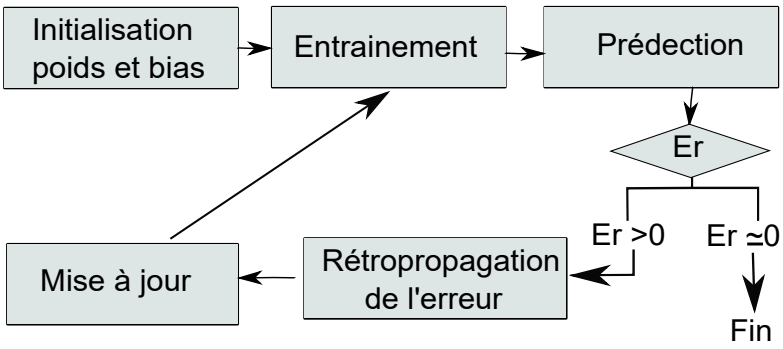

FIGURE 4.37 – Les grandes étapes de l'algorithme de rétropropagation

2 - Calculer $x_j^{(n)} = f^{(n)}(\sum_k w_{jk}^{(n)} \times x_k^{(n-1)} + b_j^{(n)}) = f(h_j^{(n)})$, tels que n soit le numéro de la couche, $f^{(n)}$ la fonction d'activation dans la couche n; $x_k^{(n)}$ le résultat du neurone j dans la couche n et $x_k^{(0)} = x_k$ sont les entrées du réseau de neurones.
3 - Calculer l'erreur $E_j = \frac{1}{2}(y_j - x_j^{(sortie)})^2$ avec y est la sortie recherchée.
4 - Si l'erreur est proche de zéro, on s'arrête sinon, nous passons à l'étape (5).
5 - Calculer les dérivés partielles de l'erreur par rapport aux $w_{jk}^{(n)}$ et $b_j^{(n)}$ en commençant par la grande valeur de n vers la petite.
6 - Mettre les poids et le bais à jour :

$$w_{jk}^{(n)} = w_{jk}^{(n)} + \lambda \frac{\partial E}{\partial w_{jk}} x_j^{(n-1)}.$$

où λ est le taux d'apprentissage.
7 - Recommencer par l'étape (2).

2. **Un réseau de neurones récurrent** est un réseau de neurones qui contient des cycles dans la structure présentant des connexions récurrentes. Il permet la conservation de l'information passée comme une sorte de mémoire. Nous l'illustrons par le schéma suivant (figure 4.38).
Cette architecture contient généralement deux type d'entrées pour un seul type de sortie au niveau de chaque neurone. Parmi les deux entrées il y en a une pour les données et une autre transmettant l'information du neurone voisin de la couche telle une mémoire. La sortie est unique mais à double intérêt. Elle sert à la sortie classique pour la couche suivante ainsi que de nouvel état de la mémoire pour le neurone voisin.
Réseau récurrent à mémoire court et long termes (LSTM) : est un type de réseau de neurones récurrents basé sur l'algorithme de rétropropagation à travers le temps. Il consiste à mémoriser des évènements passés [51]. Il est basé sur une architecture séquentielle et atypique décomposée en portes qui est très utilisé pour les problèmes de prédiction. Le LSTM est composé par des unités appelés les blocs mémoires. Chaque bloc mémoire contient une input gate, forget gate et output gate. Pour chaque neurone nous avons trois entrées ; une pour les données $x(t)$, l'autre appelée état de la cellule (mémoire) $h(t-1)$, ainsi que l'état caché $C(t-1)$ et deux sorties $h(t)$ et $C(t)$. Nous détaillons son

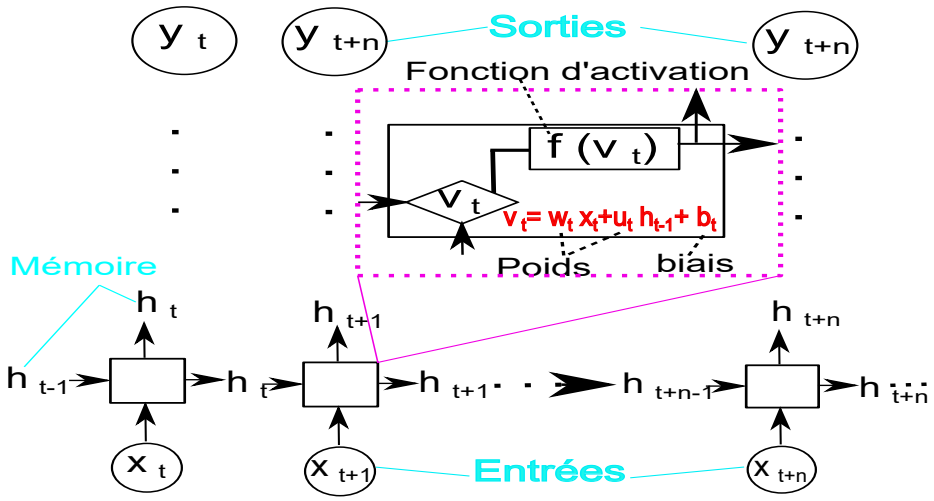

FIGURE 4.38 – Réseau de neurones récurrents

fonctionnement étape par étape :
Premièrement, l'étape d'entrée (input gate) est une étape d'extraction de l'information de la nouvelle donnée du neurone précédent. Elle est présentée par l'équation (4.39).

$$i(t) = \sigma(W_{i,1.}x(t) + W_{i,2.}h(t-1) + b_i). \tag{4.39}$$

tels que $W_{i,1.}, W_{i,2.}$ les poids, b_i le biais et σ une fonction sigmoïde.
Deuxièmement, l'étape de l'oubli (forget gate) est une étape de triage de l'information entre l'état caché et la nouvelle entrée de données dans l'optique de mémoriser les informations importantes. Elle est donnée par les deux équations (4.40) et (4.41). La mise à jour est effectuée à travers l'équation (4.41).

$$f(t) = \sigma(W_{f,1.}x(t) + W_{f,2.}h(t-1) + b_f). \tag{4.40}$$

$$\tilde{C}(t) = \tanh(W_{c,1.}x(t) + W_{c,2.}h(t-1) + b_c). \tag{4.41}$$

Troisièmement, l'étape de sortie (output gate) qui est une étape de calcul des sorties par (4.42) et (4.44) du neurone pour l'état caché par les équations suivantes :

$$C(t) = f(t) \times C(t-1) + i(t) \times \tilde{C}(t) \tag{4.42}$$

$$O(t) = \sigma(W_{o,1.}x(t) + W_{o,2.}h(t-1) + b_o). \tag{4.43}$$

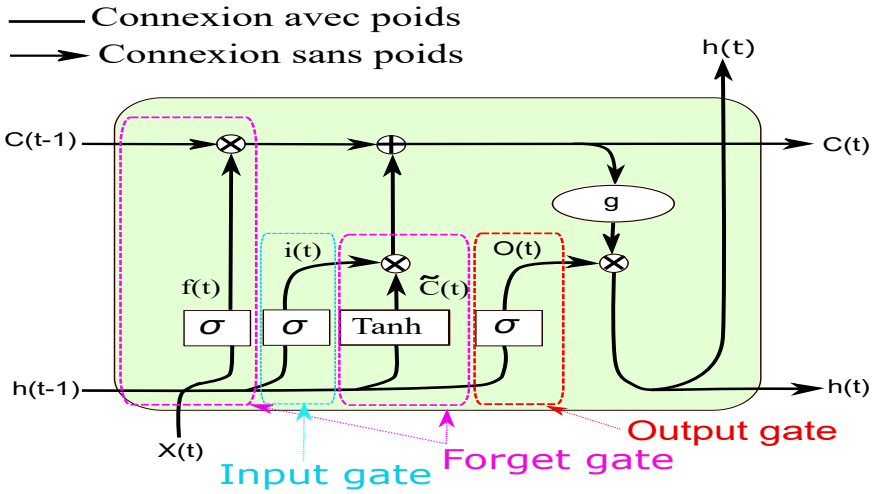

FIGURE 4.39 – Architecture interne d'une unité de réseau neuronal récurrent à mémoire à court et à long termes LSTM

$$h(t) = g(C(t)) \times O(t). \tag{4.44}$$

Les paramètres W et b représentent les poids et les biais respectivement et g est la fonction d'activation. La structure de ces portes est représentée par la figure 4.39.

Nous proposons de combiner 2 types de réseaux de neurones (MLP et LSTM) pour améliorer la prévision de la consommation d'eau.

Hybride de MLP et LSTM est la combinaison de deux types de réseaux de neurones LSTM et MLP utilisés pour prédire la consommation d'eau. Un modèle LSTM pour prédire les volumes d'eau horaires pendant les jours de semaine de fonctionnement [19]. Le deuxième modèle est un MLP pour prédire la consommation le week-end. Compte tenu de la nature du bâtiment à l'étude, la prédiction pendant les jours de week-end n'est que de petites fuites d'eau continues et des gouttes d'eau ingérables. Les paramètres des deux modèles sont sélectionnés sur la base d'une analyse empirique. Les données d'entrée sont divisées en deux ensembles, l'ensemble des jours de la semaine et l'ensemble des jours de week-end. Chaque ensemble comprend 7 semaines et il est divisé en deux sous-ensembles, l'un pour adapter et entraîner les paramètres et l'autre pour évaluer la prédiction. Les sous-ensembles d'entrainement consistent en 6 semaines représentant $5 \times 6 \times 24$ et $2 \times 6 \times 24$ heures respectivement pour les jours de semaine et les jours de week-end. Les sous-ensembles de test représentent une semaine de $2 \times 6 \times 24$ et $2 \times 1 \times 24$ heures respectivement pour les jours de semaine et les jours de week-end. Le nombre de neurones et le nombre de couches cachées ont été variés et différentes fonctions d'activation (comme sigmoïde, ReLu, Tanh ...) ont été testées pour MLP et LSTM. À la fin, les architectures des deux ANN sont configurées avec les paramètres du tableau 4.2.

La consommation prédite par l'ANN est sur la figure 4.40. Le modèle ANN est composé de deux modèles : Le MLP qui sert à prédire la consommation les jours de week-end et le LSTM est

TABLE 4.2 – Paramètres de LSTM et MLP pour la prévision de la consommation d'eau respectivement en semaine et en week-end

	LSTM	MLP
Nombre de couche cachée	2	1
Nombre de neurone	100/100	100
Fonction d'activation	relu/relu	relu

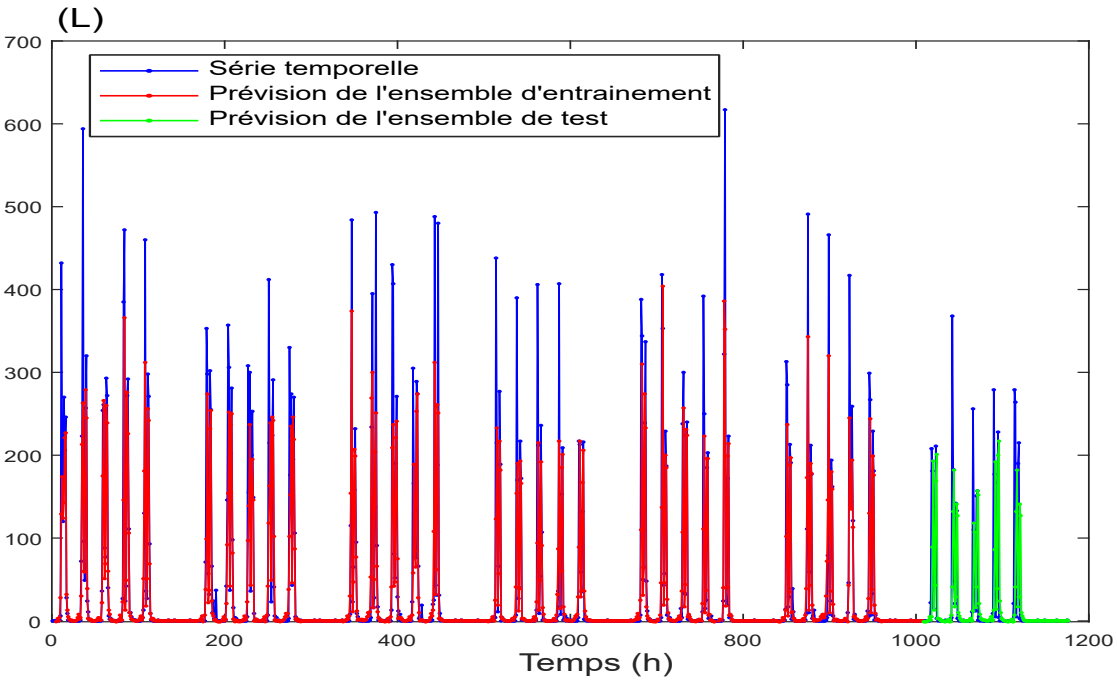

FIGURE 4.40 – Prévisions des résultats avec le modèle ANN composé du LSTM pour les jours de la semaine et du MLP pour les jours du week-end

utilisé pour prédire la consommation d'eau les jours de la semaine.

4.3.5 Modèles hybrides

Les modèles hybrides sont des modèles construits par la combinaison de plusieurs modèles individuels avec les paramétres déjà définis (Modèle déterministe, réseau de neurones, SARIMA et Holt-Winters) pour créer un meilleur modèle de prédiction des processus de consommation et pour obtenir une plus petite erreur dans la consommation prévue. Ces modèles hybrides se composent des éléments suivants :

1. *Hybride 1* combine les quatre modèles individuels.
2. *Hybride 2* combine le modèle déterministe de séries chronologiques et le modèle ANN.
3. *Hybride 3* combine le modèle déterministe de séries chronologiques et le modèle SARIMA.
4. *Hybride 4* combine le modèle déterministe de séries chronologiques et le modèle de

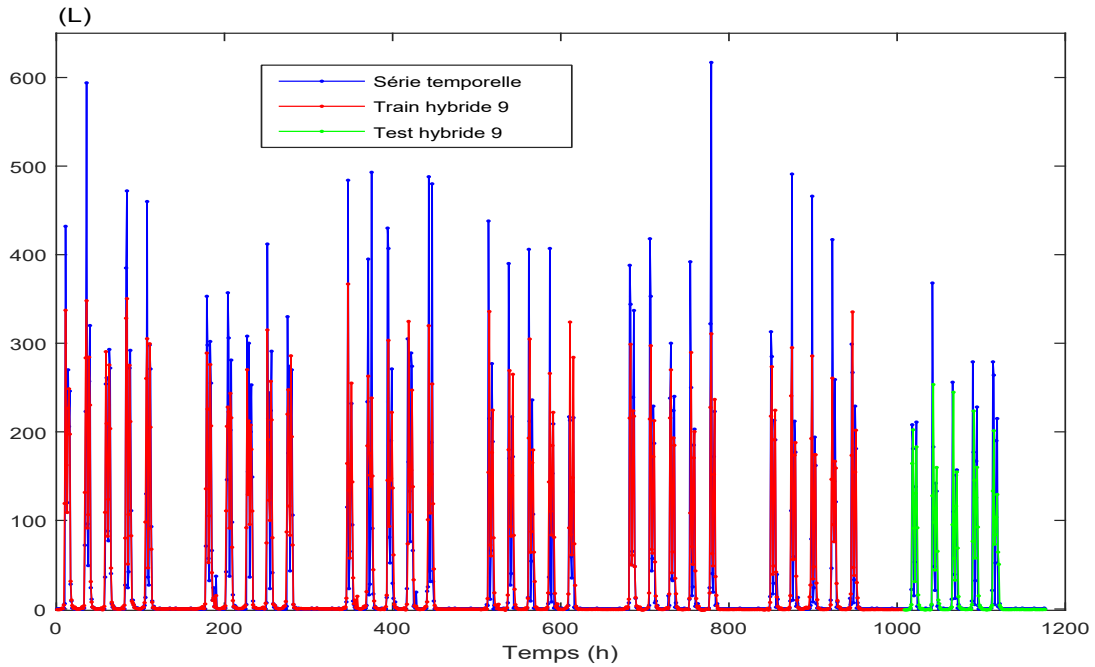

FIGURE 4.41 – Modèle hybride de modèle déterministe, ANN et le modèle de Holt-Winters

Holt-Winters.

5. *Hybride 5* combine les modèles de ANN et SARIMA.
6. *Hybride 6* combine les modèles de ANN et Holt-Winters.
7. *Hybride 7* combine les modèles de Holt-Winters et SARIMA.
8. *Hybride 8* combine le modèle déterministe de séries chronologiques, les modèles ANN et SARIMA.
9. *Hybride 9* combine le modèle déterministe de séries chronologiques, les modèles ANN et Holt-Winters.
10. *Hybride 10* combine le modèle déterministe de séries chronologiques, les modèles SARIMA et Holt-Winters.
11. *Hybride 11* combine les modèles de Holt-Winters, ANN et SARIMA.

Les résultats des modèles hybrides sont obtenus en faisant la moyenne de la combinaison des modèles individuels.
La figure 4.41 montre les résultats obtenu par le modèle hybride 9.

4.3.6 Prévision hebdomadaire des séries chronologiques

La prévision de la consommation d'eau
Les données de séries chronologiques pour la consommation horaire d'eau présentées dans l'exemple 4.15 et les résultats de prévision de la consommation d'eau sont détaillées pour

chaque méthode. La précision de tous les modèles est évaluée par l'erreur quadratique moyenne (RMSE) [52] calculée avec :

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t - X_t^*)^2} \tag{4.45}$$

où X_t est la valeur souhaitée et X_t^* la valeur prédite par chaque modèle.

Table 4.3 – Comparaison des modèles proposés

	Train RMSE	**Test RMSE**	**Mean RMSE**
Déterministe	57.71	35.75	46.73
ANN	51.19	76.42	63.8
SARIMA	59.3	57.19	58.24
Déterministe non paramétrique	52.19	17.13	34.66
Holt-Winters	56.74	44.99	50.82
Hybride 1	43.25	40.65	41.95
Hybride 2	43.76	50.89	47.32
Hybride 3	49.31	41.47	45.39
Hybride 4	49.17	35.07	42.13
Hybride 5	42.18	48.92	45.55
Hybride 6	42.66	43.01	42.59
Hybride 7	55.94	50.78	53.36
Hybride 8	41.88	43.89	42.88
Hybride 9	42.19	40.07	41.13
Hybride 10	49.28	41.51	45.40
Hybride 11	44.89	43.58	44.23

Le tableau 4.3 comprend les valeurs RMSE pour les ensembles d'entrainement et de test ainsi que leurs valeurs moyennes. Dans ce tableau, Train RMSE reflète les capacités générales des modèles sur les données d'apprentissage. Alors que Test RMSE montre les capacités d'interpolation des modèles sur les données de test.
A l'aide des résultats donnés dans le tableau 4.3 on peut voir que le modèle hybride 9 représenté dans la figure 4.41 et qui combine les modèles déterministes, le modèle de Holt-Winters et le modèle hybride composé du LSTM et du MLP donne des prédictions plus précises avec le minimum d'erreur par rapport aux autres modèles hybrides. Mais la plus petite erreur dans le tableau est donnée par le modèle déterministe non paramétrique. Donc, le résultat du modèle déterministe estimé par les moyennes mobiles est efficace en améliorant la précision de la prévision de la consommation d'eau. L'article [53] a résumé les résultats de la prévision de la consommation d'eau au restaurant universitaire.

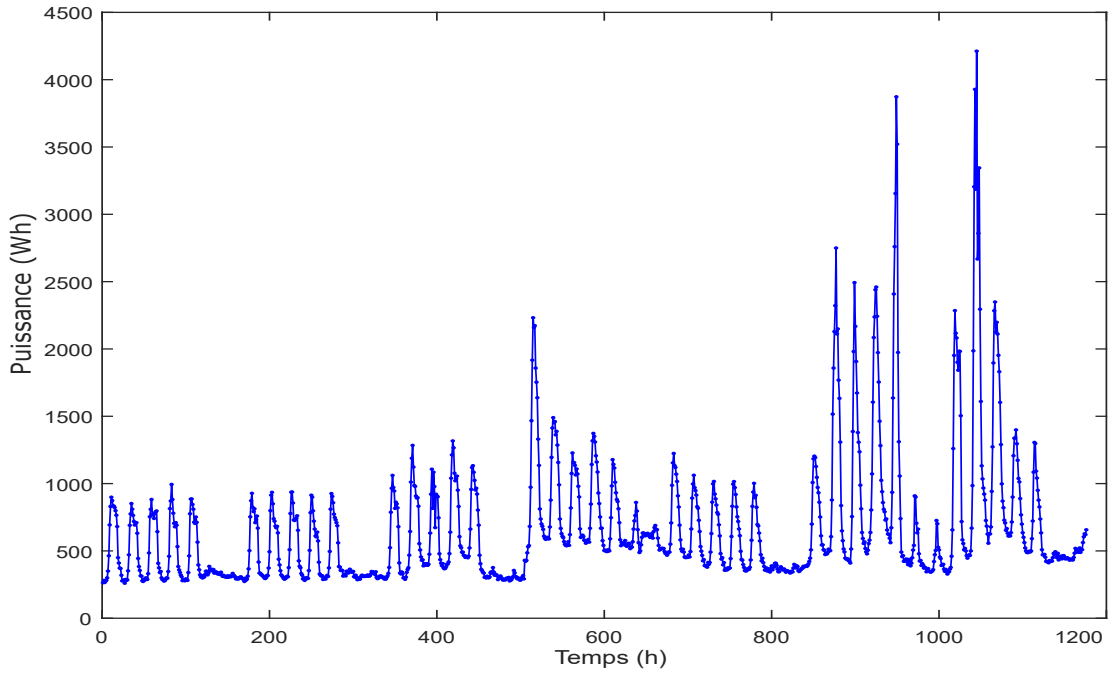

FIGURE 4.42 – Série chronologique de la consommation horaire d'électricité.

La prévision de la consommation d'électricité

Maintenant, pour montrer les performances des modèles appliqués aux données sur l'eau, nous les appliquons à d'autres bases de données comme les données de consommation d'électricité. Ainsi la série temporelle représentée par la figure 4.42 contient 7 semaines de données échantillonnées en heure de consommation électrique dans un bâtiment tertiaire. Nous prenons 6 semaines pour l'entrainement et 1 semaine pour le test.

Les trois critères de choix de modèle déterministe représentés dans la figure 4.43 sélectionne le modèle multiplicatif.

L'estimation de la tendance par la méthode des moindres carrés dans le modèle déterministe multiplicatif est donnée par :

$$Z_t = 0,39t + 422,52.$$

La prédiction avec le modèle déterministe paramétrique est fournie avec le modèle multiplicatif et elle est représentée par la figure 4.44. Bien que l'estimation non paramétrique et la prévision du modèle déterministe multiplicatif soient représentés dans la figure 4.45.

La figure 4.46 représente les résultats du modèle SARIMA$(1,1,0)(0,1,0)_{168}$, qui est donné par :

$$(0.38\ B)(I-B)(I-\ B^{168})X_t = \epsilon_t. \tag{4.46}$$

Un modèle LSTM est défini pour la consommation d'énergie horaire dans la figure 4.47. Ce modèle prend une entrée et une seule sortie. La normalisation est appliquée pour simplifier le calcul au niveau du modèle et pour réduire l'espace de variation de chaque élément de la série chronologique. Les entrées sont définies avec une fenêtre glissante qui couvre l'ensemble des

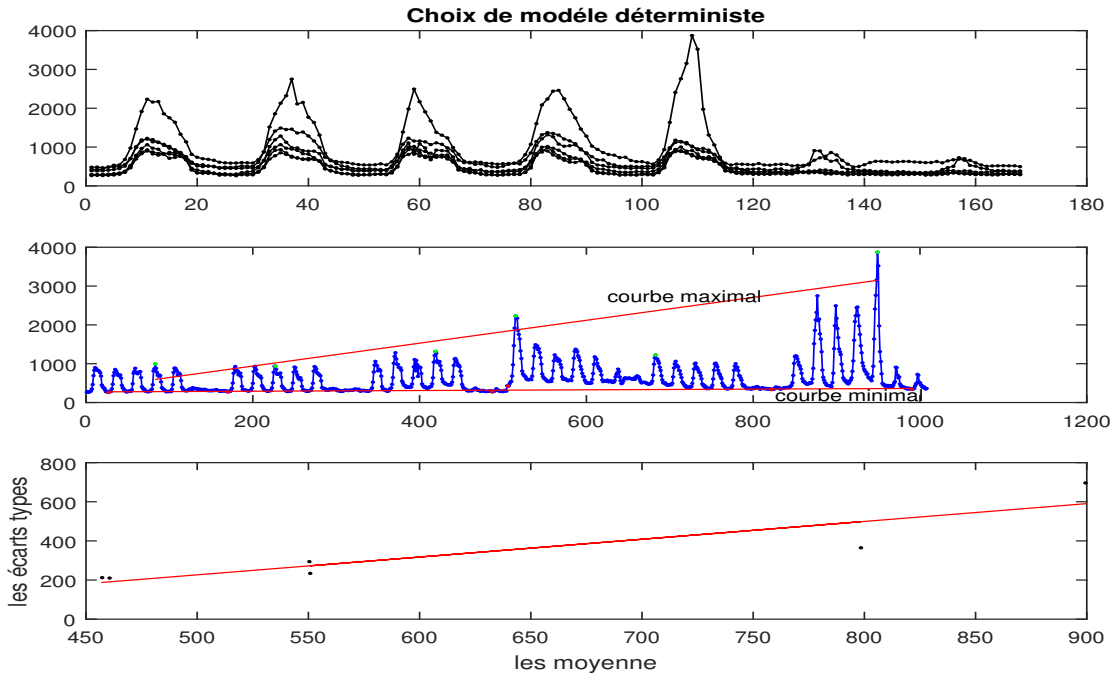

FIGURE 4.43 – Les grandeurs utilisées pour faire un choix de modèle déterministe pour la série temporelle de l'électricité.

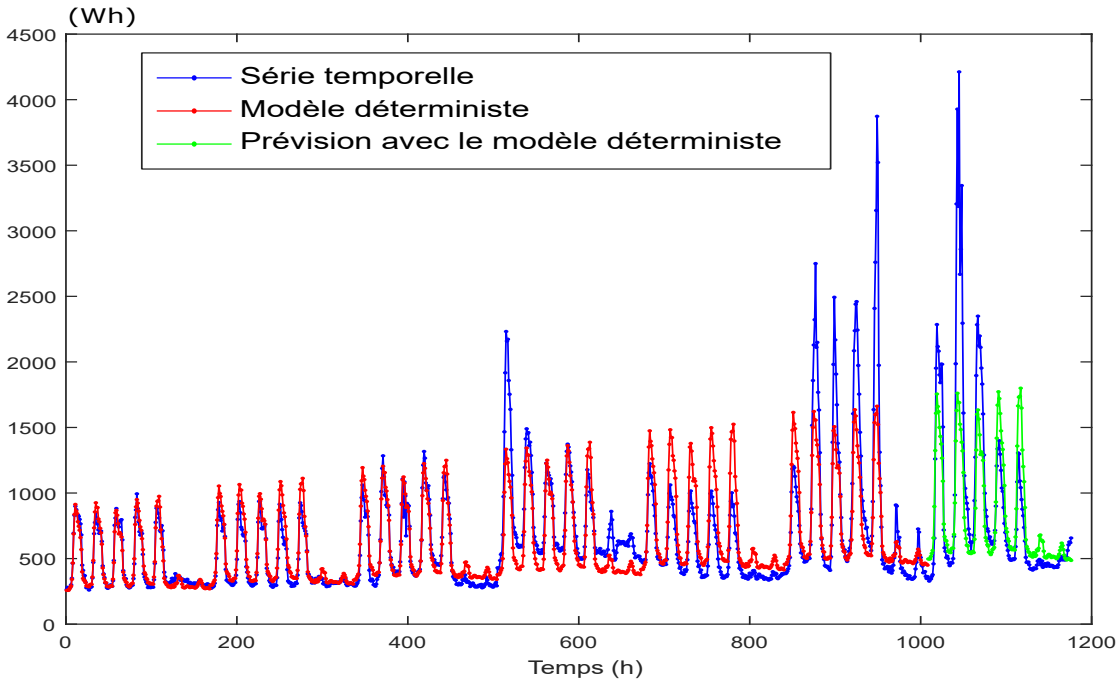

FIGURE 4.44 – Résultats de prédiction du modèle déterministe multiplicatif des prévisions de l'électricité.

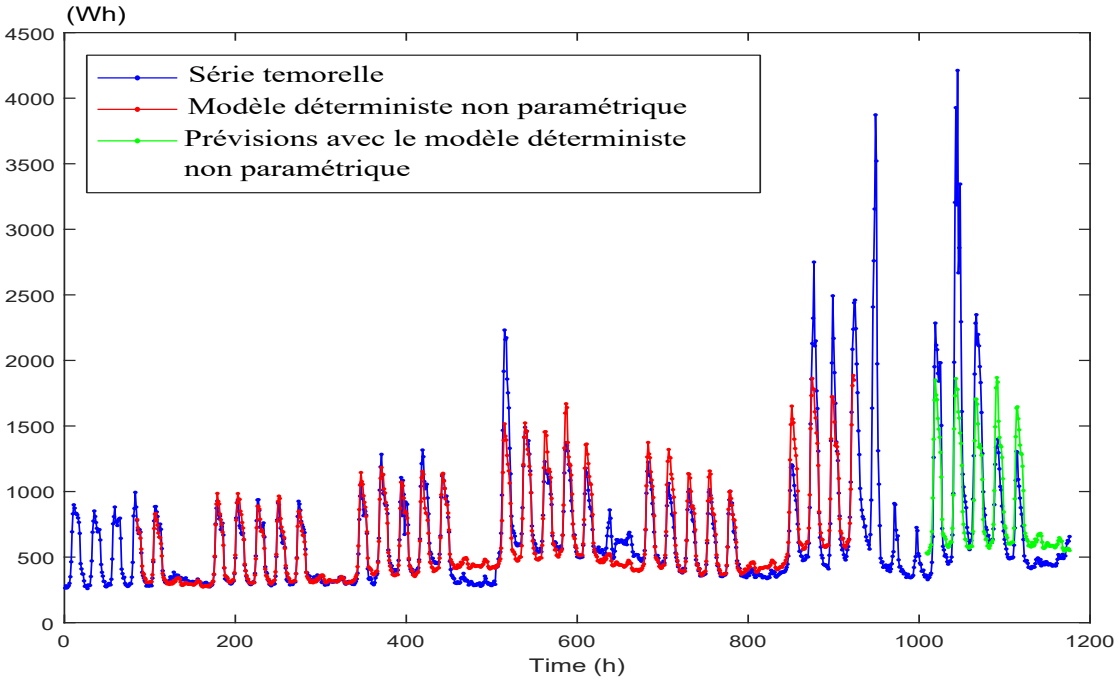

FIGURE 4.45 – Résultats du modèle multiplicatif déterministe non paramétrique pour la prévision de la consommation d'électricité.

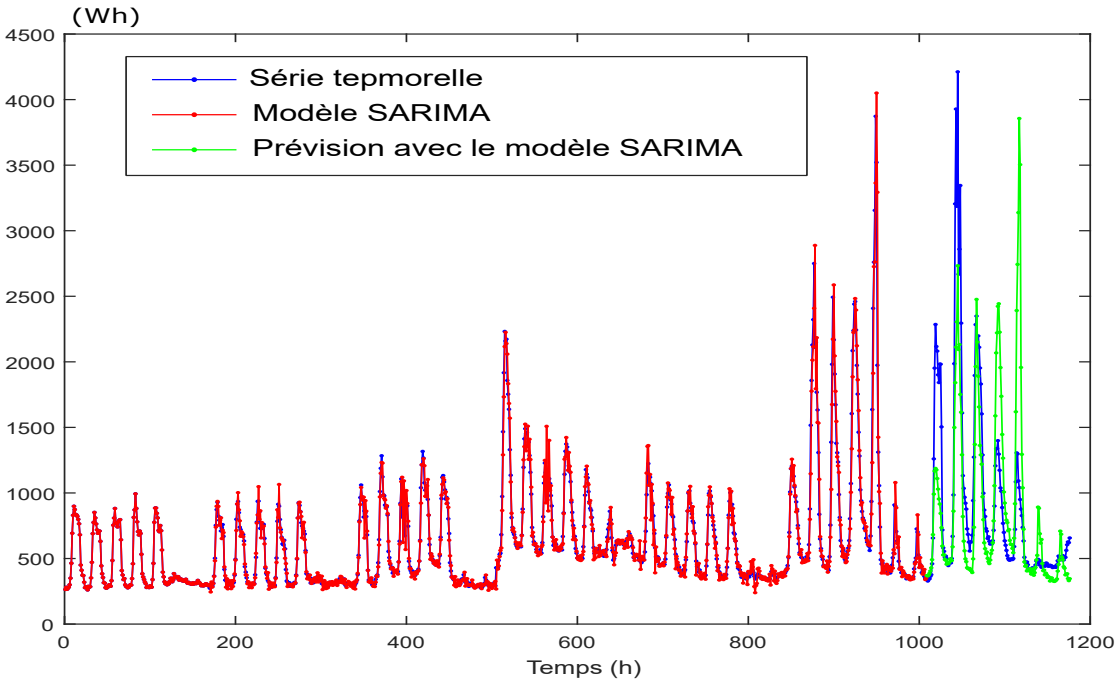

FIGURE 4.46 – Résultats de prévisions de puissance du modèle SARIMA (1,1,0) (0,1,0) [168].

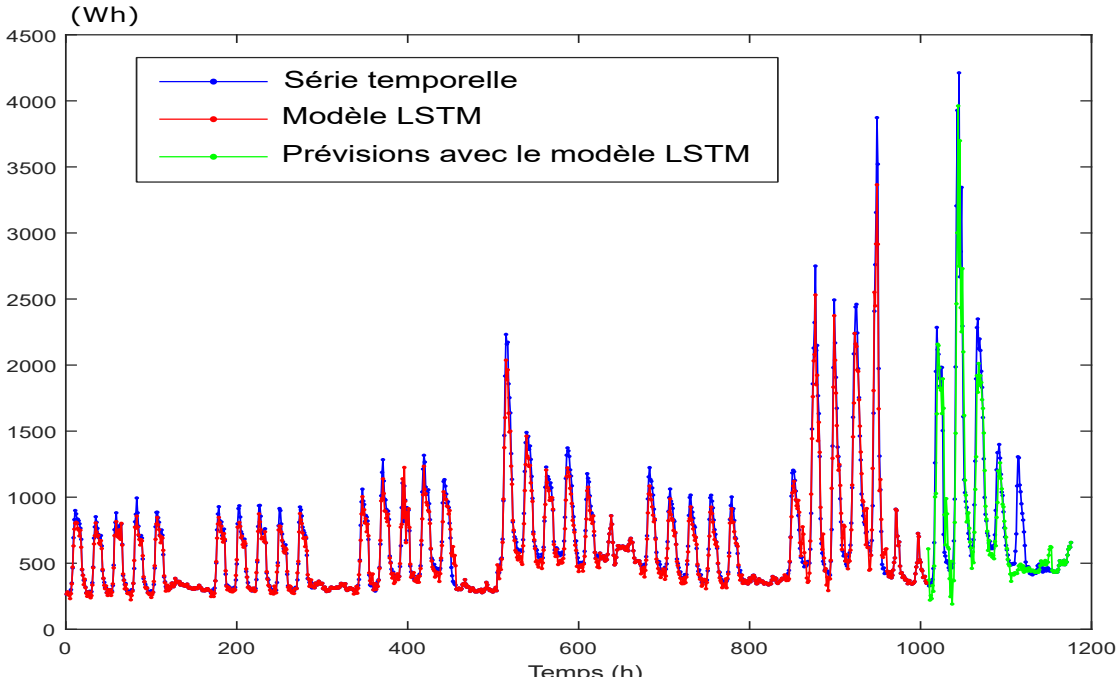

FIGURE 4.47 – Résultats du modèle ANN pour les séries chronologiques de consommation électriques.

données. Il est composé d'une unité LSTM avec 100 neurones et une fonction d'activation ReLu.
Les paramètres des modèles Holt-Winters multiplicatifs représentés par la figure 4.48 sont donnés par : $\alpha = 0.93,\ \beta = 0,\ \gamma = 1$.
La figure 4.49 représente les résultats de l'application du modèle hybride 5 à la prévision de la consommation électrique.
Le tableau 4.4 représente les erreurs de prévision de la série temporelle électrique avec different méthodes.
Selon les résultats dans le tableau 4.4, le modèle hybride 5 qui combine les ANN et le modèle SARIMA offre la plus petite erreur de prévision. Nous nous en servons donc pour prédire la consommation électrique.

4.4 Modèlisation des courbes de charge journalières

La modélisation des courbes de charge est une tâche difficile en raison de la diversité des courbes de charge pour des jours sélectionnés qui représente à la fois la non-coïncidence de la consommation et la variété illimitée des caractéristiques des utilisateurs.
Nous avons utilisé les méthodes numériques dans la section 4.2 pour construire la courbe de charge de la consommation d'eau du 25/05/2018 et la courbe de charge électrique du 28/01/2020. Maintenant, nous trouvons les mêmes courbes de charge en utilisant les séries temporelles.

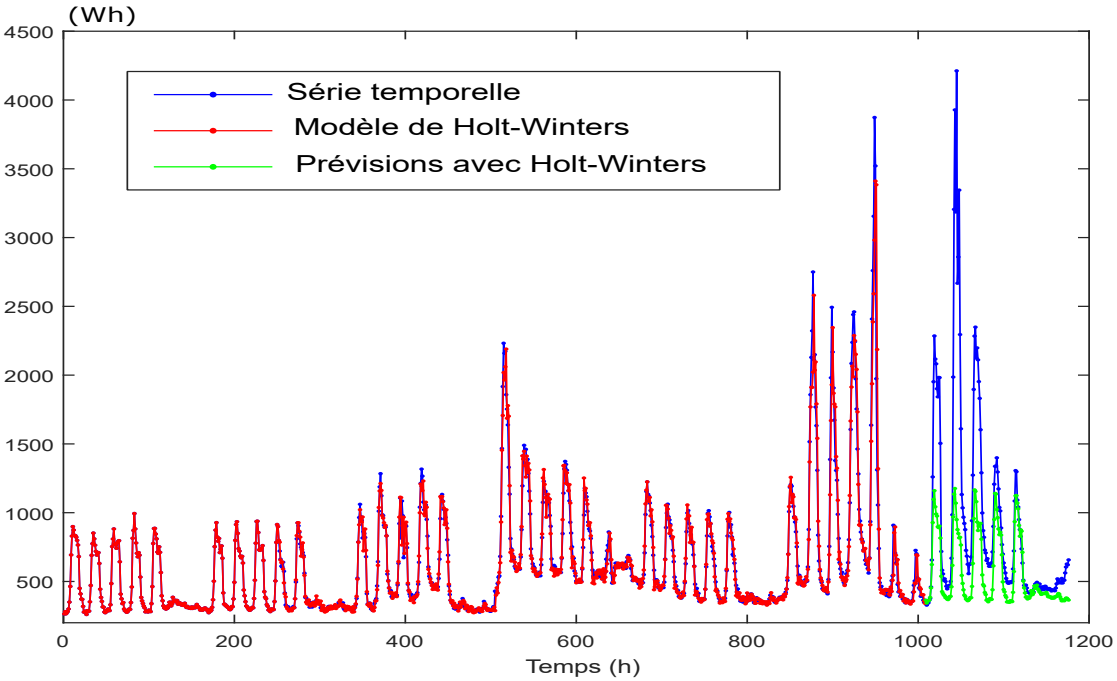

FIGURE 4.48 – Résultats des prévisions par le modèle multiplicatif Holt-Winters pour les séries chronologiques d'électricité.

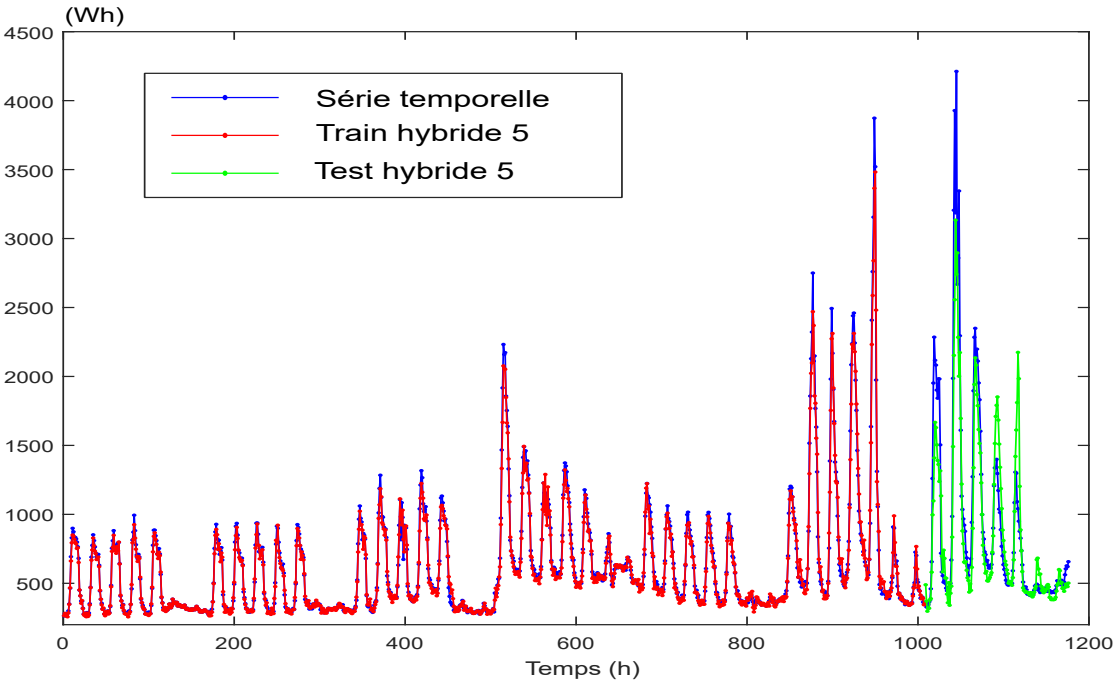

FIGURE 4.49 – Résultats des prévisions d'électricité par le modèle hybride 5 des modèles ANN et SARIMA

TABLE 4.4 – Évaluation prévisionnelle des modèles ANN, SARIMA, déterministes et hybrides avec la mesure RMSE

	Électricité RMSE		
	Train RMSE	**Test RMSE**	**Mean RMSE**
Déterministe	213.79	461.49	337.64
ANN	75.57	311.03	193.3
SARIMA	90.28	579.54	334.91
Déterministe non paramétrique	130.67	477.57	304.12
Holt-Winters	95.78	644.68	370.24
Hybride 1	81.01	406.24	243.62
Hybride 2	125.31	310.3	217.8
Hybride 3	118.13	491	304.57
Hybride 4	128.34	537.03	332.68
Hybride 5	56.17	325.62	190.90
Hybride 6	62.29	409.57	235.93
Hybride 7	86.34	539.61	312.97
Hybride 8	89.24	354.03	221.64
Hybride 9	96.39	406.12	251.26
Hybride 10	99.16	507.52	303.34
Hybride 11	62.29	396.03	229.16

4.4.1 Applications

1- La série temporelle de la consommation d'eau contient des données échantillonnées en heure sur 78 jours du 19 janvier 2018 au 17 juin 2018 sauf les week-ends, les vacances universitaires et les jours fériés dans un restaurant universitaire.
Dans les modèles avec des résidus des séries temporelles, nous prenons 80 % des données échantillonnées par heure pour l'entraînement des modèles et 20 % pour les tests.
La figure 4.50 représente l'estimation paramétrique du modèle de prédiction déterministe de la première CdC dans la partie test le 25 mai 2018.
La figure 4.51 représente la CdC estimée par le modèle non paramétrique classique le même jour.
Le modèle SARIMA $(2,0,2)(2,1,1)_{24}$ est le meilleur modèle qui minimise les critères d'information (AIC et BIC). Il est représenté dans la figure 4.52 et il se définit par l'équation (4.47).

$$(0.12\ B - 0.54\ B^2)(0.04\ B + 0.12\ B^2)(I - B^{24})X_t = (-0.19\ B + 0.36\ B^2)(-0.76\ B)\epsilon_t. \quad (4.47)$$

La figure 4.53 illustre la courbe de charge prédite par la méthode de Holt-Winters additive avec les paramètres $\alpha = 0.002$, $\beta = 0.09$ et $\gamma = 0.39$.
Les prédictions faites par le modèle qui combine entre le modèle déterministe et SARIMA

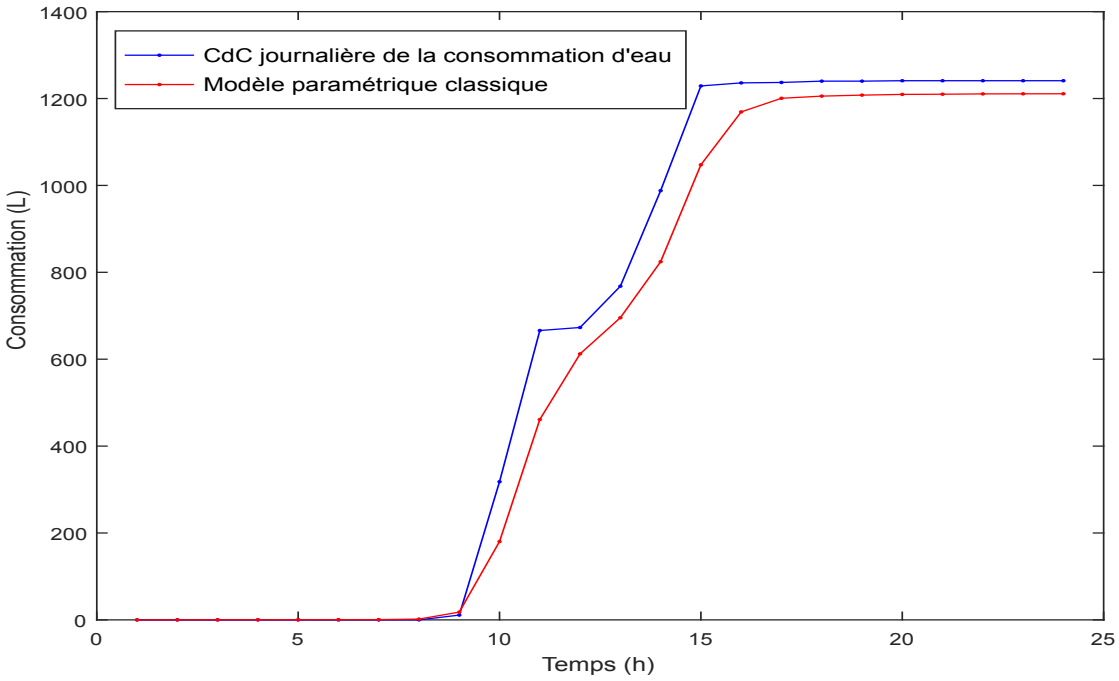

FIGURE 4.50 – L'estimation de la courbe de charge d'eau par le modèle paramétrique classique estime la CdC d'eau

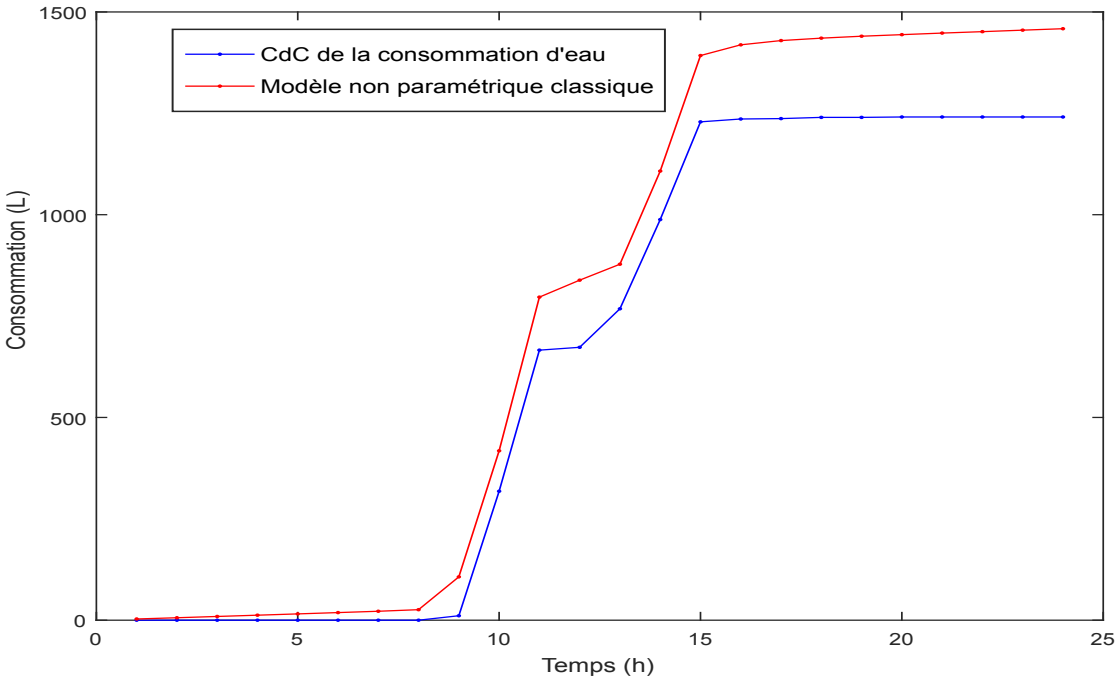

FIGURE 4.51 – L'estimation de la courbe de charge d'eau par le modèle non paramétrique classique.

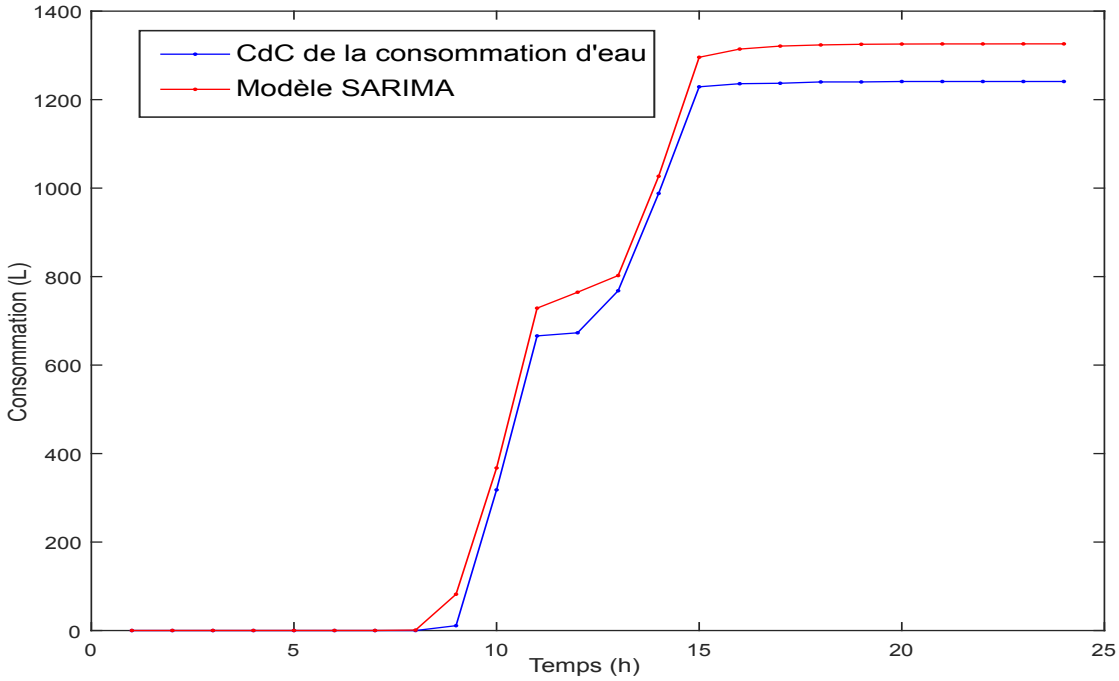

FIGURE 4.52 – La courbe de charge d'eau estimée par le modèle SARIMA

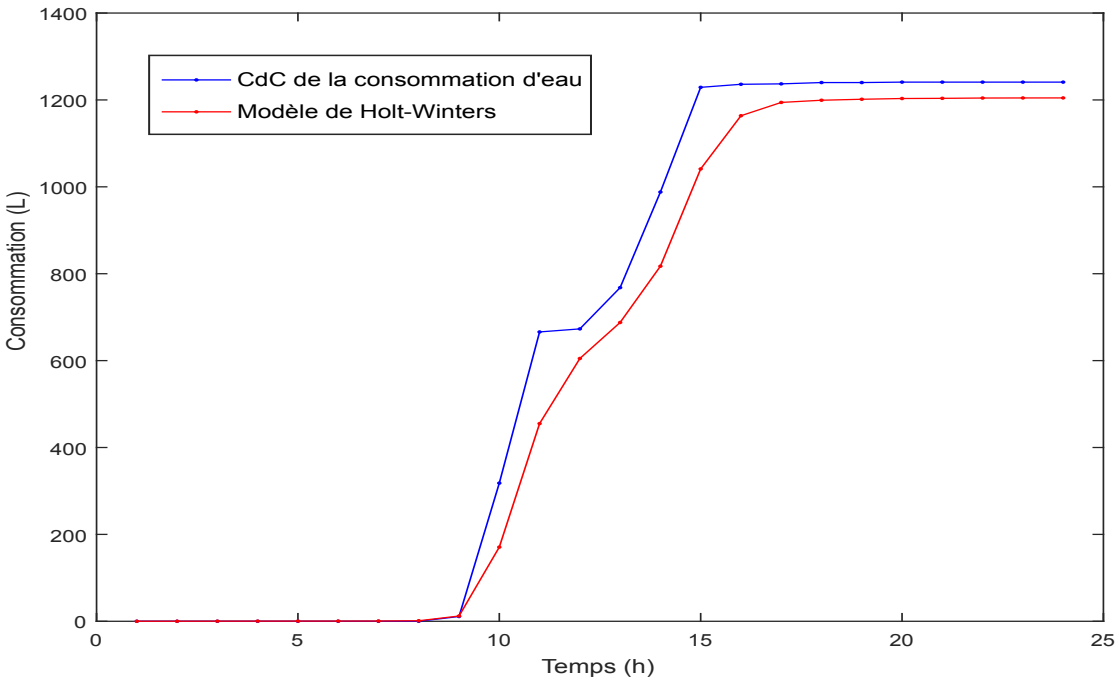

FIGURE 4.53 – La courbe de charge d'eau estimée par le modèle de Holt-Winters

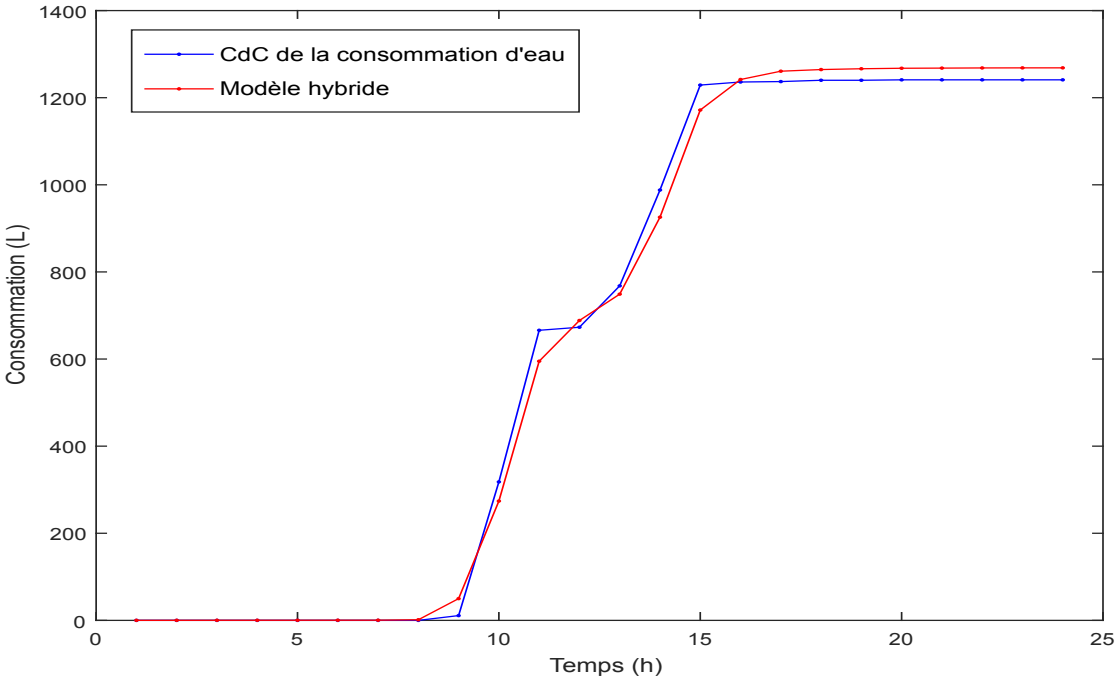

FIGURE 4.54 – La courbe de charge d'eau estimée par le modèle hybride

$(2,0,2)(2,1,1)_{24}$ sont données par la figure 4.54.

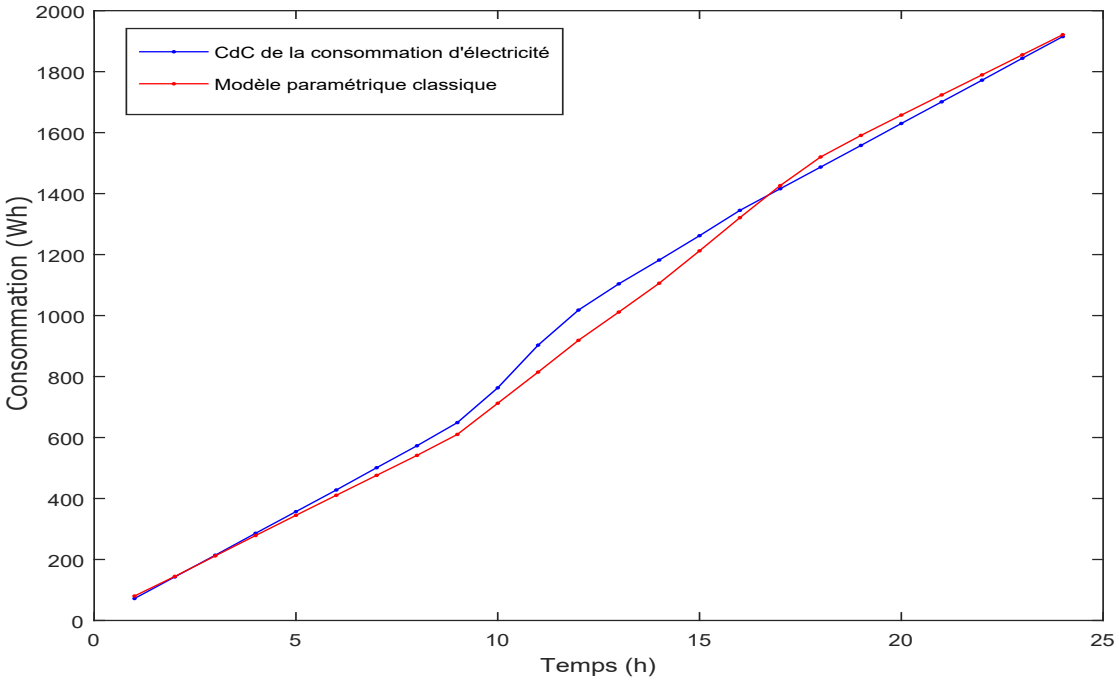

FIGURE 4.55 – CdC électrique estimée par le modèle paramétrique classique

2- Pour la puissance électrique nous avons 42 courbes de charge électrique qui clarifient la puissance consommée dans un bureau des doctorants. La figure 4.55 représente la prédiction du modèle déterministe additif pour estimer paramétriquement la première courbe de charge dans la partie test le 28 janvier 2020.
La figure 4.56 représente la prédiction du modèle non paramétrique classique le même jour.
Le modèle SARIMA $(1,1,1)(2,1,0)_{24}$ représenté sur la figure 4.57 est défini par l'équation (4.48).

$$(-0.62\ B)(0.65\ B + 0.29\ B^2)(I - B)(I - B^{24})X_t = (0.16\ B)\epsilon_t. \tag{4.48}$$

La figure 4.58 représente la courbe de charge prédite par la méthode de Holt-Winters additif des paramètres $\alpha = 0.94$, $\beta = 0.01$ et $\gamma = 1$.
La figure 4.59 représente le modèle hybride des modèles classiques paramétrique et SARIMA $(1,1,1)(2,1,0)_{24}$.

4.4.2 Discussion

Dans cette partie, nous allons évoquer quelques avantages et inconvénients de chaque méthode utilisée pour estimer la courbe de charge et comparer ces méthodes à l'aide de 3 critères (erreur, temps de calcul et nombre de paramètres).

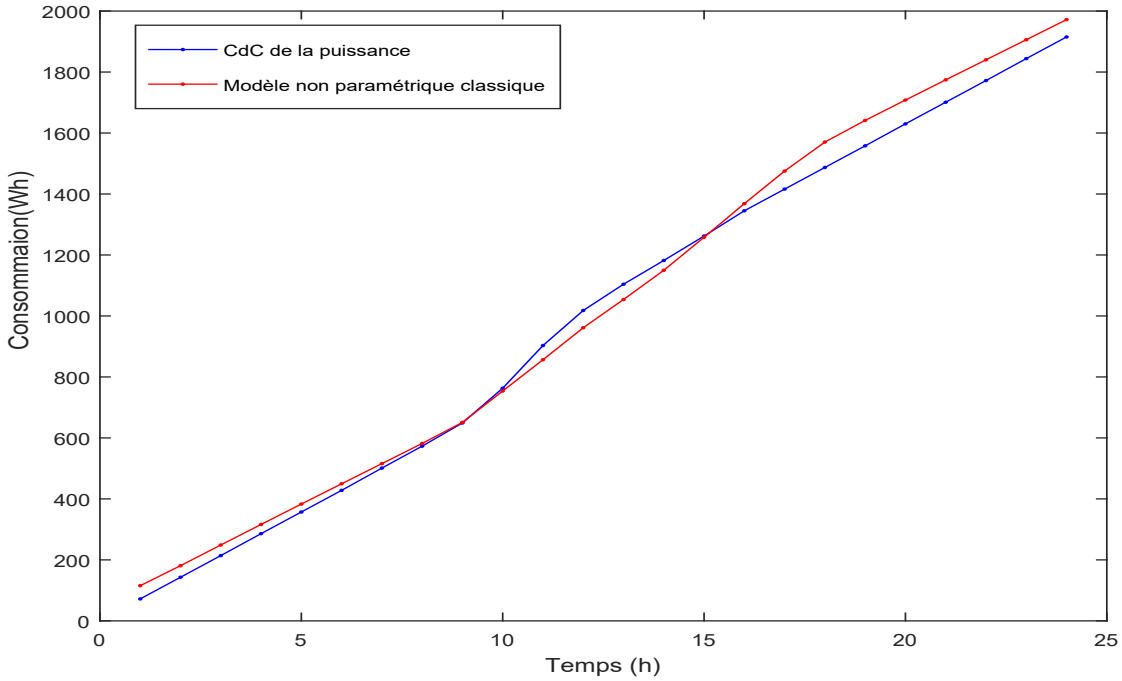

FIGURE 4.56 – CdC électrique estimée par le modèle non paramétrique classique

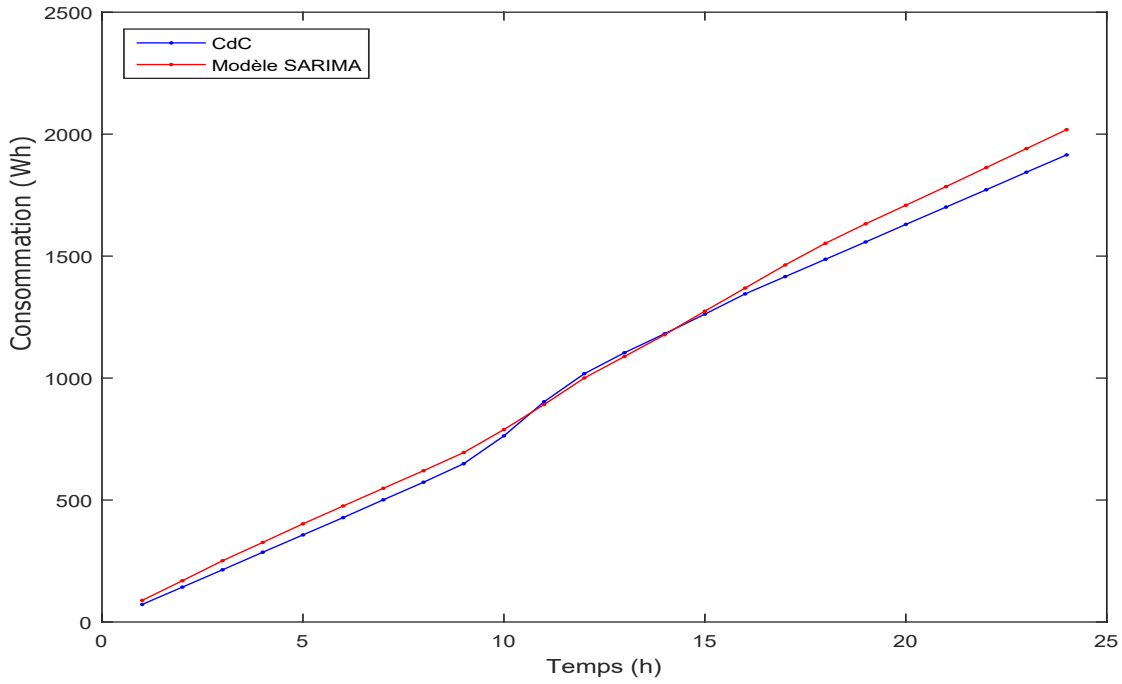

FIGURE 4.57 – CdC électrique estimée par le modèle SARIMA

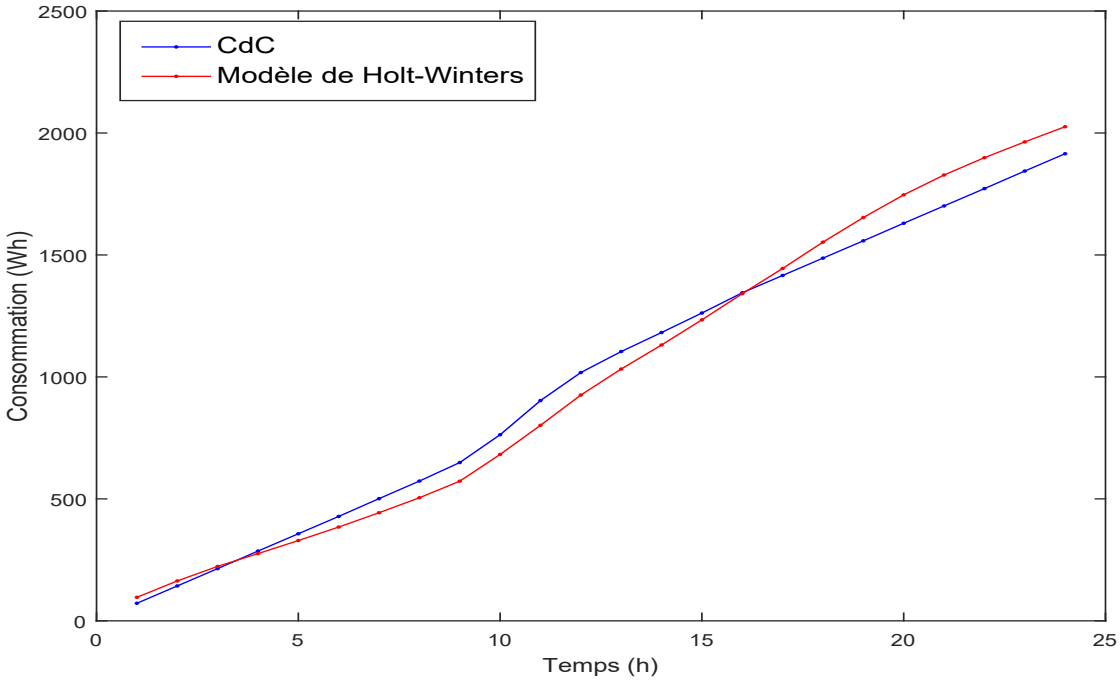

FIGURE 4.58 – CdC électrique estimée par le modèle de Holt-Winters

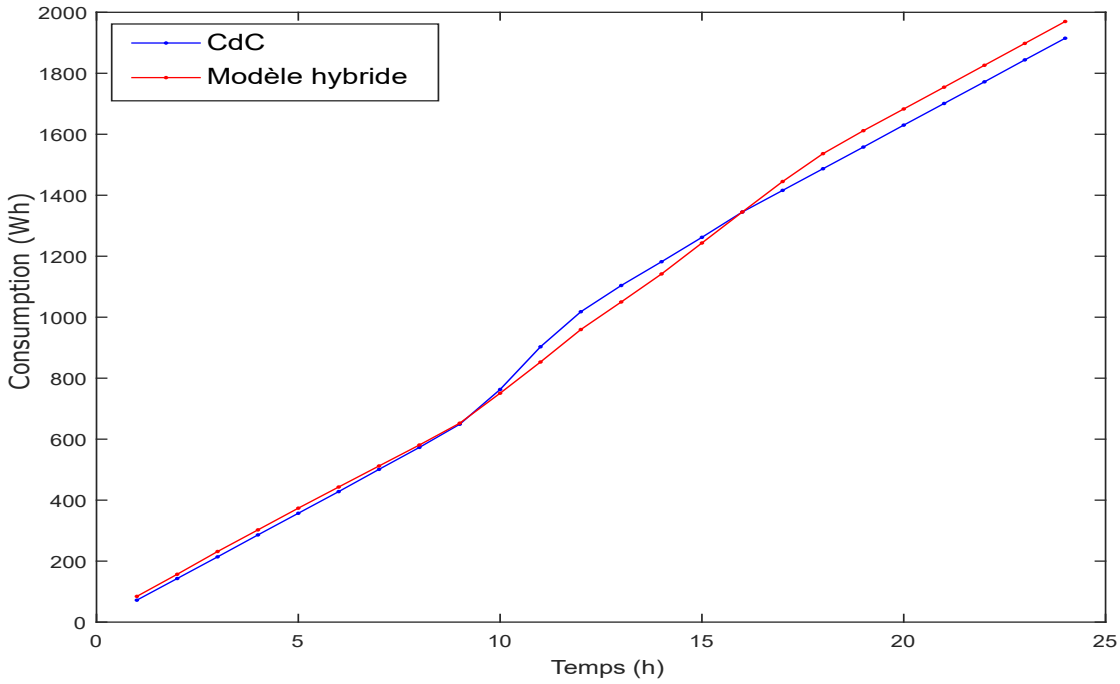

FIGURE 4.59 – CdC électrique estimée par le modèle hybride

L'interpolation de Lagrange est relativement simple et le polynôme d'interpolation de Lagrange est unique et facile à interpréter mais ne converge pas forcément vers la courbe de charge aux extrêmes selon le phénomène de Runge (voir figures 4.1 et 4.2). L'un des inconvénients de l'interpolation de Lagrange est que le polynôme est de degré n, donc les manipulations peuvent devenir lourdes si nous avons beaucoup de points. De plus, si nous ajoutons ou supprimons un point d'interpolation, nous le recalculerons tous. Malgré une certaine facilité d'estimation des nombreux paramètres d'un polynôme des moindres carrés, il est relativement simple à mettre en œuvre, le choix de l'ordre polynomial est erroné. L'interpolation au sens de Tchebychev fait partie des solutions qui permettent d'éviter les phénomènes d'interpolation de type Runge voire figure 4.4. Mais, même cela n'évite pas forcément le phénomène d'oscillations (voir figure 4.4). En outre, il faut beaucoup de temps pour trouver les points de Tchebychev, puis calculer les coefficients polynomiaux. Parmi les autres solutions, nous mentionnons l'interpolation par morceaux par splines cubiques. Cependant, si la pente entre deux nœuds est bien supérieure à 1 alors l'estimation des paramètres de la spline peut diverger. L'avantage des courbes de Bézier composites est que la modification d'un point ne déplace pas toute la courbe. Cependant, le modèle n'est pas optimal. Le principal avantage des méthodes paramétriques classiques est qu'elles fournissent un modèle facile à interpréter. La modélisation non paramétrique s'adaptera à de nombreuses séries chronologiques sans modification, tandis que pour les modèles paramétriques, il peut être difficile de choisir le bon modèle. L'avantage du modèle SARIMA est qu'il prend en compte la variation des données et la saisonnalité mais il n'est pas facile de trouver leurs paramètres. Le modèle de Holt-Winter se caractérise par sa simplicité conceptuelle et sa facilité de mise à jour. Cependant, la sélection des coefficients avec le modèle de Holt-Winter n'est pas toujours facile. Par contre, la prédiction est de moins bonne qualité que le modèle SARIMA. Pour les modèles résiduels, la combinaison des modèles individuels améliore la prédiction et minimise l'erreur qui est notée dans les tableaux 4.5 et 4.6.

Selon le tableau 4.5, le modèle journalier adaptatif à la CdC journalière de la consommation d'eau au restaurant universitaire est le modèle hybride qui combine le modèle déterministe et SARIMA$(2,0,2)(2,1,1)_{24}$.

Les résultats donnés dans le tableau 4.6 qualifient le modèle d'interpolation spline cubique car il représente la courbe de charge journalière d'électricité dans le bureau des doctorants avec une erreur minimale et un temps d'exécution raisonnable.

L'estimation par densité de la CdC d'eau avec des données échantillonnées en heure donne une grosse erreur car 24 points ne permettent pas d'estimer les paramètres gaussiens dans le modèle de mélange gaussien, par contre, les données échantillonnées en minute donnent de bons résultats, en particulier l'estimation du noyau qui réduit l'erreur à 0,13 %.

Table 4.5 – Comparaison des modèles proposés appliqués aux données sur l'eau.

	RMSE	Temps de calcul	Nombre de paramètres
Interpolation de Lagrange	0	1.38 seconds	24
Modèle des moindres carrés linéaire	202.77	0.13 seconds	2
Modèle des moindres carrés polynomiale	8.15	6.91 seconds	23
Interpolation de Tchebychev	449.28	36.44 seconds	24
Spline cubique	0	0.044 seconds	23*4
Courbe de Bezier	77.84	0.20 seconds	18
Modèle paramétrique classique	76.98	4.63 second	2
Modèle non paramétrique classique	81.31	3.95 seconds	0
Modèle SARIMA	61.40	37.81 seconds	13
Modèle de Holt-Winters	142.75	0.03 seconds	3
Modèle hybride	30.20	42.44 seconds	15

Table 4.6 – Comparaison des modèles proposés appliqués aux données électriques.

	RMSE	Temps de calcul	Nombre de paramètres
Interpolation de Lagrange	0	1.71 seconds	24
Modèle des moindres carrés linéaire	37.45	0.12 seconds	2
Modèle des moindres carrés polynomial	6.68	1.82 seconds	11
Interpolation de Tchebychev	6.79	50.8 seconds	24
Spline cubique	0	0.005 seconds	23*4
Courbe de Bezier	14.64	1.10 seconds	25
Modèle paramétrique classique	43.47	8.4 seconds	2
Modèle non paramétrique classique	48.73	5.21 seconds	0
SARIMA	54.30	2.43 seconds	10
Holt-Winters	75.86	0.08 seconds	3
Modèle hybride	37.24	10.83 seconds	12

5 Détection des fuites d'eau

5.1 Introduction

La conservation des resources en eau et la prévention de son gaspillage sont des éléments fondamentaux pour les besoins de l'humanité. Un des problèmes crutiaux du gaspillage de cette resource vitale est lié aux fuites. La fuite d'eau est un évènement qui a tendance à se produire dans des moments inattendus. Le plus important est de pouvoir intervenir suffisamment tôt et rapidement afin de limiter les dégâts et de minimiser le coût des réparations.
Dans la majorité des cas, les fuites ne sont pas détectées assez tôt, notamment lorsque les dommages sont mineurs. Les réseaux de distribution d'eau (WDN) sont caractérisés par de nombreux nœuds et un grand nombre de branches. L'identification des tuyaux de fuite est donc une tâche très difficile. De plus, un débit constant, petit et diffus ne peut pas être détecté par les instruments de mesure conventionnels, d'autant plus que les données de consommation ne sont généralement enregistrées et transmises que sur une longue période. Cela peut entraîner une perte d'eau drastique [54]. En Europe, 11% de la population européenne et 17% de son territoire ont été affectés par la pénurie d'eau d'après l'estimation de la commission européenne [16].
L'émergence du paradigme de l'internet des objets (IoT) a ouvert de nouvelles perspectives dans plusieurs domaines [55], [56]. Recemment, des capteurs améliorés ont été développés [57]. Ils peuvent être utilisés pour détecter ces fuites. L'importance de détecter les fuites consiste à préserver les ressources en eau, à éviter les dommages consécutifs dans le WDN et à limiter la demande en eau. De plus, une fuite d'eau peut affecter la qualité de l'eau en introduisant des infections dans le WDN et avoir des conséquences importantes sur la santé et la sécurité de la population [14].
Dans ce chapitre, nous présentons plusieurs méthodes de détection des fuites d'eau : la détection par la CdC maximale, le Minimum night flow (MNF), l'approche basée sur la dérivée et l'indicateur de fuite d'eau basé sur l'approche de densité temporaire de consommation (WLICTD). Les résultats de l'approche WLICTD sont publiés dans [58].
Les données de consommation d'eau au restaurant universitaire contiennent deux fuites d'eau illustré par la figure 5.1. La fuite 1 est représentée par les courbes de charge colorées en rouge et en bleu. Elle a commencé le 17 janvier 2018 vers 22 : 57 et elle s'est poursuivie jusqu'au

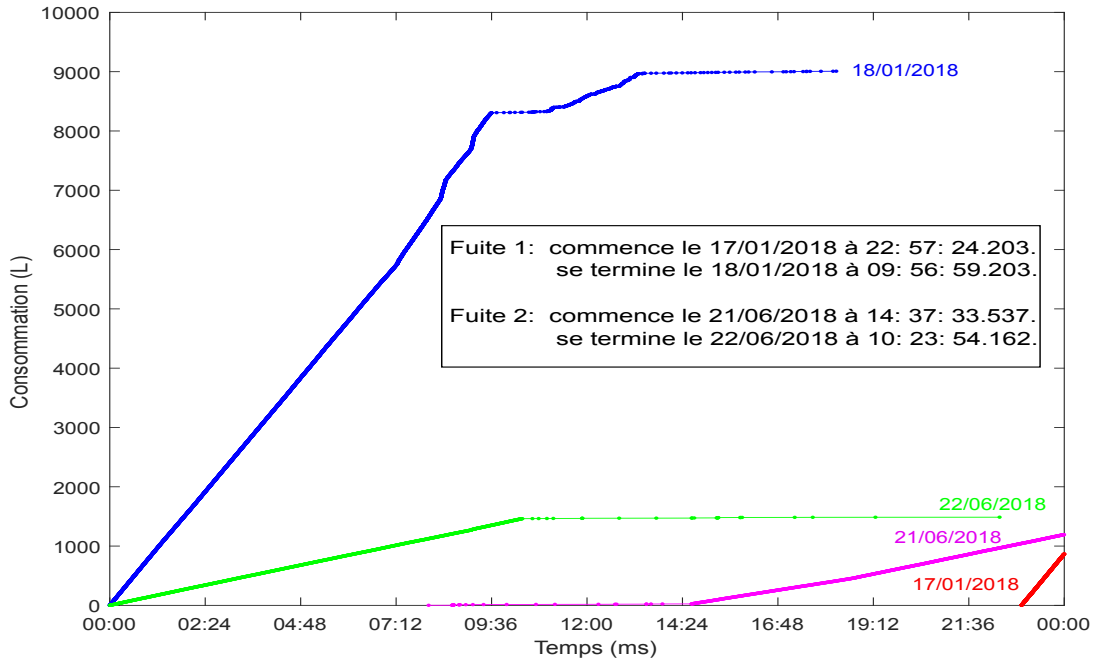

FIGURE 5.1 – Les courbes de charge journalières de la consommation d'eau dans le restaurant universitaire avec des fuites.

lendemain, 18 janvier 2018, à 9 : 56. La fuite 2 a commencé le 21 juin 2018 à 14 : 37 et elle a été détectée le 22 juin 2018 à 10 : 23. Elle est représentée avec les courbes en rose et en vert.

5.2 Densité temporaire de la consommation pour la détection des fuites d'eau en temps réel

Nous présentons un nouvel indicateur de la détection des fuites d'eau appelé WLICTD (Water Leakage Indicator based on the Consumption Temporary Density) [58].
WLICTD peut distinguer une consommation anormale (une situation de fuite) d'une consommation normale (consommation quotidienne). L'indicateur proposé a été évalué à l'aide d'un ensemble de données réelles de consommation d'eau obtenues à partir d'un compteur intelligent installé au restaurant universitaire. Les résultats révèlent la force et l'importance de l'indicateur proposé, toutes les situations de fuite d'eau sont détectées dans une période pratique.

5.2.1 Densité temporelle

Dans le but de la détection des fuites d'eau nous utilisons la densité temporelle inspirée de [59] et [60] pour analyser et mesurer la progression d'événements aléatoires dans le temps. Pour

chaque instance t_i, la densité temporelle $D(t_i)$ est définie par l'équation (5.1).

$$D(t_i) = \lambda^{t_i - t_{i-1}} D(t_{i-1}) + 1 \tag{5.1}$$

Avec t_i qui représente l'instant actuel, et t_{i-1} qui représente le dernier instant de la mise à jour de la densité D ; $\lambda \in]0,1[$ est un paramètre appelé facteur d'évanouissement qui correspond à la vitesse à laquelle la consommation diminue lorsque l'écart de temps est grand. La valeur de λ est ajustée par des tests empiriques par rapport à la contrainte de minimisation du temps de détection de la fuite d'eau ; $D(t_i)$ est la densité à l'instant t_i.
Grâce à cette fonction, il est possible de mesurer combien de litres d'eau ont été consommés au cours d'une période de temps. La densité D est définie par récurrence. Donc il est important de mentionner la densité à l'instant t_0 qui correspond à l'instant d'initialisation de la densité. La densité $D(t_0)$ est nulle, puisqu'il n'y a pas de consommation en début de journée $D(t_0) = D(0) = 0$. Lors de la consommation d'un litre d'eau à l'instant t_i, la densité est mise à jour en ajoutant 1 à l'ancienne densité puisque en chaque instante t_i nous consommons un litre d'eau. Comme cette fonction dépend du temps écoulé entre le temps de la dernière mise à jour de la densité et le temps actuelle donc, elle donne une vue globale du profil de consommation d'eau. Ainsi, si la consommation d'eau est régulière, la densité est faible, tandis que si la consommation est anormale, la densité est élevée.
Les figures 5.2 et 5.3 présentent deux exemples de l'évolution de la densité temporelle pour la consommation d'eau au cours d'une journée normale (c'est-à-dire une journée sans fuite d'eau). Nous observons qu'une consommation élevée implique une densité supérieure à 1 puisque la grande consommation implique une consommation successive et rapide ce qui signifie que les écarts de temps sont petits, et comme $\lambda \in]0,1[$, cela implique que $\lambda^{t_i - t_{i-1}} > 0$, ce qui signifie que $D(t_i) > 1;\ i > 1$. Alors que si les écarts de temps sont importants (ce qui signifie que la consommation est lente), alors $\lambda^{t_i - t_{i-1}} = \exp(t_i - t_{i-1}) \ln \lambda$ se rapproche de zéro donc la densité est égale à 1.

5.2.2 Algorithme WLICTD

L'algorithme 2 décrit le programme utilisé pour détecter les fuites d'eau en fonction de la densité temporelle.
Tout d'abord, nous déterminons le seuil qui est égal à la période maximale où les densités supérieures à 1 dans tous les jours normaux qui sont représentés par les flèches rouges sur la figure 5.4 pour un exemple de la densité temporelle un jour normal (24 janvier 2018).

$$seuil\ 1 = \max_i(\triangle^*(t_i)) = \max_i(t_{k_i} - t_{L_i}) \tag{5.2}$$

Tels que $D(t_{k_i}) = D(t_{L_i}) = 1$ et $\forall j \in]k_i,\ L_i[,\ D(t_j) > 1$.
Deuxièmement, nous calculons la densité pour chaque instant t_i en utilisant :

$$\begin{cases} D(t_0) & = 0 \\ D(t_i) & = 0,5^{\triangle t_i} D(t_{i-1}) + 1 \end{cases}$$

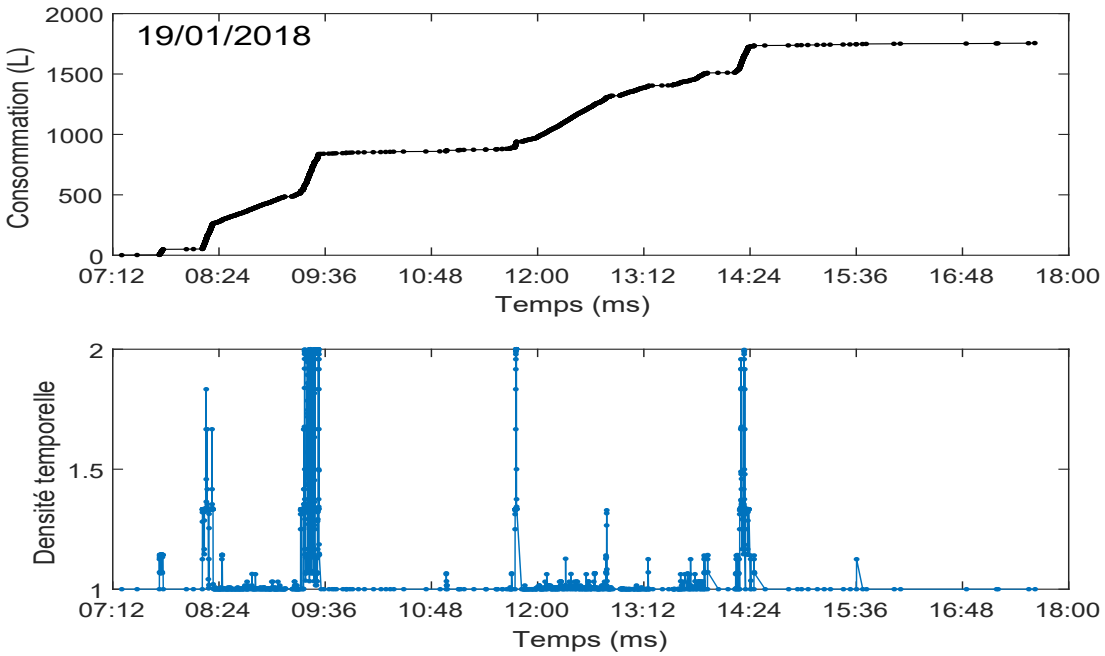

FIGURE 5.2 – La densité temporelle de la consommation d'eau au cours d'une journée typique avec $\lambda = 0,5$ et sa courbe de charge.

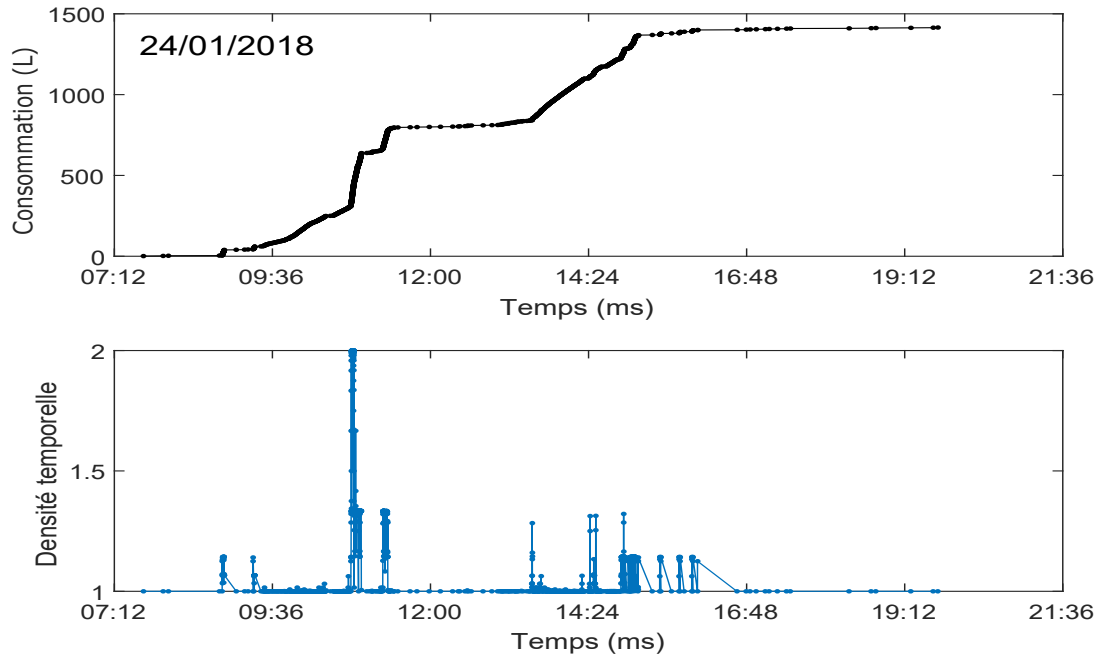

FIGURE 5.3 – La densité temporelle et la courbe de charge de la consommation d'eau durant un jour normal.

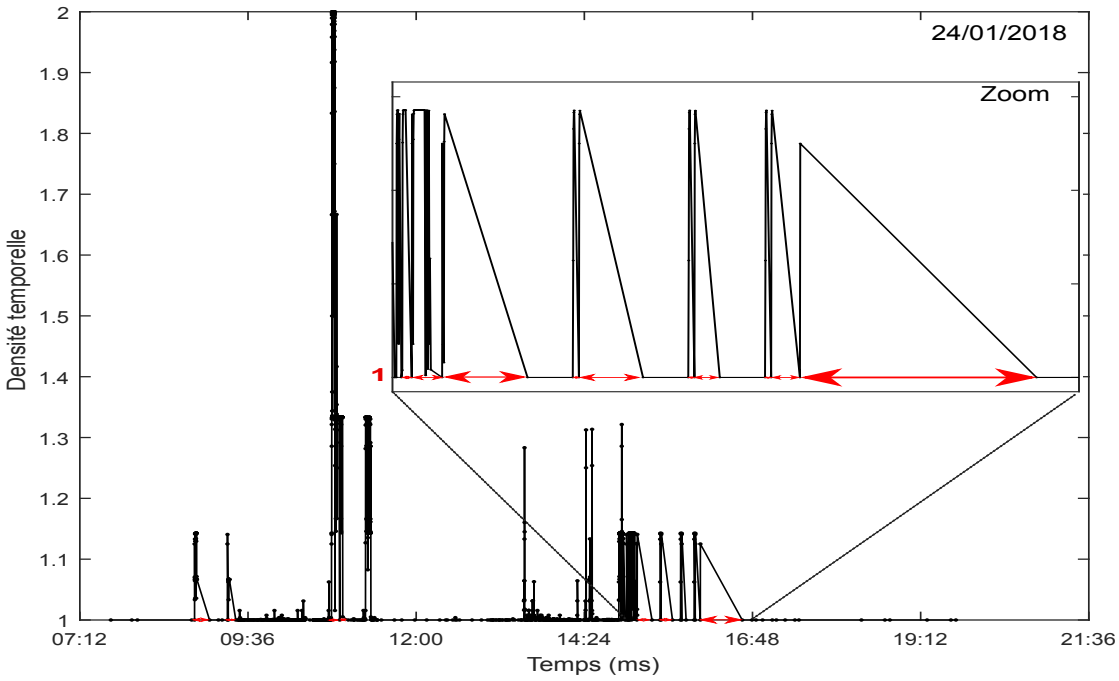

FIGURE 5.4 – Le seuil de la densité temporelle de la consommation d'eau au 24/01/2018.

Enfin, la détection de la fuite d'eau est donnée par deux tests de comparaison, le premier montre une consommation élevée avec une densité supérieure à 1, et le second confirme la continuité de cette forte consommation avec un dépassement du seuil 1.

Algorithm 2 WLICTD (Water Leakage Indicator Consumption Temporary Density)

1- Déterminer le seuil 1.
2- Détection des fuites.
for À chaque instant **do**
  Calculer la densité temporelle $D(t_i)$ à l'instant t_i
  if $D(t_i) \neq 1$ **then**
    if $\triangle t_i > seuil\ 1$ **then**
      Détection de fuite.
      Break
    end if
  end if
end for

5.2.3 Application

Nous visons maintenant à appliquer l'algorithme précédent à notre ensemble de données dans le restaurant universitaire.
Grâce à l'algorithme WLICTD, nous avons réussi à détecter toutes les fuites d'eau dans un délai

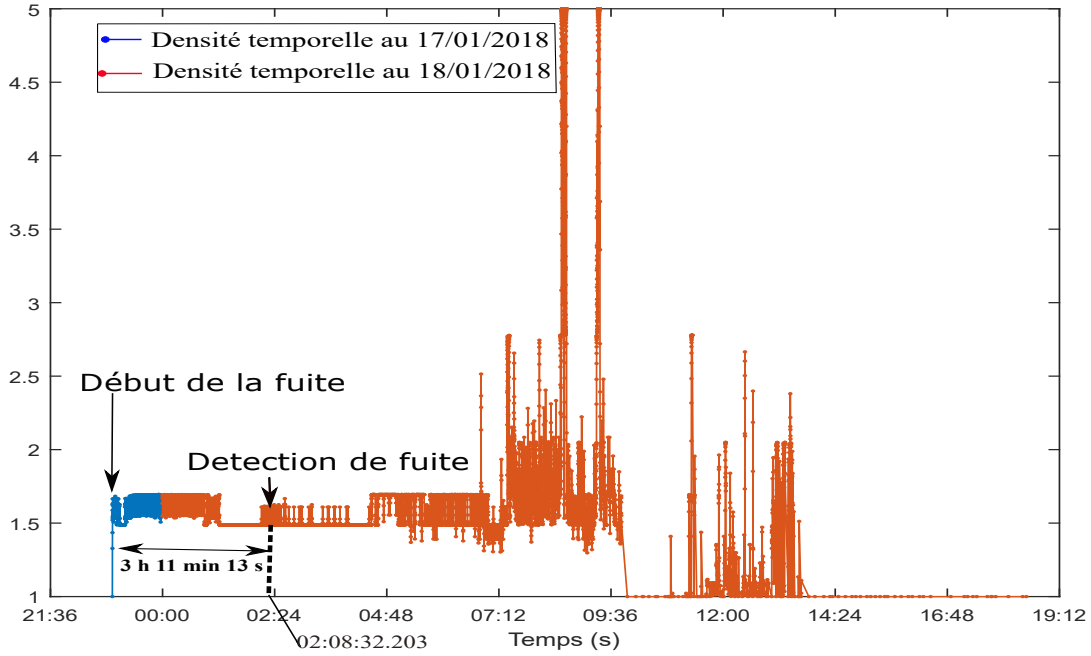

FIGURE 5.5 – La densité temporelle de la consommation d'eau pendant les jours de la première fuite.

d'environ 3 heures.
Le choix du paramètre intégré dans le calcul de la densité temporelle λ est une tâche très importante, par exemple pour nos données la valeur $\lambda = 0,1$ ne permet pas de détecter les fuites d'eau puisque la densité temporelle est égale à 1 pour toutes les instants. De plus, la valeur $\lambda = 0.99$ permet de détecter les fuites d'eau après 14 h 33 min 09 s. Ainsi, le temps de détection est très long. Il s'avère que les valeurs de λ entre 0.4 et 0.9 donnent de bons résultats (petit temps de détection). Il faut faire plus d'expériences pour fixer une valeur "universelle" λ ou utiliser l'apprentissage en profondeur et l'intelligence artificielle pour la fixer dans différentes situations.
La figure 5.5 illustre la densité temporelle en fonction du temps en secondes pour deux jours anormaux (17 janvier 2018 et 18 janvier 2018). Notre approche nous permet de détecter cette fuite le 18 janvier 2018 à 02 : 08 : 32.203. La figure 5.6 présente la densité temporelle en fonction du temps en secondes dans deux jours anormaux le 21 juin 2018 et le 22 juin 2018 respectivement. L'approche WLICTD détecte cette fuite le 21 juin 2018, à 17 : 48 : 42.537.

5.3 Détection des fuites d'eau par la courbe de charge maximale

La courbe de charge maximale exprime le modèle de la plus grande consommation d'eau durant une période spécifique. Nous observons que l'axe de temps de la courbe de charge n'est pas équidistant, contrairement à l'axe de consommation. Mais pour construire la courbe de charge maximale, nous avons besoin d'avoir le même pas dans l'axe de temps. Nous échantillonnons

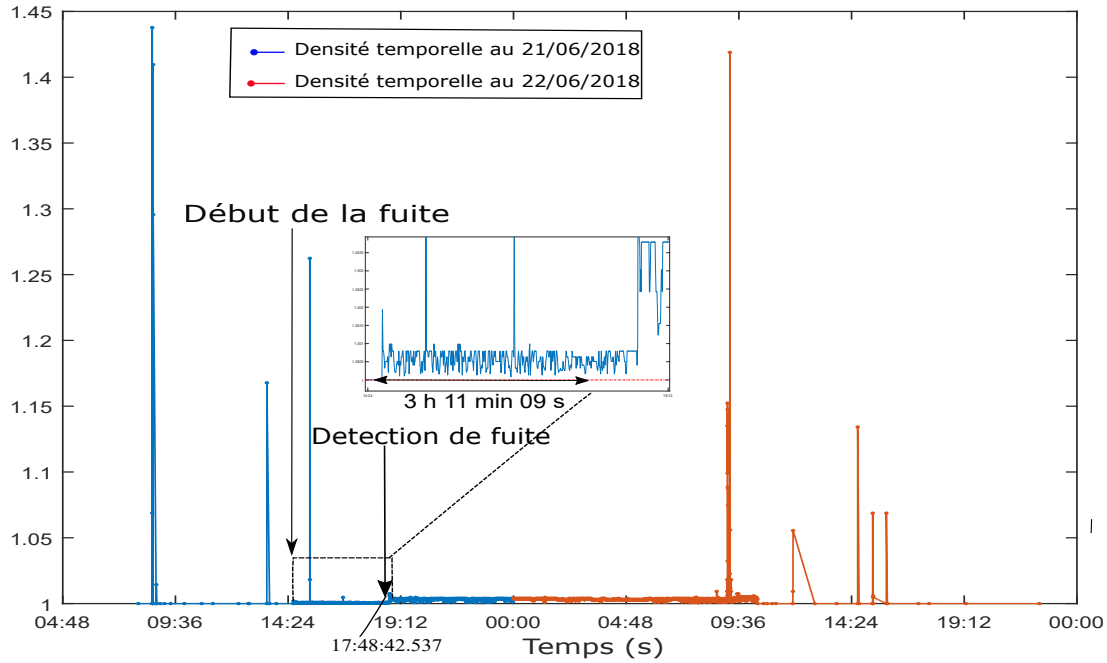

FIGURE 5.6 – La densité temporelle de la consommation d'eau pendant les jours de la deuxième fuite.

donc les données. La courbe maximum est construite à partir d'une base de données échantillonnée en minutes $C_j(i),\ i = 1:1440$ en utilisant toutes les courbes de charge normales j. Pour calculer la consommation maximal, on utilise l'équation (5.3)

$$C_{max}(i) = \max_j C_j(i),\ i = 1,...,1440. \tag{5.3}$$

tels que i représente la minute, j exprime le jour et $C_j(i)$ est le nombre de litres d'eau consommés de minuit à la minute i le jour j.
Pour détecter la fuite d'eau, il faut que la courbe de charge d'une journée anormale dépasse la courbe de charge maximale. C'est-à-dire la consommation par minute supérieure à la consommation maximale dans la même minute et nous ajoutons un intervalle d'une demi-heure où la courbe de charge est supérieure à la courbe maximale pour confirmer la fuite.
La figure 5.7 explique la détection des fuites d'eau en utilisant la courbe de charge maximale représentée en noir. Les autres courbes sont les courbes de charge échantillonnées en minute des jours des deux fuites d'eau. Nous comparons la consommation maximale $C_{max}(i)$ et la consommation $C_j(i)$ à l'un des jours anormaux présentés dans la figure 5.7. Dès que la consommation dépasse la consommation maximale, nous détectons la fuite. La courbe maximum permet de détecter la première fuite d'eau le 18 janvier 2018 à 12h30 et la fuite 2 le 22 juin 201 à 00h30.

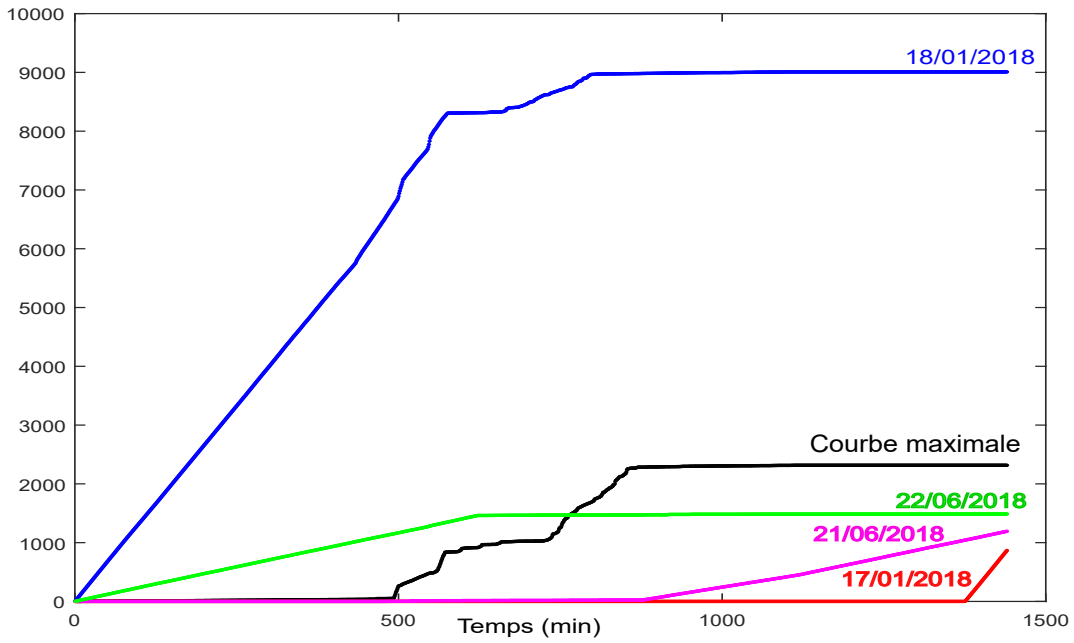

FIGURE 5.7 – La courbe de charge maximal et les courbes de charge des jours de fuites.

5.4 Détection des fuites d'eau par la dérivée

Le débit (Flot) est une fonction qui représente l'écoulement de la consommation d'eau à chaque instant. Il est connu en mathématique par la dérivée
L'utilisation de seuil connu dans la littérature qui est défini par la valeur maximale de la dérivée dans les jours normaux sur notre base de données ne permet pas de détecter les fuites d'eau. Tandis que l'utilisation de l'approche en bas permet de détecter la fuite d'eau après 110 minutes sur une base de données échantillonnées en minute. Mais l'approche ne fonctionne pas avec les données brutes.

5.4.1 Approche

Pour détecter les fuites d'eau par la dérivée, nous utilisons les étapes suivantes :

1. Échantillonner les données en minute.
2. Calculer la dérivée $F_j(i), i = 1:1440$ pour les jours normaux j.

$$F_j(i) = \frac{C(i) - C(i-1)}{i - (i-1)} = C(i) - C(i-1)$$

   Avec $C(i)$ le nombre de litres d'eau consommés à la minute i. $C(i-1)$ la consommation à l'instant $i-1$.

3. Définir un seuil :

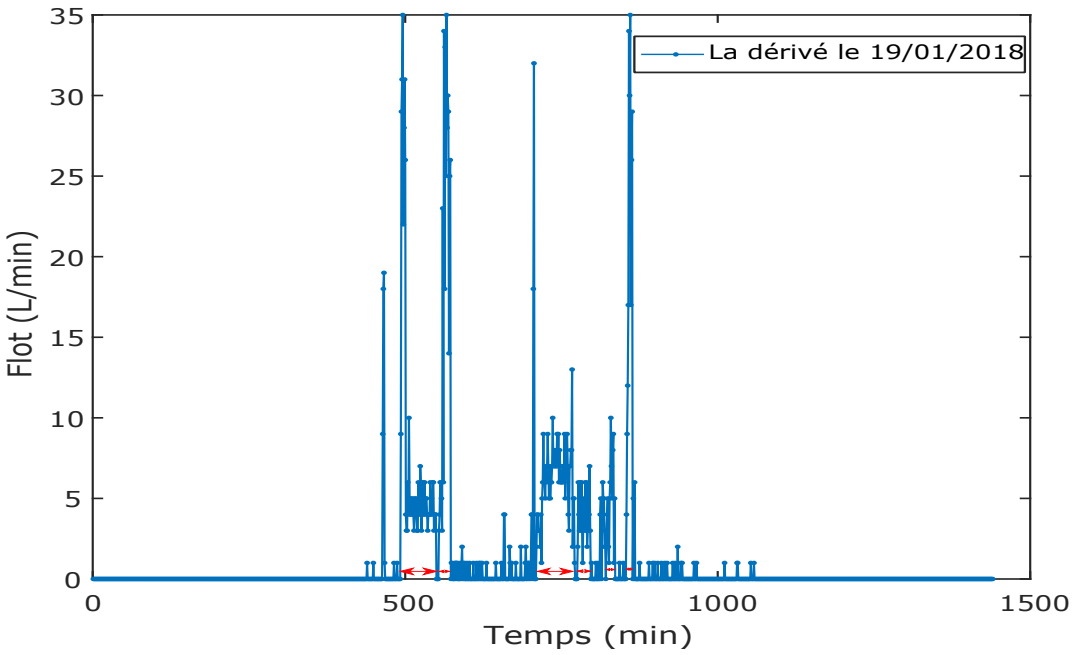

FIGURE 5.8 – Exemple de la fonction débit dans un jour normal

Tous d'abord, nous calculons les périodes $\triangle_i^*$ où les dérivées sont non nulles, elles sont indiquées par les flèches rouges dans la figure 5.8.
C'est à dire, $\triangle_i^* = t_{l_i} - t_{k_i}; k_i < l_i, F(t_{k_i}) = F(t_{k_i}) = 0, \forall\, j \in]k_i,\ l_i[; F(t_j) \neq 0$
Ensuite, nous définissons le seuil qui est déterminé par le maximum de ses périodes.

$$seuil\ 2 = \max_i(\triangle_i^*). \tag{5.4}$$

La figure 5.8 représente la fonction de débit dans un jour normal (19/01/2018), où les flèches en rouges représente les $\triangle_i^*$.

4. Détecter des fuites à l'aide d'un seuil
   La dernière étape consiste à tester tout d'abord le flot calculé à l'instant i avec la valeur 0 si ce dernier est non nul, nous comparons l'écart de temps $\triangle_i$ au seuil 2. Si cette période déplace le seuil alors il y a une fuite d'eau. Sinon, nous répétons la dernière étape avec l'instant $i+1$.

L'algorithme 3 résume les étapes de l'approche basée sur la dérivée.

5.4.2 Résultat

L'algorithme 3 permet de détecter les deux fuites au bout de 110 minutes puisque la fuite 1 est détectée le 18 janvier 2018 à 00h46 et la fuite 2 est détectée le 21 juin 2018 à 16h28.
La figure 5.9 illustre la fonction d'écoulement dans les deux jours de fuite 1. La fonction d'écoulement de la fuite 2 est représentée sur la figure 5.10.

Algorithm 3 Flot

1- Échantillonner les données en minutes.
2- Calculer la dérivée dans les jours normaux.
3- Définir un seuil 2.
4- Détection de fuite.
for Chaque jour j **do**
  for $i = 1:1440$ **do**
    Calculer $F_j(i)$ à la minute i
    if $F_j(i) \neq 0$ **then**
      if $\triangle_i > seuil\ 2$ **then**
        détection de fuite.
        Break
      end if
    end if
  end for
end for

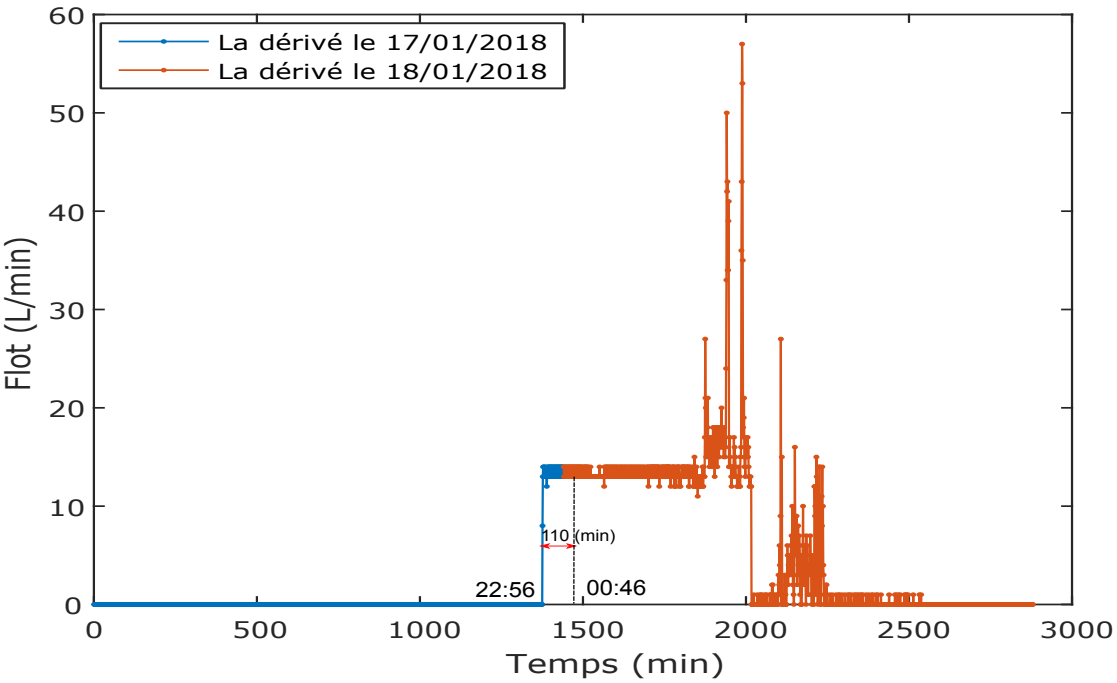

FIGURE 5.9 – Résultat de l'application de l'approche qui se base sur la dérivée pour détecter la fuite 1

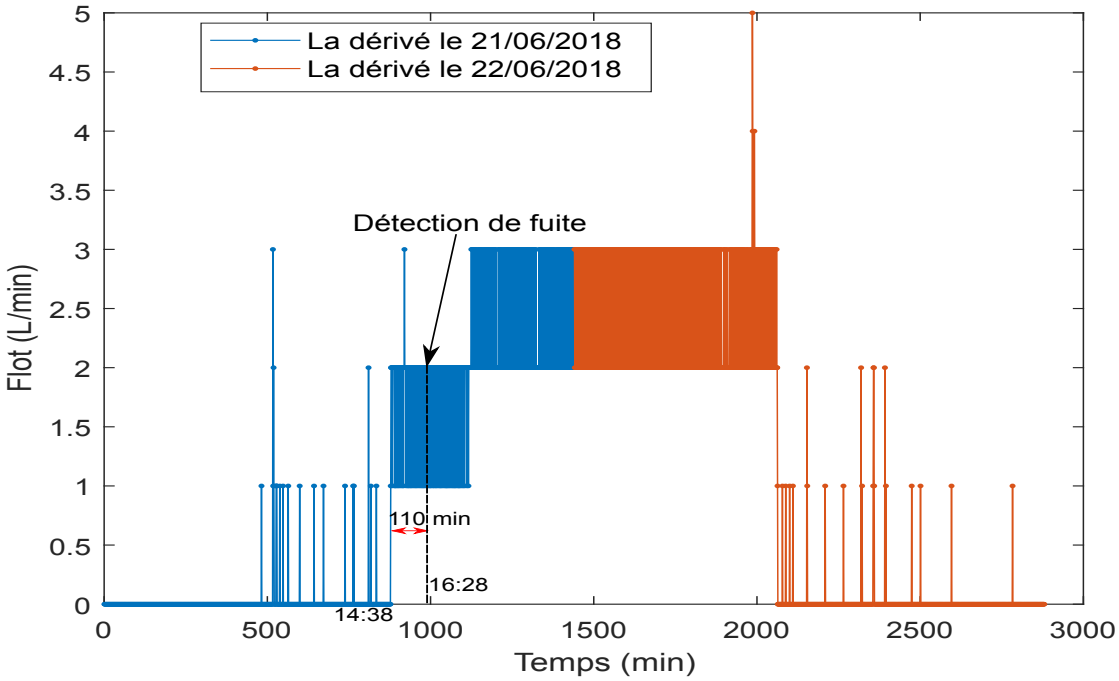

FIGURE 5.10 – Résultat de la détection de la fuite 2 par la dérivée

5.5 Minimum night flow

Le Minimum Night Flow est un seuil défini pour une partie isolée où la demande en eau est généralement faible. MNF est basé sur la fonction d'écoulement dans la partie isolée [61]. Pour déterminer le MNF il faut compter sur plusieurs jours normaux [62]. La méthode MNF est utilisée pour détecter les fuites d'eau pendant la nuit. La nuit équivaut à la période sans consommation qui est représentée par la partie 1 du chapitre 1. MNF est le nombre maximum de litres d'eau consommés dans la partie 1 pendant les jours normaux. En d'autres termes, c'est la valeur maximale du flot dans la partie 1. Pour tous i de la partie 1, et pour tous les jours normaux j, le MNF est définie par :

$$MNF = \max_{i,j} F_j(i) \tag{5.5}$$

En utilisant des données échantillonnées en minute qui représentent la consommation d'eau dans le restaurant universitaire, nous trouvons : MNF = 2 L/min.
La consommation à la minute 1378 du 17 janvier 2018 a dépassé le MNF (voir figure 5.11). Donc, la fuite 1 est détectée par la MNF le 17 janvier 2018 à 22h58. La fonction de flot du 21 juin 2018 définie dans la partie 1 de la figure 5.12 est supérieure à la MNF à 20 $h = 1200$ min.

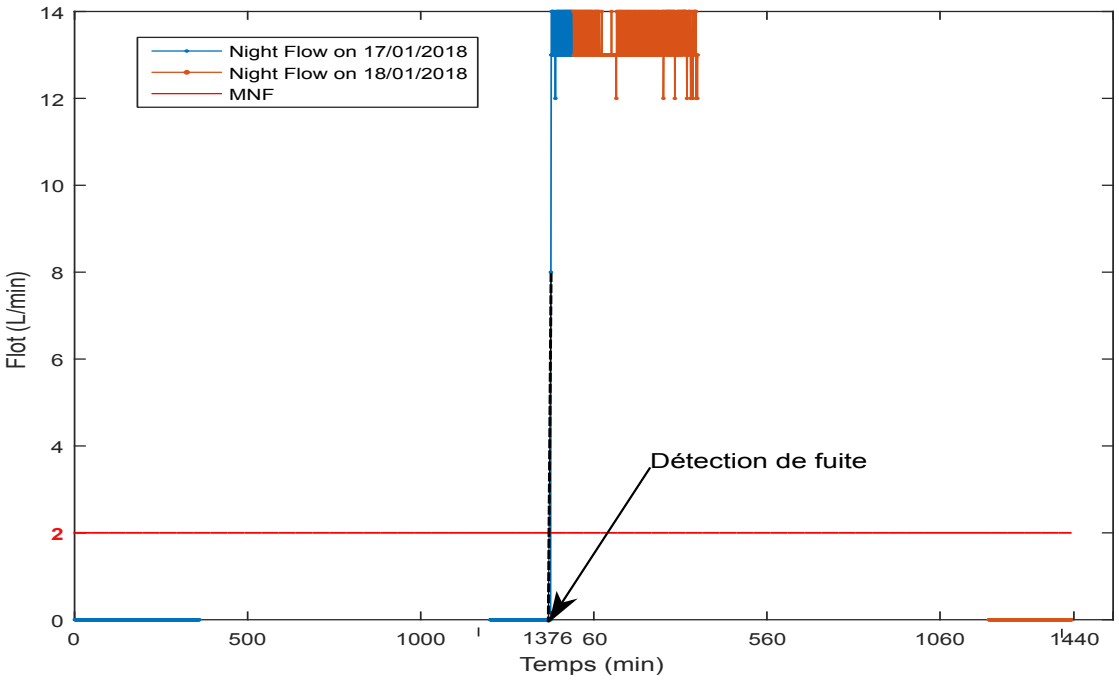

FIGURE 5.11 – Détection de la fuite 1 par le MNF.

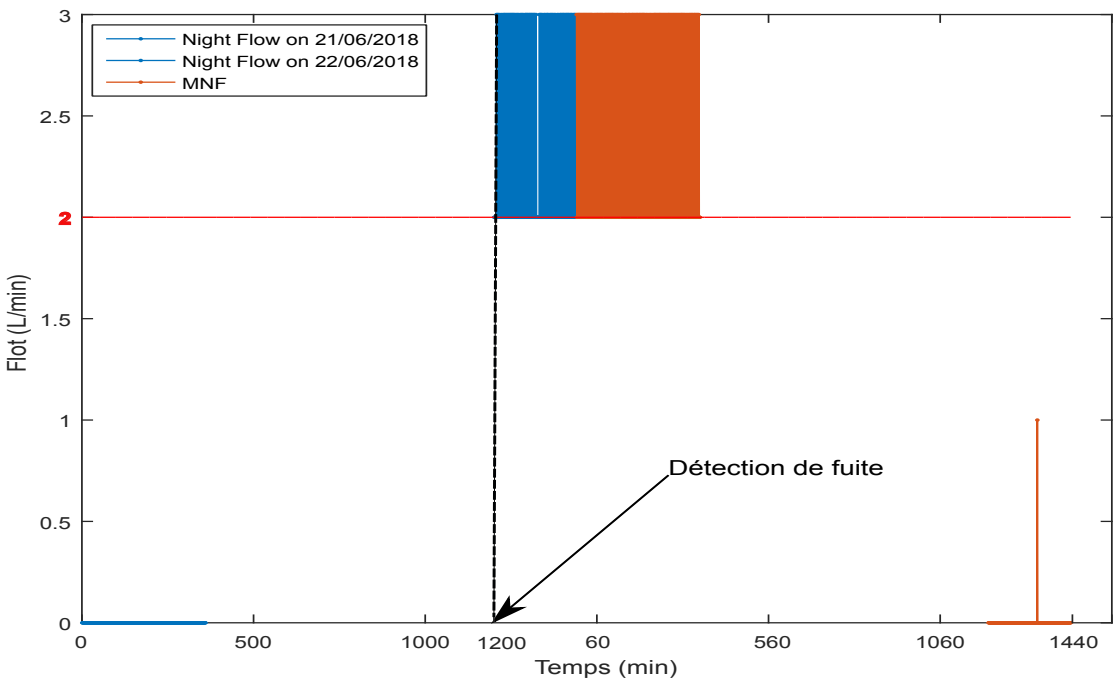

FIGURE 5.12 – Détection de la fuite 2 par le MNF.

5.6 Comparaison

La période de détection des fuites est le critère utilisé pour comparer entre les méthode de détection des fuites.
Le tableau 5.1 représente les périodes de détection des fuites d'eau par courbe de charge maximale, MNF, WLICTD et l'approche basé sur la dérivé. À partir des résultats du tableau 5.1, nous observons que la fuite 1 a été détectée par le MNF au bout de 25 secondes, par la courbe maximale après une heure 35 minutes, par le débit après une heure 50 minutes, et par WLICTD après 3 heures 11 minutes et 13 secondes. La fuite 2 a également été détectée après 1 heure 50 minutes par flot, après 3 heures 11 minutes 20 secondes par WLICTD, après 4 heures 22 minutes par le MNF, et après 9 heures 52 minutes par la courbe maximale.
Nous remarquons que le MNF détecte la fuite 1 très rapidement car la fuite commence pendant la nuit et le seuil MNF est testé dans cette partie. Bien que la fuite 2 commence à 14 : 37 pendant la partie 2, le MNF met beaucoup de temps à la détecter en attendant que la partie 1 commence. De plus, si le début de la fuite approche à 6h 00 du matin, le temps de détection du MNF augmentera puisque le MNF n'a pas dans sa zone d'application (la nuit). Donc la meilleure méthode qui a détectée la fuite 2 est l'approche basée sur la dérivée appliquée sur les données échantillonnées en minute. Sur la base des données brutes, la meilleure approche qui a détectée la fuite 2 est WLICTD. La détection avec une approche basée sur la densité temporelle donne l'assurance que toutes les fuites d'eau sont détectées sur les données brutes dans un délai raisonnable et que le temps n'est pas perdu à chercher une échelle d'échantillonnage permettant ou non de détecter les fuites d'eau par d'autres moyens.

Table 5.1 – Comparaison des différentes approches pour détecter des fuites d'eau

	Fuite 1	**Fuite 2**
WLICTD	3 h 11 min 13 s	3 h 11 min 20 s
Courbe maximal	1 h 35 min	9 h 52 min
MNF	25 s	4 h 22 min
Flot	1 h 50 min	1 h 50 min

5.7 Détection des fuites d'eau le week-end

L'approches WLICTD est applicable les jours ouvrables mais pour les jours de week-end cela ne fonctionne pas car WLICTD est basée sur la densité temporelle qui dépend des données de temps mais en jours de week-end normaux il n'y a pas de consommation au restaurant universitaire. Il en va de même pour la dérivée. Ainsi pour détecter les fuites d'eau le week-end, nous avons défini un seuil représentant la consommation maximale un week-end normal qui est égal à 3 litres d'eau dans le restaurant universitaire. La figure 5.13 explique la détection des fuites d'eau le week-end, sur la base d'un seuil affiché en rouge, et la courbe bleue représente la courbe de charge du 24 mars 2018.

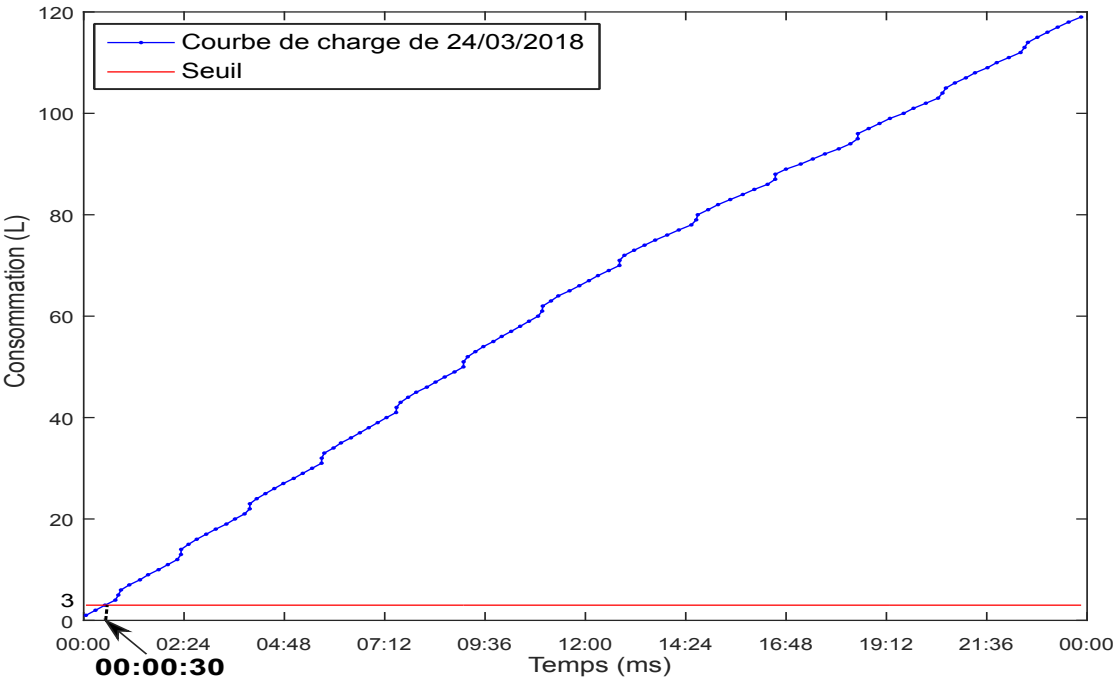

FIGURE 5.13 – La détection de fuite d'eau un jour de week-end.

5.8 Conclusion

La détection des fuites d'eau par des méthodes classiques prend beaucoup de temps. L'utilisation de la courbe maximale [20] pour détecter les fuites d'eau est fiable dans les périodes où il n'y a pas beaucoup de consommation comme les nuits au restaurant universitaire. De même, nous pouvons utiliser le MNF pour détecter les fuites de la nuit [14] et [20]. Nous affirmons que la nouvelle approche basée sur la densité temporelle de consommation est très efficace sur la base de données brutes. Cette densité fournit une description fiable du comportement des utilisateurs en temps réel, c'est-à-dire un profil de consommation d'eau dans le temps qui permet de reconnaître les activités anormales telles qu'une consommation d'eau élevée ou une fuite d'eau. WLICTD est capable de détecter les fuites d'eau en 3 heures environ. Cette approche permet de réduire le temps de réparation et donc de minimiser les dommages potentiels causés par la fuite. L'approche MNF reste toujours la meilleure pour les périodes de nuit. Il s'avère que l'approche basée sur la dérivée est très fiable pendant les périodes de consommation importante, cela comparé aux autres approches pour détecter les fuites d'eau sur les données échantillonnées en minute.

6 Conclusion et perspectives

Conclusion générale

Ce travail a porté sur la modélisation et l'exploration des données. Le contenu de la thèse s'articule autour des aspects théoriques et appliqués de l'évaluation et de la modélisation des données de consommation. Les objectifs réalistes que nous cherchons à atteindre sont nombreux. Nous avons cherché à comprendre les données et à extraire un maximum d'information sur la façon de consommer et le comportement des consommateurs. Après avoir compris le mode de consommation, nous avons voulu généraliser le profil de consommation à l'aide de modèles mathématiques. Nous voulions également prédire des consommations en s'appuyant sur les séries temporelles. La préservation des resources en eau étant un élément important de sa gestion, il est donc important de détecter les fuites d'eau le plus tôt possible, nous avons proposé un algorithme original WLICTD un modèle hybride qui permettent de racourcir considérablement le temps de détection. Ces objectifs nécessitent une grande quantité de données dont les problèmes de stockage sont également discuté dans ce travail.
À travers cette thèse, nous visons à résoudre les problèmes évoqués en introduction, qui sont :

- Analyser les données de la consommation d'eau et d'électricité.
- Réduire ces données.
- Trouver le meilleur modèle qui estime les courbes de charge.
- Détecter les fuites d'eau.

Les contributions utilisées pour résoudre ces problématiques sont différentes. Pour comprendre les données, il est nécessaire de les analyser. Et cette analyse que nous avons abordée à travers des études statistiques, qui nous ont permis de donner une vision globale de l'aspect quantitatif de la consommation à différentes périodes et de différencier les jours de vacances et des jours ouvrables, même les jours où se sont produits les évènements les plus inhabituels, par exemple, les grandes fuites d'eau. Les courbes de charge nous ont permis d'examiner en profondeur les données afin de pouvoir différencier les jours normaux des jours anormaux. Elles permettent également l'observation des fuites d'eau (fuite d'une petite quantité d'eau).
Cette analyse nécessite de grandes quantités de données, ce qui nous a permis d'envisager la

manière de stockage des données et d'aborder le problème de stockage des données. Pour cela, nous avons utilisé des méthodes paramétriques (la régression linéaire) et non paramétriques (l'échantillonnage et le regroupement) de la réduction numérosité des données. Ensuite, nous avons suggéré la méthode d'apprentissage avec la carte auto-organisatrice pour réduire les données qui a donné de bons résultats.
En termes de prédiction de la consommation horaire d'eau et d'électricité, les modèles hybrides montrent leurs performances par rapport aux modèles individuels. Dans un premier temps, des modèles de prévision individuels sont étudiés. Ainsi, le modèle déterministe de séries temporelles est considéré. Ensuite, un modèle stochastique basé sur les résidus a été étudié en plus des modèles de lissage exponentiel. Les résidus jouent un rôle très important en termes d'analyse de séries temporelles. Cependant, les modèles ne traitent que la partie linéaire de la série temporelle. Ces modèles ne traitent pas des corrélations non linéaires au niveau des résidus. Par conséquent, nous choisissons des approches neuronales pour déterminer la non-linéarité de la série temporelle. Deux réseaux de neurones sont étudiés, MLP et LSTM pour prédire la consommation horaire. Nous avons évalué les modèles proposés avec une mesure permettant le calcul de l'erreur RMSE. Ainsi, pour la base de données sur l'eau, l'erreur minimale a été donnée par le modèle hybride 9 qui mélange le modèle déterministe, le modèle de Holt-Winters et la combinaison des modèles LSTM et MLP. Alors que le modèle hybride 5 qui combine le modèle SARIMA et LSTM est le meilleur modèle pour la base de données sur l'électricité. S'appuyer sur ces deux modèles permet de déterminer des prévisions horaires de consommation d'eau et d'électricité dans un bâtiment tertiaire.
Après avoir appliqué plusieurs modèles qui estiment la courbe de charge journalière de la consommation d'eau et d'électricité, nous avons conclu que la convergence du polynôme d'interpolation de Lagrange et de l'approximation des moindres carrés vers la courbe de charge n'est pas assurée lorsque le nombre de points d'interpolation (et donc le degré du polynôme) est amené à tendre vers l'infini. Le théorème de Faber établit l'existence de fonctions pour lesquelles l'erreur augmente avec le nombre de nœuds (voir le phénomène de Runge pour la fonction $f(x) = \frac{1}{1+x^2}$). Pour résoudre ce problème de divergence de l'interpolation polynomiale, on remplace nos points par les points Tchebychev. Mais le polynôme d'interpolation de Tchebychev ne corrige pas complètement l'effet Runge et le polynôme ne converge pas aux extrémités, nous utilisons donc une interpolation par morceaux de spline cubique ou de la courbe de Bézier. La prédiction du modèle de courbe de Bézier est plus complexe. Par conséquent, le modèle spline cubique donne de bons résultats si nous n'avons pas trop de variations dans nos données, ce qui est le cas avec nos données d'électricité. Compte tenu de l'effet résiduel de nos données de consommation d'eau, nous qualifions le modèle hybride du modèle paramétrique classique et SARIMA parmi les modèles avec résidus comme le meilleur modèle de la courbe de charge journalière de consommations d'eau.
Lorsque nous avons examiné les données sur l'eau, nous avons remarqué la présence de fuites d'eau, qui est l'un des plus gros problèmes auxquels les ressources en eau sont confrontées. C'est ce qui nous a fait aborder la détection des fuites d'eau. Pour détecter les fuites d'eau, nous avons proposé une nouvelle approche basée sur la densité temporelle appelée WLICTD. Cette approche permet de détecter rapidement les fuites ainsi que d'utiliser des données brutes en temps réel

sans les modifier. Nous avons également utilisé la fonction de débit pour détecter les fuites d'eau qui a donné les meilleurs résultats pendant la période active (période de consommation importante) de la journée, mais uniquement avec un échantillonnage de données en minutes. Aussi, l'approche dite MNF montre toujours ses performances pendant la nuit (période de faible consommation) dans un bâtiment tertiaire par rapport à la détection avec la courbe de charge maximale.

Perspectives

Les perspectives de ce travail concernent les directions suivantes :

— Reprendre l'étude et tester les méthodes présentées dans cette thèse sur d'autres jeux de données.

— Dans le cadre de la modélisation des courbes de charge, nous proposons de généraliser les modèles de courbes de charge quotidienne pour tous les jours à l'aide de réseaux de neurones.

— Pour l'avenir, nous avons l'intention d'étudier les performances de notre indicateur dans des environnements complexes tels que les maisons intelligentes où le comportement des utilisateurs est très variable et fluctuant. De plus, nous visons à exploiter les techniques d'intelligence artificielle afin d'améliorer l'algorithme WLICTD grâce à une meilleure sélection du paramètre λ utilisé dans la fonction de densité temporelle.

— Il est fortement recommandé de détecter les anomalies de consommation électrique.

— Pour une bonne compréhension du comportement des utilisateurs, il est souhaitable de séparer les différents appareils par les données de consommation disponibles.

— Pour des besoins écologiques, travailler sur les problèmes de réductions et de stockage des données, ainsi que sur le comportement des consommateurs.

Liste des publications

Article dans un journal international avec comité de lecture

[57] Ticherahine, A., Bourebia, S., Wira, P. et Makhlouf, A., Consumption Temporary Density for the Detection of Water Leakages in Real-time, *International Journal of Computers and Communications*, 68-74, 2019.

Conférence internationale avec comité de lecture

[52] Ticherahine, A., Boudhaouia, A., Wira, P. et Makhlouf, A., Time series forecasting of hourly water consumption with combinations of deterministic and learning models in the context of a tertiary building, *International Conference on Decision Aid Sciences and Application (DASA) (pp. 116-121)*, IEEE, 2020.

Article soumis à un journal international

- Ticherahine, A., Boudhaouia, A., Wira, P. et Makhlouf, A., Hourly prediction in a time series of water and electricity consumption in the context of a tertiary building using hybrid models. submitted to Applied Mathematics and Computation, soumis en 2021.

Articles en préparation

- Ticherahine, A., Makhlouf, A. et Wira, P., Combination of a flow-based approach and minimum night flow for the detection of water leakage, soumis en 2021.

- Ticherahine, A., Wira, P. et Makhlouf, A., Comparison of parametric descriptors for modeling load curves, soumis en 2021.

Bibliographie

[1] M. Firat, M. E. Turan, and M. A. Yurdusev, "Comparative analysis of neural network techniques for predicting water consumption time series," *Journal of hydrology*, vol. 384, no. 1-2, pp. 46–51, 2010.

[2] C. Zhong, T. Guo, Z. Jiang, X. Liu, and X. Chu, "A hybrid model for water level forecasting : A case study of wuhan station," in *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, 2017, pp. 247–251.

[3] M. E. Banihabib and P. Mousavi-Mirkalaei, "Extended linear and non-linear auto-regressive models for forecasting the urban water consumption of a fast-growing city in an arid region," *Sustainable Cities and Society*, vol. 48, p. 101585, 2019.

[4] Y. Maruyama and H. Yamamoto, "A study of statistical forecasting method concerning water demand," *Procedia Manufacturing*, vol. 39, pp. 1801–1808, 2019.

[5] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Systems with applications*, vol. 37, no. 1, pp. 479–489, 2010.

[6] S. BuHamra, N. Smaoui, and M. Gabr, "The box–jenkins analysis and neural networks : prediction and time series modelling," *Applied Mathematical Modelling*, vol. 27, no. 10, pp. 805–815, 2003.

[7] D. Kofinas, N. Mellios, E. Papageorgiou, and C. Laspidou, "Urban water demand forecasting for the island of skiathos," *Procedia Engineering*, vol. 89, pp. 1023–1030, 2014.

[8] W. Deng, G. Wang, X. Zhang, Y. Guo, and G. Li, "Water quality prediction based on a novel hybrid model of arima and rbf neural network," in *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, 2014, pp. 33–40.

[9] G. Huang and L. Wang, "Hybrid neural network models for hydrologic time series forecasting based on genetic algorithm," in *2011 fourth international joint conference on computational sciences and optimization*, 2011, pp. 1347–1350.

[10] T. Fang and R. Lahdelma, "Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system," *Applied energy*, vol. 179, pp. 544–552, 2016.

[11] J. Farajzadeh, A. F. Fard, and S. Lotfi, "Modeling of monthly rainfall and runoff of urmia lake basin using "feed-forward neural network" and "time series analysis" model," *Water Resources and Industry*, vol. 7, pp. 38–48, 2014.

[12] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The m4 competition : 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020.

[13] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *International Journal of Forecasting*, vol. 36, no. 1, pp. 75–85, 2020.

[14] R. Puust, Z. Kapelan, D. Savic, and T. Koppel, "A review of methods for leakage management in pipe networks," *Urban Water Journal*, vol. 7, no. 1, pp. 25–45, 2010.

[15] R. Mamlook and O. Al-Jayyousi, "Fuzzy sets analysis for leak detection in infrastructure systems : a proposed methodology," *Clean technologies and environmental policy*, vol. 6, no. 1, pp. 26–31, 2003.

[16] B. Farley, S. Mounce, and J. Boxall, "Field testing of an optimal sensor placement methodology for event detection in an urban water distribution network," *Urban Water Journal*, vol. 7, no. 6, pp. 345–356, 2010.

[17] S. R. Mounce, A. Khan, A. S. Wood, A. J. Day, P. D. Widdop, and J. Machell, "Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system," *Information Fusion*, vol. 4, no. 3, pp. 217–229, 2003.

[18] S. R. Mounce and J. Machell, "Burst detection using hydraulic data from water distribution systems with artificial neural networks," *Urban Water Journal*, vol. 3, no. 1, pp. 21–31, 2006.

[19] A. Boudhaouia and P. Wira, "Water consumption analysis for real-time leakage detection in the context of a smart tertiary building," in *2018 International Conference on Applied Smart Systems (ICASS)*, 2018, pp. 1–6.

[20] ——, "Power and water consumption monitoring with IoT devices and machine learning methods in a smart building," S. V. Philippe Hamman, Ed. Presses Universitaires de Strasbourg, 2019, vol. 346.

[21] I. Szilagyi and P. Wira, "Ontologies and semantic web for the internet of things-a survey," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016, pp. 6949–6954.

[22] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, "Towards the implementation of iot for environmental condition monitoring in homes," *IEEE sensors journal*, vol. 13, no. 10, pp. 3846–3853, 2013.

[23] P. Wira, "Energy management and consumption analysis with machine learning techniques," *International Workshop on Optimization in Logistics and Industrial Applications (IWOLIA)*, vol. 13, no. 10, 2019.

[24] A. Laouafi, "Contribution à la modélisation de la courbe de charge électrique par des techniques intelligentes," Thèse de Doctorat, Université 20 Août 1955, Skikda, 2017.

[25] M. El Guedri, "Caractérisation aveugle de la courbe de charge électrique : Détection, classification et estimation des usages dans les secteurs résidentiel et tertiaire," Thèse de Doctorat, Université Paris Sud-Paris XI, 2009.

[26] C. Hajjar, "Cartes auto-organisatrices pour la classification de données symboliques mixtes, de données de type intervalle et de données discrétisées," Thèse de Doctorat, Supélec, 2014.

[27] K. Atkinson and W. Han, *Elementary Numerical Analysis*, 3rd ed. Wiley, 2003.

[28] J. Shen, T. Tang, and L.-L. Wang, *Spectral methods : algorithms, analysis and applications*. Springer & Business Media, 2011, vol. 41.

[29] J.-P. Grivet, *Méthodes numériques appliquées pour le scientifique et l'ingénieur*. EDP Sciences, 2009.

[30] E. G. da Silva, "Méthodes et analyse numériques," 2007.

[31] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[32] J.-P. Becar and J. Vareille, "Des courbes et surfaces" bézier" une histoire de géométrie polaire brûlante d'actualité," in *XVIIème colloque national de la recherche en IUT, CNRIUT'11.*, 2011.

[33] A. Boudhaouia, "Analyse, classification et prédiction de consommation d'eau et d'électricité par des techniques de machine learning," Thèse de Doctorat, Université de Haute Alsace, Mulhouse, France, 2021.

[34] A. El Attar, "Estimation robuste des modeles de melange sur des donnees distribuées," Thèse de Doctorat, Université de Nantes, 2012.

[35] A. Gramacki, *Nonparametric kernel density estimation and its computational aspects*. Springer, 2018.

[36] L. Agnes, "Séries chronologiques." Cours de master 1, Université de Toulouse le Mirail, 1996.

[37] J. J. Commandeur and S. J. Koopman, *An introduction to state space time series analysis*. Oxford university press, 2007.

[38] Y. Caumel, *Probabilités et processus stochastiques*. Springer, 2010.

[39] Y. Aragon, *Séries temporelles avec R : Méthodes et cas*, reprint ed. Springer, 2011.

[40] M. Falk, F. Marohn, R. Michel, D. Hofmann, M. Macke, B. Tewes, and P. Dinges, "A first course on time series analysis : examples with sas," 2006.

[41] W. W. William and S. Wei, "Time series analysis : univariate and multivariate methods," *USA, Pearson Addison Wesley, Segunda edicion. Cap*, vol. 10, pp. 212–235, 2006.

[42] Q. Mao, K. Zhang, W. Yan, and C. Cheng, "Forecasting the incidence of tuberculosis in china using the seasonal auto-regressive integrated moving average (sarima) model," *Journal of infection and public health*, vol. 11, no. 5, pp. 707–712, 2018.

[43] S. BuHamra, N. Smaoui, and M. Gabr, "The box–jenkins analysis and neural networks : prediction and time series modelling," *Applied Mathematical Modelling*, vol. 27, no. 10, pp. 805–815, 2003.

[44] H. Liu, C. Li, Y. Shao, X. Zhang, Z. Zhai, X. Wang, X. Qi, J. Wang, Y. Hao, Q. Wu *et al.*, "Forecast of the trend in incidence of acute hemorrhagic conjunctivitis in china from 2011–2019 using the seasonal autoregressive integrated moving average (sarima) and exponential smoothing (ets) models," *Journal of infection and public health*, vol. 13, no. 2, pp. 287–294, 2020.

[45] A. V. Metcalfe and P. S. Cowpertwait, *Introductory time series with R*. Springer, 2009.

[46] S. Wheelwright, S. Makridakis, and R. J. Hyndman, *Forecasting : methods and applications*. John Wiley & Sons, 1998.

[47] G. Shmueli and K. C. Lichtendahl Jr, *Practical time series forecasting with r : A hands-on guide*. Axelrod Schnall Publishers, 2016.

[48] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with exponential smoothing : the state space approach*. Springer & Business Media, 2008.

[49] S. S. Haykin *et al.*, *Neural networks and learning machines*, 3rd ed. Prentice Hall, 2009.

[50] D. Zhang and S. Lou, "The application research of neural network and bp algorithm in stock price pattern classification and prediction," *Future Generation Computer Systems*, vol. 115, pp. 872–879, 2021.

[51] A. Boudhaouia and P. Wira, "Comparison of machine learning algorithms to predict daily water consumptions," in *2021 IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS)*, 2021, pp. 1–6.

[52] S. Saigal and D. Mehrotra, "Performance comparison of time series data using predictive data mining techniques," *Advances in Information Mining*, vol. 4, no. 1, pp. 57–66, 2012.

[53] A. Ticherahine, A. Boudhaouia, P. Wira, and A. Makhlouf, "Time series forecasting of hourly water consumption with combinations of deterministic and learning models in the context of a tertiary building," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 2020, pp. 116–121.

[54] K. B. Adedeji, Y. Hamam, B. T. Abe, and A. M. Abu-Mahfouz, "Leakage detection and estimation algorithm for loss reduction in water piping networks," *Water*, vol. 9, no. 10, p. 773, 2017.

[55] S. Dey, A. Roy, and S. Das, "Home automation using internet of thing," in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2016, pp. 1–6.

[56] H. Suo, J. Wan, C. Zou, and J. Liu, "Security in the internet of things : a review," in *2012 international conference on computer science and electronics engineering*, vol. 3, 2012, pp. 648–651.

[57] J. Spiegel, "Nouvelle stratégie de collecte de données pour les compteurs d'eau communicants," Thèse de Doctorat, Université de Haute Alsace, Mulhouse, France, 2019.

[58] A. Ticherahine, S. Bourebia, A. Makhlouf, and P. Wira, "Consumption temporary density for the detection of water leakages in real-time," *International Journal of Computers and Communications*, vol. 13, pp. 68–74, 2019.

[59] G. Roudiere and P. Owezarski, "A lightweight snapshot-based ddos detector," in *2017 13th International Conference on Network and Service Management (CNSM)*, 2017, pp. 1–7.

[60] S. Bourebia, H. Laghmara, B. Hilt, F. Drouhin, S. Bindel, J. Ledy, J.-P. Lauffenburger, and P. Lorenz, "A belief function-based forecasting link breakage indicator for vanets," *Wireless Networks*, vol. 26, no. 4, pp. 2433–2448, 2020.

[61] M. Eugine, "Predictive leakage estimation using the cumulative minimum night flow approach," *Am. J. Water Resour*, vol. 5, no. 1, pp. 1–4, 2017.

[62] T. Al-Washali, S. Sharma, F. Al-Nozaily, M. Haidera, and M. Kennedy, "Modelling the leakage rate and reduction using minimum night flow analysis in an intermittent supply system," *Water*, vol. 11, no. 1, p. 48, 2019.