Analyse, classification et prédiction de consommation d'eau et d'électricité par des techniques de machine learning

Aida Boudhaouia Miled

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)

Analyse, classification et prédiction de consommation d'eau et d'électricité par des techniques de machine learning

Thèse

préparée par

Aida BOUDHAOUIA

présentée pour obtenir le grade de
Docteur de l'Université de Haute Alsace
Discipline : Informatique

soutenue publiquement le 07 Janvier 2022 devant le jury composé de :

Pr. Yacine Ouzrout, Université Lumière Lyon 2 (rapporteur)
Dr. HDR Franck Gechter, Université Bourgogne Franche Comté (rapporteur)
Pr. Sidi Mohammed Senouci, Université de Bourgogne Franche-Comté (examinateur)
Pr. Dominique Knittel, Université de Strasbourg (examinateur)
Pr. Patrice Wira, Université de Haute Alsace (directeur de thèse)

Remerciements

Je tiens à remercier mon directeur de thèse Pr. Patrice Wira pour la confiance qu'il m'a accordé en me proposant ce sujet de thèse et de m'accepter parmi son équipe. Qu'il trouve ici toute ma gratitude pour la qualité de son encadrement et sa disponibilité.

Je remercie mes collègues stagiaires, doctorants, enseignants, techniciens et personnels de l'IRIMAS et de l'IUT de mulhouse pour leurs sympathie et convivialité.

Un immense merci aux Pr. Yacine Ouzrout et Dr. Franck Gechter pour l'immense honneur qu'ils m'ont fait en acceptant la lourde tâche de rapporter mon travail.

Je tiens également à remercier Pr. Sidi Mohammed Senouci et Pr. Dominique Knittel d'avoir accepté d'être membres examinateurs du jury de ma thèse.

Une thèse a été pour moi un défi personnel, une occupation supplémentaire et un sacrifice qui reste bien gravé pour toute ma vie comme point culminant de mon parcours.

Mon cher Mari Faîçal, tu m'as poussé à rêver et d'arriver aux étoiles. Je n'arriverais jamais à te remercier pour la réussite que j'ai eu avec toi.

Un gros câlin pour mes filles Manar, Yara et Maya ainsi que mon petit bout de chou Firas pour leurs patience et encouragements indéfectibles. Vous êtes ma puissance et mon refuge.

Enfin, je souhaite remercier toute ma famille, mes parents pour leur soutien moral sans failles ainsi que ma sœur Afef et mes frères.

"La femme a une puissance singulière qui se compose de
la réalité de la force et de l'apparence de la faiblesse."
— Victor Hugo —

Abstract

The objective of this thesis is to study users behavior as water and electricity consumers. This is achieved from data collected by smart meters and available in the form of time series of irregularly spaced events in time that represent the instants of consumption at a single measurement point. These data are also translated into the form of Load Curves (Cdc) generally sampled. After a pre-treatment, the data is used for the supervised classification of daily water quantities with different Machine Learning (ML) algorithms. Then, a new approach was proposed to determine the probability density function of the daily water consumption of users in order to reduce data and optimize computer resources. This has facilitated the development of an algorithm for detecting small as well as large water leaks. Many supervised ML approaches have been applied and individually tested for the classification of electrical devices according to the harmonic current characteristics they generate. Secondly, we have proposed hybrid approaches that combine the previous methods in order to obtain highly efficient classification results. Finally, in the context of consumption forecasting, hybrid models that combine deterministic and learning approaches have been implemented and tested to predict the next moments of consumption or the quantities of consumption expected in the future hours with much better performance.

Key words: Smart consumption meter, load curve, time series, data analysis, classification, clustering, prediction, machine learning, artificial neural networks, SARIMA, LSTM.

Résumé

L'objectif de cette thèse consiste à étudier le comportement des utilisateurs en tant que consommateurs d'eau et d'électricité. Ceci est réalisé à partir de données collectées par des compteurs intelligents et disponibles sous la forme de séries temporelles d'événements irrégulièrement espacés dans le temps qui représentent les instants de consommation (1 litre ou 1 kWh). Ces données sont également traduites sous la forme de Courbes de charge (Cdc) généralement échantillonnées par minute ou par heure. Dans ce cadre, une plateforme de monitoring de bâtiments a été développée ex nihilo afin de collecter des données qui sont issues d'un seul point de mesure regroupant de multiples usages et plusieurs usagers. Après un pré-traitement, les données servent à la classification supervisée des CdC journalières d'eau avec différents algorithmes de Machine Learning (ML). Puis, une nouvelle approche a été proposée pour déterminer la fonction de densité de probabilité de la consommation d'eau journalière des utilisateurs. Le but consiste à réduire les données afin d'optimiser les ressources informatiques dans l'étape de classification. Ceci a favorisé la mise au point d'un algorithme de détection des petites ainsi que des grandes fuites d'eau. De nombreuses approches de ML supervisées ont été appliquées et testés individuellement pour la classification des appareils électriques selon les caractéristiques harmoniques de courant qu'ils génèrent. Il s'agit des K plus proches voisins, des arbres de décision ou encore les cartes auto-organisatrices. Dans un second temps, nous avons proposé des approches hybrides qui combinent les méthodes précédentes de ML afin d'obtenir des résultats de classification très performants. Enfin, dans le contexte de la prévision des consommations, pas moins de onze modèles hybrides qui combinent des approches déterministes et par apprentissage ont été implémentées et testées. Elles sont basées sur une modélisation de séries temporelles avec l'approche SARIMA (Seasonal Autoregressive Integrated Moving Average), la méthodologie de Box et Jenkins, les réseaux de neurones artificiels multicouches et le LSTM (Long Short Term Memory). Elles ont permis de prédire les prochains instants de consommation ou les quantités de consommation prévues dans les futures heures avec des performances bien meilleures qu'avec les approches individuelles.

Mots clés : Compteur intelligent de consommation, courbe de charge, série temporelle, analyse de données, classification, clustering, prédiction, apprentissage automatique, réseaux de neurones artificiels, SARIMA, LSTM.

Table des matières

Liste des figures

Liste des tableaux

1 Introduction générale

"La vérité est pareille à l'eau qui prend la forme du vase qui la contient."
— Ibn Khaldoun —

1.1 Contexte de la thèse

Des nouveaux défis environnementaux s'imposent dans le contexte de la gestion des ressources énergétiques et hydrauliques. Afin de relever ces défis, il faut envisager des développements innovants en particulier dans l'utilisation et la préparation au quotidien de ces ressources qui n'existent qu'en quantités limitées. Les nouvelles « technologies digitales » permettent de mieux appréhender la distribution et la gestion des ressources dans les bâtiments résidentiels et tertiaires. Toutefois, les avancées technologiques dans le domaine du numérique offrent des possibilités accrues de recueil des données permettant de surveiller de manière plus fine l'usage de l'eau ou de l'électricité, et de détecter plus rapidement des anomalies dans les réseaux (fuites, gaspillage, fraudes, etc).

Dans notre étude, nous abordons la question du monitoring en temps réel des réseaux de distribution d'eau et d'électricité qui représente une continuité de la thèse de Julien Spiegel préparée à l'IRIMAS [1] et qui portait sur le développement d'un compteur transmettant des données en temps réel.

De ce fait, cette thèse, qui s'intitule « **Analyse, classification et prédiction de consommation d'eau et d'électricité par des techniques de machine learning** », s'inscrit dans le contexte de l'analyse des données des réseaux intelligents (en anglais « Data Analytics of Smart Grids ») [2] afin d'analyser des masses de données recueillies via les compteurs communicants et d'en extraire de l'information.

L'IoT [3] est une extension de l'architecture internet représentée par un réseau d'appareils connectés contenant de l'électronique, des logiciels, des capteurs, des actionneurs et une connectivité permettant à ces éléments d'interagir et d'échanger des données [4]. Cette technologie permet de rendre la collecte des données automatique via internet. Elle a permis d'exploiter

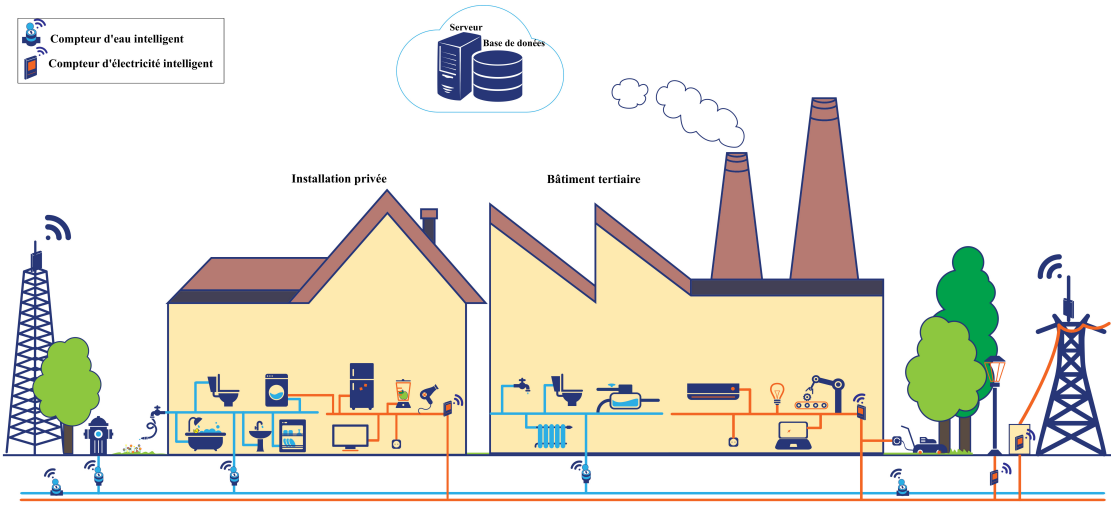

Figure 1.1: Contexte général des compteurs d'eau et d'électricité intelligents au sein des bâtiments tertiaire et privé

des stratégies AMR (Automatic Meter Reading) permettant de collecter des données de consommation d'une façon automatique et de concevoir des infrastructures de comptage avancé AMI (Advanced Metering Infrastruction). Elle permet de minimiser le temps que l'humain doit passer à gérer le système. Le concept IoT [5] vise à rendre l'architecture du système d'information plus importante pour la mise en œuvre de bâtiments intelligents d'une manière plus répandue à partir des compteurs d'eau et d'électricité intelligents. Puis, il s'agit d'explorer ces données afin de mieux cerner le comportement des utilisateurs et de monitorer le processus de consommation des ressources dans une logique d'optimisation et de développement durable. Ces données, représentées sous forme de séries temporelles et courbes de charge (Cdc) [6] représentent une grandeur qui décrit l'évolution de la consommation au cours du temps et qui sont la base pour la classification, le clustering, l'analyse et la prévision. Nous projetons dans ce cadre de recourir aux techniques de Machine Learning (ML) [7] qui ont connu une avancée importante dans la classification des données et la prédiction de la consommation future. Nous proposons différents systèmes permettant aux consommateurs de connaître leur consommation en temps quasi réel ainsi que de recevoir un message d'alarme en cas de fuite d'eau en aval de leurs propres compteurs intelligent. De ce fait, les utilisateurs peuvent prendre davantage conscience de leur consommation et éviter les gaspillages. L'analyse et la compréhension de la consommation d'eau et d'électricité permettra de se projeter vers plusieurs perspectives de recherche. On peut envisager d'autres types de consommations telles que celle de gaz.

La Fig. 1.1 représente le contexte général de différents compteurs d'eau et d'électricité intelligents pour la surveillance et la gestion de ces deux ressources au sein d'un bâtiment tertiaire et d'une installation privée.

1.2 Problématiques et objectifs de la thèse

Le nombre croissant d'appareils utilisant l'eau et l'électricité, dans les bâtiments résidentiels, industriels ainsi que commerciaux, favorise une grande consommation de ces ressources. Ils contribuent également à la polyvalence de l'utilisation quotidienne et saisonnière de l'eau dans les réseaux électriques et les réseaux de distribution d'eau (WDN) [8]. Néanmoins, le développement des stratégies innovantes pour les réseaux de distribution d'eau présente un retard considérable par rapport aux énormes études et développements réalisés sur les réseaux électriques intelligents. De ce fait, cette thèse se focalise plus sur les réseaux de distribution d'eau dans le but de les rendre intelligents.

Une surveillance continue et une analyse des données de consommation quotidienne sont judicieuses car elles apportent aux utilisateurs une bonne compréhension de leurs consommations et un contrôle plus efficace de leurs ressources. Un accès en temps réel aux données de consommation permet aux fournisseurs d'avoir une vue à jour des aspects techniques, de sécurité et de risque sur leur infrastructure [8].

Dans ce contexte, nous nous concentrerons plus particulièrement sur l'analyse des consommations d'eau et d'électricité. Pour atteindre cet objectif, une solution consiste à collecter les profils de consommation par des Cdc et des séries temporelles des consommations. Malgré la nature numérique des appareils de mesure utilisés un traitement et une discrétisation s'imposent pour harmoniser l'échelle de temps. Les données qui pourront être analysées dans ce contexte sont les index et les évènements de la consommation.

Ainsi les objectifs visés sont :

- La modélisation de la consommation d'eau et d'électricité à travers des Cdc,
- Le regroupement des usagers en classes, chaque classe étant caractérisée par un profil d'usage (profil global et/ ou hebdomadaire (niveau bâtiments), profil journalier (normale/ anormale, jours d'activités/ weekends, vacances, fériés),
- La détection des changements d'habitudes de consommation à l'échelle d'un consommateur, également à des fuites d'eau ou à des phénomènes anormaux (à l'exception des fêtes, canicules, etc),
- La prévision à court terme de la consommation d'un usager ou d'une classe de consommateurs.

Ainsi la Fig. 1.2 résume les contributions réalisées dans cette thèse. Il n'existe pas de jeu de données de consommation d'eau et d'électricité accessible, avec un échantillonnage fin et des événements d'anomalie étiquetés. Par conséquent, nous proposons, dans cette thèse, de traiter les problématiques en suivant le processus de recueil des données. De ce fait, la quantité des données utilisées pour chaque étape est variable.

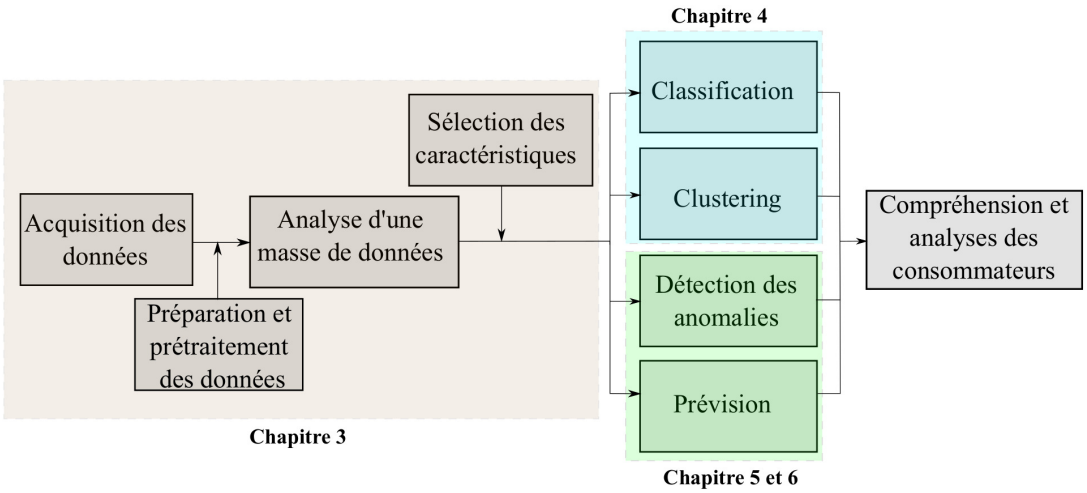

Figure 1.2: Contributions de la thèse

1.3 Publications scientifiques

Les travaux de recherche menés dans cette thèse ont été validés par les publications suivantes.

Posters scientifiques et présentations orales

- [9] **Aida Boudhaouia**, Patrice Wira, «Collecting and analyzing data from autonomous IoT sensors with application to smart buildings», Upper Rhine Cluster for Sustainability Research (URCforSR 2018): International Conference, Strasbourg, France, Septembre 24-25, 2018.
- [10] **Aida Boudhaouia**, Patrice Wira, «Detection et classification de comportements utilisateurs par apprentissage machine des consommations», Journée doctorale des Sciences Exactes, Université de Haute Alsace, Mulhouse, 12 Juin 2020.

Articles de conférences nationales

- [11] **Aida Boudhaouia**, Julien Spiegel, et Patrice Wira, «Recueil et exploitation de données issues de capteurs connectés dans un contexte de télémonitoring pour batiments intelligents», Congrès National de la Recherche des IUT (CNRIUT 2018), Aix-en-Provence, 7-8 juin 2018.

Articles de conférences internationales

- [12] **Aida Boudhaouia**, Patrice Wira, «Water consumption analysis for real-time leakage detection in the context of a smart tertiary building», IEEE, International Conference on Applied Smart Systems (ICASS'18), Medea, Algeria, p. 1-6, 24-25 November 2018.

- [13] Anissa Ticherahine, **Aida Boudhaouia**, Patrice Wira et Abdenacer Makhlouf, «Time series forecasting of hourly water consumption with a hybrid model in the context of a tertiary building», IEEE, International Conference on Decision Aid Sciences and Applications (DASA 2020), Bahrein, 8-9th November 2020.
- [14] **Aida Boudhaouia**, Patrice Wira, «Comparison of machine learning algorithms to predict daily water consumptions», IEEE, International Conference on Design & Test of integrated micro & nano-Systems (DTS 2021), Sfax, Tunisia, 07-10 Jun 2021.

Chapitre d'ouvrage

- [15] **Aida Boudhaouia**, Patrice WIRA, «Power and water consumption monitoring with Iot devices and machine learning methods in a smart building», in Sustainability Research in the Upper Rhine: Concepts and Case Studies, P. Hamman et S. Vuilleumier (eds), vol. 346, Presses Universitaires de Strasbourg, 2019.

Articles de revues internationales

- [16] **Aida Boudhaouia**, Patrice Wira, 2021, « A real-time data analysis platform for short-term water consumption forecasting with machine learning », Forecasting, 3(4), p. 682-694, 2021.
- [17] Anissa Ticherahine, **Aida Boudhaouia**, Patrice Wira et Abdenacer Makhlouf, «Hourly prediction of time series with deterministic and learning models-Application to water and electricity consumption in a tertiary building», Applied Mathematics and Computation, 2021 (Soumis).
- [18] **Aida Boudhaouia**, Patrice WIRA, «Probabilistic and learning approaches for real-time classification of time series in the context of water consumption», Applied Artificial Intelligence, 2021 (Soumis).

1.4 Organisation du manuscrit

Ce rapport de thèse est composé de sept chapitres qui sont structurés comme suit :

- Le **chapitre 2 "Algorithmes Machine Learning pour l'analyse de données de consommation"** présente un état de l'art sur les algorithmes ML. L'enjeu majeur dans cette thèse est le fait d'exprimer l'étroite symbiose entre les approches : classification, détection d'anomalie et prévision des consommations ainsi que de trouver la précision la plus accrue en ces approches. Tout cela permet de modéliser, analyser et exploiter des masses des données recueillies via les compteurs communicants et d'en extraire de

l'information. La majorité des études antérieures analysées ont adopté une précision horaire ou moindre (annuelle, mensuel, hebdomadaire, journalière) de la consommation accompagnée par d'autres indicateurs tels que la température, l'humidité, etc. Nous proposons toutefois dans cette thèse, ces techniques ML qui sont appliquées sur des données des consommations réelles issues à partir d'un seul point de mesure général.

- Le **chapitre 3 "Collecte et pré-traitement des données de consommation"** contient une description d'une infrastructure IoT au sein des bâtiments intelligents. Le système de lecture automatique et les compteurs intelligents (d'eau et d'électricité) sont abordés. La plate-forme d'acquisition des données est également présentée. Les données de consommation initialement collectées à partir de cette plate-forme sont sous la forme d'index permettant de les représenter sous forme des Cdc de consommations qui correspondent à une courbe cumulée, de débit, de la puissance moyenne, etc. Des évènements datés des consommations sont également collectés, ils correspondent à des séries temporelles des instants de consommations des litres d'eau ou des Wh de l'énergie électrique. Toutes ces données sont enregistrées sur un serveur dans une base des données à une fréquence variable inégalement espacée dans le temps. Elles seront synthétisées en Cdc journalières, hebdomadaires, mensuelles et annuelles de consommation ainsi qu'en différentes séries temporelles chronologiques. Dans ce chapitre, un pré-traitement des Cdc est proposée et une analyse d'intégrité et un échantillonnage sont proposés.

- Le **chapitre 4 "Classification automatique selon les profils des consommations"** expose notre contribution en classification automatique supervisée et non supervisée des profils de consommation d'eau ainsi que les appareils électriques. Cette classification a comme but d'explorer les classes dont la consommation semble normale et anormale.

  Tout d'abord, une classification non supervisée des Cdc de la consommation d'eau est réalisée selon deux manières différentes : directe et indirecte. La première consiste à utiliser les Cdc de la consommation d'eau sous sa forme brute recueillie du serveur. La seconde correspond à une utilisation des Cdc échantillonnées. Cette classification est réalisée avec deux approches : une classification hiérarchique ascendante et une classification avec des cartes auto-organisatrices. Une analyse probabiliste des Cdc journalières d'eau est réalisée permettant d'extraire une portion de la Cdc journalière ayant la plus d'informations. Cette étude nous permet d'identifier des périodes de consommation significative avec la fonction de densité de probabilité (PDF). Elle offre la possibilité de classifier les Cdc de la consommation d'eau plus rapidement en minimisant les nombres des données pour chaque Cdc journalière qui va permettre de détecter le plus tôt possible les comportements anormaux et les classifier plus rapidement.

  Ensuite, une classification des signatures des appareils électriques est réalisée avec trois algorithmes ML qui sont les MLP, KNN et DT. Des approches ensemblistes sont également proposées. Elles sont basées sur le principe OAA (One Against All) qui présente une classe par rapport à toutes les autres classe et l'ECOC (Error-Correcting Output Codes) permettant de corriger l'erreur de la classification. Ces deux approches ont permis

d'améliorer la classification en assurant la précision et la généralisation selon le compromis biais/variance. Une analyse plus approfondie sur les caractéristiques des appareils électriques est abordée avec un algorithme ensembliste basé sur des arbres de décision (DT) à gradient boosté. Cette analyse permet d'extraire les caractéristiques des appareils électriques selon la nature de leurs signatures harmoniques.

- Le **chapitre 5 "Détection des anomalies en temps réel et prévision des courbes de charge"** présente la contribution proposée pour détecter les anomalies liées à la consommation d'eau en temps réel. En effet, une solution complète non intrusive pour la détection des fuites d'eau est proposée. Elle s'appuie sur des Cdc et des mesures extraites effectuées sur une partie unique et centralisée d'un réseau de distribution. En effet, un algorithme de détection des fuites est proposé et qui se base sur le débit minimum nocturne (MNF), la période maximale sans consommation nulle (PWNC) et la Cdc maximale qui représente le plafond supérieur qu'une consommation ne devrait pas le dépasser. Cet algorithme a permis de détecter les petites ainsi que les grandes fuites.

  À partir de l'identification de la consommation anormale, plusieurs modèles de prévision sont proposés pour la prédiction horaire des Cdc de consommation. Ces modèles exploitent les Cdc de consommation d'eau et d'électricité permettant d'identifier un ou plusieurs modèles fournissant des précisions plus accrues et plus proche de la réalité. Nous comparons des modèles de prévision individuels à court terme de la consommation d'eau et d'électricité représentés par la décomposition déterministe en composantes essentielles de séries temporelles, le modèle SARIMA et les modèles de réseaux de neurones ; multi couches MLP et avec mémoire LSTM. La prévision offre la possibilité d'explorer la consommation en ces ressources au futur et les préparer au quotidien. Plusieurs modèles individuels et avec des combinaisons hybrides sont proposés pour la prévision afin d'identifier le modèle fournissant la précision la plus accrue.

- Le **chapitre 6 "Prévision de séries temporelles de consommations"** présente une autre contribution qui propose de prédire les séries temporelles chronologiques de consommation en termes d'évènements datés en milliseconde et en seconde. Nous allons appliquer les mêmes modèles prédictifs qui correspondent à la décomposition déterministe en composantes essentielles de séries temporelles, le modèle SARIMA et les modèles de réseaux de neurones ; multi couches MLP et avec mémoire LSTM. Une comparaison au niveau des modèles prédictifs sera proposée permettant de choisir le modèle qui fourni la précision la plus performante. Une autre comparaison de la prédiction de Cdc avec celle des séries temporelles de consommation sera également discutée afin d'étudier les avantages et les inconvénients de chaque approche.

2 Algorithmes Machine Learning pour l'analyse de données de consommation

2.1 Introduction

Actuellement, une nouvelle révolution industrielle transforme la production en une « industrie 4.0 » [19]. Elle se traduit par l'intégration des technologies numériques à tous les niveaux des chaines de production et de logistique. L'objectif de cette démarche est de collecter des informations pour permettre une supervision et une gestion quasiment en temps réel de l'ensemble d'un site de production. Disposer de données nouvelles, complètes et à jour ouvre de nouvelles perspectives aux entreprises sur différents aspects et au-delà même de la production, puisque cela peut concerner par exemple la gestion des flux entrants et sortants, les aspects de ressources humaines, la relation clients, la prospections de futurs clients, etc. En conséquence, des quantités de données massives sont collectées et tout l'enjeu réside dans le fait de pouvoir les exploiter de manière efficace. Dès lors, les approches classiques d'analyse de données semblent arriver à leur limite. D'autres techniques telles que les algorithmes ML représentent sur solutions appropriées [20]. Il s'agit là d'un ensemble d'algorithmes très génériques capable de s'adapter et de prendre en compte toutes les données pour fournir une analyse spécifique [21]. Ces algorithmes se basent ni sur des modèles ni sur des types des données particuliers.

Dans ce chapitre, un état de l'art sur ces algorithmes ML est abordé pour la classification, la prédiction ainsi que la détection des anomalies des consommations. Dans le contexte de la classification des données de consommation, une étude des techniques ML déjà existantes est proposée. Il existe peu de techniques de détection des anomalies de consommation capables de fonctionner dans tous les contextes. Nous abordons les techniques de détection des anomalies qui utilisent des données agglomérées de consommation. Dans le contexte de la prédiction, les approches ML dédiées à la prédiction des consommations seront abordées de manière détaillée. Les algorithmes et les modèles utilisés seront évalués avec des mesures statistiques. Les algorithmes seront explorés afin de choisir les plus adaptés à notre problématique, à nos données et à notre contexte.

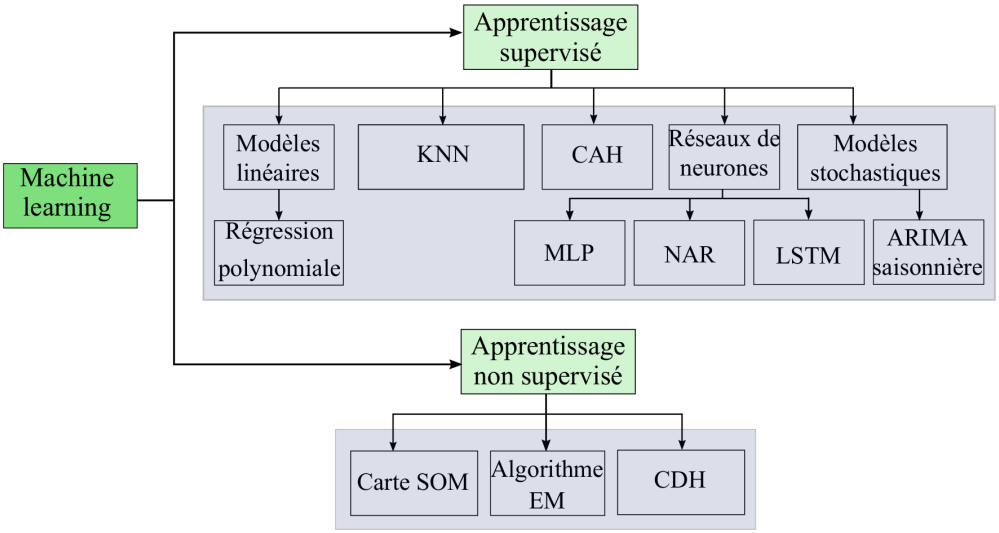

Figure 2.1: Algorithmes ML pour l'apprentissage supervisé et non supervisé utilisés dans le cadre de cette thèse

2.2 Utilisation des données par les algorithmes ML

La famille des algorithmes ML est un ensemble d'algorithmes très génériques qui a pris un essor considérable avec l'explosion des données collectées massivement. Les algorithmes s'appuient sur l'ensemble des données pour apprendre, classifier, analyser et prédire. Ce processus d'intelligence artificielle prend tout son sens avec l'arrivée des masses des données et se déroulent en trois étapes [22]. Chaque étape est réalisée sur une partition bien définie des données qui sont les données d'entrainement, les données de validation et données de test.

La première étape est l'entrainement de l'algorithme qui est exécuté sur l'ensemble des données d'entrainement pour produire des résultats qui seront comparés à des valeurs désirées. Selon cette comparaison, les différents paramètres de l'algorithme peuvent être ajustés, par une modification afin qu'ils soient valables sur l'ensemble des données de l'entrainement [23].

La seconde étape est la validation opérée sur un ensemble de données utilisées pour fournir une évaluation impartiale de l'algorithme mis en œuvre, déjà adapté sur des données, et pour régler les paramètres de modélisation afin d'améliorer l'ajustement du modèle. Il est à noter que le sous-ensemble de validation doit être différent de l'ensemble des données d'entrainement pour éviter le sur-ajustement [23] qui sera défini dans la Section 2.6.2.

Enfin, la troisième étape est celle du test dans laquelle l'algorithme est évalué avec la partie restante des données afin de fournir une évaluation finale de la modélisation et la prévision des performances. Cette étape n'influe pas sur les paramètres internes du modèle. La Fig. 2.1 illustre les algorithmes ML principalement utilisés dans cette thèse. L'évaluation des résultats fournis par les modèles ML sera discutée dans les paragraphes qui suivent.

2.3 Classification des données avec des algorithmes ML : MLP, KNN, HAC, DT et SOM

Plusieurs articles et études ont proposé des méthodes et des algorithmes pour classer, indexer, segmenter et discriminer les données de consommation qui sont représentées souvent sous forme de séries temporelles [24]. De nombreuses méthodes de classification, mesures de similarité et algorithmes ML ont été développés au cours de ces dernières années. Des nouvelles distances et stratégies ont été définies [25, 26, 27], certaines d'entre elles sont basées sur des outils ou résultats assez récents sur l'analyse des séries temporelles. Généralement, une classification est basée sur une mesure de « similarité» entre objets, mesures, grandeurs, et classes.

En lien avec notre problématique, Laspidou et al. [28] ont proposé une classification de la consommation d'eau trimestrielle de l'île de Skiathos en Grèce. Cette classification est faite avec la carte SOM afin de classifier la consommation d'eau sous différents volets à savoir le type de la résidence (privée ou tertiaire) et le nombre des occupants dans l'habitation. En outre, Wang et al. [27] ont classifié différents séries temporelles en se basant sur l'extraction de quelques mesures statistiques telles que l'auto corrélation, l'auto-similitude, etc. Ces mesures sont utilisées comme entrées pour la méthode de regroupement avec SOM. Plusieurs autres mesures ont été appliquées pour calculer la similarité des séries temporelles. La distance euclidienne (ED) et la distance dynamique de déformation temporelle (DTW) sont les mesures les plus utilisées pour calculer la ressemblance des séries chronologiques. Les méthodes de mesure de la distance dans les séries chronologiques ont été regroupées selon l'approche, la forme, les caractéristiques et le modèle comme dans [29]. Les distances ED et DTW sont toutes deux classifiées comme étant basées sur la forme.

2.3.1 Distances de mesure de la similitude

Une série chronologique S est définie comme étant une séquence de valeurs numériques exprimées par $(S_t)_{1\leq t\leq n}$ avec n observations décrivant un phénomène précis où t représente le temps. Soient deux séries temporelles X et Y définies telles que : $X = \{x_1, x_2, ..., x_{nX}\}$ et $Y = \{y_1, y_2, ..., y_{nY}\}$ qui sont inégalement espacées dans le temps avec deux tailles différentes nX et nY.

a. Distance euclidienne (ED)

La ED qui est la distance la plus connue. Elle calcule une longueur qui sépare deux points dans l'espace. Dans le cadre des séries chronologiques, l'application de cette distance nécessite d'avoir la même taille. Elle peut être calculée à partir de l'équation (2.1) :

$$ED(X, Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{2.1}$$

La ED totale entre deux séquences X et Y est la somme de la différence entre chaque paire de points de x_i et y_i de deux séries.

b. Déformation temporelle dynamique (DTW)

La DTW [25] en anglais "Dynamic Time Warping", est une distance particulière qui tient compte de la durée et de la vitesse de la série temporelle. Généralement, la DTW présente un avantage par rapport à la ED si les points des données présentent un décalage temporel défini par un avancement ou un retard. Dans ce cas, la DTW ne va pas être influencée par ce décalage et se concentre sur les formes des séries temporelles dans le temps. Cette distance est décrite dans l'équation (2.2) :

$$DTW(X,Y) = d(x_i, y_i) + (min) \begin{cases} DTW(i+1, j), \\ DTW(i, j), \\ DTW(i, j+1), \end{cases} \tag{2.2}$$

Il existe de nombreux chemins de déformation de borne exponentielle dont le but est de choisir le chemin qui minimise la distance. La variable d représente généralement la ED. La distance DTW calcule la plus petite distance entre tous les points permettant la possibilité d'une association entre un et plusieurs points. C'est une méthode qui exploite l'élasticité de l'axe temporel d'une série chronologique afin de détecter des formes similaires avec des phases différentes dans une autre série. La DTW a été appliquée dans l'extraction de séries chronologiques pour résoudre la difficulté liée à la mise en cluster [30] de séries temporelles de longueurs variables dans l'espace euclidien ou contenant des possibles similitudes déphasées.

Ces distances permettent, entre autres, de classifier des données sous forme d'un regroupement en classes. Dans le cas où nous disposons des connaissances sur le nombre des classes et si on connaît la classe à laquelle appartient chaque donnée, la classification d'un nouveau échantillon est dite supervisée. Autrement, si le nombre des classes est inconnu et si nous n'avons aucune idée sur les classes auxquelles appartiennent les données, la classification d'un nouveau échantillon est dite non supervisée. Dans ce cas on parle de regroupement et en anglais de "clustering".

2.3.2 Classification non supervisée : clustering des données avec ML

Dans le cas d'un clustering [28], le nombre des classes est souvent inconnu. De ce fait, plusieurs chercheurs tels que [31] ont proposé une approche pour estimer le nombre de clusters via la statistique des écarts. Cette statistique compare le total de la variation intra-cluster pour différentes valeurs de k avec celles attendues dans la distribution de référence nulle des données. Le nombre de cluster optimal est une valeur qui maximisera la statistique d'écart entre chaque cluter. Cela revient à vérifier qu'un échantillon x à classer devrait avoir une distance faible par rapport aux échantillons constituant le même cluster. Cependant, cet échantillon x assure une

grande distance par rapport aux échantillons constituant les autres clusters. Les étapes de la méthode statistique des écarts sont résumées dans l'algorithme 1.

Parmi les méthodes de classification la plus utilisée, nous distinguons la classification hiérarchique ascendante (HAC).

Algorithm 1 Evaluation criteria using the gap method (samples data, k_{max}) [31]

Grouping of samples data, by varying the number of clusters $k = 1, \ldots, k_{max}$
Calculation of the total intra-cluster variation W_k
Generation of reference data sets B with uniform distribution
Grouping of B by varying the number of clusters $k = 1, \ldots, k_{max}$
Calculation of the total intra-cluster variation W_{kb}
Calculation of the estimated deviation statistic: Difference of the deviation from the value W_k observed relative to its expected value W_{kb} under the null hypothesis $Cluster(K) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}^*) - log(W_k)$
Calculation of standard deviation of statistics
Choose k such that the gap statistic is within a standard deviation of the difference in $k+1$ $Cluster(K) \geq Cluster(K+1) - \sigma_{K+1}$

a. Classification hiérarchique ascendante (HAC)

La classification hiérarchique est l'une des approches les plus largement utilisées pour la classification non supervisée, mais elle reste limitée à de petits ensembles des données en raison de sa complexité de calcul quadratique ($o(n^2)$). La HAC est appliquée d'une manière agglomérative. La HAC part des individus seuls que l'on regroupe en sous-ensembles, qui sont à leur tour regroupés, et ainsi de suite. Cette technique permet en une classification non supervisée pour identifier les clusters.

La HAC consiste à calculer une matrice de distance (ED ou DTW) de chaque paire des données, par exemple une paire de Cdc. L'algorithme commence par n classes constituant les singletons tels que chaque entrée constitue une série de valeurs. Il cherche, à chaque étape, à former des classes par une agrégation de chaque paire de série de valeurs les plus proches en distance du regroupement précédent. Les groupements successifs sont représentés sous la forme d'un arbre binaire appelé dendrogramme [32]. Les étapes de la méthode de la classification hiérarchique sont résumées dans l'algorithme 2.

b. La carte auto-organisatrice SOM

La carte SOM est un algorithme d'auto-organisation qui a été proposé par Kohonen [22]. Cet algorithme projette l'espace des données à classer $\mathfrak{D}$ sur un espace discret appelé carte $\mathfrak{C}$. Elle est constituée par un ensemble de neurones inter-connectés. Les neurones de cette carte sont organisés dans une seule couche qui peut être considérée comme une grille multidimensionnelle sans couche de sortie. Les neurones l'organisent pour refléter la distribution des données.

Algorithm 2 HAC

Distance = (X(x_i,:)-Y(y_j,:))
Matrice Distance = squareform(Distance) ▷ Rebuild the distance matrix
repeat
  New groupe = identify and group the two classes closest to the previous grouping partition
  Update Matrice Distance by replacing the two classes grouped by the new New groupe
  Calculate Distance of New groupe with each of the other classes
until Aggregation into a single class

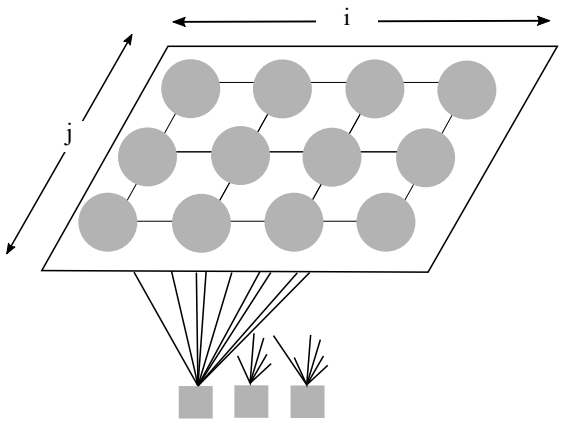

Figure 2.2: Architecture générale d'une carte auto-organisatrice SOM

La Fig. 2.2 montre l'exemple d'une grille de la carte SOM. La position d'un neurone au sein de la grille est donnée avec ses indices i et j dans un repère 2D. Cette position est fondamentale et permet de définir la notion de voisinage. Chaque neurone possède des poids qui définissent sa position dans l'espace des entrées. L'apprentissage consiste à mettre à jour les poids qui vont engendrer une mise à jour de la position de chaque neurone.

2.3.3 Classification supervisée des données avec ML

a. Le perceptron multicouche (MLP) pour une classification supervisée

Un réseau neuronal [33] est un système inspirée du fonctionnement des neurones biologiques qui a une propension naturelle à stocker des connaissances expérientielles et à les rendre utilisables. Il présente les connaissances acquises par le réseau à travers un processus d'apprentissage qui peut être réalisé via un séquençage d'informations ou un cycle récurrent.

En général, un réseau de neurones est constitué d'un ensemble de neurones inter-connectés 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. Le réseau MLP est le réseau de neurone le plus utilisé. Cela est du à sa capacité d'apprendre de l'ensemble des données d'entraînement et à son efficacité à résoudre des problèmes de

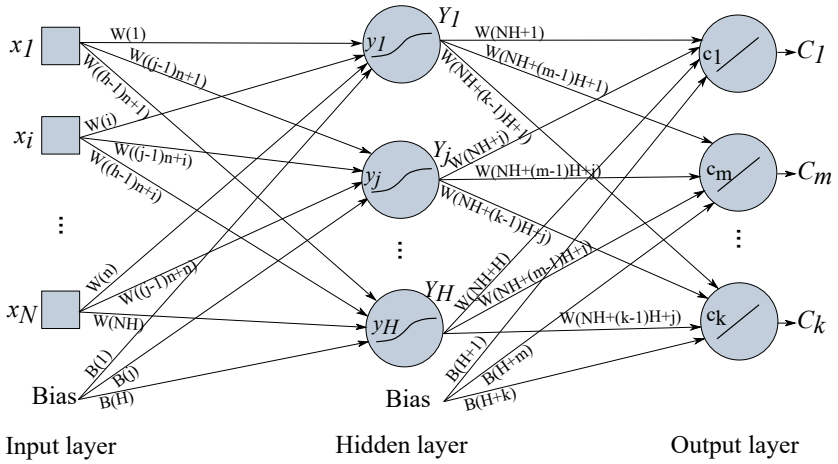

Figure 2.3: Architecture générale d'un réseau de neurone MLP

classification non linéaire et de reconnaissance de formes. La structure et le principe du MLP sont détaillés ci-dessous.

Les neurones du réseau MLP sont structurés en couches (les couches cachées et la couche de sortie). Un réseau MLP a une couche d'entrée, une couche de sortie et une ou plusieurs couches cachées. Il n'y a pas de neurone artificiel dans la couche d'entrée du MLP. Les données entrent dans le réseau via la couche d'entrée, traversent les couches cachées et finissent par sortir du réseau par la couche de sortie. L'architecture de ce réseau est illustrée par la Fig. 2.3.

Les sommes pondérées des neurones sont calculées comme suit :

$$y_j = \sum_{i=1}^{N} (W((j-1)N+i) \times x_i) + B(j), \tag{2.3}$$

avec *W((j-1)N+i)* le poids de la i^{th} entrée au j^{th} neurone caché. *B(j)* est le biais de la j^{th} neurone caché avec $j = 1, 2, ..., H$. La sortie de chaque neurone caché est alors donnée en appliquant une fonction d'activation sigmoïde par (2.4) par exemple, et la classe d'appartenance est obtenu via l'équation (2.5), avec $m = 1, 2, ..., k$ classes. Des fonctions d'activation linéaires sont appliquées pour chaque neurone de sortie.

$$Y_j = sigmoid(y_j) = \frac{1}{(1 + exp(-y_j))} \tag{2.4}$$

$$C_m = \sum_{i=1}^{h} (W(NH + (m-1)H + j) \times S_j) + B(H + m), \tag{2.5}$$

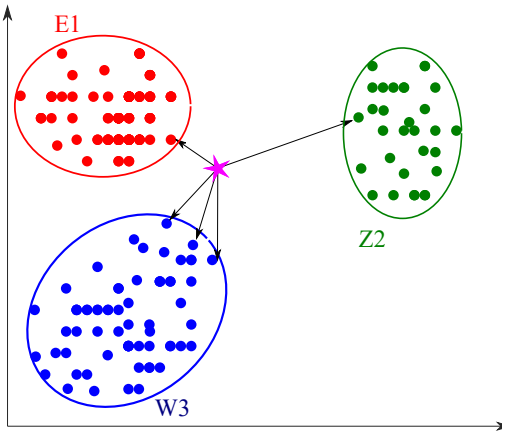

Figure 2.4: Exemple de classification supervisée avec le modèle KNN

b. K-plus proches voisins KNN

La méthode des k-voisins les plus proches, également connue sous le nom de KNN [34], consiste à classer des instances inconnues en fonction de leurs distances par rapport aux points des données (voisins) constituant des échantillons antérieurs. Cette méthode parcourt tous les cas disponibles et classe des nouveaux cas en fonction d'une mesure de similarité. Généralement, la ED est employée comme mesure de distance. Elle consiste à classifier une instance inconnue en fonction de sa distance par rapport à des points constituant un échantillon d'apprentissage connu a priori appelés voisins. Les résultats de cette méthode sont représentés par la Fig. 2.4.

L'algorithme du KNN consiste à construire une structure des données P pour un point d'entrée q. On a un point $x \in P$ qui minimise $ED(x,q)$ peut être trouvé rapidement, avec ED à partir de la distance euclidienne définie dans l'équation (2.1) avec :

$$ED\left(\left(e_1, e_2, ..., e_j\right)_E, \left(w_1, w_2, ..., w_j\right)_W, \left(z_1, z_2, ..., z_j\right)_Z\right) \tag{2.6}$$

Ensuite le plus proche voisin de la nouvelle instance des données q est donné par :

$$ED\left(q, x_s\right) = min_i ED\left(x, x_i\right) \tag{2.7}$$

Ainsi on choisit le k plus proche voisin et la classe de l'échantillon q est déterminé en fonction de la classe majoritaire parmi les k plus proches voisins de l'échantillon q.

Le choix du nombre de voisins k est libre, avec $k \in \mathbb{N}$ il est très utile de choisir k impair pour éviter toute égalité entre le choix. Choisir des grandes valeurs pour k réduit le bruit de la classification mais rend les frontières entre les classes moins distinctes. Afin de trouver un compromis, plusieurs méthodes peuvent être considérées pour avoir le bon choix. On peut citer les exemples de la validation croisée et la minimisation de l'erreur de classification [22].

2.4 Détection des anomalies de consommation

Détecter des données anormales est une tâche primordiale en apprentissage automatique et pour la fouille des données. Dans le contexte de la consommation d'eau et d'électricité, ces données anormales peuvent souvent refléter un comportement inhabituel en consommation qui se manifeste par un changement des habitudes de consommation, une fuite d'eau, un appareil défectueux, un problème au niveau de la canalisation, un sur-éclairage, des équipements inefficaces, etc. En effet, la mise en œuvre de cette détection permet d'améliorer considérablement les performances des algorithmes ML et ainsi comprendre le comportement des données associées dans le temps. La détection des anomalies est largement utilisée dans d'autres domaines d'application comme la détection des fraudes par carte de crédit bancaire, l'assurance ou la santé, les télécommunications [35], la détection d'intrusion dans la cybersécurité [36], etc.

Dans le contexte de la consommation d'eau et d'électricité, une anomalie peut être définie de plusieurs manières différentes. Par exemple, dans le cas d'un bâtiment tertiaire (un supermarché, école, administration, etc), la consommation en électricité et en eau est généralement très faible pendant les jours fériés et les weekends par rapport aux autres jours de la semaine [37, 38]. Ainsi, on peut considérer anormale une consommation, pendant ce type de jours, qui ressemble à une consommation d'un jour de travail.

Dans le cadre de notre travail de thèse, toute consommation peut être qualifiée d'anormale ou non anormale uniquement après l'avoir comparée aux données historiques ou/et avoir constaté cette anomalie sur le terrain. Selon [37, 39], les anomalies sont globalement classées en trois types : les anomalies ponctuelles, les anomalies collectives et les anomalies contextuelles. Une anomalie de points est définie lorsqu'une observation individuelle est considérée comme anormale par rapport au reste des données. Par exemple, un point dans une Cdc qui peut être aussi considéré comme pic critique injustifié. Une anomalie séquentielle ou collective est identifiée lorsqu'une séquence d'observations est anormale par rapport au reste des données. Par exemple, une Cdc qui s'écarte des autres observations et ayant un comportant différent par rapport à la majorité des données. Cependant, une anomalie contextuelle est constatée lorsqu'une observation est considérée comme normale par rapport à un contexte mais pas dans un autre contexte. Nous pouvons citer le cas d'un comportement de consommation en semaine par rapport au weekend. La compréhension, l'identification et la résolution de la consommation d'eau et d'électricité anormale dans les bâtiments peuvent conduire à des économies de ces ressources et à la détection d'appareils défectueux.

Les compteurs intelligents qu'ils soient d'eau ou d'électricité offrent la possibilité de détecter efficacement [40] un appareil laissé en fonctionnement par erreur, une panne de compresseur dans un réfrigérateur ou de robinet d'eau laissé partiellement en position ouverte, l'installation et l'utilisation d'un nouvel appareil, des invités en visite pendant une longue période (période des fêtes), un changement de saison et un climat exceptionnellement trop chaud/froid. Un modèle de consommation d'énergie pourrait également être impliqué afin de détecter des phénomènes rares et imprévus d'une activité malveillante comme le vol d'électricité [37] ou d'eau.

La classification des données est importante, pour permettre la recherche de similarités et les anomalies parmi les données de consommations. Concrètement, le but de cette tâche est d'apprendre, à partir d'un ensemble des données, un modèle reflétant au mieux la "normalité" pour détecter comme anormale toute autre donnée s'écartant significativement de ce dit modèle.

Parmi les recherches d'anomalies de la consommation d'eau [41, 42, 43, 44], nous trouvons le plus souvent la détection des fuites et les grands pics de consommation. Dans ce contexte, plusieurs approches ont été proposées. Citons par exemple [45] qui a proposé de détecter et d'estimer des fuites sur la base du débit mesuré dans les réseaux de distribution d'eau. Cette analyse est réalisée par le calcul du débit des nœuds au niveau de la canalisation, puis un contrôle est effectué avec une valeur de tolérance prédéfinie. Avec cette approche, une méthode pour identifier des tuyaux critiques, ayant une probabilité plus élevée à avoir des fuites d'eau, a été proposée. Une autre approche [46] basée sur la logique floue a permis de détecter les fuites dans l'infrastructure d'un pipeline. Cette approche s'appuie sur la typologie (matériau, longueur, diamètre et âge des tuyaux), et l'environnement (demande, topographie et pressions). Les auteurs ont également intégré d'autres paramètres qui représentent les conditions de fonctionnement (taille de la population, logements et caractéristiques socio-économiques, niveau de vie). Ces auteurs ont résolu le problème des fuites d'eau mais leur approche sera inadéquate une fois que le réseau de distribution d'eau est modifié. Une autre étude [47] a proposé d'utiliser des capteurs de détection de mouvement, des thermostats, des caméras, etc. Cependant, la mise en œuvre de ces équipements est très peu pratique et très couteuse et peut atteindre 75% du coût total de l'installation. Par conséquent, nous optons pour une détection des anomalies de consommation avec une approche simple qui ne nécessite qu'un seul point de mesure et qui est moins couteuse. Les données à utiliser sont représentées sous forme de Cdc des consommations synthétisées sous forme des séries temporelles chronologiques.

2.5 Prévision de consommation

La prévision [48, 49, 50, 51] est un sujet très répandu et ses applications se retrouvent dans de nombreux domaines : dans la finance pour la prédiction de l'évolution des cours de la bourse, en météorologie pour la prédiction à court terme de la température, pour la prédiction du nombre de réservations d'un vol, etc. Dans le contexte de cette thèse, les processus des consommations sont manipulés comme des séries temporelles. Il existe un large éventail de méthodes et modèles pour prévoir les valeurs futures d'une série chronologique. Ces méthodes sont classées en deux grandes catégories. La première, représente les méthodes déterministes, qui consiste à supposer que la série chronologique est parfaitement prévisible à l'aide de ses valeurs passées structurées en différentes représentations mathématiques. La deuxième catégorie qui regroupe les modèles stochastiques sont identiques aux méthodes déterministes mais considèrent également une structure aléatoire probabiliste de corrélation non nulle [52, 53]. La prévision [54] des consommations est énumérée sur une variété d'horizons temporels à très court-terme de quelques minutes à une heure, à court-terme de quelque heures à une semaine, à moyen-terme de quelques semaines à un an et à long-terme pour prédire plusieurs années [52].

Le choix d'un modèle est basé sur certaines caractéristiques, comme par exemple le type des données à fournir au modèle, le temps de calcul requis, ainsi que l'horizon temporel à envisager. Néanmoins, très peu de travaux de recherches se sont confrontés à des installations individuels [55, 56] et à une résolution accrue des données.

Quelques travaux de recherches qui ont traité la prévision des consommations avec des modèles simples appelés individuels. Nous citons par exemple [57] qui se base sur un modèle des chaines de Markov non homogènes permettant de connaître la dynamique de la consommation d'eau journalière. Ce modèle est capable de prédire les comportements de la consommation d'eau journalière en fonction des facteurs exogènes tels que le climat, le type du jour, etc. Une autre étude [58] traite la prédiction de la demande en eau aux échelles hebdomadaire et horaire avec un modèle autorégressif développé sur la base d'une composante périodique sur les données de séries chronologiques. Cette prédiction est basée sur des modèles à une multitude de périodes.

La plupart de ces études se concentrent sur la prévision de la consommation en introduisant d'autres paramètres et en utilisant différents modèles prédictifs selon la nature des données d'entrées et les objectifs visés. En effet, nous constatons que l'horizon des prévisions fournies dépendent énormément des bases des données d'entrées aux modèles. Ces bases des données ont généralement une résolution annuelle, saisonnière, mensuelle, hebdomadaire, journalière ou horaire. Pour notre cas, nous projetons de fournir des prévisions plus précises avec une base des données des consommations à résolution en milliseconde avec quatre modèles qui sont le modèle déterministe, SARIMA, MLP et LSTM.

2.5.1 Modèle déterministe basé sur la décomposition de la série temporelle

Un modèle de série temporelle est une équation précisant la façon dont les composantes représentées par la tendance, la saisonnalité et le bruit s'articulent les unes par rapport aux autres pour constituer cette série chronologique. Nous supposons qu'un modèle déterministe, représenté par une série temporelle X_t, est décrit selon une fonction temporelle et une variable ε_t représentant l'erreur.

Une analyse statistique consiste à exprimer la série temporelle de la consommation en fonction de ses composantes essentielles. Ces dernières sont définies avec : une composante fondamentale qui représente l'évolution générale du phénomène étudié caractérisé par la tendance, un mouvement saisonnier qui reflète les variations saisonnières, et une composante aléatoire représentant des variations accidentelles imprévisibles assimilé à un résidus [51, 52].

La tendance Z_t est une fonction monotone qui traduit l'évolution vers le haut, croissant, ou vers le bas, décroissant, du niveau moyen du phénomène par rapport au temps. La tendance générale d'une série temporelle est évaluée selon les deux extrémités X_1 et X_n. Elle peut être linéaire ou non linéaire.

Ensuite, la saisonnalité S_t représente une période connue p qui se répète de façon plus ou

moins identique tout au long de la période T. Dans notre cas d'étude, la consommation d'eau (éventuellement d'électricité) peut contenir une saisonnalité journalière, hebdomadaire, etc.

Finalement, la composante résiduelle ϵ est une variation accidentelle qui représente la partie non structurée de la série temporelle. Elle est modélisée par une suite de variables aléatoires ϵ_t, $t = 1, 2, ..., n$ centrées, non corrélées et de même variance (bruit blanc).

Selon [59], il existe trois schémas de décomposition du modèle de la série temporelle :

1. Le schéma de décomposition additif qui représente la somme de trois composantes. L'équation (2.8) décrit cette somme.

$$X_t = Z_t + S_t + \varepsilon_t \tag{2.8}$$

   L'amplitude des variations saisonnières est constante autour de la tendance. Ce type de schéma constitue un modèle additif linéaire où les changements au fil du temps sont systématiquement de la même variation.

2. Le schéma de décomposition multiplicatif de sorte que la série X_t est le produit de trois composantes :

$$X_t = Z_t \times S_t \times (1 + \varepsilon_t) \tag{2.9}$$

   L'amplitude des variations saisonnières varie. Ce type de schéma constitue un modèle multiplicatif non linéaire, quadratique ou exponentiel. Les changements augmentent ou diminuent avec le temps.

3. Le schéma de décomposition est mixte, de sorte que la série X_t combine l'addition et la multiplication des éléments saisonniers. Dans ce cas, nous pouvons supposer, par exemple, que la composante saisonnière agit de façon multiplicative, alors que les fluctuations irrégulières sont additives. Cette décomposition est définie par l'équation (2.10) suivante :

$$X_t = Z_t \times S_t + \varepsilon_t \tag{2.10}$$

où Z_t est la tendance, S_t représente la composante saisonnière de la série.

Pour identifier le schéma de décomposition du modèle déterministe [60], nous pouvons nous baser soit sur une méthode graphique soit sur une méthode analytique. Chaque méthode est appliquée selon des techniques différentes définies comme suit :

A) Méthode graphique :

   i. Technique du test de bande

Sur le graphique de la série temporelle, nous traçons une droite passante par les minima de la courbe de la série et une autre droite passant par les maxima. Si les deux droites sont parallèles, on est en présence d'un modèle additif. Le cas contraire, on est en présence d'un modèle multiplicatif.

ii. Technique du profil

Nous traçons des courbes en superposition sur celle de la série temporelle. Si les différentes courbes sont à peu près parallèles, alors on est en présence d'un modèle additif. Le cas contraire, on est en présence d'un modèle multiplicatif.

B) Méthode analytique :

i. Technique du tableau Buys et Ballot

La méthode consiste à calculer d'abord les moyennes et les écarts types de la série temporelle pour chaque période s. Ensuite,une estimation des coefficients de la droite de moindres carrés $\bar{X}_t$ sur les nuages de points de l'abscisse qui représente la moyenne et l'axe d'ordonnée qui représente l'écart type σ_t, on obtient l'équation (2.11) :

$$\sigma_t = \alpha \bar{X}_t \tag{2.11}$$

$\alpha = 0$, signifie que la pente des moindres carrées est égale (ou très proche) de 0 alors on déduit que le modèle est additif. Sinon, $\alpha \neq 0$ signifie que la pente des moindres carrées n'est pas nulle. Alors, on peut dire que le modèle est multiplicatif.

Après avoir déterminé et choisi le type du modèle, additif ou multiplicatif, nous procédons par la suite à estimer chaque paramètre afin de construire un modèle de série temporelle pour la prévision future. L'estimation des paramètres est effectuée par la recherche de la tendance et de la saisonnalité décrite comme suit :

- La tendance, qui définie l'évolution de la série par rapport au temps, est estimée avec la méthode de Mayer [52] ou la méthode des moindres carrés [60]. Dans notre cas, nous avons choisi la méthode des moindres carrés basée sur l'équation d'une droite : $Z_t = at + b$.
- La saisonnalité est déterminée à partir des coefficients c_i. Chaque coefficient est calculé par l'équation suivante :

$$c_i = \frac{1}{p} \sum_{j=0}^{n} R_{i+j\times p}, \tag{2.12}$$

avec $i = 1, ..., p.$, R_t représente le rapport de correction de la série par rapport à la tendance. Si la série suit une décomposition multiplicative, donc $R_t = \frac{X_t}{Z_t}$, sinon $R_t = X_t - Z_t$.

Les coefficients de la saisonnalité normalisés sont identifiés par :

- $\hat{c}_i = c_i \frac{1}{p} \sum_{i=1}^{p} c_i$ si le modèle est multiplicatif,

– $\hat{c}_i = c_i - \frac{1}{p}\sum_{i=1}^{p} c_i, i = 1, ..., p$, dans le cas où le modèle est additif.

2.5.2 Modèle stochastique saisonnier autorégressif à moyenne mobile intégrée (SARIMA)

Le modèle de moyenne mobile autorégressive intégrée avec saisonnalité, ou SARIMA, est une approche stochastique pour la modélisation des données de séries temporelles univariées qui peuvent contenir des composantes de tendance et saisonnalité. Elle est une approche efficace pour la prévision de séries chronologiques, bien qu'elle nécessite une analyse minutieuse et une expertise du domaine afin de configurer les paramètres. Cependant, il n'est pas facile de les identifier. Pour faciliter la tâche, les processus stochastiques utilisent des fonctions de différentiation, d'auto-corrélation et d'auto-corrélation partielle pour aider à identifier un modèle acceptable. Parmi ces modèles, nous distinguons le modèle SARIMA [49].

Dans les modèles stochastiques, ε_t est une variable aléatoire représentant le bruit d'un processus aléatoire [60, 61]. Un modèle ARIMA saisonnier ou SARIMA s'écrit avec la notation suivante : SARIMA$(p, d, q)(P, D, Q)_S$.

Pour configurer un modèle SARIMA, il faut définir six paramètres. Les trois premiers paramètres servent à décrire les éléments de la tendance de la série qui sont :

$$\begin{cases} 1.\ p\ : \text{l'ordre de l'autorégressive.} \\ 2.\ d\ : \text{l'ordre de la différence de la série.} \\ 3.\ q\ : \text{l'ordre de la moyenne mobile.} \end{cases}$$

En plus des paramètres de la tendance, le modèle SARIMA nécessite une deuxième configuration de trois autres paramètres décrivant la saisonnalité S :

$$\begin{cases} 1.\ P\ : \text{l'ordre de la partie autorégression saisonnière.} \\ 2.\ D\ : \text{l'ordre de la différence saisonnière.} \\ 3.\ Q\ : \text{l'ordre de la moyenne mobile saisonnière.} \end{cases}$$

Le modèle SARIMA est un modèle non déterministe [62] adapté à une série périodique et stationnaire par différentiation et il est définie par :

$$\Phi(B)F(B)(I-B)^d(I-B^s)^D X_t = \theta(B)\theta'(B^s)\varepsilon_t \tag{2.13}$$

où $\Phi(B)$, $F(B)$, $\theta(B)$, $\theta'(B)$ sont des polynômes de degré respectivement p, P, q, Q, B l'opérateur

de retard avec $B^d X_t = X_{t-d}$ et ε_t le bruit avec :

$$\begin{cases} .\nabla^d = (I-B)^d \\ .\nabla_s^d = (I-B^s)^d \\ .\Phi(B) = I - \varphi_1 B - ... - \varphi_P(B^p), (\varphi_1, ..., \varphi_p) \in \mathbb{R}^p \ et \ \varphi_P \neq 0 \\ .\Phi'(B) = I - \varphi'_1 B - ... - \varphi'_P(B^P), (\varphi'_1, ..., \varphi'_p) \in \mathbb{R}^P \ et \ \varphi'_P \neq 0 \\ .\Theta(B) = I + \theta_1(B) + ... + \theta_q(B)^q, (\theta_1, ..., \theta_q) \in \mathbb{R}^q \ et \ \theta'_q \neq 0 \\ .\Theta(B) = I + \theta'_1(B) + ... + \theta'_Q(B)^Q, (\theta'_1, ..., \theta'_Q) \in \mathbb{R}^Q \ et \ \theta'_Q \neq 0 \end{cases} \tag{2.14}$$

La méthode *Box-Jenkins* [63] permet la modélisation probabiliste et non déterministe de la série temporelle basée sur une méthode heuristique pour majorer et chercher les paramètres p et q du modèle SARIMA. La méthode comporte essentiellement cinq étapes principales [64] qui sont :

1. La transformation de la série : c'est la première étape qui consiste à préparer la série temporelle avec la recherche du paramètre d. Si la série présente une composante saisonnière, alors nous commençons à l'éliminer pour la rendre stationnaire. On identifie la saisonnalité, puis la stationnarité de la série en commençant par D qui élimine la périodicité ensuite d qui rende la série stationnaire. Selon la représentation de l'auto-corrélation partielle, nous pouvons détecter la stationnarité de la série temporelle. En effet, la série est considérée stationnaire si sa fonction d'auto-corrélation décroit rapidement et se rapproche de zéro.

   Une autre méthode pour vérifier la stationnarité de la série temporelle est celle du test d'hypothèse de non stationnarité de Dickey-Fuller augmenté (ADF) [53]. Elle vérifie la stationnarité par un rejet de l'hypothèse nulle du test au seuil de 5%.

   Par la suite, dans le cas où la série temporelle est non stationnaire, nous appliquons une dérivation simple d fois jusqu'à la série devient stationnaire.

2. L'identification des paramètres p, q, P, Q à partir d'une sélection du modèle SARIMA. La sélection est réalisée en se basant sur le critère d'information Aikaike (Akaïke Information Criterion : AIC) où le critère d'information Bayésien (Bayesian Information Criterion : BIC). Ces deux critères se reposent sur la notion de vraisemblance. Pour choisir le modèle le mieux adapté on procède à une minimisation de l'un des critères. On a :

   $$AIC = -2\log(L(\theta)) + 2v \tag{2.15}$$
   $$BIC = -2\log(L(\theta)) + nv \tag{2.16}$$

   avec v le nombre de paramètres. θ est le vecteur qui contient les paramètres p et q. n est le nombre des observations et L est la vraisemblance [65]. La fonction de l'auto-corrélation et la fonction d'auto-corrélation partielle constituent d'autres moyens pour identifier et choisir les paramètres.

3. L'estimation des modèles par la recherche des coefficients des polynômes Φ, F, θ, θ' du

modèle SARIMA par la méthode des moindres carrées classique, la méthode de Yule-Walker ou la méthode du maximum de vraisemblance. La méthode la plus utilisée est celle du maximum de vraisemblance.

4. La validation du modèle par l'analyse des résidus qui devraient vérifier certaines propriétés statistiques qui sont défini comme suit :
   - la normalité vérifiée par la représentation de l'histogramme des résidus et le graphe de la loi normale ou par le graphe de quantile-quantile (qq-plot).
   - l'absence d'auto-corrélation vérifié par les tests paramétriques de Box et Pierce (1970) et de Ljung et Box (1978) ou par les tests non paramétriques de retournement et le test de monotonie [64].
   - l'homoscédasticité qui est identifiée par le test de White (1980) ou le test de Engel (1982) [64].
5. La prévision de la série qui va confirmer le choix du modèle est faite selon une simulation sur un ensemble d'échantillons de test. La formule de prévision est donnée par :

$$X_{t+k} = \sum_{i=1}^{p} \alpha_i \bigtriangledown^D X_{t+k-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t+k-i} \tag{2.17}$$

2.5.3 Réseau de neurones multi couche (MLP)

Un MLP (ou un réseau de rétropropagation) [66], tel qu'il est défini dans la Section 2.3, dans le contexte de la classification supervisée, est utilisé de la même façon dans le contexte de la prévision.

Ce réseau est composé d'une couche d'entrée, d'une couche de sortie et d'une couche cachée inter-connectées avec des poids initialisés aléatoirement. Comme le montre la Fig. 2.3 qui représente la topologie typique d'un réseau de neurones MLP. Autrement dit, la couche d'entrée ne représente que l'étape d'association de l'entrée avec la couche cachée de traitement.

Le processus d'apprentissage de réseau de neurone MLP se compose de deux grandes étapes [22, 67]. La première étape consiste en une propagation directe de l'information à partir de la couche d'entrée, vers les couche cachée jusqu'à la couche de sortie. La deuxième étape consiste en une comparaison entre le résultat trouvé avec le réseau de neurone et celui attendu. Lorsque la différence est grande, l'erreur sera rétropropagée et sera distribuée à chaque nœud de chaque couche. Une mise à jour des poids est réalisée en se basant sur l'optimisation du gradient descendant. La mise à jour des poids est faite avec l'équation (2.18) définie par :

$$\omega_{new} = \omega - \alpha \frac{\delta j}{\delta w} \tag{2.18}$$

Le processus de correction pondérée continuera à mettre à jour les poids du modèle avec les autres données de la base d'apprentissage. La formation du réseau se poursuivra jusqu'à ce que

l'erreur finale atteigne une plage acceptable ou atteigne une valeur prédéterminée de temps d'apprentissage.

Le vecteur d'entrée du MLP se compose des échantillons de séries chronologiques précédents : $X(t) = [x_1, x_2, ..., x_i, .., x_n]$. Soit $h = [h_1, h_2, ..., h_j, .., h_m]$ le vecteur représentant les neurones de la couche cachée et son processus d'apprentissage est décrit comme suit :

$$\begin{cases} h_j = f(net_j), \\ net_j = \sum_{j=1}^{m} v_{ij} x_i, \end{cases} \tag{2.19}$$

où m représente le nombre des neurones dans la couche cachée.

Le vecteur de sortie est représenté par $\hat{y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_k, .., \hat{y}_l]$ et son processus d'apprentissage est donné par l'équation suivante :

$$\begin{cases} \hat{y}_k = f(net_k), \\ net_k = \sum_{j=1}^{m} w_{jk} h_j, \end{cases} \tag{2.20}$$

l est le nombre des neurones dans la couche de sortie. $V = [v_1, v_2, ..., v_j, .., v_m]$ et $W = [w_1, w_2, ..., w_k, .., w_l]$ représentent respectivement les vecteurs de poids entre la couche d'entrée et la couche cachée et les poids entre la couche cachée et la couche de sortie qui sont mise à jour avec l'équation (2.18).

2.5.4 Réseau récurrent à mémoire court et long terme (LSTM)

Le Long Short Term Memory (LSTM) [68, 69], plus explicitement réseau de neurones récurrents à mémoire court-terme et long terme, est un type particulier de réseaux des neurones récurrents (RNN) [68, 70]. Dans un RNN, chaque neurone ou unité de traitement est capable de maintenir l'état interne ou la mémoire afin de conserver les informations associées à l'entrée précédente. Cette fonctionnalité est particulièrement importante dans de nombreuses applications liées aux séries chronologiques. L'idée principale derrière ce type d'architecture de réseau neuronal est la prise en compte du temps. Le nom de ce réseau neuronal est dérivé du fait que ces types de réseaux fonctionnent de manière récursive. Une opération est effectuée pour chaque élément d'une séquence dont sa sortie dépend de l'entrée courante et des opérations précédentes. Ceci est accompli en réutilisant une sortie du réseau à l'instant t avec l'entrée réseau à l'instant $t + 1$ (c'est-à-dire que la sortie de l'étape précédente est combinée avec la nouvelle entrée dans le nouvel étage). Ces cycles permettent l'existence d'informations d'une étape à l'autre.

Le LSTM [71] est considéré comme le modèle le plus adéquat pour la prévision des données stochastiques et bruitées telles que les données liées à la consommation qui dépendent du temps. Cette utilisation est en adéquation avec la capacité de déduire les pics et les oscillations des courbes des consommations en fonction du temps.

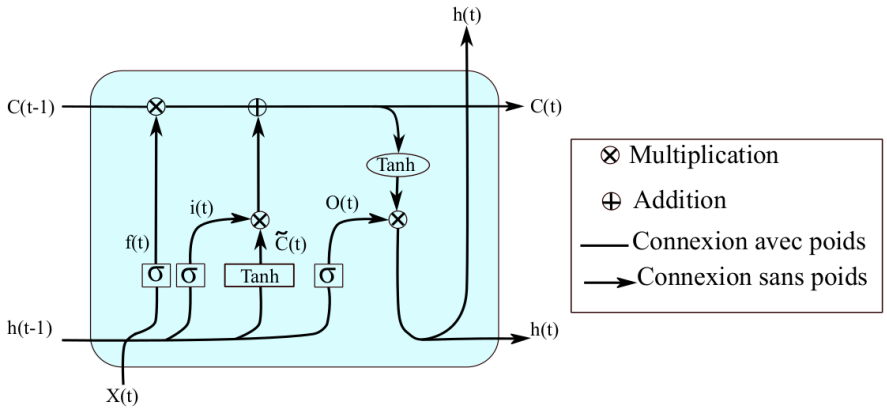

Figure 2.5: Architecture interne d'une unité du réseau de neurones récurrents à mémoire court-terme et long terme LSTM

Le LSTM est basé sur la rétro-propagation à travers le temps (BPTT) [70] pour calculer les poids. Ce RNN possède des modules contenant 4 couches qui communiquent dans une structure spéciale. Ils sont composés par des unités appelées les blocs mémoire. Chaque bloc mémoire contient une porte d'entrée "input gate", une porte de sortie "output gate" et une porte d'oublie "forget gate". L'architecture interne de l'unité LSTM est présentée par la Fig. 2.5. Chaque porte est représentée par une équation. La porte d'entrée $i(t)$, comme la présente l'équation (2.21), consiste à transmettre la sortie de l'étape *(t-1)* et l'entrée x à travers une autre couche sigmoïdale.

$$i(t) = \sigma\left(W_i.[h(t-1), x(t)] + b_i\right) \tag{2.21}$$

Une fonction tangente hyperbolique est appliquée sur les données d'entrées et la sortie de l'étape précédente pour créer un vecteur d'une nouvelle valeur $\tilde{C}(t)$ à ajouter comme état interne. La mise à jour de l'état interne est effectuée à travers l'équation (2.22).

$$\tilde{C}(t) = tanh\left(W_c.[h(t-1), x(t)] + b_c\right) \tag{2.22}$$

La porte oublie $f(t)$ est une sigmoïde qui prend en entrée la sortie h de l'étape *(t-1)* et l'entrée x à l'instant t. Elle est représentée par l'équation (2.23).

$$f(t) = \sigma\left(W_f.[h(t-1), x(t)] + b_f\right) \tag{2.23}$$

Enfin, la porte de sortie $O(t)$, décrite par l'équation (2.25) est basée sur l'état de la cellule $C(t)$ définie par l'équation (2.24). Cet état est obtenu en appliquant une tangente hyperbolique.

$$C(t) = f(t) \times C(t-1) + i(t) \times \tilde{C}(t) \tag{2.24}$$

Cette porte de sortie définie par l'équation (2.25) est obtenue à partir de l'application d'une

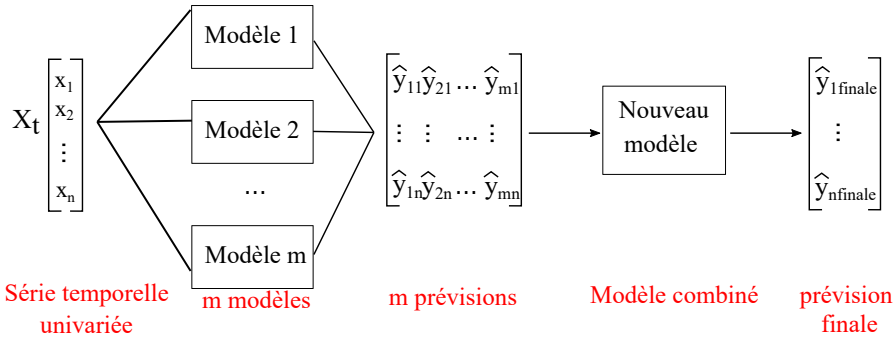

Figure 2.6: Principe général de la combinaison des modèles prédictifs

sigmoïde à l'entrée $x(t)$ et la sortie de l'étape précédente.

$$O(t) = \sigma(W_o.[h(t-1), x(t)] + b_o) \tag{2.25}$$

W et b représentent respectivement les poids et les biais aux différents niveaux dans le bloc mémoire LSTM.

2.5.5 Modèles hybrides pour la prévision des consommations

Pour améliorer les performances, plusieurs activités de recherche se focalisent sur la prévision des consommations avec différentes combinaisons de modèles. Citons à titre d'exemple, l'approche hybride construite avec un réseau de neurones (ANN) et un modèle de moyenne mobile intégrée auto-régressive (ARIMA) qui a été proposé dans [72] pour la prédiction des séries chronologiques afin de prédire des données annuelles : les données des taches solaires, les données du lynx canadien et le taux de change de la livre sterling.

Dans un travail de recherche [73], les auteurs ont choisi les réseaux neuronaux de corrélation en cascade (CCNN), les réseaux neuronaux de régression généralisée (GRNN) et les réseaux neuronaux Feed Forward (FFNN) pour prédire la consommation mensuelle d'eau. Une autre étude [74] propose deux modèles de prédiction quotidiens, l'ARIMA et le NARX (Nonlinear Auto-Regressive Exogenous) pour prédire la consommation d'eau urbaine. De plus, [75] a présenté une approche hybride qui combine ARIMA et ANN pour prédire la demande mensuelle en eau urbaine et le modèle additif Hault-Winters est choisi pour la prévision trimestrielle. Toutefois, on peut trouver de nombreux autres exemples de méthodes hybrides. En effet, [48] a présenté un nouveau modèle hybride d'ARIMA et de réseau des neurones à fonction de base radiale (RBF-NN) sur des données hebdomadaires. En outre, une autre étude [76] propose la méthode de régression multiple corrigée par la formule de régression linéaire pour prédire la consommation d'eau du Japon. Selon les auteurs [76], cette méthode de régression n'a pas pu améliorer la précision de la prévision d'eau.

Parmi les combinaisons des modèles possibles nous distinguons le stacking (ou dit parfois

blending) qui consiste en la moyenne d'un ensemble de modèles hybrides. L'idée principale de la construction de ce modèle prédictif en combinant différents modèles est illustrée par la Fig. 2.6. Dans cette méthode, une simple moyenne des prévisions en provenance des modèles individuels est réalisée.

2.6 Évaluation des performances et des capacités des approches ML

2.6.1 Mesures statistiques

Des mesures statistiques qui représentent des indicateurs clés de performance, permettent de fournir des informations sur les performances et le comportement d'un modèle donné. Dans le contexte de ML, ces mesures sont appelées des fonctions de coût [22]. Parmi ces mesures, nous trouvons l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (MSE), la racine de l'erreur quadratique moyenne (RMSE), l'erreur en pourcentage absolu moyen (MAPE), l'erreur absolue relative (RAE) et la racine de l'erreur quadratique relative (RRSE) [23]. Le RMSE est la mesure la plus utilisée pour comparer différents modèles de classification, ce qui donne une mesure quantitative simple et transparente de la différence entre l'entrée et la cible [77]. Le MAPE compte parmi les mesures les plus utilisées pour comparer des modèles de prévision. La représentation mathématique des différents indicateurs est donnée par les équations suivantes :

$$MSE(L^2, Wh^2) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{2.26}$$

$$RMSE(L, Wh) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \tag{2.27}$$

$$MAPE(\%) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \tilde{y}_i)}{y_i}\right| \times 100 \tag{2.28}$$

2.6.2 Compromis biais/variance

Les modèles ML peuvent présenter un biais ou une variance dans leurs résultats (de classification, prédictions, etc.). Ils sont soumis à un compromis biais/variance [22].

Un algorithme biaisé génère des classifieurs individuels simples qui ne prédisent pas parfaitement l'ensemble d'entrainement des données. Ceci ne peut pas exprimer l'ensemble des données pour anticiper l'ensemble de test qu'il ne connait pas. De ce fait, ces modèles sont moins sensibles aux données d'apprentissage et ils sont plus général et donc à priori ils sont mieux adaptés aux futures prédictions. Ce phénomène est appelé le sous-apprentissage (connu aussi à travers l'expression sous-ajustement, underfitting en anglais). À l'inverse, un algorithme a une grande variance est un algorithme très sensible aux jeux des données. En changeant ces derniers, il modifie radicalement sa prédiction. Par conséquent, si l'erreur sur l'ensemble d'entrainement

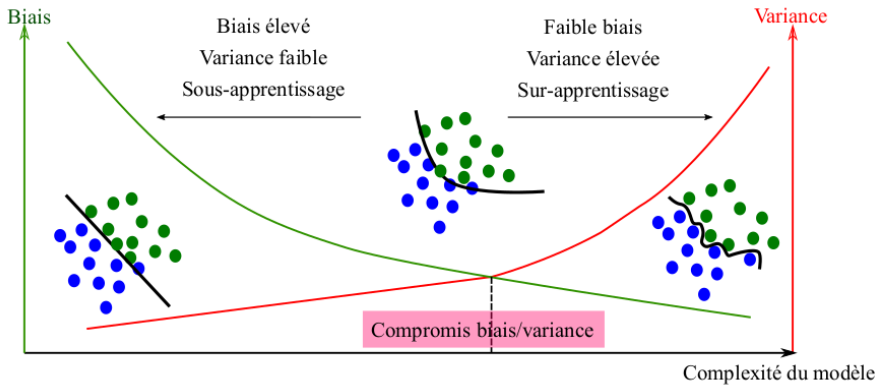

Figure 2.7: Complexité du modèle et compromis entre la dualité de biais/variance

est faible, son application sur un nouveau jeu des données, l'erreur a le risque d'être grande. Ce phénomène est appelé le sur-apprentissage (connu aussi sur-ajustement, overfitting en anglais). La Fig. 2.7 décrit ces deux phénomènes : le sur-apprentissage et le sous-apprentissage avec le positionnement optimal des modèles qui devront être choisi.

Mathématiquement, une prédiction de y avec un modèle définit avec la fonction f en se basant sur les données x peut être obtenue par l'équation suivante :

$$Y = F(X) + \epsilon \tag{2.29}$$

où ϵ représente l'erreur.
Les erreurs représentées dans (2.30), peuvent être décomposées en biais et variance :

$$\begin{aligned} Erreur(x) &= [E[\hat{f}(x)] - f(x)\ ]^2 + E[\ (\hat{f}(x) - E[\hat{f}(x)])^2\ ] + \sigma_e^2 \\ Erreur(x) &= Biais^2 + Variance + ErreurIrréductible \end{aligned} \tag{2.30}$$

2.7 Conclusion

Ce chapitre est consacré aux algorithmes ML pour la classification et la prédiction de données de consommation. Pour la classification des données, le calcul de la similarité a été réalisé toujours avec une distance qui peut être la ED, DTW, Hamming, etc. Les algorithmes de vérification de la similarité peuvent avoir différents principes tels que le KNN, DT, HAC, SOM, MLP, etc.
Dans le cadre de notre application, nous allons tester d'abord les modèles les moins complexes jusqu'aux modèles qui nécessitent plus du temps et de mémoire. Nous allons tester ces modèles dans différentes conditions avec différentes quantités de données. Pour la prévision, nous avons choisi de varier les modèles de manière que chacun traite une spécificité particulière des données qui seront considérées comme étant des séries temporelles. Le modèle déterministe est une technique simple qui permet de chercher un modèle mathématique qui caractérise la série temporelle. Vu la complexité des données des consommations, ce modèle constitue une

exploration permettant d'analyser la structure interne en terme de tendance, de saisonnalité et de bruit. Nous retenons également le modèle stochastique SARIMA, qui à son tour traite les données sous forme de séries temporelles selon des variations périodiques en fonction d'une saisonnalité avec un caractère aléatoire. C'est un modèle qui modélise l'aspect stochastique de la série mais il ne peut pas prendre en compte la nature non linéaire de la série. Le modèle déterministe et SARIMA sont des approches efficaces pour la modélisation et la prévision de séries chronologiques, bien qu'ils nécessitent une analyse approfondie de la nature et de la structure de la série. Pour contourner cet inconvénient, les réseaux de neurones présentent une alternative puissante pour la prévision des séries temporelles. Un réseau de neurone MLP permet de prendre en charge les aspects non linéaires caractérisant la série temporelle des consommations. Le LSTM est un autre réseau de neurone qui dépasse la capacité d'un MLP pour modéliser la série temporelle avec des fonctions non linéaires en introduisant une mémoire des entrées précédentes.
D'autres algorithmes ont été testés tels que le support vecteur machine (SVM) [22], NAR, CNN, ARIMA [48] n'ont pas abouti à des bons résultats. Tous les modèles construits dans cette thèse seront décrits et évalués selon la dualité biais/variance. Ces algorithmes permettent au mieux de prendre en charge les comportements complexes et les dynamiques des consommations d'eau et d'électricité. Pour ce fait, nous avons varié et diversifié les techniques ML et autres modèles. Dans le chapitre qui suit, nous proposons une description détaillée du processus de collecte des données, la plateforme intelligente proposée ainsi qu'une démarche de prétraitement des données collectées. Ces techniques ML seront utilisées et appliquées sur des données des consommations issues d'un seul point de mesure et d'une manière non intrusive.

3 Collecte et pré-traitement de données de consommation

3.1 Introduction

Ce chapitre présente une mise en œuvre complète d'une stratégie de collecte des données basée sur des compteurs intelligents. Nous proposons une plateforme de comptage avancée qui repose sur les technologies IoT. Dans notre cas d'étude, une installation de plusieurs compteurs intelligents a été déployé dans différents types de bâtiments : tertiaires et privés. Cette technologie rend la collecte automatique et centralise toutes les données dans une base des données. Cette plate-forme IoT consiste en une infrastructure AMI (Advanced Metering Infrastructure) permettant une surveillance continue de la consommation d'une manière non intrusive et à partir d'un seul point de collecte général. Les données de consommation d'eau et d'électricité sont transmises à un serveur Web dans des conditions de fonctionnement presque instantanées.

Les données recueillies d'un point de mesure seront représentées sous forme de Cdc et de séries temporelles. Des anomalies peuvent apparaitre lors de la collecte des données qui provoquent un manque des données. Afin de pouvoir exploiter cette masse des données, il est primordial de synthétiser les profils des consommations en Cdc, journalière, horaire. Ce chapitre sera consacré à l'étude et la préparation de ces données afin de les explorer pour étudier les profils de consommation et les comportements des utilisateurs.

3.2 Stratégies AMR pour la lecture automatisée des compteurs

3.2.1 Le concept de l'IoT

Une très grande partie des IoT sont conçus pour la domotique et contribuent à ce qu'on appelle les bâtiments intelligents ou «Smart building» [2]. Un système domotique connecte généralement les périphériques à un concentrateur central. Les utilisateurs peuvent contrôler le système, dans notre cas le bâtiment, via internet avec des applications sur tablette, ordinateur de bureau,

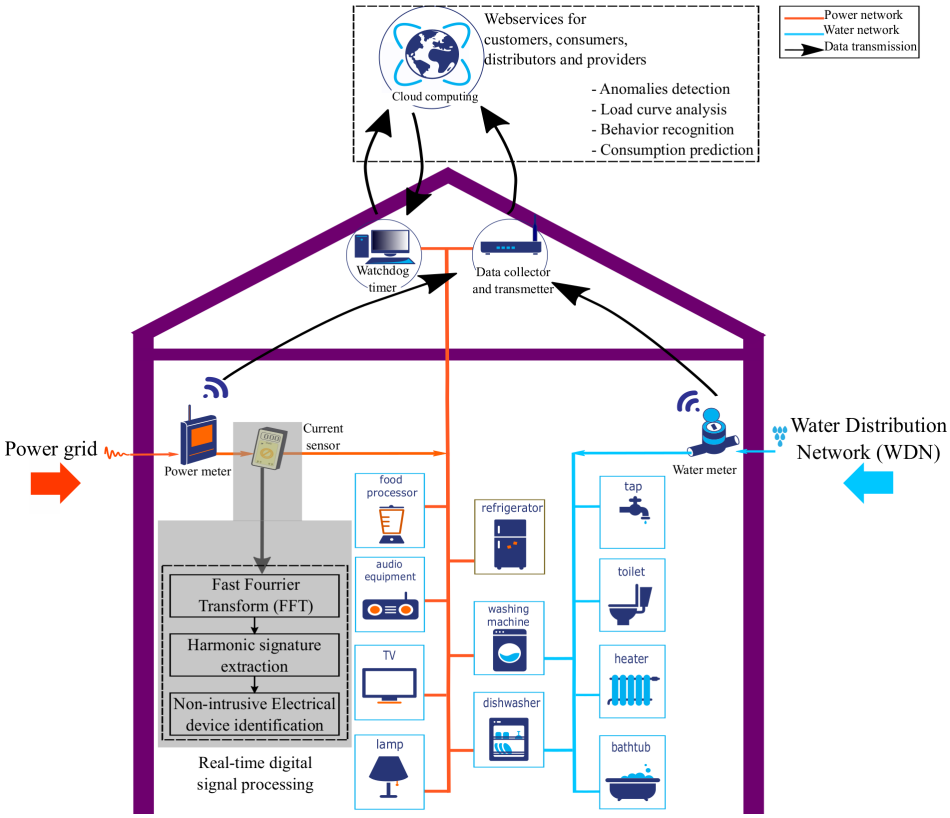

Figure 3.1: Architecture générale de la plateforme IoT pour une lecture automatique des compteurs d'eau et d'électricité au sein d'un bâtiment intelligent

téléphone mobile à travers des interfaces Web [78]. Une véritable interopérabilité des communications est nécessaire : les normes matérielles et logicielles doivent être totalement compatibles avec l'architecture ouverte d'internet afin de fournir des interfaces compatibles et accessibles.

Les dispositifs IoT élargissent le concept de la domotique sur plusieurs aspects citons entres autres l'éclairage, le chauffage et la climatisation, les systèmes de média et de sécurité, les compteurs intelligents (eau, électricité, gaz ...), les capteurs (température, humidité, luminosité ...) [79, 80], caméras de surveillance, actionneurs ... [81].

La croissance de l'IoT s'explique par les progrès importants réalisés dans les technologies intégrées. Ce processus est associé à des réductions de coûts substantielles et à l'émergence d'une nouvelle génération d'équipements et d'appareils intelligents. En étant connectés en permanence à internet, ces objets connectés fournissent naturellement une quantité des données importantes et variées. En outre, les données générées fournissent potentiellement de nouveaux services et applications en ligne aux utilisateurs [82, 83]. Le concept IoT vise à rendre l'infrastructure du système d'information nécessaire à la mise en œuvre des bâtiments intelligents sans être immersive et envahissante [5].

Les IoT sont de plus en plus utilisés en milieu urbain [4]. Des capteurs et des compteurs appropriés basés sur les technologies IoT peuvent être facilement placés sur des bâtiments et autres lieux d'une ville. Par exemple, l'évaluation de l'intégrité structurelle peut être obtenue en surveillant les contraintes mécaniques du bâtiment à l'aide de capteurs de vibrations et de déformation sans fil. Les niveaux de pollution peuvent être estimés à l'aide de capteurs atmosphériques extérieurs et la sécurité liée aux activités humaines dans les bâtiments privés et publics peut être calculée à l'aide des algorithmes avancés associés à un ensemble de capteurs intérieurs hétérogènes.

Les compteurs qui sont des dispositifs IoT sont capables d'enregistrer et de transmettre instantanément des évènements datés correspondants à de petites quantités d'énergie ou d'eau consommées. Les évènements de ces compteurs fournissent des informations complètes sur le moment de la consommation et permettent de mieux comprendre le comportement des utilisateurs [84]. Aujourd'hui, l'IoT et le traitement de l'information associées font des compteurs intelligents une solution abordable, non seulement pour les clients à forte consommation d'énergie mais aussi pour des ménages.

Les données issues de compteur dans un intervalle de quelques minutes à quelques secondes permettent par exemple de définir et de mettre à jour les profils existants, c'est-à-dire les Cdc, et de créer de nouvelles courbes correspondantes à la consommation de chaque client. L'architecture générale de la plateforme développée repose sur les technologies ouvertes d'internet notamment liées aux IoT. Elle est représentée par la Fig. 3.1.

Aujourd'hui et avec les stratégies AMR, certaines villes peuvent lire les compteurs une fois par semaine, une ou deux fois par jour [85]. De toute évidence, l'AMR permet aux fournisseurs de services publics d'économiser le coût des déplacements périodiques vers chaque emplacement physique pour lire un compteur. Un autre avantage est que la facturation peut être basée sur des consommations en temps quasi réel. Ces mesures sont appelées informations à haute résolution en temps opportun. Ces informations précises, peuvent être associées à une analyse, offrent la possibilité aux fournisseurs de services publics et aux clients à mieux contrôler l'utilisation et la production de l'énergie électrique, de gaz ou de la consommation d'eau.

3.2.2 Compteurs intelligents

Les compteurs intelligents [86] sont des appareils de mesure à impulsions améliorés avec un circuit électronique intégré et un micro-contrôleur ou un microprocesseur avec des capacités de traitement intégrées. En effet, ces compteurs intelligents hébergent la mise en œuvre de stratégies appropriées pour leur propre gestion d'énergie, traitement de mesures, compression des données, transmission de messages et interaction locale afin de répondre à des événements spécifiques avec l'utilisateur et l'environnement [87]. Il s'agit d'un système de collecte centralisé par communication internet (avec ou sans fil). Ainsi, le système est capable de collecter des données beaucoup plus fréquentes et d'une précision accrue, permettant une collecte en temps

réel des données de consommation. Toutes les données de tous les compteurs intelligents sont centralisées dans une base des données unique hébergée sur des serveurs distants.

L'infrastructure physique qui prend en charge la base des données est fournie à distance par un fournisseur de cloud qui assure également la maintenance et la disponibilité. Les données des compteurs intelligents sont stockées dans le format bien connu SQL (Structured Query Language), qui est particulièrement utile pour traiter des données structurées dans lesquelles il existe des relations entre différentes entités / variables des données hétérogènes. Des fonctionnalités informatiques sont également disponibles sur le cloud pour l'analyse en ligne et le traitement en temps réel des données. La plate-forme d'acquisition des données est flexible et modulaire car elle repose sur une architecture IoT avec des formats des données ouverts et des protocoles de communication standards.

La mise en œuvre d'un comptage avancé par une AMR est un processus complexe qui pose de nombreux problèmes technologiques. Nous avons mis en place des compteurs d'eau et d'électricité intelligents pour centraliser automatiquement les données de consommation dans une base des données unique d'eau et d'électricité. L'analyse des données permet une surveillance efficace et précise de la consommation d'eau et d'énergie à différents niveaux. L'enjeu de notre travail de thèse se concentre sur des méthodologies génériques permettant d'analyser les masses des données recueillies via les compteurs intelligents et d'en extraire de l'information utile pour différents acteurs à l'instar des compagnies d'eau et d'électricité, les villes et les consommateurs.

Des compteurs supplémentaires ainsi que des capteurs de tout types peuvent être insérés dans l'architecture proposée. Des données, mesures et événements peuvent être enregistrés dans la base des données et exploités par des algorithmes pour fournir des services appropriés. Parmi ces mesures nous pouvons citer les mesures environnementales telles que la température extérieure et intérieure, l'humidité, la luminosité, la qualité de l'air ... et aussi des événements liés aux activités d'un individu. La Fig. 3.2 illustre les aspects matériels et logiciels de collecte, de stockage et d'analyse des données associées. L'ensemble a permis la création d'un réseau entre les compteurs intelligents pour centraliser les informations et pour pouvoir les diffuser aux gestionnaires, clients et autres parties, tels que les distributeurs et les fournisseurs.

3.3 Acquisition des données de consommation

La solution adoptée pour la collecte des données est non intrusive [88] permettant de collecter des données à partir d'une seule partie centralisée d'un réseau de distribution. La base des données contient essentiellement des données fournies par les compteurs intelligents d'eau et/ou d'électricité. Le processus de la collecte des données jusqu'à l'élaboration de la Cdc de la consommation journalière brute puis échantillonnée est représenté par le diagramme de la Fig. 3.3.

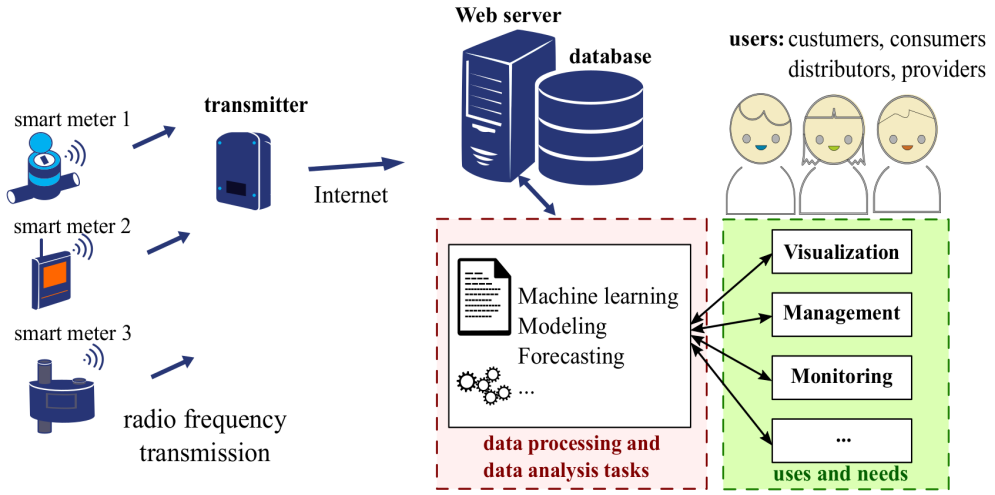

Figure 3.2: Principe proposé de l'Advanced Metering Infrastructure (AMI)

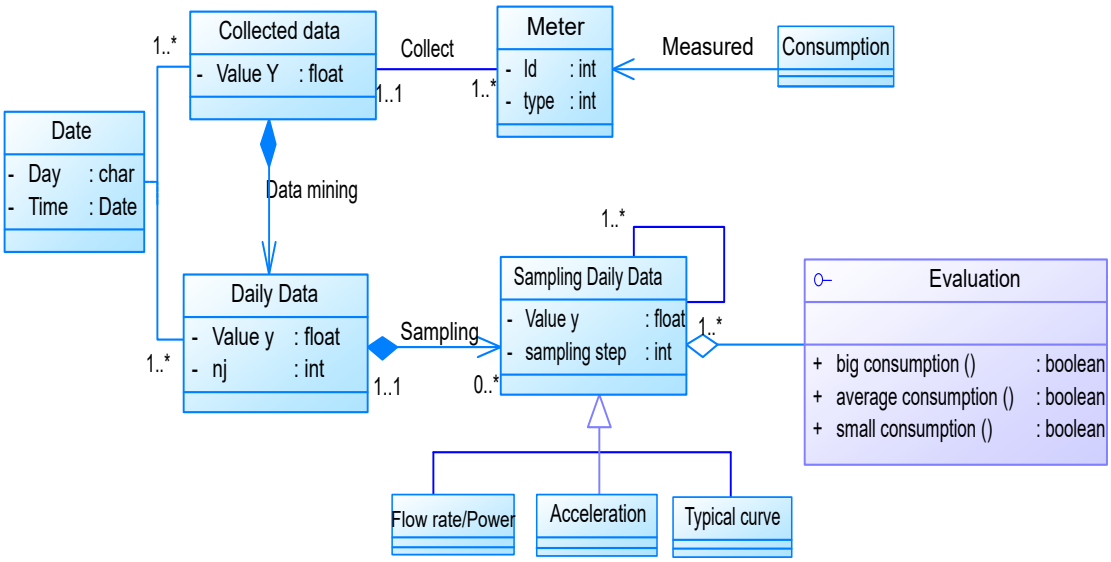

Figure 3.3: Diagramme de classe général de la spécification fonctionnelle du suivi des consommations

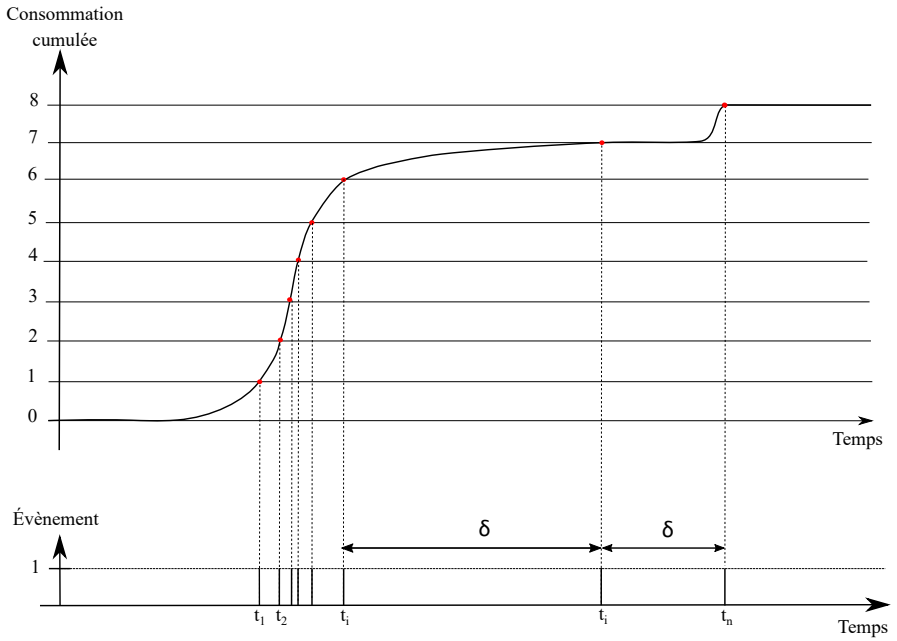

Figure 3.4: Illustration des données brutes générées (points rouges) dès qu'une variation est détectée par le capteur reliés par une interpolation

3.3.1 Évènement de consommation

Un litre, respectivement un Wh, consommé correspond à un évènement. La nouvelle stratégie proposée dans [1] offre la possibilité de transmettre toutes les données brutes mesurées. Cela est vrai pour la collecte des données de la consommation d'eau ainsi que la consommation électrique. Les données brutes sont définies comme étant des évènements horodatés selon les instants $t_1, t_2, , t_i, ..., t_n$. Ces événements sont générés dès qu'une variation d'un litre, respectivement un Wh, dans la grandeur mesurée est détectée.

Dans le cadre d'un compteur d'eau, cette variation est fixe et correspond à l'écoulement d'eau selon la taille du compteur et des caractéristiques du capteur. Dans notre cas, un compteur DN15 présente une précision d'écoulement d'un litre. Cependant, l'écoulement d'eau avec un plus grand compteur (DN100) est de 10 litres [1]. De ce fait, plus le débit est élevé, plus les événements sont rapidement générés et la variation temporelle Δ est petite. La succession des événements est illustrée sur la Fig. 3.4 avec une présentation cumulée de la consommation ainsi que ses horodatages $t_1, t_2, , t_i, ..., t_n$. La date t_n correspond à l'instant pendant lequel le $n^{ième}$ évènement a eu lieu.

Dans le contexte de la consommation électrique, les compteurs sont caractérisés par une précision de 0,1 Wh et de 1 Wh. Dans ce cadre, des scripts ont été développés et adaptés pour la collecte de sorte que chaque Wh consommé sera transmis au serveur et sera sauvegardé dans la base des données afin d'être exploité. Formellement, les évènements horodatés représentent une suite de valeurs numériques représentant la variation temporelle Δ. Cette dernière est

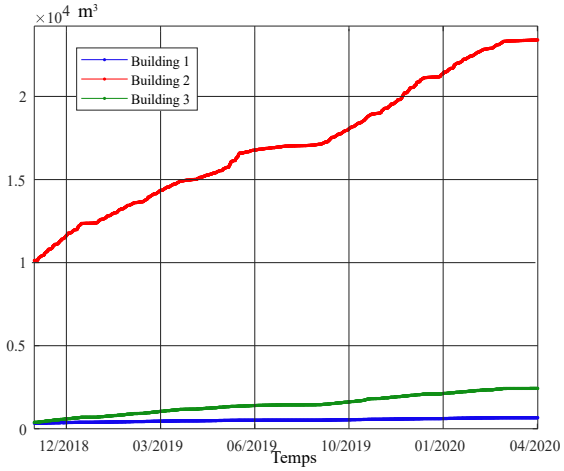

Figure 3.5: Cdc générales de la consommation d'eau cumulée correspondant à trois bâtiments tertiaires

appelée série chronologique, ou chronique, ou série temporelle. Elle est une suite d'observations chiffrées, ordonnées dans le temps selon l'apparition de chaque évènement :

$$\begin{aligned} \delta(y_1(t_1), y_2(t_2) = \delta(y_2(t_2), y_3(t_3) = ... = \delta(y_i(t_i), y_{i-1}(t_{i-1})) = ... = \delta(y_{T-1}(t_{T-1}), y_T(t_T)) = 1, \\ X = \{\delta(t_1, t_2), ..., \delta(t_i, t_{i-1}), ...\delta(t_T, t_{T-1})\}, \end{aligned} \tag{3.1}$$

3.3.2 Courbe de charge (Cdc) des consommations

Les consommations mesurées dans cette étude conduisent à des Cdc de puissance électrique et de volume d'eau au cours du temps. Ces Cdc décrivent l'évolution au cours du temps de ces consommations. La Fig. 3.5 représente trois Cdc générales de la consommation d'eau dans trois installations différentes.

Une Cdc générale désigne une grandeur numérique qui décrit l'évolution au cours du temps de la consommation d'eau ou d'électricité d'un système donné [6]. L'AMI proposé par [1] offre la possibilité de transmettre l'index de la consommation d'eau. Cet index est la valeur affichée sur le compteur comme sur un totalisateur de la consommation d'eau illustré par la Fig. 3.6. La Fig. 3.7 illustre un exemple des données de la consommation d'eau brute consommée le 19/01/2018 à partir des évènements datés générés par le compteur d'eau sous forme d'une Cdc et sa Cdc relative re-construite à partir des index transmis sur le serveur. Cette courbe re-construite présente un décalage variable Δ dans le temps pour chaque donnée collectée.

$\Delta_j(t_i)$, défini dans l'équation (3.2), est l'index représentant la consommation totale d'eau en litres depuis l'installation du compteur. Cet index s'incrémente en fonction du temps jusqu'à

Figure 3.6: Compteur d'eau servant de totalisateur de la consommation d'eau avec un affichage d'un index mécanique

l'instant t_i du jour j avec :

$$\Delta_j(t_i) = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (f_j(t_i)), \tag{3.2}$$

avec f est le débit d'eau en un instant i. $j = [1, ..., N]$ est une valeur qui identifie le jour et N représente le dernier jour d'observation. Le cumule de consommation est fournie par un compteur et représente la consommation globale en litre ou en Wh. La Cdc de l'électricité est un mélange additif linéaire des puissances appelées de différents usages électriques pendant une durée donnée. Les mesures proviennent d'un seul compteur intelligent qui génère un évènement daté pour chaque Wh ou kWh consommé.

La Fig. 3.5 illustre des Cdc de la consommation d'eau de trois bâtiments tertiaires sur une période de 16 mois (de décembre 2018 à avril 2020). $\Delta_j(t_i)$ représente le nombre cumulé de litres consommés depuis l'installation du compteur. $f_j(t_i)$ représente le débit moyen pour l'eau et la puissance moyenne pour l'électricité au jour j entre deux instants, t_i et t_{i-1} avec :

$$f_j(t_i) = y_j(t_i) - y_j(t_{i-1}). \tag{3.3}$$

Le nombre de litres consommés en une seule journée jusqu'à l'instant t_i est donné par (3.4) :

$$y_j(t_i) = \sum_{i=1}^{n_j} (f_j(t_i)), \tag{3.4}$$

chaque $y_j(t_i)$ est le mélange additif linéaire de la consommation d'eau de différents appareils (robinets, lave-vaisselle, toilettes ...).

Les données de consommation d'eau ne sont rapportées qu'à certains instants spécifiques

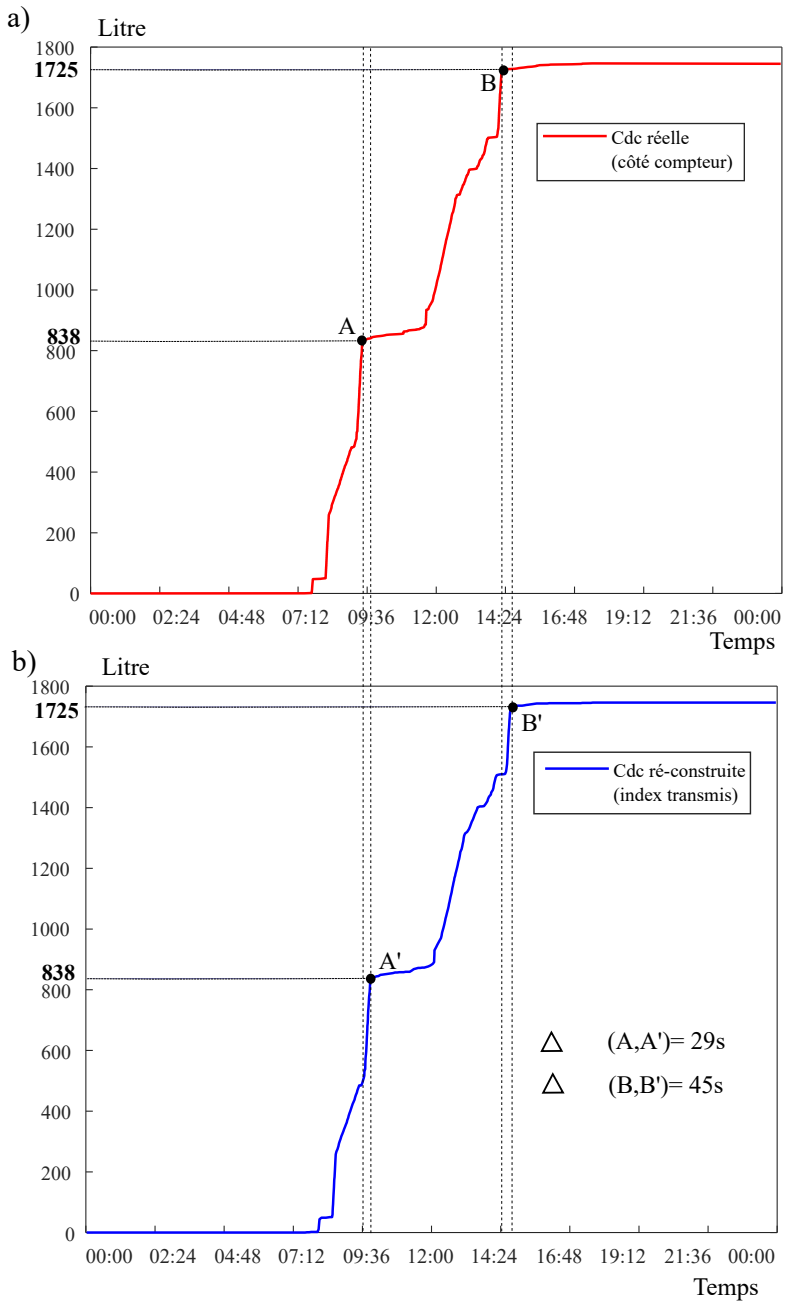

Figure 3.7: a) Représentation des données de la consommation d'eau brute à partir des évènements datés générés par le compteur d'eau sous forme d'une Cdc et b) la Cdc re-construite à partir des index transmis sur le serveur avec un décalage variable Δ dans le temps pour chaque donnée

identifiés par l'architecture interne du compteur intelligent, notés $[t_1, ..., t_{nj}]$ selon sa stratégie de communication [1]. La plateforme proposée de recueille des données de consommations est basée sur la récolte des évènements. Autrement dit, pour chaque Wh consommé, l'index s' incrémente et le nouveau index sera sauvegardé sur le serveur en fonction du temps. Les instants d'observation diffèrent en général d'un jour à l'autre. Les n_j mesures du jour j sont regroupées dans un vecteur Y_j :

$$Y_j = [y_j(t_1), y_j(t_2), ..., y_j(t_{n_j})]^T. \tag{3.5}$$

Chaque y représente un index qui est le volume d'eau (respectivement le nombre de Wh) consommé au cours de chaque t_i. Les données récupérées de la plateforme sont inégalement espacées dans le temps [89, 90]. Par conséquent, un échantillonnage est proposé par la suite afin de rendre la série uniformément espacée avec une précision d'une minute, d'une heure ou d'un jour selon l'objectif visé.

La courbe cumulée globale de la consommation d'eau peut être vue comme étant un ensemble de vecteurs de la consommation journalière successive d'une période donnée :

$$\Delta = [Y_1, Y_2, ..., Y_j, ..., Y_N], \tag{3.6}$$

Y_1 et Y_N représentent respectivement le vecteur du premier jour jusqu'au vecteur du dernier jour d'observation. Chaque Y_j est censé être un jour de mesures complètes. Cette formalité fournit une spécification fonctionnelle pour exprimer le processus des données de consommation d'eau. Un re-échantillonnage régulier de ces données est utilisé pour obtenir des données « intra journalières » faciles à manipuler. En effet, le vecteur $Y_j(t_i)$ pourrait être ré-échantillonnée sur un vecteur d'un nombre prédéfini de valeurs en fonction du temps (minute, heure, jour, semaine ...) pour créer un vecteur d'index journalier échantillonné Cj. Les équations (3.2) et (3.4) sont représentées respectivement par les classes « Données collectées » et « Données quotidiennes » de la Fig. 3.3.

3.4 Plate-forme d'acquisition des données sur le Web et profils étudiés

Des compteurs intelligents ont été insérés dans différents bâtiments. Un seul compteur surveille la consommation globale du réseau électrique du bâtiment et l'autre lit la consommation globale d'eau du WDN [91] du bâtiment. Les données de ces deux compteurs intelligents sont collectées en permanence dans des conditions de fonctionnement réelles. Les données historiées sur plus d'un an ont été stockées dans une base des données. Le processus d'acquisition de ces données est illustré dans la Fig. 3.8. Les données servent à répondre à des objectifs et des exigences qui sont décrits dans le Tableau 3.1, selon chaque utilisateur. Nous distinguons divers profils d'utilisateurs qui peuvent être le fournisseur de la ressource, le distributeur, le sous-distributeur et les consommateurs finaux pour l'eau [1] et l'électricité.

Table 3.1: Objectifs et spécifications des fournisseurs, distributeurs, sous-distributeurs et consommateurs

	Eau	**Électricité**
Fournisseurs	* Vendre une eau de bonne qualité * Veiller et préserver la qualité de l'eau potable * Prévenir des éventuelles pollutions * Répondre à la demande croissante en eau	* Vendre une électricité de bonne qualité * Répondre à la demande croissante en électricité * Produire une électricité moins carbonée * Détecter et surveiller la distorsion d'électricité
Distributeurs	* Assurer la distribution complète de l'eau avec le minimum de perte * Réduire et limiter les fuites dans le réseau de distribution * Détecter la détérioration de la canalisation et des systèmes vieillissants * Gérer et maintenir le réseau de distribution * Réduire les coûts associés à l'utilisation de systèmes complémentaires pour surveiller la qualité d'eau * Réduire les pertes d'eau non facturées (NRW)	* Assurer la distribution complète de l'électricité * Détecter la détérioration des équipements électriques et des systèmes vieillissants * Gérer et maintenir le réseau de distribution * Réduire les coûts associés à l'utilisation de systèmes complémentaires pour surveiller la qualité d'électricité
Sous-distributeurs	* Gérer l'installation des compteurs et des abonnements * Protéger les données et la vie privée des consommateurs * Détecter et réduire les anomalies dues aux erreurs de comptage * Réduire les pertes d'eau non facturées (NRW) * Moderniser le réseau pour garantir sa sécurité et sa performance	* Gérer l'installation des compteurs électriques et des abonnements * Protéger les données et la vie privée des consommateurs * Détecter et réduire les anomalies dues aux erreurs de comptage * Moderniser le réseau électrique pour garantir sa sécurité et sa performance
Consommateurs	* Réduire les factures et les consommations en eau et en électricité * Surveiller et comprendre l'utilisation quotidienne des ménages * Détecter les principales parts de consommation * Adapter ou modifier sa manière de consommer l'eau et l'électricité en fonction des ménages	

Une infrastructure informatique de cloud computing permet de stocker, de gérer et de traiter les données grâce à des ressources matérielles et logicielles pouvant être physiquement distribuées dans le monde entier et utilisées à distance via Internet. Cette infrastructure est capable de recevoir les données de consommation d'eau transmises par le système de surveillance, de les valider et de les archiver dans une base des données au sein de la plateforme cloud.

Dans ce travail, nous utilisons des données collectées de la consommation d'un bâtiment tertiaire éducatif (IUT de Mulhouse) composé d'un ensemble de 8 bâtiments et un bâtiment à usage domestique. Parmi les bâtiments tertiaires, nous distinguons un restaurant (RU) absorbe d'une consommation journalière répétitive d'eau hors weekend, jours fériés et vacances. Pendant les jours travaillés, il y'a une grande consommation due à la présence de plus de 1300 personnes. Cependant, pendant les weekends et les vacances universitaires, la consommation est assimilée à celle d'un bâtiment privé étant donnée qu'un immeuble de fonction est habité par une dizaine de personnes.

La nouvelle stratégie de communication des compteurs d'eau intelligents [92] qui a été proposée et développée afin de compresser et de transmettre les données avec une très haute résolution est proposée par [1]. Cette stratégie permet de reconstituer les temps de consommation de chaque litre et d'améliorer la consommation d'énergie du compteur lors de la transmission des données. Ces dernières sont transmises toutes les 5 minutes maximum et l'intervalle de transmission est totalement adaptatif et lié à la consommation d'eau qui ne dépasse pas normalement les 30 minutes [1]. Une consommation d'eau supérieure entraîne une plus grande quantité de données. Les données sont collectées en permanence dans des conditions de fonctionnement réelles. Les données sont ainsi centralisées et stockées dans une base de données SQL [12]. Les données de la consommation d'eau sont structurées dans des trames de données qui comportent entre autre l'index qui est le volume d'eau consommé, et la date de chaque litre consommé en milliseconde appelée évènement daté [1]. La collecte des données haute résolution d'utilisateurs individuels, fournit un vaste corpus d'information sur lequel des modèles ML peuvent être basés [93].

Enfin, ces modèles traitent les données de consommation grâce à des algorithmes appropriés qui extraient des caractéristiques utiles pour les utilisateurs et les fournisseurs d'eau. En ce qui concerne les utilisateurs, la plateforme cloud calcule le volume d'eau consommé par un utilisateur individuel au cours des dernières 24 h et la série temporelle correspond au débit à un pas de temps donné (par exemple, une heure).

Les données peuvent également être utilisées pour d'autres développements, tels que la conception d'analyses et de services, la formation et le test d'algorithmes ML et le déploiement de technologies de réseau intelligent supplémentaires. Les détails et les explications techniques de la plateforme réalisée avec le compteur d'eau et d'électricité sont détaillées dans les annexes A.2.1 et A.2.2.

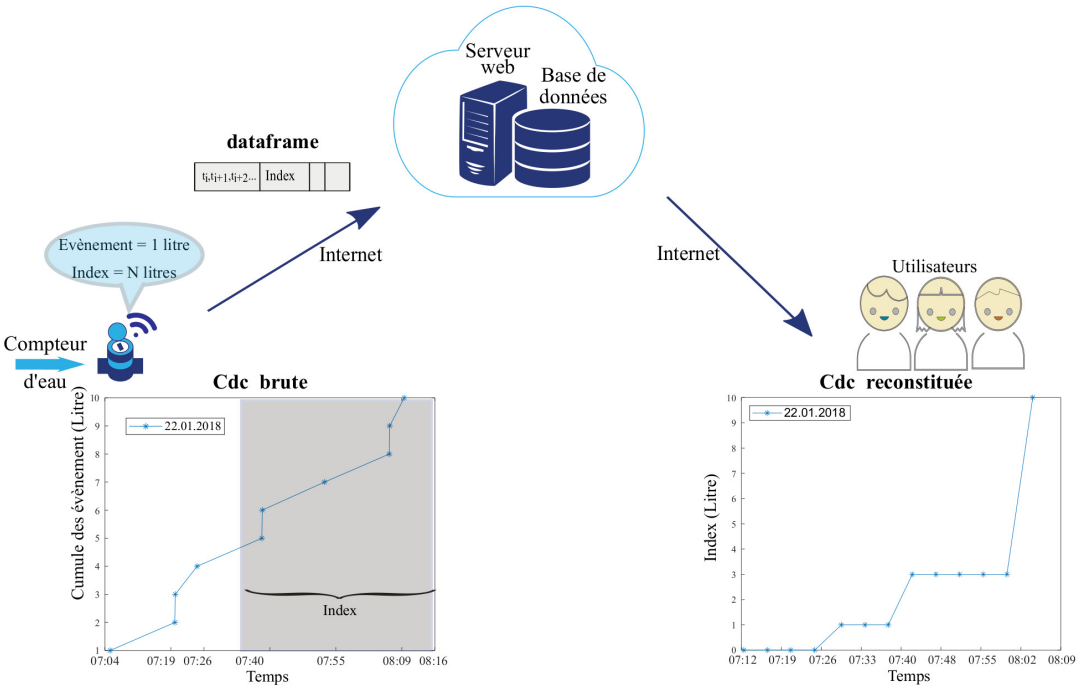

Figure 3.8: Processus d'acquisition des données de la consommation à partir du compteur intelligent communicant d'eau dans une base des données sur le cloud et la ré-construction en Cdc

3.5 Pré-traitement et analyse des Cdc

3.5.1 Échantillonnage des Cdc

Échantillonner une Cdc revient à prendre un certain nombre de points régulièrement espacés dans le temps. Cela permet d'analyser la consommation d'eau à des instants réguliers et précis de la journée. Nous considérons les consommations journalières Y_j en les ré échantillonnant à chaque minute. Cela conduit à une Cdc quotidienne qui consiste en $e = 1440$ valeurs qui correspondent à chaque minute d'une journée qui commence toujours à 0 litre :

$$C_j = [y_j(1), ..., y_j(e)]^T. \tag{3.7}$$

Les Fig. 3.9 et Fig. 3.10 représentent quelques Cdc d'eau et d'électricité de la consommation journalière ré échantillonnées en minute (représentées en (a)). Nous pouvons également représenter un vecteur de débit d'eau et la puissance moyenne d'électricité F_j de taille $e = 1440$ comme étant la dérivée de la Cdc C_j (représentées en (b)).

$$F_j = [0, (C_j^{'})]^T. \tag{3.8}$$

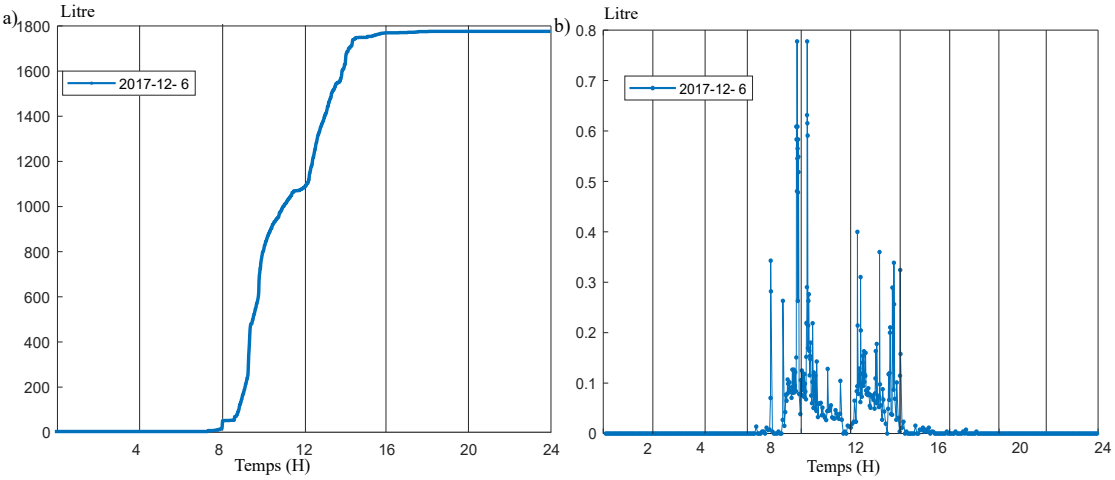

Figure 3.9: a) Cdc de la consommation d'eau cumulée échantillonnée en minute b) le débit moyen de la consommation d'eau journalière l/s

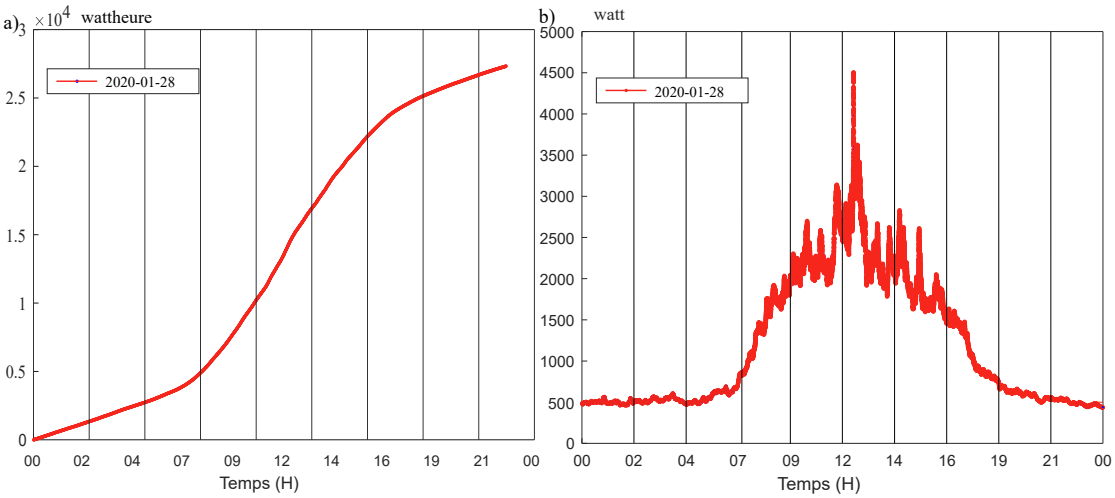

Figure 3.10: a) Cdc de la consommation d'électricité cumulée échantillonnée en minute b) la puissance moyenne de la consommation d'électricité journalière W

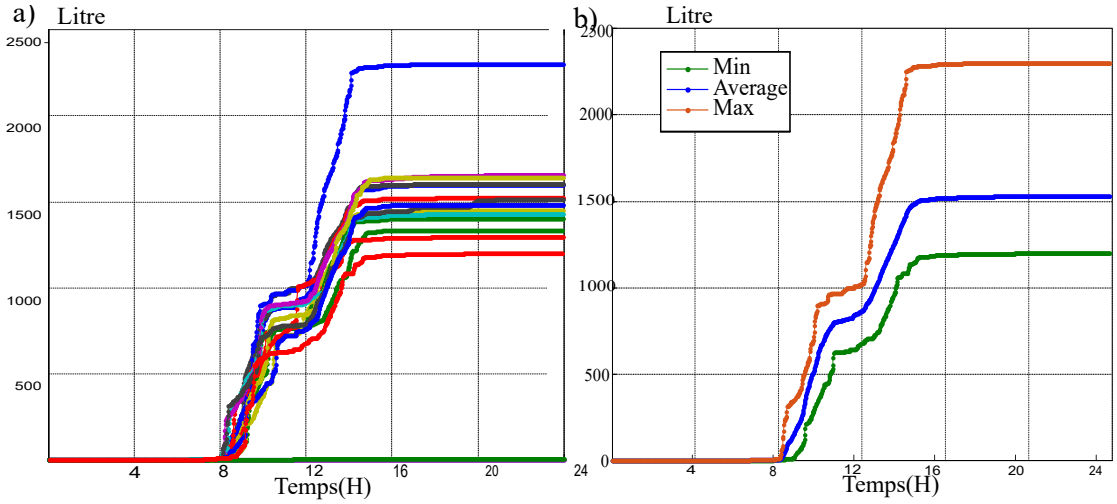

Figure 3.11: Cdc journalière de la consommation d'eau : a) les Cdc des lundis depuis décembre 2017 jusqu'à avril 2018 et b) les Cdc moyennes, minimales et maximales associées

3.5.2 Caractéristiques des consommations et Cdc références

Plusieurs caractéristiques de consommation d'eau doivent être définies afin de servir de référence pour distinguer les consommations normales des consommations anormales. On définit donc C_{avg} comme étant une Cdc journalière moyenne sur une période Per :

$$C_{Per}^{avg} = avg(C_1, C_2, ..., C_p), \tag{3.9}$$

où Per est un ensemble de p jours. De la même manière, nous pouvons définir C_{Per}^{max} et C_{Per}^{min} qui sont respectivement la Cdc maximale et la Cdc minimale, c'est-à-dire les valeurs maximales et minimales pour chacune des 1440 minutes des jours sur la période considérée. Ces caractéristiques sont des courbes représentatives qui peuvent servir de référence dans l'analyse des consommations d'eau.

Par exemple, la Fig. 3.11 montre les Cdc quotidiennes des lundis sur une période de 5 mois et les courbes moyennes, minimales et maximales résultantes. Ainsi, la consommation d'eau d'un jour j, C_j, peut être comparée à C_{Per}^{avg} où $Per = Weekday = \{Lundi, ...Vendredi\}$. Les samedis et dimanches sont des jours qui seront traités différemment des autres jours de la semaine avec Per = {Samedi, Dimanche}.

Notre cas d'étude de la consommation d'eau concerne un bâtiment tertiaire. Il s'agit là d'un restaurant universitaire où l'utilisation est caractérisée par des aspects habituels et répétitifs de la consommation d'eau. Dans les 5 premiers jours de la semaine, les consommations sont très similaires mais sont très différentes des jours de weekend où il n'y a pas d'activité comme le montre la Fig. 3.12.

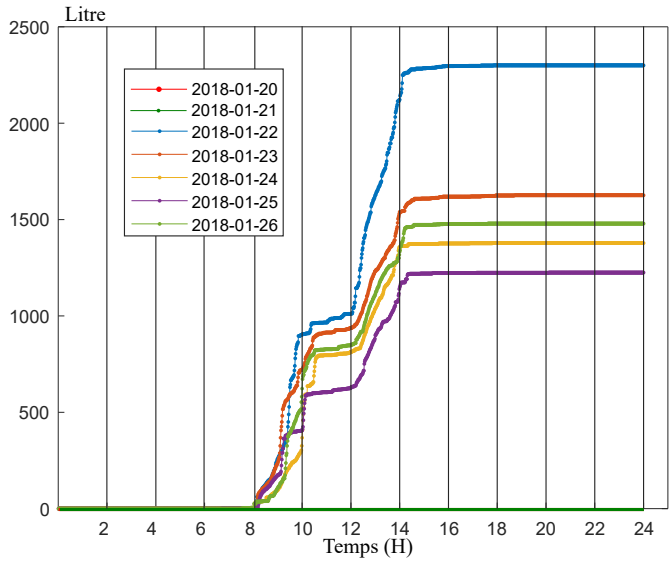

Figure 3.12: Exemple d'une semaine de consommation d'eau avec des Cdc journalières avec 5 Cdc de consommation pendant la semaine (lundi à vendredi) et 2 Cdc avec des valeurs nulles pendant le weekend

3.5.3 Vérification de l'intégrité et reconstitution des données manquantes

Un pré-traitement [94] sur les données est réalisé pour deux objectifs : le contrôle d'intégrité et l'interpolation des données. Cette interpolation offre la possibilité de préparer les données pour une classification ou prédiction de la consommation. Elle permet de reconstituer les Cdc qui représentent un manque des données.

Une interpolation des données est optée en cas d'une interruption lors de la collecte des données qui peut engendrer un manque. Les données réelles dans les différentes applications peuvent présenter communément des données ou des plages des données manquantes ou incomplètes. La Fig. 3.13 illustre un exemple de manque des données au cours de la journée.

L'interpolation des données est adaptée selon la nature du jour (travaillé ou weekend) ainsi que le type du bâtiment (tertiaire, privé). Parmi les méthodes d'interpolation nous citons l'interpolation spline cubique, linéaire ...

Une interpolation spline cubique [95] est une fonction qui est définie par morceaux constituée d'un polynôme avec k nœuds : $x_1 ... x_k$. f est une fonction qui est définie sur chaque intervalle $[x_i, x_i + 1]$. En posant $\beta_i = f(x_i)$ et $\delta_i = f"(x_i)$:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \tag{3.10}$$

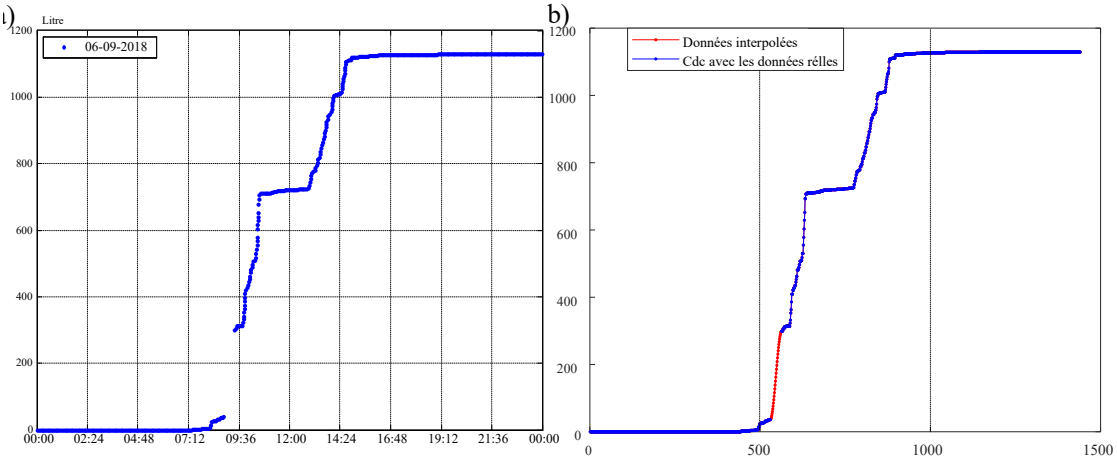

Figure 3.13: a) Cdc journalière représentant un manque des données, b) détermination du manque des données à partir d'une interpolation splines

avec

$$a_j^-(x) = \frac{x_{i+1} - x}{h_i} \tag{3.11}$$

$$a_j^+(x) = \frac{x - x_i}{h_i} \tag{3.12}$$

$$h_i = x_{i+1} - x_i \tag{3.13}$$

$$c_j^- = \frac{1}{6}\left[\frac{(x_{i+1} - x)^3}{h_i} - h_i(x_{i+1} - x)\right] \tag{3.14}$$

$$c_j^+ = \frac{1}{6}\left[\frac{(x - x_i)^3}{h_i} - h_i(x - x_i)\right] \tag{3.15}$$

En appliquant cette interpolation, nous pouvons déterminer les données manquantes. La Fig. 3.13, b représente une interpolation spline cubique.

L'interpolation linéaire [95] consiste à approcher une fonction f par la fonction $\bar{f}$ telle que $\bar{f}(x_a) = y_a$ et $\bar{f}(x_b) = y_b$ avec :

$$\bar{f}(x) = \frac{y_a - y_b}{x_a - x_b}x + \frac{x_a y_b - x_b y_a}{x_a - x_b} \tag{3.16}$$

Une méthodologie de pré-traitement des données est proposée et présentée dans la Fig. 3.14. Le processus du pré-traitement est déclenché une fois que nous avons récupéré les données de consommation. Les Cdc références seront possible à définir à partir de la base de données collectées. Avec ce pré-traitement, nous pouvons aussi rafraichir les Cdc références.

Avec le processus de pré-traitement proposé dans la Fig. 3.14, nous avons défini en entrée, une série temporelle brute Y_j qui représente une Cdc d'un jour j, et est caractérisée par

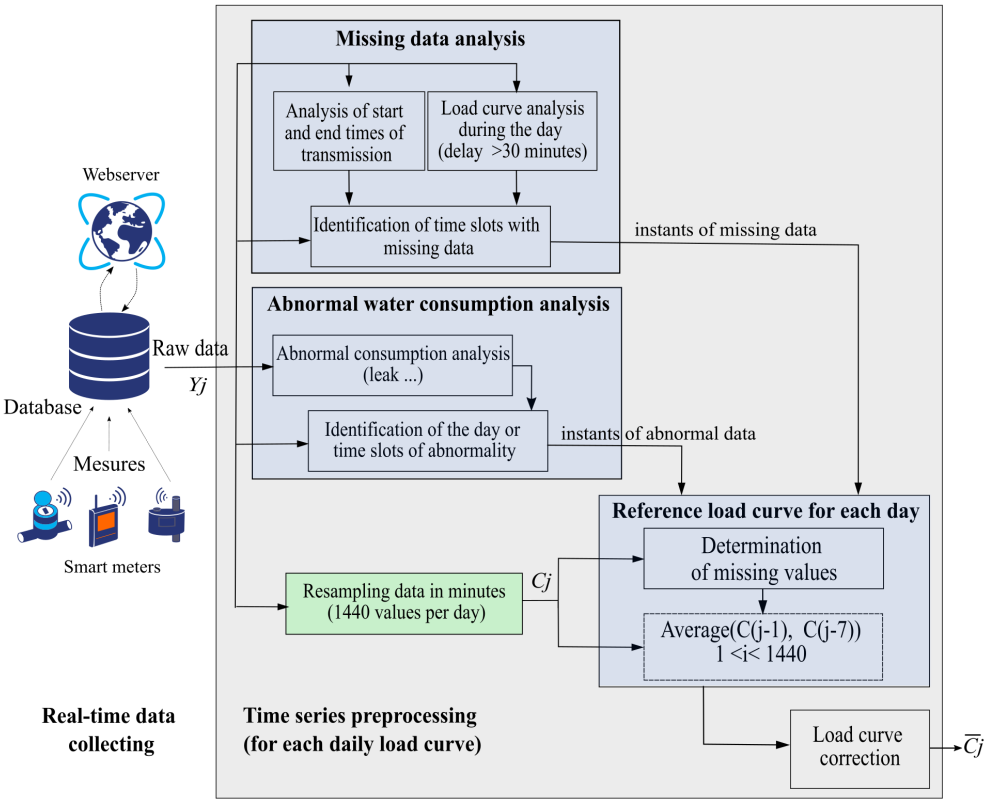

Figure 3.14: Pré-traitement des séries chronologiques pour chaque Cdc journalière

l'équation (3.17).

$$S_j = Y_j = \left[y(t_1), y(t_2), ..., y(t_i), ..., y(t_{n_j})\right] \tag{3.17}$$

Étant donné qu'une prévision de la consommation d'eau est visée avec une précision d'une heure, l'identification du manque des données ne peut être détectée qu'avec un échantillonnage encore plus accru, c-à-d inférieure à une heure, imposé par l'architecture interne de la plateforme. Par conséquent, un échantillonnage en minutes est effectué pour détecter les zones du manque des données et identifier la consommation anormale. Avec un échantillonnage en minute, nous obtenons une Cdc avec *n=1440* données qui correspondent à chaque minute d'une journée décrite avec (3.18) qui correspond à l'équation (3.7) définie antérieurement.

$$C_j = \left[y(t_1), y(t_2), ..., y(t_i), ..., y(t_{n=1440})\right] \tag{3.18}$$

Après un pré-traitement, nous obtenons une série temporelle $\bar{C}_j$ qui représente une Cdc de la consommation d'eau échantillonnée en minutes et sans manque de données. Ce processus de pré-traitement est répété à chaque jour.

Des programmes de surveillance ont été mis en place pour obtenir davantage d'informations

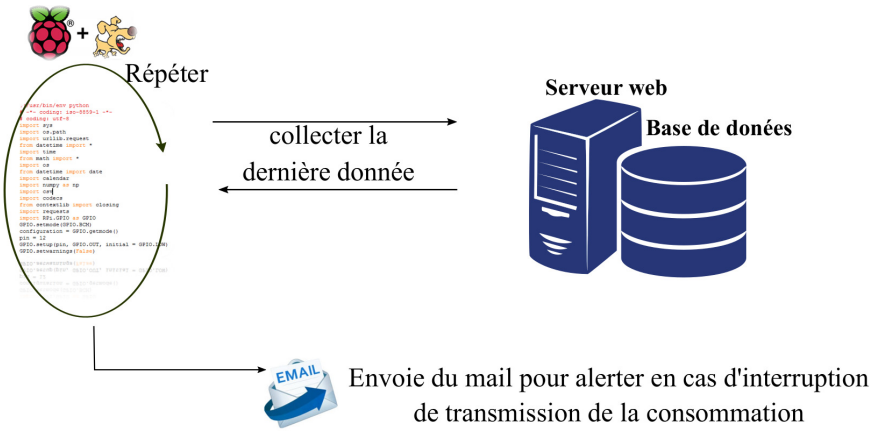

Figure 3.15: Schéma descriptif d'un watchdog mis en œuvre sur un Raspberry PI pour surveiller la transmission instantanée des données de consommation

sur la couverture radio, le nombre de trames reçues et pour tester si un manque des données correspond à une absence de communication de 30 minutes, selon l'architecture proposée, par le serveur ou le compteur d'eau.

La Fig 3.15 représente une description du fonctionnement d'un programme de surveillance "watchdog" mis en œuvre sur un Raspberry PI pour surveiller la transmission instantanée de donnée de consommation.

3.6 Jeux des données pour la classification et la prévision

Nous avons construit plusieurs jeux de données pour la classification et la prévision de la consommation d'eau et d'électricité.

3.6.1 Bases de données de la consommation d'eau

Cinq bases de données de la consommation d'eau sont proposées comme étant des entrées aux différents algorithmes d'apprentissage. Une description de chaque base est donnée brièvement ci-dessous :

- **BaseT1-Classification** :
  - id du compteur : id=2005 cf. Annexe A.2.2
  - du 01 Décembre 2017 au 08 Avril 2018
  - consommation moyenne d'environ 977 l, un minimum de 0 l et un maximum de 9786 l par jour
- **BaseT1-Détection** :

  - id du compteur : id=2005 cf. Annexe A.2.2
  - du 01 Décembre 2017 au 08 Avril 2018
  - consommation moyenne d'environ 977 l, un minimum de 0 l et un maximum de 9786 l par jour

- **BaseT1-prévision** :
  - id du compteur : id=2005 cf. Annexe A.2.2
  - du 03 Septembre 2018 au 30 Juin 2019
  - consommation moyenne d'environ 783.87 l avec une consommation maximale de 2639l et minimale de 0l par jour

- **BaseP1-prévision** :
  - id du compteur : id=2007 cf. Annexe A.2.2
  - du 01 Octobre 2018 au 31 Décembre 2018
  - consommation moyenne d'environ 187.11 avec une consommation maximale de 489 l et minimale de 0 l par jour

- **BaseP1-évènement-prévision** :
  - id du compteur : id=2007 cf. Annexe A.2.2
  - d'Octobre 2018 au Décembre 2018
  - comporte 2000 données

3.6.2 Bases de données de la consommation d'électricité

Trois bases de données de la consommation électricité sont proposées comme étant des entrées aux différents algorithmes d'apprentissage. Une description de chaque base est donnée brièvement ci-dessous :

- **BaseT2-Classification** : c'est la base des données utilisée pour la classification de 8 appareils électriques différents. Chaque échantillon de cette base de donnée représente un appareil électrique décrit par la signature harmonique du courant mesuré. La signature harmonique est constituée des amplitudes et des angles des composantes harmoniques suivantes obtenues par une transformée de Fourier : 1 (fréquence fondamentale) et 3, 5, 7, 9, 11, 13 et 15. Chaque appareil est ainsi décrit par 16 caractéristiques. Elles ont déjà été utilisées dans une étude antérieure [96]. Afin d'étudier la tâche de classification au plus loin, cette base de donnée est "dérivée" en quatre sous-bases contenant chacune 10 000 échantillons. Ces variantes sont définies de la manière suivante :
  - la sous-base "normale" contient des signatures harmoniques brutes obtenues directement avec une transformée de Fourier du courant ;

  - la sous-base "bruitée" est constituée des 16 caractéristiques de la signature harmonique mais entachées de 10% de bruit blanc ;
  - la sous-base de données "trafiquée" contient des signatures quelques peu trafiquées dans le sens où une seule des 16 caractéristiques est pertinente et donne directement l'information de la classe à laquelle l'appareil appartient ;
  - la sous-base de données "aléatoire" contient des signatures harmoniques aléatoires autrement dit il est extrêmement difficile d'établir une correspondance ou un lien entre les 16 termes et la classe à laquelle l'appareil appartient

- **BaseT2-prévision** :
  - id du compteur : id=206 cf. Annexe A.2.1
  - du 25 Novembre 2019 au 09 Février 2020
  - consommation journalière moyenne de 14 581 Wh par jour d'un ensemble de 7 semaines qui sont choisies pendant cette période

- **BaseT2-évènement-prévision** :
  - id du compteur : id=11 cf. Annexe A.2.1
  - du 01 Avril 2020 au 03 Avril 2020
  - constituée de 2000 données

- **BaseT3-évènement-prévision** :
  - id du compteur : id=206 cf. Annexe A.2.1
  - du 31 Mai 2021 au 04 Juin 2021
  - comporte 36 300 données

3.7 Conclusion

Dans ce chapitre, nous avons proposé une plateforme d'acquisition des données de consommation d'eau et d'électricité. La solution adoptée dans cette thèse est la collecte des données d'une manière non intrusive qui a permis une collecte des données à partir d'une seule partie centralisée d'un réseau de distribution. Les données récoltées sont l'index, qui représente le volume d'eau ou de l'énergie électrique consommé, et les évènements datés, qui représentent les instants de consommation de chaque litre d'eau et de chaque Wh. Les profils des consommations ont été représentés sous forme des Cdc journalières décrivant l'évolution de la consommation en fonction du temps. Ces données constituent des séries temporelles qui seront utilisées pour la classification, le clustering, l'analyse et la prévision.

Nous avons pu constater des pertes de données lors de la collecte. Cela correspond essentiellement à des interruptions lors de la sauvegarde dans les bases de données sur le cloud, dues

généralement à des problèmes réseaux, ou à des coupures électriques. De ce fait, des scripts ont été mis en place afin de signaler les cas d'interruption ou de d'un mauvais fonctionnement du compteur. Ces données ont été analysées dont le but est d'obtenir des Cdc des consommations échantillonnées et représentables permettant de conduire à des Cdc références. Les courbes références, C^{avg} C^{max} et C^{min}, offrent la possibilité d'une comparaison entre les profils journaliers et d'évaluer certaines consommations.

Une vérification de l'intégrité des données a été proposée. Elle consiste en une vérification d'un éventuel manque des données. Cela comporte une étape d'identification puis une autre étape de reconstitution des données à partir d'une interpolation selon le type de la consommation et du bâtiment. Pour la consommation d'un bâtiment tertiaire, nous avons proposé une interpolation spline cubique pour les jours travaillés. En cas de manque pendant les weekends, une interpolation linéaire a été retenue pour déterminer les données manquantes. Pour la consommation d'un bâtiment privé, nous avons proposé une interpolation par la moyenne de la consommation du jour d'avant (j-1) et du même jour de la semaine d'avant (j-7).

Les données de la consommation d'eau et d'électricité seront considérées comme des entrées pour tous les algorithmes ML développés dans les chapitres qui suivent afin de résoudre les différents problèmes posés dans cette thèse à savoir le clustering, la classification, la prévision des consommations et la détection des anomalies.

4 Classification automatique selon les profils de consommation

4.1 Introduction

Ce chapitre est focalisé sur la classification supervisée et non supervisée, appelée en anglais « clustering », des données représentées sous forme de Cdc et de séries temporelles des consommations. À partir des Cdc de la consommation d'eau collectées, nous ne pouvons réaliser aucune interprétation du comportement ou de la dynamique de ces courbes en fonction du temps. De ce fait, nous proposons, comme alternative d'explorer des données, de regrouper les Cdc de la consommation d'eau en classes avec un clustering dans le but de réduire le nombre des données en gardant que celles qui comportent plus d'informations. Nous proposons également une analyse probabiliste de ces Cdc journalières avec la PDF permettant d'identifier les périodes pendant lesquelles il y a eu de consommations significatives. De ce fait, la réduction des données va permettre, entre autres, de détecter le plus tôt possible les comportements anormaux qui se manifestent souvent par des grandes ou des faibles consommations. La détection des anomalies sera réalisé par une classification des Cdc.

Une analyse de la consommation de l'électricité a été mise en œuvre et a permis de séparer différents appareils à partir du courant de charge [96]. À partir de cela , nous disposons d'un ensemble des données décrivant huit appareils électriques par leurs signatures harmoniques. Notre objectif est de proposer un modèle de classification supervisée capable d'identifier et de classifier les signatures électriques de chaque type d'appareils dans des conditions de fonctionnement réelles. Différentes méthodes ont été testées pour classifier ces appareils : un perceptron multicouche MLP, un arbre de décision DT ainsi que la méthode de KNN. La classification avec ces méthodes sera ensuite comparée avec d'autres approches ensemblistes comme OAA et ECOC. Une exploration des appareils électriques sera réalisée à partir des signatures harmoniques permettant de simplifier et optimiser les modèles de classification en minimisant le nombre des caractéristiques en entrée pour la classification.

Table 4.1: Distance et corrélation de quelques Cdc de la consommation d'eau et les courbes de débit calculée respectivement avec la Cdc moyenne et la courbe de débit moyen

Jour (j)	(C^{avg})		(F^{avg})	
	$corr(C_{Per}^{avg}, C_j)$	$dist(C_{Per}^{avg}, C_j)$	$corr(F_{Per}^{avg}, F_j)$	$dist(F_{Per}^{avg}, F_j)$
2018-04-02	86,39	3.8e+04	-1.47	84.11
2018-04-03	99,95	4.6e+03	43.61	1.3e+02
2018-04-04	99,88	1.3e+03	46.93	1.5e+02
2018-04-05	99.72	3.1e+03	49.66	1.5e+02
2018-04-06	99,58	6.9e+03	51.94	1.5e+02

4.2 Approche probabiliste pour la classification des Cdc de la consommation d'eau

4.2.1 Méthode proposée pour le clustering des Cdc de la consommation d'eau

Dans cette section, une classification non supervisée des Cdc de la consommation d'eau de la base de données "BaseT1-Classification" est proposée avec des approches simples pour chercher la ressemblance d'une Cdc avec des courbes références.

a. Calcul statistique pour l'exploration des Cdc

Tout d'abord, une exploration des Cdc est réalisée avec le coefficient de corrélation de Pearson. Cela est réalisé avec la base de données "BaseT1-Classification" définie dans le paragraphe 3.6. Nous avons retenu que 90 jours dont la consommation soit différente de 0. Un échantillonnage des Cdc est réalisé de sorte que chaque Cdc est composée de 1440 données par jour.

La corrélation $corr(C_{Per}^{avg}, C_j)$ et la distance $dist(C_{Per}^{avg}, C_j)$ sont deux outils pour mesurer la ressemblance et la similitude de deux vecteurs. Dans ce contexte, ces deux vecteurs sont une Cdc journalière et une Cdc de référence respectivement C_j et C_{Per}^{avg}.

La corrélation est donnée par :

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_y}, \tag{4.1}$$

et la distance est calculée de la manière suivante :

$$cov(X, Y) = \frac{1}{N} \sum_{n=1}^{p} \sum_{j=1}^{q} n_{ij} (x_i - \overline{X})(y_j - \overline{Y}). \tag{4.2}$$

où $cov(X, Y)$ est la covariance entre X et Y, σ_x, σ_y, $\overline{X}$ et $\overline{Y}$ sont respectivement l'écart type et les valeurs moyennes de X et Y. Per représente les jours travaillés de cette base "BaseT1-Classification".

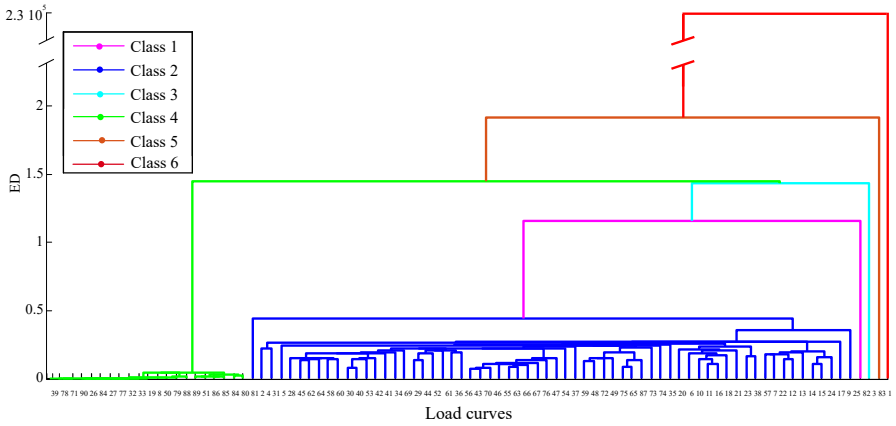

Figure 4.1: Dendrogramme issu du HAC des Cdc avec la ED

Les résultats pour quelques jours sont présentés dans le Tableau 4.1.

De toutes les corrélations avec C_j, on peut constater qu'une seule est inférieure à 90%. Le jour correspondant (02/04/2018) présente donc une distance plus élevée par rapport aux autres jours. Pendant ces jours, la consommation d'eau était différente et cela est également confirmé par la corrélation et la distance calculées avec F_j qui est le débit moyen journalier f_j échantillonné en minute.

Cette méthode permet la recherche des jours dont le profil de consommation constitue une différence avec la Cdc référence. Elle est simple, mais elle ne sera pas très performante dans le cas où la base de données comprend un nombre important de Cdc. En outre, l'interprétation des résultats constitue son inconvénient majeur à cause de la diversification et le choix de la Cdc référence. Pour cette raison, les algorithmes ML représentent une solution alternative pour palier à ces problèmes.

b. Clustering de Cdc de la consommation d'eau avec HAC

Dans cette section, nous cherchons à construire des groupes de Cdc de la consommation d'eau. Pour ce faire nous procédons à une classification non supervisée dont les classes ne sont pas connues à l'avance. Trois méthodes de regroupement direct et indirect sont proposées afin de regrouper les Cdc en classes. Le clustering direct est un clustering de Cdc brutes (non échantillonnées). Cependant, le regroupement indirect transforme les données collectées à intervalle variable en données échantillonnées. Pour procéder à ce type de clustering, il est nécessaire d'initialiser un nombre de classes à l'avance. Pour ce faire, une méthode statistique des écarts est appliquée pour choisir le nombre de clusters nécessaires.

Nous avons choisi un nombre de classes maximal k_{max} =10 comme initialisation de l'algorithme 1. L'application de cet algorithme a fourni un nombre des classes optimal qui vaut 6. Par conséquent, le clustering est paramétré avec 6 classes.

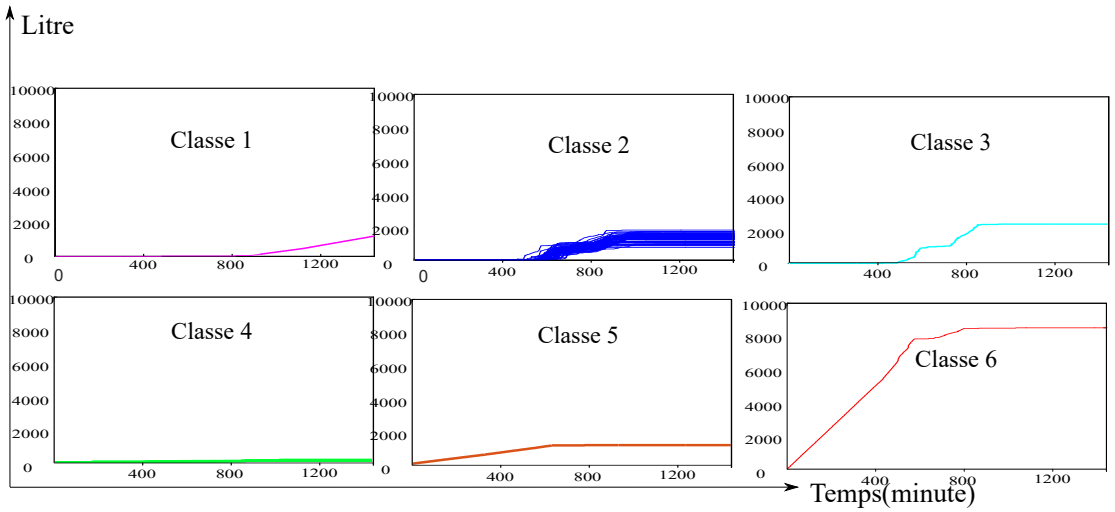

Figure 4.2: Cdc classées avec l'approche HAC utilisant la ED avec 6 classes

- Clustering HAC des Cdc avec ED
  Pour réaliser un HAC avec ED, les Cdc devraient être de même longueur. Par conséquent, les Cdc à classer seront échantillonnées.

  La base de données "BaseT1-Classification" reflète le comportement des utilisateurs de consommation d'eau, nous savons qu'il n'y aura pas de consommation très importante en moins d'une minute. Pour cette raison, nous choisissons une granularité d'une minute, de sorte que les Cdc d'un profil donné soient correctement échantillonnées dans le temps. Ces Cdc sont échantillonnées à 1440 données pour chaque jour (1440 minutes). Ceci permet de calculer la ED plus facilement.

  La Fig. 4.1 montre le dendrogramme obtenu en appliquant l'algorithme HAC. Selon cette figure, la classe 2 représente la majorité des Cdc, avec une part de 71,11 %. Avec ce regroupement nous pouvons conclure que les courbes de consommation quotidienne avec le comportement le plus normal sont regroupées dans cette classe. Les classes 1, 5 et 6 représentent les Cdc dans lesquelles il y a des fuites constatées (sur terrain).

  La Fig. 4.2 illustre les classes des Cdc de la consommation d'eau avec la ED. Nous remarquons que cette technique de clustering a permis d'identifier lisiblement les Cdc de la consommation d'eau ayant un comportement différent. Celui-ci est traduit par des ED très grandes par rapport au reste des classes. Les Cdc de la consommation d'eau pendant les weekends sont également regroupées dans une même classe, qui est la classe 4.

- Clustering des Cdc de la consommation d'eau avec HAC basé sur la DTW
  La DTW permet de se concentrer sur les différents comportements de consommation au cours de la journée, ainsi que sur le séquencement de chaque litre consommé tout en réduisant les différences et les décalages temporels des Cdc. En effet, la distance DTW réduit l'importance des différences temporelles et se concentre sur le comportement et

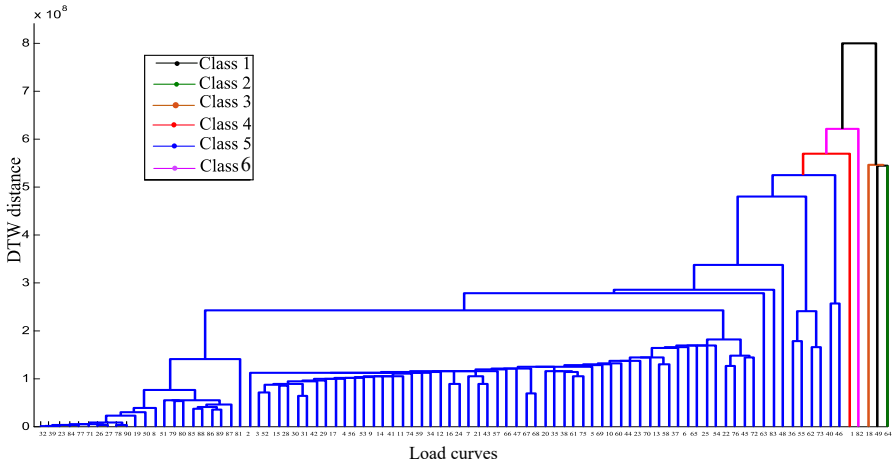

Figure 4.3: Dendrogramme issu du HAC des Cdc avec la distance DTW

son organisation dans la chronologie. Cette technique offre la possibilité de procéder au clustering sans échantillonnage des données ou de vérifier si les différents Cdc journalières ont le même nombre des données.

La Fig. 4.3 montre le dendrogramme obtenu en appliquant l'algorithme HAC avec la DTW. L'application de cette distance a fourni une représentation différente de la ED. En effet, nous constatons que la distance DTW se concentre sur la forme et les inclinaisons internes pour chaque Cdc. Les jours à très faible consommation sont considérés parmi les jours travaillés de la semaine.

En comparant le clustering HAC avec la ED et DTW, la première distance a bien réussi à regrouper les Cdc des weekends dans une même classe différente à celle des Cdc de jours travaillés.

c. Clustering de Cdc de la consommation d'eau avec la carte SOM

Une autre méthode de classification des Cdc cumulées consiste à utiliser le principe d'une SOM [97, 22]. Cette méthode est très peu utilisée pour le clustring particulièrement avec les séries de valeurs [98] qui sont dans notre cas d'étude des Cdc de la consommation d'eau. À l'inverse du clustering avec la HAC, le SOM est une méthode d'apprentissage automatique qui reste toujours opérationnelle avec un nombre important de Cdc. Ce réseau de neurones est composé d'une seule couche de neurones artificiels qui est une représentation discrétisée de l'espace d'entrée des échantillons d'apprentissage et sont donc appelés une carte. L'apprentissage consiste à déplacer la position des neurones en utilisant une fonction de voisinage pour préserver les propriétés topologiques de l'espace d'entrée. En cas d'une classification, chaque neurone correspond également à une classe et les données sont donc regroupées par similarité.

L'implémentation d'une SOM [15] a été réalisée avec Matlab [97].

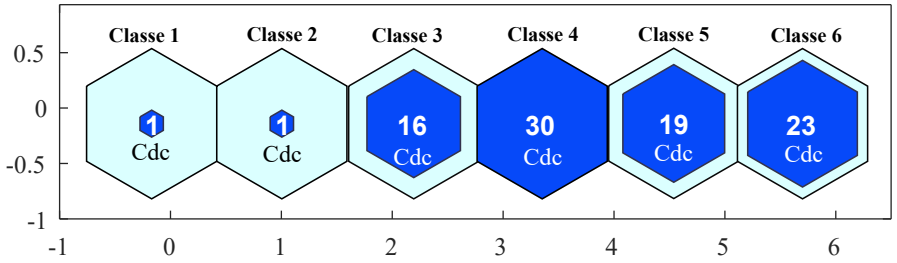

Figure 4.4: Résultats du clustering des Cdc de la consommation d'eau journalière avec un apprentissage SOM avec 6 neurones

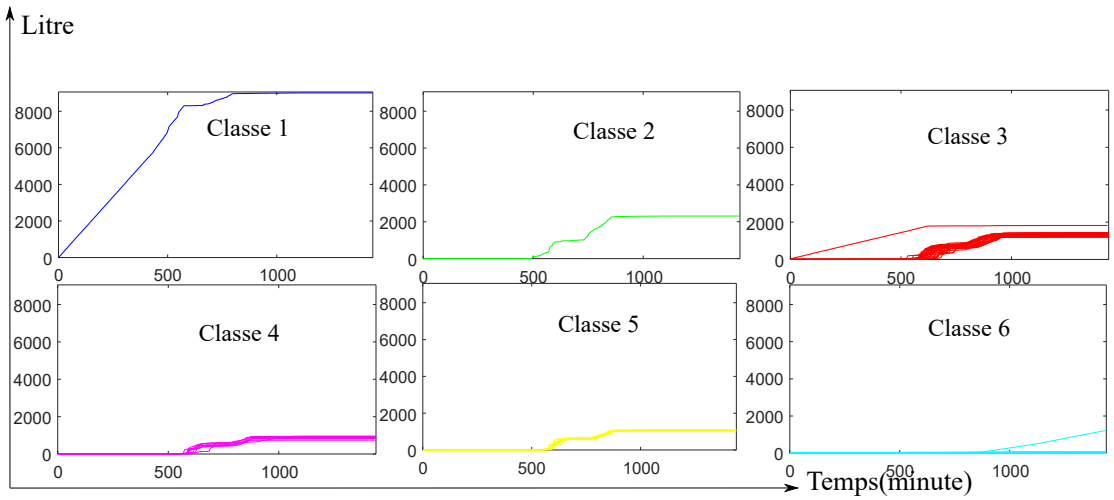

Figure 4.5: Cdc classées en 6 classes avec la carte auto organisatrice SOM

Table 4.2: Résultats du clustering des Cdc de la consommation d'eau à partir de leur dispersion dans les différentes classes avec les 3 méthodes

	ED	DTW	SOM
Classe 1	82	49	1
Classe 2	3	64	3
Classe 3	2,4,5,6,7,9,10,11,12,13, 14,15,16,17,18,20,21,22, 23,24,25,28,29,30,31,34, 35,36,37,38,40,41,42,43, 44,45,46,47,48,49,52,53, 54,55,56,57,58,59,60,61, 62,63,64,65,66,67,68,69, 70,72,73,74,75,76	18	2,4,12,13,15,17,22,24, 29,31,36,44,52,56,59, 83
Classe 4	8,19,26,27,32,33,39,50, 51,71,77,78,79,80,81,84, 85,86,87,88,89,90	1	5,6,7,9,10,11,14,16,18, 20,21,23,28,30,34,35,37, 38,40,41,42,45,53,57,58, 60,61,62,64,69
Classe 5	83	2,3,4,5,6,7,8,9,10,11,12, 13,14,15,16,17,19,20,21, 22,23,24,25,26,27,28,29, 30,31,32,33,34,35,36,37, 38,39,40,41,42,43,44,45, 46,47,48,50,51,52,53,54, 55,56, 57,58,59,60,61,62, 63,65,66,67,68,69,70,71, 72,73,74,75,76,77,8,79, 80,81,83,84,85,86,87,88, 89,90	25,43,46,47,48,49,54,55, 63,65,66,67,68,70,72,73, 74,75,76
Classe 6	1	82	8,19,26,27,32,33,39,50, 51,71,77,78,79,80,81, 82,84,85,86,87,88,89,90

Dans notre cas, une carte SOM est composée de 6 neurones, c'est-à-dire correspondant à 6 classes, est utilisée pour élaborer des classes des Cdc échantillonnées avec 1440 points de la même base de données "BaseT1-Classification". Les jours avec une consommation d'eau similaire sont donc regroupés en classes et le résultat est fourni par les Fig. 4.4 et Fig. 4.5.

4.2.2 Évaluation et comparaison du clustering des Cdc

Le Tableau 4.2 résume la dispersion des Cdc de la consommation d'eau dans les 6 classes avec les différentes méthodes du clustering telles que pour les trois méthodes, les Cdc sont identifié par un numéro de 1 à 90.

Table 4.3: Coefficient de corrélation cophénétique et temps d'exécution total pour le calcul de la similitude des Cdc avec le clustering HAC avec la ED et le DTW

	ED(Cdc)	DTW(Cdc)	ED(Cdc(I))
Coefficient de corrélation cophéntique	0.98	0.48	0.99
Temps total du clustering (s)	30.23	48.76 10	23.24

La mesure de la similarité entre les échantillons est l'étape essentielle pour une classification précise et efficace des Cdc. La mesure de similitude construit une valeur absolue de similitude entre chaque paire de séquences. Une bonne classification est définie en minimisant la distance intra-cluster et en maximisant la distance inter-cluster.

Le coefficient de corrélation cophénétique c [99] est une méthode objective de comparaison de plusieurs dendrogrammes. Ce coefficient est un indicateur de la qualité d'une classification non supervisée [100] pour un HAC. Elle est obtenue à partir de la distance cophénétique qui est la différence entre deux observations représentées dans le dendrogramme par la hauteur des branches se terminant dans le même groupe. C'est aussi la distance entre deux groupes contenant ces observations avant qu'elles ne soient réunies. Toutes ces distances définies entre les deux Cdc forment une matrice triangulaire. La qualité d'un arbre de classification peut se résumer par un coefficient de corrélation cophénétique c entre les valeurs de la matrice de distances initiales. La meilleure classification a un coefficient plus proche de 1. Ce coefficient permet donc de comparer les représentations graphiques de HAC à travers les distances cophénétiques induites par la classification non supervisée à l'aide d'une seule valeur.

$$c = \frac{\sum_{i<j}(Distance(i,j) - \bar{d})(c(i,j) - \bar{c})}{\sqrt{(\sum_{i<j}(Distance(i,j) - \bar{d})^2 (c(i,j) - \bar{c})^2)}} \tag{4.3}$$

avec *Distance(i, j)*: représente la distance entre x_i et x_j comme donnée dans la matrice de dissimilarité, *c(i, j)* est la distance cophénétique entre x_i et x_j, $\bar{d}$ est la moyenne des *Distance(i, j)* et $\bar{c}$ est la moyenne de *c(i, j)*.

La Tableau 4.3 présente le coefficient de corrélation cophénétique calculé pour l'ED de la Cdc quotidienne complète et DTW et le temps nécessaire pour calculer les distances Cdc et la construction du dendrogramme HAC.

Selon le coefficient cophénétique, nous retenons le clustering effectué avec ED. Nous comparons maintenant le clustering avec cette distance et celle obtenue avec la carte SOM. Graphiquement, il est bien clair dans la Fig. 4.5 qu'il existe des Cdc dans la classe 3 et 5 qui sont totalement différentes aux restes des Cdc de la même classe. Cependant, selon la Fig. 4.2, le HAC avec ED a bien séparé ces courbes et les a mis dans des classes différentes.

Selon plusieurs études [101], les réseaux de neurones s'avèrent très largement la meilleure

solution pour le clustering. Cela parait intéressant dans le cas où le nombre des classes est important. En outre, la SOM reste relativement stable et robuste par rapport aux valeurs aberrantes et aux dimensions non pertinentes des données.

Notre cas d'étude concerne un nombre relativement petit constitué d'un ensemble de 90 Cdc et un nombre relativement restreint de classes égale à 6. Par conséquent, nous avons retenu le clustering réalisé avec le HAC et ED.

Le clustering est appliqué sur des Cdc de la consommation d'eau échantillonnées en minute. Le processus de la classification n'est peut alors être lancé que lorsque toute les 1440 données sont disponibles. Cela est considéré comme une contrainte dans le cas où une anomalie (fuite d'eau, manque de la consommation ...) aura lieu. Pour faire face à cette contrainte, nous proposons d'explorer les Cdc afin de les classifier avec un seul sous-ensemble de données relatives à chaque jour.

La consommation journalière d'eau est un processus aléatoire avec des mesures bruitées sans aucune connaissance sous-jacente. Pour faire face à cette incertitude, les modèles probabilistes sont les plus adéquats. De plus, le processus de consommation présente des schémas récurrents sur différentes périodes. Une première méthode est proposée qui consiste à exploiter l'architecture de la plateforme d'eau. Cela revient à analyser les instants t_i des index récupérés. La deuxième méthode consiste à analyser la PDF. Cette exploration permet d'analyser le profil de consommation d'eau en minute. Il s'agit là d'une solution permettant le contrôle en temps réel afin d'éviter toutes anomalies et une intervention plus rapide.

4.2.3 Analyse de la consommation d'eau selon les instants de transmission de données

Une analyse de la consommation d'eau à partir des Cdc journalière est réalisée afin d'explorer la période pendant laquelle la consommation est importante. Comme décrit précédemment, l'architecture interne du compteur définit un temps de transmission des données inégalement espacées. Les instants d'envoi sont influencés par le volume d'eau consommé. Avec une grande consommation d'eau, les données sont récupérées avec des laps de temps $\delta(t_i)$, plus petit. Inversement, avec une petite consommation d'eau, les données sont récupérées avec des laps de temps $\delta(t_i)$, plus grands. Mathématiquement, cette analyse qui est basée sur le $\delta(t_i)$ minimum (Min) et maximum (Max) de chaque jour, est décrite dans l'algorithme 3. Avec cet algorithme,

Algorithm 3 Zone divided (Cdc, $\delta(t_i)$)

if $Min(\delta(t_i)) \leqslant \delta(t_i)$ **and** $\delta(t_i) \leqslant \frac{[Max(\delta(t_i)) - Min(\delta(t_i))]}{3}$ **then**
    Existing water consumption (Contained consumption)
else if $Max(\delta(t_i)) \geq \delta(t_i)$ **and** $\delta(t_i) \geq 2 \times \frac{[Max(\delta(t_i)) - Min(\delta(t_i))]}{3}$ **then**
    Null or a little water consumption
end if

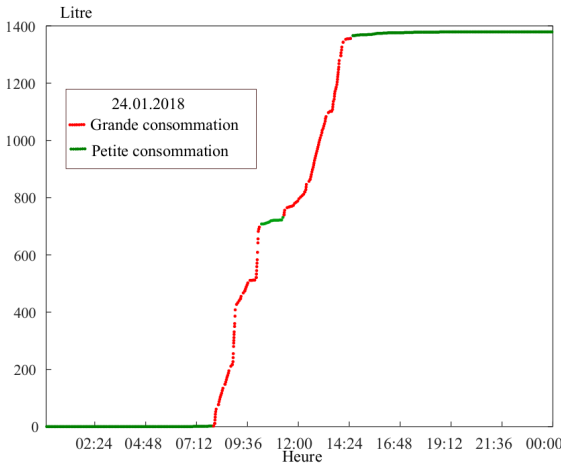

Figure 4.6: Décomposition automatique de la Cdc de consommation d'eau en zones

la grande consommation est évaluée par rapport aux instants de réception des index de la part de la plateforme intelligente. Cela consiste à comparer les écarts de réception des données. Un petit écart explique qu'il y a eu une consommation d'eau continue. La grande consommation se manifeste par les deux pentes dans une Cdc journalière représentées par une couleur rouge dans la Fig. 4.6.

Cette analyse simple permet de visualiser graphiquement la Cdc en deux zones de consommation. La grande consommation est pendant la journée de 8h jusqu'à 11h et de midi jusqu'à 15h. Cependant la petite consommation est de minuit jusqu'à 8h, environ de midi et le soir de 15h jusqu'à minuit. Cette technique simple a permis d'identifier plusieurs périodes de consommation pendant une journée. Nous cherchons une seule période dont la consommation figure dans toutes les Cdc de tous les jours ayant la consommation la plus importante. Pour cette raison, nous avons opté pour une analyse probabiliste plus sophistiquée qui permet d'attribuer pour chaque période de consommation une probabilité. La fonction de densité de probabilité PDF est capable d'explorer le processus de consommation d'eau et de l'exprimer sous forme de probabilités avec un estimateur de noyau gaussien.

4.2.4 Optimisation de données selon la PDF

Afin d'estimer la PDF [102] d'une quantité d'eau consommée pendant une période donnée t, nous utilisons un estimateur de densité KDE (Kernel Density Estimator) [103, 104]. Un KDE est un modèle de densité non paramétrique avec un type particulier de fonction de noyau $k_h()$ appelé noyau de lissage qui satisfait les conditions suivantes :

$$\begin{cases} \int K_h(t)\,dt = 1, \\ \int tK_h(t)\,dt = 0, \\ \int t^2 K_h(t)\,dt > 0, \end{cases} \tag{4.4}$$

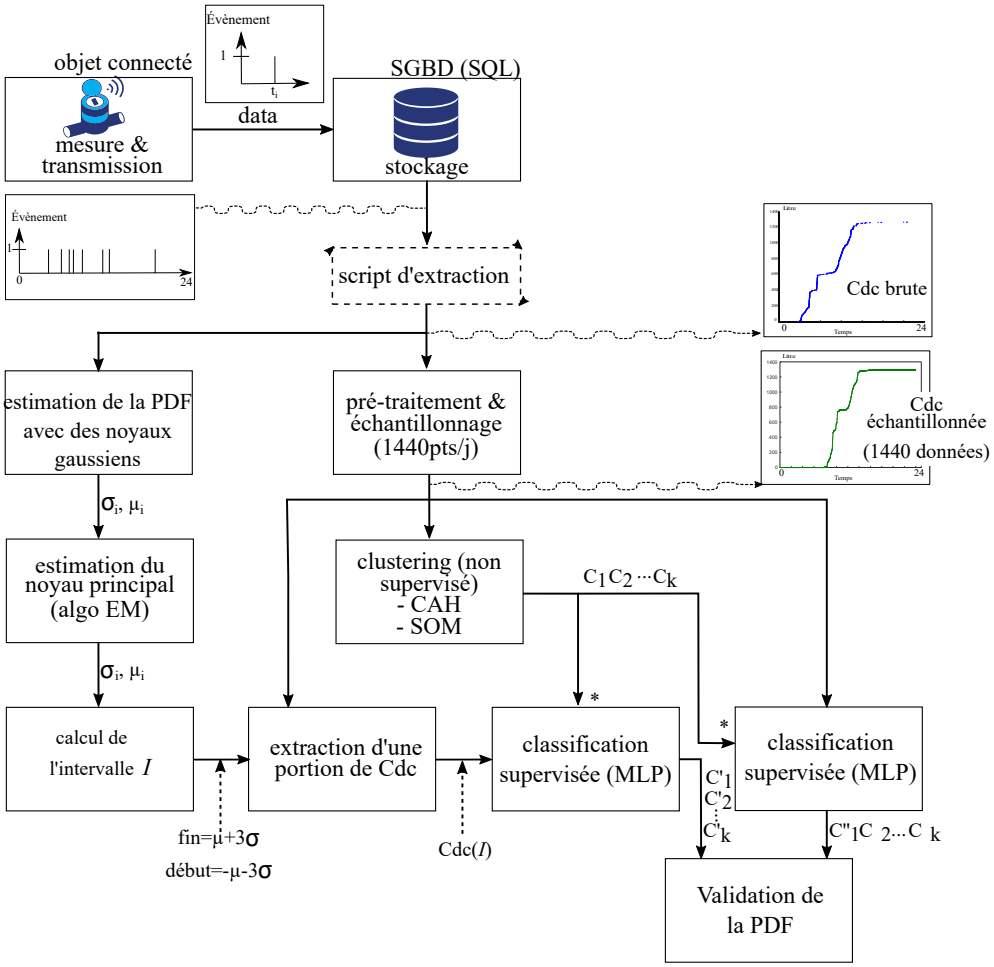

Figure 4.7: Approche proposée pour la classification des Cdc de la consommation d'eau journalière avec un nombre des données restreint pour estimer la PDF

avec $h > 0$ est la bande passante du noyau qui permet de contrôler le lissage de la fonction de densité PDF [105]. Cette dernière peut être estimée en calculant la probabilité en chaque point :

$$\hat{p}(t) = \frac{1}{N} \sum_{i=1}^{N} K_h(t - t_i) \tag{4.5}$$

La méthode KDE [106] ne dépend d'aucune connaissance préalable de la distribution des données. Il s'agit donc d'une estimation utilisant uniquement les données [103]. Pour notre cas d'étude, l'objectif de la PDF est de trouver la probabilité de consommation d'eau à chaque instant et de la représenter sous forme d'une courbe gaussienne. La Fig. 4.7 présente notre démarche qui est décrite comme suit. À partir d'une Cdc qui représente une série temporelle de la consommation d'eau journalière, nous appliquons la fonction PDF afin de représenter les distributions gaussiennes. Nous appliquons l'algorithme EM pour la recherche des paramètres des gaussiennes.

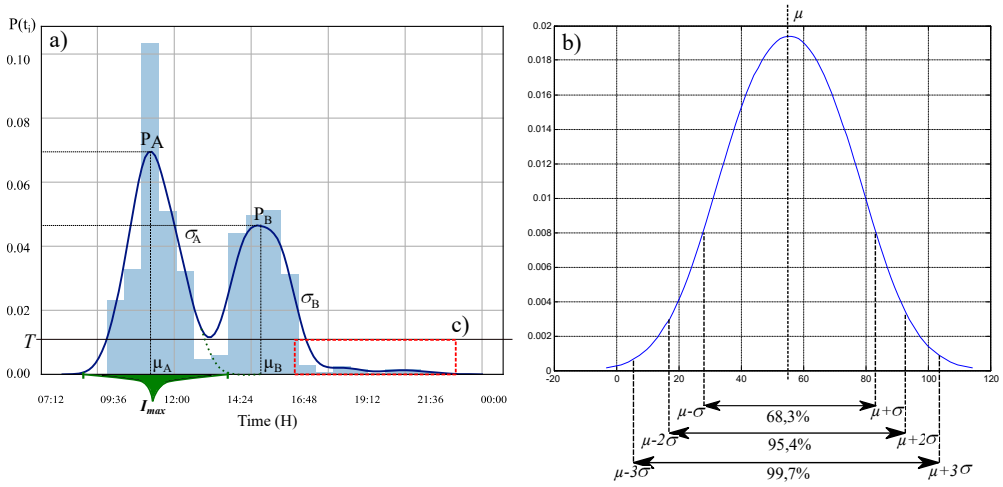

Figure 4.8: a) PDF avec noyau gaussien pour identifier le seuil T, b) intervalle de confiance d'une distribution gaussienne et , c) les bruits représentés par de petites gaussiennes

La PDF permet alors d'identifier, pour chaque jour, quelques litres "indexés" au début et à la fin de la journée, représentant les fuites d'eau nocturnes [12]. Ces fuites sont considérées comme étant un bruit par rapport à la consommation d'eau brute des utilisateurs. Par conséquent, un seuil T est choisi empiriquement permettant d'identifier que les gaussiennes significatives. La Fig. 4.8.c illustre un exemple de petites formes de gaussiennes représentant quelques litres d'eau. Le seuil T est définit par :

$$I = \{\hat{p}(t) \geq T)\}, with\ 0 \leq T \leq 1 \tag{4.6}$$

Nous supposons, dans cette étude, que la PDF présente deux Gaussiennes qui sont les plus significatives. Afin d'identifier les intervalles de densité I, l'algorithme EM est appliqué pour trouver les paramètres gaussiens μ et σ.

4.2.5 Identification des paramètres de la PDF

L'algorithme EM [107] est un algorithme itératif permettant de trouver le maximum de vraisemblances des paramètres des modèles probabilistes lorsque le modèle dépend des variables latentes non observables. Cet algorithme offre la possibilité de chercher les paramètres du mélange gaussien identifié avec la PDF. Ces paramètres sont définis par μ et σ qui sont respectivement les centres des gaussiennes (l'espérance) et la distribution des données (la variance). L'analyse et l'exploration probabiliste permettent de récupérer au moins la période la plus importante durant chaque journée, d'identifier plusieurs intervalles à l'intérieur de cette période, si nécessaire, de n'importe quelle taille, avec n'importe quelle granularité et fournir un mappage isomorphe entre les intervalles et les probabilités. Après l'identification des paramètres gaussiens ayant la probabilité la plus importante par rapport au mélange de distributions, un calcul de l'intervalle de normalité (confiance) est effectué tel qu'il est présenté sur la Fig. 4.8, b. Cet intervalle indique

la période pendant laquelle la consommation est la plus concentrée. L'intervalle de normalité pris pour avoir près de 99% de I est :

$$[\mu - 3\sigma, \mu + 3\sigma] \tag{4.7}$$

4.2.6 Application de la PDF aux Cdc de la consommation d'eau pour la réduction de données

a. Courbes de charge d'entrée

Le but de cette étude est de générer des estimations PDF pour la consommation enregistrée par un compteur intelligent individuel. Ces estimations peuvent faciliter la prise de décision en aidant les consommateurs à identifier et à minimiser leur consommation excessive d'eau, en particulier pendant les périodes de pointe identifiées avec le KDE. Pour les fournisseurs, ces estimations peuvent être utilisées pour développer des stratégies de tarification innovantes en fonction du temps d'utilisation pour leurs consommateurs cibles. Sur une période d'environ 5 mois, une sélection de 90 Cdc de consommation d'eau quotidienne est effectuée. Cette sélection tient compte de la nature du bâtiment et de l'objectif de la PDF à appliquer par la suite. En effet, le bâtiment étudié contient une consommation d'eau uniquement pendant les jours de la semaine hors week-end et jours fériés. De plus, la PDF explore chaque Cdc de consommation journalière afin d'en déduire la PDF la consommation. Les données d'entrée pour la recherche de la PDF sont les Cdc de la consommation d'eau journalière. Pour avoir une représentation plus simplifiée du temps cumulé en millisecondes de $\delta(t_i)$.

b. Paramètres utilisées et classification non supervisée avec ML

L'analyse des différents Cdc fournit une représentation répétitive et dominante d'un mélange de deux gaussiennes pendant les jours de la semaine, qu'ils sont représentés sur la Fig. 4.8, a. Le but de cette distribution est d'identifier, pour chaque Cdc journalière, l'intervalle pendant lequel il y a la plus forte probabilité d'avoir de l'eau consommée. Puis une généralisation des intervalles par rapport à la totalité de toutes les Cdc pour trouver l'intervalle I.

L'exécution de l'algorithme EM a permis de déterminer les paramètres de deux gaussiennes A et B. En fonction de σ, il sera possible d'identifier la forte probabilité. Si $\sigma_A < \sigma_B$, nous pouvons déduire que la gaussienne A a la probabilité la plus élevée. Sinon, la gaussienne B a la probabilité la plus élevée. Dans le cas où σ de deux gaussiennes sont égales, le raisonnement sera basé sur la première gaussienne. En effet, les σ de la première gaussienne présentent souvent (plus de 91,67%) la petite ou la même valeur que la variance de la seconde gaussienne B. Par conséquent, la première gaussienne a la plus forte probabilité de consommation d'eau.

Après avoir déterminé les paramètres des courbes gaussiennes σ et μ il ne reste plus qu'à déterminer l'intervalle I de la probabilité la plus élevée. Cet intervalle est déterminé à partir

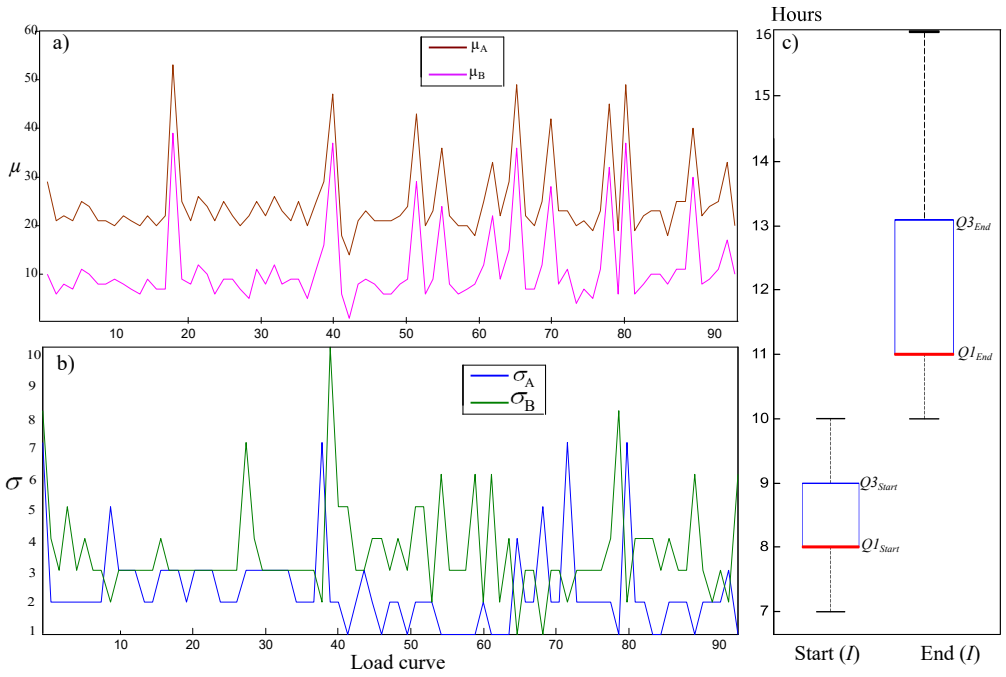

Figure 4.9: Paramètres gaussiens a) Les centres, b) les variances de deux Gaussiennes A et B et c) Box-plot de Cdc (I) des intervalles horaires de la période I

de l'intervalle de normalité de la première gaussienne I_{max} représenté dans la Fig. 4.8, a. Il est identifié à partir de l'intervalle de confiance de la gaussienne avec (4.7). Pour chaque Cdc, un intervalle Cdc(I_{max}) est identifié. La Fig. 4.9, c, montre la dispersion et la centralité de la distribution des valeurs avec l'heure de début et de fin de la consommation d'eau pour chaque jour.

À partir de la Fig. 4.9, c, nous définissons l'intervalle I avec : **I=[Start(I)=Q1**$_{start}$**,End(I)=Q3**$_{End}$**]** de telle sorte que Q1 et Q3 sont défini à partir des quartiles des heures arrondies. Dans notre étude de cas, on obtient I=[*8h, 13h*] qui représente 300 données (minutes).

4.2.7 Classification supervisée de la consommation d'eau en temps réel avec le réseau MLP

Dans cette section, une classification non supervisée est fournie en utilisant la ED. Ce regroupement est effectué uniquement sur une partie de chaque Cdc représentée par l'intervalle I avec 300 données.

Les résultats obtenus avec ce clustering sont représentés sur la Fig. 4.10 b). Les résultats du clustering de Cdc journalières complètes (1440 donnée) sont représentés sur la Fig. 4.10 a). Selon les résultats de l'évaluation du clustering du Tableau 4.3, le coefficient de corrélation cophénétique c calculé avec la ED(Cdc(I)) avec un nombre des données est égale à 0,99. Cette valeur a été améliorée par rapport au coefficient c calculé avec la distance ED(Cdc) avec les

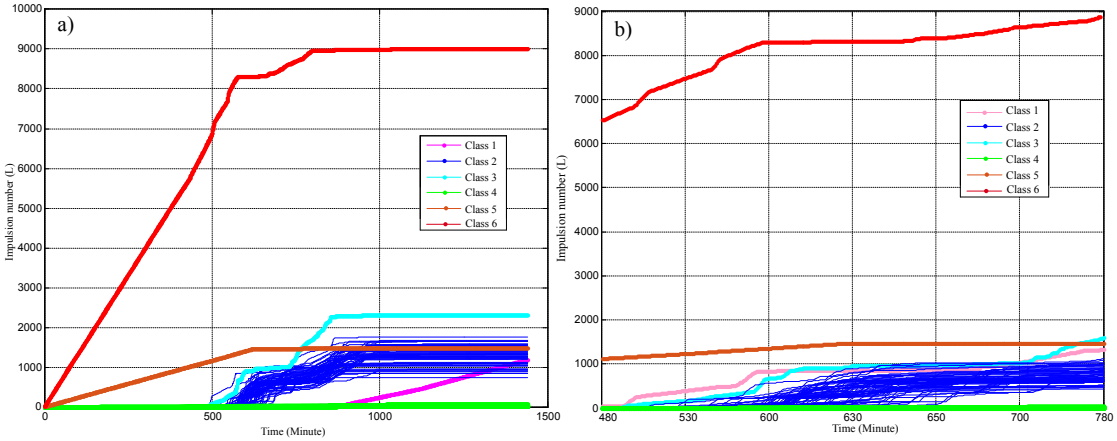

Figure 4.10: a) Résultat de la classification des Cdc journalières échantillonnées avec 1440 données et b) résultats de la classification des Cdc sur la période I échantillonnées en 300 données

1440 données pour chaque jour qui était 0,98. En effet, le temps nécessaire pour la construction des clusters a diminué d'environ 7s.

a. Paramètres de la classification supervisée des Cdc avec le modèle MLP

Dans cette section nous nous intéressons à classifier les Cdc représentées que sur la période I en utilisant le modèle MLP décrit dans la Section 2.3. Les entrées du MLP sont la consommation échantillonnée représentée par 300 données. Chaque donnée représente un nombre de litres consommés à une minute donnée de l'intervalle I. Nous avons testé plusieurs configurations pour le MLP et nous avons gardé à la fin celle qui a fournit la meilleur classification. Cette configuration consiste en 12 neurones pour la couche cachée et les neurones de sortie représentent les 6 classes représentées par $k = 6$. La fonction d'activation des neurones cachés est une sigmoïde. Nous avons divisé les données comme suit. L'ensemble d'entrainement est composé de 70% des Cdc qui sont choisi au hasard. Le reste représente l'ensemble de test.

En conservant les mêmes paramètres, ce MLP va nous servir à classifier les Cdc journalières complètes. À travers cette classification, nous aurons la possibilité d'évaluer le clustering proposé dans la Section 4.2 ainsi que d'analyser l'impact de la réduction des données des Cdc de la consommation d'eau par la comparaison des résultats de la classification.

b. Évaluation de la classification des Cdc : hors ligne et en temps réel

En appliquant le MLP sur la classification des Cdc quotidiennes complètes et les portions Cdc(I), nous avons obtenu les résultats du Tableau 4.4. Selon ces résultats, nous pouvons constater tout d'abord que le clustering des Cdc avec la technique HAC construit avec la ED est acceptable. Cela peut être confirmer avec les taux de classification MLP(Cdc) qui est 99.92%. Cependant,

Table 4.4: Résultats de la classification MLP sur Cdc et Cdc(I)

	MLP(Cdc)	MLP(Cdc(I))
Taux d'entrainement (%)	99.92	99.97
Taux de test (%)	55.89	100
Temps de calcul (s)	468.53	41,44

le modèle MLP n'a pas pu généraliser sa classification qui peut être justifié par la complexité des Cdc. Nous avons constaté également que la classification sur la période I de 300 données a fourni des résultats meilleurs et plus rapides pour l'ensemble d'entraînement et de test par rapport à la totalité des 1440 données. Les résultats de la classification des Cdc avec un nombre des données réduit sont 99.97% pour l'ensemble d'entrainement et 100% pour l'ensemble de test. De plus, le temps nécessaire pour l'apprentissage et le test de classification avec les données complètes était 468.53s qui a été réduit à 41.44s pour les Cdc avec 300 données.

Nous avons lancé un premier script de classification des Cdc de la consommation d'eau avec un MLP qui est destiné à s'exécuter en temps réel. Ce script permet de recueillir à chaque minute la consommation d'eau consommée pendant l'intervalle I. Ensuite, il classe chaque Cdc(I) avant la fin de la journée avec 300 données entrées pour chaque classification. Un deuxième script fonctionne avec le même principe que le premier sauf qu'il collecte les 1440 données journalières échantillonnées en minutes. La classification avec la Cdc complète ne peut être réalisée qu'une fois le jour est terminé.

4.3 Classification supervisée des appareils électriques avec ML

Les courbes de charge cumulées résultantes représentent la consommation globale et les appareils connectés en aval dans le réseau électrique sont inconnus à l'avance. Cette configuration est appelée surveillance de charge non intrusive (Non-Intrusive Load Monitoring, NILM) [96, 108]. Elle consiste à mesurer des données de consommation agrégées et à utiliser des algorithmes intelligents pour identifier la nature des équipements du consommateur [88, 96]. Dans cette étude, nous allons classifier des appareils électriques au sein d'un bâtiment tertiaire.

4.3.1 Identification des signatures harmoniques de courant

Avec les algorithmes ML [22], il est possible de détecter dans les Cdc certains modèles d'utilisation de l'électricité. Cependant, il est très difficile de déduire avec précision quels appareils consomment ou non d'électricité. De même il est difficile de comprendre comment les appareils consomment de l'énergie. La toute première contribution dans ce domaine provient de l'approche de la surveillance de la charge non intrusive NILM [88]. Le défi ici consiste à obtenir des informations détaillées sur le courant et la tension.

Les algorithmes de désagrégation requièrent une signature de charge complète et la bonne

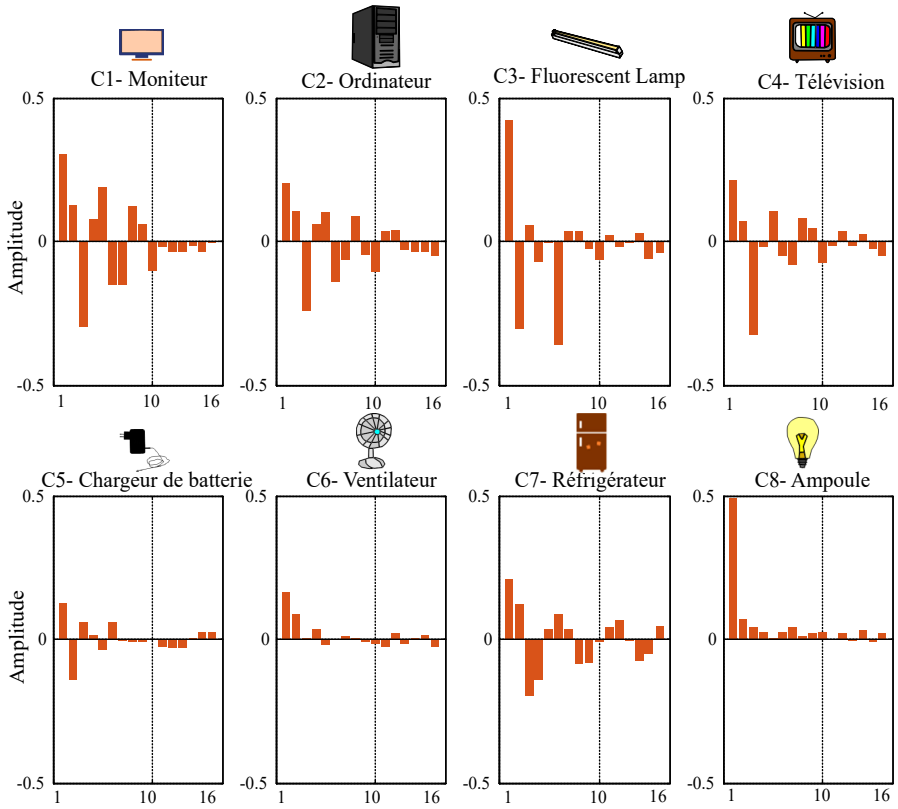

Figure 4.11: Composants de courant harmonique de 8 appareils

combinaison de charges uniques en fonction de la consommation d'énergie réelle. Une méthode NILM basée sur des réseaux de neurones artificiels, c'est-à-dire une approche ML, a été proposée dans [96]. Cette approche consiste à utiliser un réseau de neurones pour apprendre et à approximer avec des mises à jour en temps réel de la transformée de Fourier rapide [97] du signal mesuré. Ce réseau a permis d'identifier l'amplitude et la phase de chaque harmonique d'ordre supérieur du courant mesuré sur une ligne électrique. Une harmonique est définie selon [96] comme une composante d'une onde périodique ayant une fréquence qui est un multiple entier de la fréquence fondamentale de la ligne électrique. Une mesure de courant haute fréquence est nécessaire, elle correspond à un échantillonnage inférieur à chaque milliseconde de $\frac{1}{10^{-5}}$. Après la convergence de l'apprentissage, les signatures harmoniques sont efficacement extraites et sont caractéristiques de la distorsion du signal du courant de la charge. Ces signatures sont composées des amplitudes des 16 premières composantes harmoniques d'ordre supérieur les plus significatives du courant et sont utilisées en tant qu'entrées d'un algorithme de classification. Ces données sont issues d'une analyse et un traitement du signal électrique en temps réel présentée dans la Fig. 3.1 (avec le bloc en gris).

Dans les expériences réalisées dans [96], seules les harmoniques impaires, du fondamental jusqu'à la 15ème harmonique sont considérées comme des amplitudes significatives. Par conséquent, pour les appareils pris en compte, le vecteur de caractéristiques est de dimension

16 [109]. La Fig. 4.11 montre les signatures harmoniques de 8 appareils électriques.

Cette méthode NILM, qui a été complètement implémentée avec des algorithmes ML, est capable d'identifier et de classifier les signatures électriques de chaque type d'appareils dans des conditions de fonctionnement réelles.

4.3.2 Données d'entrée et modèles de classification des appareils électriques

a. Données d'entrées : signature électrique des appareils

Les consommations des appareils qui sont caractérisées par les signatures électriques sont mesurées dans différentes conditions que ce soit bruitées, normale trafiquée ou aléatoire. La base des signatures globale comporte 40 000 échantillons. Chaque échantillon est une signature d'un appareil qui est représentée par 16 caractéristiques qui ne sont que les entrées des classificateurs supervisés. Une sélection aléatoire de 70% de la base de données globale pour former le modèle et 30% pour le tester a été faite.

Les appareils sont étiquetés de la manière suivante avec 8 classes : C1 = moniteur, C2 = ordinateur, C3 = lampe fluorescente, C4 = télévision, C5 = Chargeur de batterie, C6 = ventilateur, C7 = réfrigérateur, C8 = ampoule.

À partir de la base de données des signatures des appareils électriques, nous voulons construire un modèle fournissant un taux de classification le plus élevé possible en vérifiant le compromis biais variance défini dans Section 2.6.2. Dans ce qui suit, nous n'allons présenter que les modèles ayant fournis les meilleurs résultats de classification. Les modèles de classification individuels sont la DT, MLP et KNN. Le MLP et le KNN seront utilisés avec deux différentes configurations de paramètres pour chacun.

b. Modèle pour la classification automatique des signatures des appareils électriques

Les appareils électriques ont été classifiés avec trois algorithmes ML supervisés : l'arbre de décision DT, le MLP et le KNN.

- Perceptron multicouche (MLP) : modèle 1

  Un premier modèle de MLP pour la classification des amplitudes harmoniques constituant les signatures des appareils électriques est proposé. En tant que MLP typique, ce réseau de neurones utilise plusieurs neurones sigmoïdes dans la couche cachée et des neurones linéaires dans la couche de sortie. La valeur du seul neurone dans la couche de sortie du réseau MLP fourni le numéro de la classe (1,2, ..., 8). L'architecture du modèle est représentée par la Fig. 4.12 a) avec MLP, modèle 1.

- Perceptron multicouche (MLP) : modèle 2

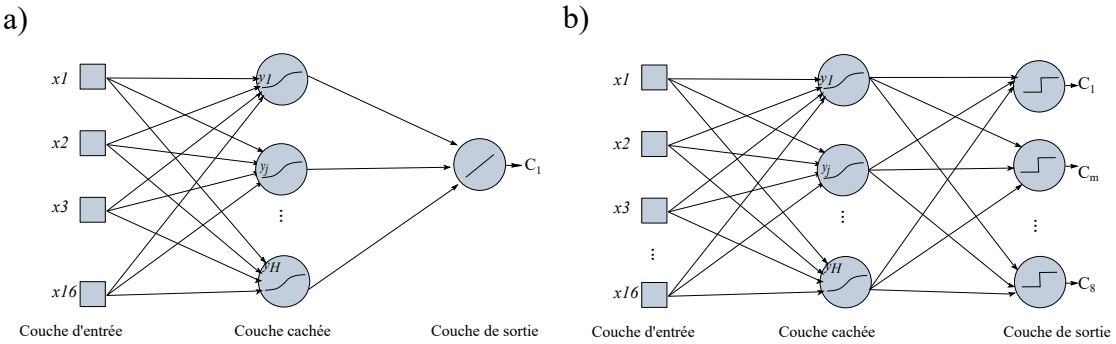

Figure 4.12: a) MLP, modèle 1 b) modèle 2

Le deuxième modèle proposé de MLP pour la classification des amplitudes harmoniques constituant les signatures des appareils électriques est représenté par la Fig. 4.12 b. En tant que MLP typique, ce réseau de neurones utilise plusieurs neurones sigmoïdes dans la couche cachée et des neurones linéaires dans la couche de sortie. Le nombre de neurones dans la couche de sortie du réseau MLP dépend du nombre d'appareils que nous voulons identifier à partir du signal d'entrée. Dans notre cas, le MLP possède 8 neurones de sortie. Chaque neurone de sortie, fournira les valeurs 0 ou 1 qui correspondent aux états ON ou OFF de chaque appareil. Par exemple, pour une sortie (1,0,0,0,0,0,0,0) correspond à C 1 qui est un moniteur. Ce MLP est donc un modèle à sortie binaire appelé ici MLP modèle 2.

- Arbre de décision (DT)

  L'arbre de décision de type CART [110] est utilisé qui comporte une racine, des branches et des feuilles. C'est un arbre binaire qui est conçu à partir d'une division descendante. À chaque division un échantillon est testé et un sous-arbre est choisi. Les feuilles de la DT représentent les 8 classes d'appareils.

- Le k-plus proche voisins (KNN)

  La méthode des KNN, consiste à classer des nouvelles signatures des appareils en fonction de leurs distances par rapport aux plus proche voisins constituant des échantillons antérieurs. Nous avons comparé deux configurations de KNN avec modèle 1, k=1 et modèle 2, k=10. La distance utilisée pour calculer le plus proche voisin est la distance euclidienne dans un espace de 16 dimensions.

4.3.3 Classification automatique des appareils électriques avec les modèles individuels

Le processus de classification comprend deux phases : l'apprentissage des données et le test des échantillons en 8 classes prédéfinies des appareils électriques. Pour le modèle MLP avec ces deux configurations, nous avons varié le nombre de neurones dans la couche cachée (de 2 à 25 neurones). Les résultats de la classification avec les modèles 1 et 2 du MLP sont évalués par les taux de classifications représentés dans le Tableau 4.5. Dans ce tableau, le MLP de la configuration

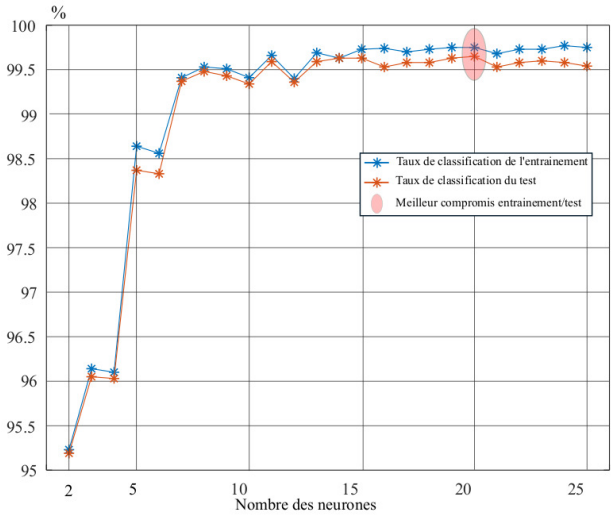

Figure 4.13: Exemple de compromis de taux de classification entrainement/test

du modèle 1 avec 17 neurones dans la couche cachée a fourni les meilleurs performances en entrainement 98,60% et en test 98,19%. Cependant, le MLP de la configuration du modèle 2 avec 20 neurones dans la couche cachée a fourni les performances les plus acceptables en entrainement 99,29% et en test 99,23%. Nous avons retenu ces résultats des taux de classification qui représentent un meilleur compromis biais/variance malgré que nous pouvons distinguer quelques résultats plus performants en classification. L'identification de la configuration retenue est recherchée avec l'algorithme 4. Ce dernier prend comme entrées les taux de classification de deux sous-ensembles (entrainement et test) en adéquation avec le nombre des neurones correspondant. Un exemple du meilleur compromis est identifié dans la Fig. 4.13.

Algorithm 4 Compromis_classification (nb-train, nb-test, rate-train, rate-test)

```
1:  sort [nb-train, rate-train]                          ▷ sorting by rate-train
2:  sort [nb-test, rate-test]                            ▷ sorting by rate-test
3:  if rate-train=100 and rate-test=100 then
4:      if nb-train= nb-test then
5:          best configuration = nb-train
6:      end if
7:  else
8:      find max(rate-train)                             ▷ we get nb-train
9:      find max(rate-test)                              ▷ we get nb-test
10:     if nb-train<nb-test then
11:         best configuration = nb-test
12:     else
13:         best configuration = nb-train
14:     end if
15: end if
```

Les résultats obtenus par le modèle MLP avec les deux configurations proposées peuvent être

Table 4.5: Comparaison de la précision des configurations du MLP avec le modèle 1 et le modèle 2 avec le nombre de neurones cachés de 2 à 25 pour classifier les données de la base globale des signatures des appareils

Nombre de neurones dans la couche cachée	**Taux de classification (%) des signatures harmoniques de 8 appareils électriques avec deux configurations différentes du MLP**			
	Modèle 1		**Modèle 2**	
	Train	**Test**	**Train**	**Test**
2	73.89	72.96	9.31	9.62
3	78.08	77.41	29.14	28.33
4	80.24	78.93	38.13	37.67
5	92.46	92.26	45.85	45.39
6	93.39	92.88	64.27	63.43
7	93.24	92.50	73.23	72.47
8	94.96	94.48	74.41	73.77
9	96.08	95.43	74.27	73.48
10	97.32	96.71	80.31	80.11
11	96.81	96.43	80.15	79.74
12	96.31	95.90	79.49	79.22
13	96.77	96.333	86.06	88.74
14	97.38	96.98	88.66	88.66
15	96.46	95.85	95.56	95.28
16	98.14	97.76	98.98	98.83
17	**98.60**	**98.19**	96.61	96.28
18	97.81	97.58	96.30	95.82
19	97.57	97.15	96.11	95.67
20	97.31	96.69	**99.29**	**99.23**
21	98.49	98.19	99.31	99.11
22	98.23	97.78	99.36	99.16
23	97.39	96.84	99.34	99.17
24	98.42	97.99	99.31	99.10
25	97.96	97.70	99.38	99.13

Table 4.6: Résultats des taux de classifications des signatures des appareils électriques (%) avec le DT, les deux modèles de MLP proposés et l'approche KNN

	DT	MLP		KNN	
		modèle 1, avec 17 neurones	modèle 2, avec 20 neurones	K=1	K=10
Taux de classification (%) en apprentissage	99,67	98.60	99.29	100	84.75
Taux de classification (%) en test	99.09	98.19	99.23	79.30	79.88
Temps total de calcul (s)	15	62.42	260.5	86	73

comparés à ceux des autres méthodes,le DT et le KNN. Les résultats de la classification des signatures des appareils électriques sont représentés par le Tableau 4.6.

La technique DT est la méthode la plus rapide qui permet une meilleure précision par rapport aux deux approches, à savoir MLP et KNN. Avec cette méthode, 1,8% seulement des signatures ne peuvent pas être associées à une classe prédéfinie d'appareils électriques. La DT est une méthode très simple, facile à visualiser et à interpréter. Cependant, un problème de sur-apprentissage peut apparaitre si l'arbre est trop profond [111]. En plus, la DT est moins flexible parce que la classification d'un nouveau appareil se fait en parcourant l'organisation hiérarchique défini lors de la formation. Par ailleurs, un taux d'erreur plus petit n'est pas toujours atteint de manière fiable avec la conception d'un simple classificateur.

4.3.4 Approches ensemblistes pour la classification des appareils électriques avec OAA et ECOC

Dans cette section deux approches ensemblistes seront comparés permettant chacune de classifier les appareils électriques en huit classes. Notre objectif est de déterminer une approche permettant d'améliorer les résultats de la classification obtenus avec des modèles simples qui sont détaillés dans la Section 4.3.3. De ce fait, une première approche ensembliste ECOC (Error-Correcting Output Codes) [112] et une deuxième approche OAA [113] seront testées.

a. Approche ensembliste pour la classification binaire ECOC

Cette approche consiste à utiliser un ensemble de classifieurs à base de MLP. Elle permet ainsi de déterminer une stratégie de combinaison qui regroupe les résultats issus de ces classifieurs. Cette approche est basée sur le principe ECOC (Error-Correcting Output Codes) [112] permettant de définir plusieurs classifieurs, dans notre cas des MLP, afin que chacun se spécialise à la classification d'une classe particulière. Nous pourrions alors transformer le problème de la classification multiple, dans notre cas à huit classes d'appareils, en classification binaire.

Table 4.7: Classifieurs ECOC avec les Code Classe de huit appareils électriques à classifier

Classe	Classifieur			
	f_1	f_2	f_3	f_4
C1	1	1	1	1
C2	1	1	0	1
C3	1	0	1	1
C4	0	1	0	1
C5	1	0	1	0
C6	1	1	0	0
C7	0	0	1	1
C8	0	0	0	0

Cette méthode consiste à créer pour chaque ensemble de classes un modèle qui le sépare des autres [114]. Tout d'abord, il s'agit là d'identifier les 8 classes par des chaines de caractères distinctes qui sont appelées "Code Classe" comportant chacune des 0 et 1. Pour ce faire, il faut déterminer le nombre de classifieurs avec lesquels l'apprentissage et le test ont été réalisés [113].

Soient k classes et l qui correspond à la taille du Code Classe CC qui identifie chaque classe d'appareil. La taille du Code Classe et le nombre de classifieurs devraient assurer quelques propriétés qui sont :

- la taille du Code Classe qui devrait avoir une longueur suffisante afin de pouvoir attribuer des Code Classe distinctes pour chaque classe d'appareil.
- le nombre de classifieurs qui devrait être inférieur au nombre de classes. Le cas contraire, la classification serait de type classe parmi toutes les autres classes (OAA) [113].

Pour codifier les 8 classes nous avons $3 \leq l(CC) < k$ parce que $2^3 = 8$ avec $CC \in \{0, 1\}$. Après avoir testé les valeurs 3, 4, 5, 6 comme taille du Code Classe, nous avons gardé $l(CC)$ avec 4 bits. La taille du Code Classe a conduit à 4 classifieurs qui sont f_1, f_2, f_3 et f_4 représentés dans le Tableau 4.7. Chaque classifieur permet de classifier n classes avec $1 \leq n \leq k$. Chaque classifieur va apprendre une certaine séparation des classes. Le premier classifieur f_1 apprend les classes 1, 2, 3, 5 et 6. Le deuxième classifieur f_2 apprend les classes 1, 2, 4 et 6. Le troisième classifieur f_3 apprend les classes 1, 3, 5 et 7. Le dernier classifieur f_4 apprend les classes 1, 2, 3, 4 et 7.

Nous avons alors implémenté quatre MLP permettant d'identifier n classes. Ces modèles ont été testés en variant le nombre de neurones dans la couche cachée et à la fin nous avons retenu 20, 16, 16 et 6 neurones respectivement pour les classifieurs f_1, f_2, f_3 et f_4. Chaque modèle a une seule sortie qui peut correspondre à une ou plusieurs classes.

Le test d'une nouvelle signature correspond à l'agrégation de résultats de quatre classifieurs et sera par la suite comparé avec les CC avec la distance de Hamming. Cette distance permet de comparer deux chaines de caractères de valeurs binaires. Formellement, cette distance est

définie avec $\hat{c}$ qui est la classe prédite et CC qui est la classe désirée. F est un ensemble de valeurs représentant chaque appareil par une série de quatre valeurs composées de 0 et 1. n_{ap} est le nombre de classes qui vaut 8 classes dans notre cas d'étude.

$$\begin{cases} \forall CC, \hat{c} \in F = \{0,1\} = (CC_i)_{i\in[1,n\_ap]} \\ \hat{c} = (\hat{c}_i)_{i\in[1,n\_ap]} \\ D(c,\hat{c}) = \neq \{i, CC \neq \hat{c}\} = \sum_{i=1}^{n\_ap} (CC_i \ XOR \ \hat{c}_i) \end{cases} \tag{4.8}$$

L'équation (4.8) représente la distance de Hamming qui permet de déterminer le nombre d'éléments de l'ensemble des valeurs de CC qui diffèrent de celle de $\hat{c}$.

Les résultats de cette classification du modèle MLP-ECOC sont comme suit. Le taux de classification des signatures des appareils électriques est 99.74% pour l'entrainement et 99.5% pour le test. La durée moyenne pour la formation et le test est 52.22 s.

b. Approche ensembliste pour la classification binaire OAA

Afin de comparer la classification ECOC, une deuxième approche est proposée qui se base sur la classification d'une seule classe à la fois. De ce fait, huit classifieurs MLP-OAA à sorties binaires sont proposés. Les entrées pour chaque MLP-OAA sont l'ensemble de 16 caractéristiques correspondant à chaque appareil électrique. Chaque MLP-OAA est défini avec 8 sorties qui permet d'identifier qu'une classe à la fois. Ces 8 modèles ont été évaluées 10 fois et les résultats retenus sont les moyennes. Tout d'abord, un choix de la configuration de chaque MLP-OAA en terme du nombre de neurones dans la couche cachée a été fait. Le Tableau 4.8 représente les choix de la configuration finale de 8 modèles MLP-OAA. Le choix de ce nombre est basé sur les taux de la classification de l'ensemble d'entrainement et de test. La classe $\hat{c}$ d'une nouvelle signature est obtenue à partir des sorties des 8 MLP-OAA avec l'équation suivante :

$$\hat{c} = argmax(\text{MLP-OAA}_i)_{i\in[1,8]} \tag{4.9}$$

Un compromis est recherché afin de choisir la configuration optimale en nombre de neurones dans la couche cachée par l'application de l'algorithme 4.

Selon les résultats obtenus et représentés dans le Tableau 4.8, les modèles MLP-ECOC ont été utilisés pour la classification du moniteur, l'ordinateur et la télévision qui correspondent aux C1, C2 et C4 ces appareils sont les plus difficiles à classifier par rapport au reste.

Toutefois, nous retenons que les appareils C3 et C8 sont les plus faciles à classifier. Ceci a été constaté à partir des taux de classification élevés avec l'ensemble d'entrainement et de test.

Les résultats de classification avec le modèle MLP-OAA donnent un taux de classification des signatures des appareils électriques de 99.91% pour l'entrainement et de 99.9% pour le test. Le programme de classification a durée environ 300s.

Table 4.8: Nombre optimal de neurones nécessaires et résultats de classification des appareils électriques pour chaque modèle MLP-OAA à part

	C1	C2	C3	C4	C5	C6	C7	C8
Nombre de neurones dans la couche cachée	20	20	3	19	21	20	11	2
Taux de classification en apprentissage (%)	99.81	99.75	100	99.78	100	100	100	100
Taux de classification en test (%)	99.72	99.65	100	99.73	100	100	100	100
Durée moyenne (seconde)	21.41	19,04	1.8	51.25	22.4	5.35	16.16	1.67

Les modèles MLP-OAA utilisés pour la classification des appareils électriques ont montré leurs performances par rapport au MLP du modèle 1 et 2 aussi devant le DT et le KNN. En comparant les résultats obtenus avec l'approche MLP-OAA, nous avons obtenu 99.91% el comparant par rapport à la classification avec la DT qui représente respectivement 99.67%. Le modèle MLP-OAA a dépassé aussi la performance de classification en le comparant avec le MLP modèle 2 qui a fourni un taux de classification de 99.29%. Il est à noter que les classifieurs individuels comme la DT, le MLP et le KNN permettent de résoudre le problème de la classification multi-classe et de procéder directement à la classification. Par contre, ces méthodes ne garantissent pas souvent la précision en classification. Pour remédier à ce problème, l'approche MLP-ECOC permet d'améliorer la classification et de simplifier la complexité par rapport à l'approche MLP-OAA. Soit $O(C)$ la complexité du MLP, la complexité de l'approche MLP-OAA est $8 \times O(C)$. Cependant, la complexité de l'approche MLP-OAA est $4 \times O(C)$.

En comparant nos résultats de classification des signatures de tous les appareils électriques étudiés avec d'autres résultats fournis par différents travaux de recherche [96, 109], nous pouvons prouver les performances de notre approche. À titre de comparaison, le taux de classification du moniteur est de 98,70% réalisé avec le modèle SVM à noyau RBF et de 76,86% avec un MLP binaire réalisée dans [96] qui reste inférieur à notre résultat qui vaut 99.8%. De même, les résultats de la classification du ventilateur ainsi que le réfrigérateur sont respectivement 61.5% et 88.8% [109] qui reste également inférieur à notre résultat qui atteint 100%.

À partir de la classification MLP-OAA, nous avons remarqué qu'il existe des appareils plus difficile à classifier que les autres. De ce fait, nous proposons, en se basant sur les algorithmes d'apprentissage en profondeur (deep learning) [115], d'analyser les caractéristiques des appareils électriques dans le but de réduire le nombre des entrées aux modèles de classification.

4.3.5 Réduction des caractéristiques des appareils électriques avec ML

Attribuer un score d'importance pour chaque caractéristique peut jouer un rôle important pour fournir un aperçu des données, un aperçu du modèle et la base de la réduction de la dimension.

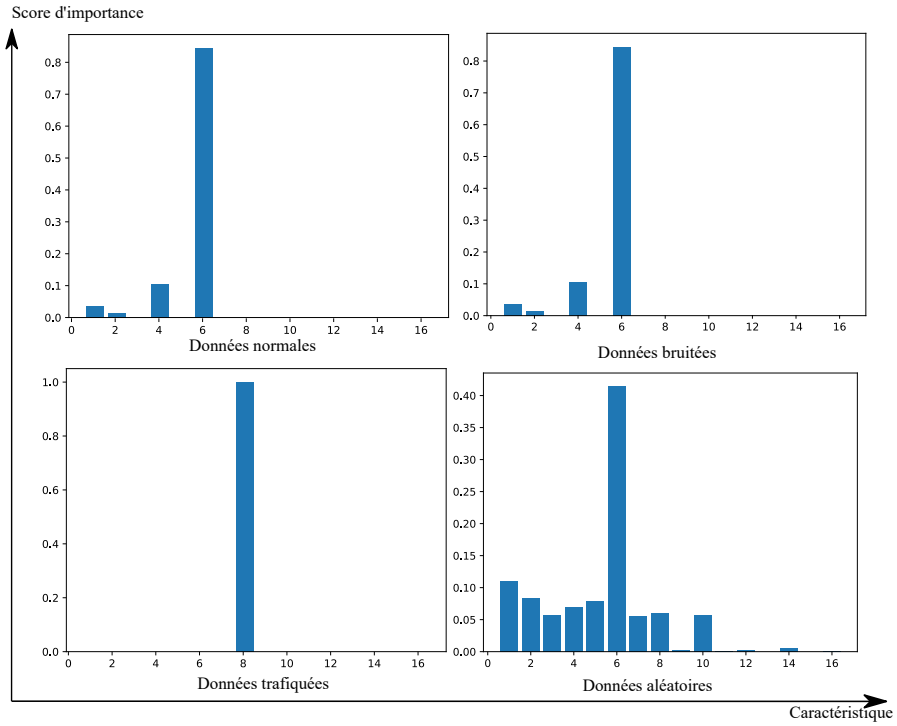

Figure 4.14: Attribution des scores d'importances aux caractéristiques des appareils électriques avec la méthode des arbres à gradient boostés

Les scores peuvent mettre en évidence les caractéristiques les plus pertinentes pour faire la distinction d'un appareil par rapport à l'autre, et inversement, les caractéristiques les moins pertinentes. Généralement, cette phase peut être interprétée par un expert du domaine et peut être utilisée comme base pour collecter des données plus nombreuses ou différentes. Ceci permet de diminuer le nombre des entrées aux modèles de classification et avoir un modèle moins complexe. Dans ce cas, nous exploitons les performances de la technique des arbres de décision pour identifier les scores d'importance de chaque caractéristique.

Dans notre contexte, un nouveau modèle représenté par des arbres de décision DT qui apprennent les données de telle sorte que plus une caractéristique est utilisée pour prendre des décisions clés, plus son importance est élevée. Cette importance est calculée explicitement pour chaque caractéristique de l'ensemble des données, ce qui permet de classer et de comparer les attributs les uns aux autres. Elle est calculée selon le principe suivant : l'importance de chaque caractéristique est calculée pour un arbre de décision unique par le nombre de fois que chaque point de partage des caractéristiques améliore la mesure de performance, pondérée par le nombre d'observations dont le nœud est responsable. Les importances des caractéristiques sont ensuite moyennées sur tous les arbres de décision. L'attribution des scores d'importance de chaque caractéristique est basée sur un modèle ensembliste de combinaison de plusieurs DT qui est l'ensemble des arbres à gradient boostés [116]. Le boosting de gradient est une technique d'apprentissage automatique qui consiste à générer chaque arbre de régression. Avec

une fonction de perte prédéfinie, chaque DT mesure l'erreur à chaque étape et la corrige à la suivante. La bibliothèque XGBoost de scikit-learn avec ses paramètres par défaut (Python) est utilisée.

Nous avons procédé à une analyse sur chaque sous-ensemble des données à part qui sont : les données normales, les données bruitées, les données trafiquées et les données aléatoires. Chaque sous ensemble est exploré afin d'extraire des caractéristiques principales d'affectation à une classe d'appareils spécifique. Cette affectation est basée sur le modèle MLP-ECOC proposé. Les résultats de l'application de ce modèle sont donnés dans la Fig. 4.14.

L'application de DT boosting de gradient a attribué des scores d'importance à chaque caractéristique des appareils électriques. À partir de ces scores nous n'avons sélectionné que les caractéristiques avec un score supérieur à 0.005 et permettant de sélectionner un nombre de caractéristiques selon le sous-ensemble des données :

- Données normales : Une sélection de 4 caractéristiques parmi les 16 initialisées au départ (1, 2, 4, 6). La classification de tous les appareils électriques n'a pas dépassé une durée d'une seconde avec un taux de classification de 100%.
- Données bruitées : Une sélection identique de la sélection des données normales de 4 caractéristiques parmi les 16 initialisées au départ (1, 2, 4, 6). La classification des tous les appareils électriques n'a pas dépassé une durée d'une seconde avec un taux de classification de 100%.
- Données trafiquées : Nous retenons que la huitième caractéristique identifiée automatiquement. La classification de tous les appareils électriques n'a pas dépassé une durée d'une seconde avec un taux de classification de 100%.
- Données aléatoires: Une sélection de dix premières caractéristiques est réalisée. La classification de tous les appareils électriques avec un nombre de caractéristique plus petit a montré son impacte positif sur les taux et la durée de classification des données des signatures des appareils électriques. Les résultats sont représentés dans le Tableau 4.9.

Pour la classification des appareils électriques, nous avons comparé trois modèles ML différents : une DT, un MLP avec deux configurations et un KNN avec deux configurations. Les résultats de la classification ont montré que le réseau de neurones avec sa deuxième configuration, modèle 2, a fourni les résultats les plus performants. Cependant, ce modèle nécessite le temps d'exécution le plus long. Pour cette raison, nous avons proposé deux modèles de classification ensemblistes MLP-ECOC et MLP-OAA. Ces modèles ont permis de transformer la classification multi classe, dans notre cas huit classes, en une classification binaire. Ces deux modèles ont permis d'améliorer les résultats de classification de l'ensemble d'entrainement et de test.

Une autre méthode est appliquée pour explorer et analyser les 16 caractéristiques de chaque

Table 4.9: Classification automatique des appareils électriques avec le modèle MLP-ECOC avant et après l'extraction des caractéristiques les plus importantes

	16 caractéristiques				10 caractéristiques			
	nombre neurones	**Train (%)**	**Test (%)**	**Durée (s)**	**nombre neurones**	**Train (%)**	**Test (%)**	**Durée (s)**
C1	17	99,3	99,05	0,5	2	99,84	99,75	0,38
C2	18	98,84	98,86	1,36	17	98,95	98,9	1,31
C3	2	100	100	0,4	2	100	100	0,2
C4	21	98,8	98,9	1,36	21	99,25	98,9	0,67
C5	7	99,96	99,9	0,94	6	100	100	0,94
C6	13	100	100	0,99	9	100	100	0,98
C7	10	99,97	100	0,9	2	100	100	0,3
C8	2	100	100	0,39	2	100	100	0,39

ensemble des appareils électriques. Cette méthode consiste à réduire le nombre de caractéristiques par l'attribution d'un score d'importance. Cette analyse a permis d'extraire un nombre de caractéristiques sans avoir son ordre harmonique ou sa nature électrotechnique. Elle a offert une classification parfaite de 100% des appareils électriques pendant un temps réduit de moins d'une seconde pour trois sous-ensembles de données qui sont la base normale, trafiquée et aléatoire. En outre, la classification des données bruitées difficiles à classifier a été amélioré avec une réduction des caractéristiques par rapport aux données initiales.

4.4 Conclusion

Dans ce chapitre, nous avons abordé la problématique de la classification non supervisée des Cdc de la consommation d'eau et la classification supervisée des appareils électriques.

Une approche probabiliste a été proposée pour découvrir automatiquement les profils de consommation d'eau journalière des utilisateurs. Cette approche a permis d'explorer automatiquement l'évolution d'un profil de consommation particulier avant la récupération complète des données sur la période considérée. Avec cette étude, nous pouvons classifier plus rapidement les Cdc journalières, en l'occurrence avant la fin d'une journée. À partir du clustering réalisée précédemment, une classification supervisée est lancée en continue et en temps réel avec MLP qui est un réseau de neurones artificiel. Le clustering et la classification supervisée des Cdc d'eau ont permis de détecter des changements d'habitudes de consommation à l'échelle d'un consommateur. Ils ont également permis de détecter au plutôt des classes de consommation "inhabituelles" dues par exemple à des fuites d'eau ou des évènements occasionnels (fêtes, canicules, etc).

Pour l'analyse des consommations d'électricité, des approches supervisées et non supervisées ont été mises en œuvre afin de séparer les différents appareils à partir du courant de charge agrégé. Cette approche non intrusive, qui a été complètement implémentée avec des algorithmes ML,

est capable d'identifier et de classifier les signatures électriques de chaque type d'équipement dans des conditions de fonctionnement réelles. Ainsi, huit différents appareils électriques ont pu être identifiés et séparés. D'autres méthodes ont également été testées pour classifier des appareils électriques : un MLP, un arbre de décision DT de type CART été mis en place ainsi que la méthode des k-voisins les plus proches KNN. Les résultats montrent que la DT a donné les meilleurs résultats avec les meilleures performances. Cependant, un problème de sur-apprentissage peut apparaitre si l'arbre est trop profond et les données sont plus nombreux. Pour palier à ce problème, nous avons proposé deux approches ensemblistes qui consistent à utiliser un ensemble de classificateurs optimaux. Celle-ci est basée sur une combinaison qui regroupe les résultats issus des MLP et les principes de ECOC et OAA. L'étude et la comparaison des méthodes classiques telles que la DT, KNN, MLP, a permis de comprendre les qualités respectives des modèles ensemblistes de type ECOC et OAA. L'approche ECOC a montré ses performances par rapport aux MLP simples, le DT et le KNN. Cependant, l'approche OAA dépasse toutes autres classifications en terme de performance. Toutefois, nous constatons que la classification avec MLP-OAA est plus complexe que celle de MLP-ECOC avec $O(k)$ qui peut être coûteuse lorsque le nombre de classes est important.

Les modèles MLP-OAA nécessitent un temps de calcul plus long et présentent une importante complexité. Par conséquent, nous avons proposé de simplifier les calculs par une diminution des caractéristiques des appareils électriques. Ceci a consisté en une analyse et une extraction des caractéristiques. À partir de 16 entrées initialisées pour chaque modèle de classification, une analyse avec un algorithme de gradient boosting basé sur un ensemble des arbres de décision a été proposé pour attribuer des scores d'importance à chaque caractéristique. Cette méthode a été testée sur quatre ensembles des données dont chacun se caractérise par un aspect différent : bruitées, aléatoires, normales et trafiquées. Ainsi, la classification avec un nombre plus réduit a amélioré les performances de classification en taux et en durée.

5 Détection des anomalies en temps réel et prévision des courbes de charge

5.1 Introduction

Les fuites d'eau sont considérées comme des consommations anormales de cette ressource qui engendrent parfois des énormes pertes et des dégâts qui peuvent même endommager la structure du bâtiment. Le défi majeur ici est de détecter instantanément cette anomalie de consommation. Les nouveaux compteurs d'eau intelligents offrent de nouvelles opportunités de surveillance et de contrôle à distance des consommations et des fuites. Ce chapitre présente une solution complète non intrusive à ces défis. Elle s'appuie sur des Cdc et des mesures extraites effectuées sur une partie unique et centralisée d'un réseau de distribution AMI. Nous abordons cette problématique de détection des fuites d'eau au sein d'un bâtiment connecté à partir d'un seul point de mesure. Nous proposons une démarche de détection des fuites d'eau en temps réel qui a été expérimentée et validée au sein d'un bâtiment tertiaire éducatif. Cette démarche prend en charge simultanément la phase de pré-traitement des données de la consommation d'eau pour la prévision future.
Afin de bien mener une prévision, la détection des anomalies de consommation est considérée parmi les étapes les pus importantes du pré-traitement des données. Cela revient à avoir une base de données qui reflète au mieux une consommation quotidienne sans anomalies. Ces dernières peuvent être dues à des fuites d'eau ou une interruption lors de la collecte des données de la consommation. Dans ce cadre, nous serons emmenés à comparer des modèles de prévision individuels à court terme des Cdc de la consommation d'eau et d'électricité. Ces modèles correspondent à la décomposition déterministe en composantes essentielles de séries temporelles, le modèle SARIMA et les modèles de réseaux de neurones ; MLP et LSTM.

5.2 Détection des anomalies liées à la consommation d'eau

Dans ce contexte, nous proposons une détection des fuites d'eau en temps réel. L'étude sera réalisée sur des ensembles des données réelles fournies par des compteurs installés dans différents bâtiments tertiaire éducatif. Sachant que les données à haute résolution de la consommation

d'eau sont disponibles presque toutes les minutes, des algorithmes plus sophistiqués peuvent être développés pour identifier les fuites éventuelles en temps réel. Nous proposons une approche qui sera testée sur un ensemble des données collectées sur une période de cinq mois. Avec le clustering des Cdc, réalisé dans la Section 4.2, nous avons pu construire une base de référence de la consommation d'eau habituelle qui reflète au mieux le vrai comportement des utilisateurs.

5.2.1 Paramètres de l'algorithme de détection des fuites d'eau

Cet algorithme permettra de détecter en temps réel les anomalies liées à la consommation d'eau en l'occurrence les fuites d'eau. Cet algorithme est basé sur les paramètres suivants :

a. Cdc journalière maximale "C_{max}"

La courbe $C_{max} = max(C_{Per})$ va servir à la détection de grandes fuites. À tout moment de la journée, elle représente un seuil maximum de consommation d'eau. Cette détection est effectuée avec une procédure qui compare l'indice instantané $y_j(t_i)$ de $C_j(t_i)$ de chaque minute à $max(y_{Per}(t_i))$ de C_{max} qui représente le seuil.

b. Débit nocturne minimum "MNF"

Le MNF va servir à la détection des fuites nocturnes. Dans les analyses des Cdc effectuées précédemment dans 4.2.3 et 4.2.4, nous avons identifié les périodes avec la consommation la plus élevée et la plus fréquente. Les anomalies peuvent avoir lieu à tout moment de la journée et peuvent être produites par un robinet qui coule, un tuyau cassé au niveau de l'installation, une machine défectueuse, etc.

Afin de mieux analyser et détecter les fuites, une autre analyse des Cdc s'impose. Il s'agit là de distinguer entre une fuite qui peut avoir lieu la journée combinée avec une consommation courante quelconque et une fuite qui peut avoir lieu la nuit où la consommation est plus au moins constante et proche de zéro. Par conséquent, nous considérons deux profils de consommation d'eau différents avec une utilisation spécifique :

- Pendant la nuit, le débit moyen d'eau est constant et proche de zéro. Cela est représentée par la Zone 1 dans la Fig. 5.1 ;
- Pendant la journée, le débit moyen est généralement différent de zéro, sauf durant les weekends, jours fériés et vacances. Cela est représentée par la Zone 2 dans la Fig. 5.1.

Cette spécificité est liée au type du bâtiment étudié. Pour une installation privée de type habitation, nous pouvons considérer que la consommation est présente à tout moment.

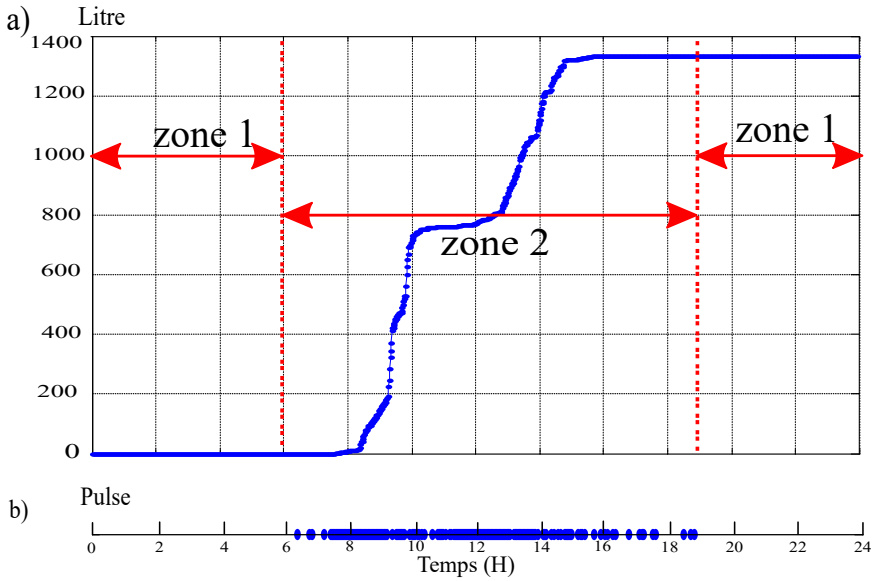

Figure 5.1: a) Division de la Cdc de la consommation d'eau en zones et b) dispersion des évènements datés pendant une journée

Théoriquement, il y a peu de consommation d'eau pendant la Zone 1. Néanmoins, il y a un très faible débit due à des pertes provenant des installations qu'on ne peut pas assimiler à des fuites. Ces pertes peuvent être modélisées par une valeur de référence appelée MNF [45]. Selon notre étude, chaque bâtiment connaît des petites fuites nocturnes qui correspondent aux fuites de fond en raison de problèmes et de dysfonctionnements des robinets et des toilettes [45].

c. Durée "$max(\mathbf{PWNC})_j$"

Cette durée correspond à la période sans consommation nulle (Period Without Null Consumption, PWNC) qui va servir à la détection de petites fuites. Nous définissons le PWNC comme étant une période sans consommation nulle, c'est-à-dire une durée pendant laquelle la consommation d'eau est sans interruption. Nous pouvons avoir plusieurs PWNC le même jour et ces périodes sont différentes d'un jour à l'autre.

Dans le cadre de la détection des fuites d'eau, nous cherchons la période maximale $max(\text{PWNC})_j$ sur une période Per de plusieurs jours. Pour chaque jour, nous calculons $max(\text{PWNC})_j$ comme le montre la Fig. 5.2. Cette valeur représente la durée maximale possible d'une consommation continue durant la journée. Dès que cette durée est dépassée, nous pouvons constater qu'une fuite a eu lieu ou que le comportement de consommation a changé.

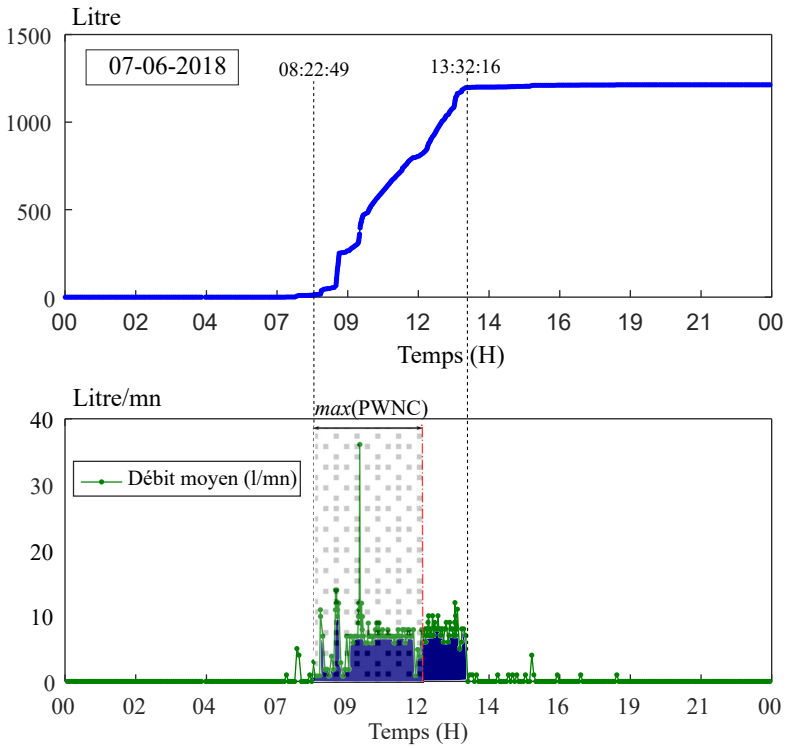

Figure 5.2: Exemple d'une Cdc de la consommation d'eau journalière et identification de la période max(PWNC)

5.2.2 Algorithme de détection des fuites d'eau en temps réel

Nous proposons un algorithme de détection des fuites pour surveiller la consommation d'eau en temps réel. Tous les paramètres précédemment définis sont impliqués dans un script. Ce dernier se produit toutes les minutes, prend en entrée la dernière valeur d'index $y_j(t_i)$ et utilise également un vecteur de paramètres définis par :

$$X = \{H,\ Start\_h,\ End\_h,\ max(y_{Per}(t_i)),\ max(\text{PWNC}),\ MNF\}$$

Dans X, $max(y_{Per}(t_i))$, max(PWNC) et le MNF, H, $Start\_h$ et End_h, sont respectivement l'instant courant exprimé en minutes, l'instant de début de la Zone 1 et l'instant de fin de la Zone 1. La procédure de détection des fuites est récapitulée par l'organigramme de la Fig. 5.3.

En résumé, nous proposons avec cette approche un algorithme de détection des fuites en temps réel. Les grosses fuites sont détectées à partir de la Cdc de consommation journalière maximale C_{Per}^{max} et les petites fuites en analysant le débit avec le MNF et max(PWNC). Cette approche sera testée avec des données réelles.

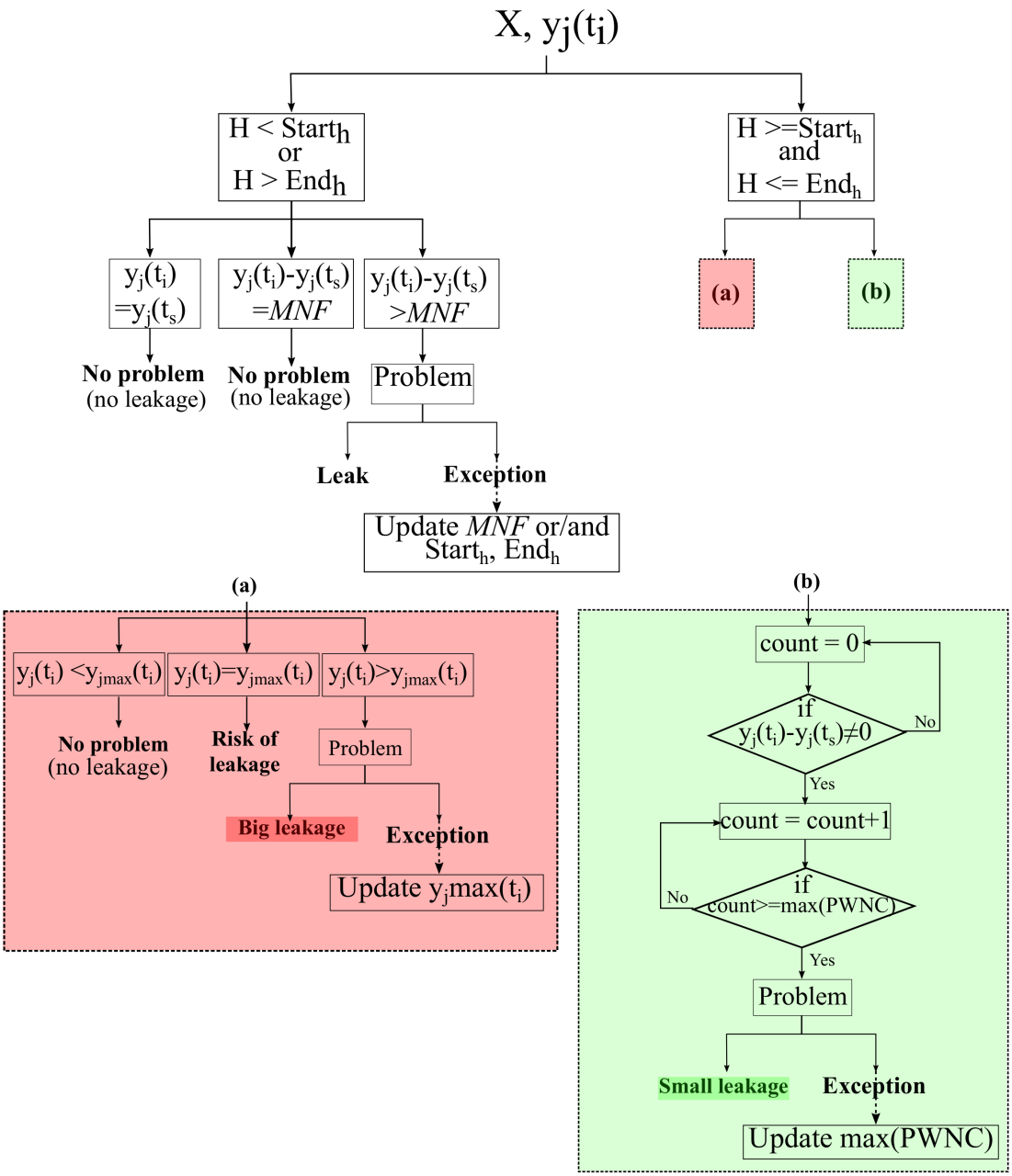

Figure 5.3: Organigramme de l'algorithme proposé pour la détection de fuite d'eau en temps réel

5.2.3 Expérimentation et résultats

Les performances de la procédure de détection des fuites ont été évaluées sur un exemple réel. Les consommations d'eau sont collectées en temps réel avec la plateforme et un compteur d'eau installés dans un RU. Ce bâtiment a une consommation moyenne d'environ 900 l / jour et ne propose que des déjeuners à midi (pas le soir). Les données ont été enregistrées sur une période de plusieurs mois, c'est-à-dire Per = 01/12/2017 au 08/04/2018, et servent à calculer les valeurs du vecteur de paramètres X définie avec la base de données "BaseT1-Détection" décrite dans la Section 3.6. La procédure utilise ces données échantillonnées avec une période d'une minute.

a. Données d'entrée et prétraitement

Le vecteur de paramètres X est déterminé à partir des données collectées. Cependant, certaines données doivent être supprimées car elles correspondent à des weekends, des jours fériés, des vacances, et à des consommations anormales passées. Cette suppression a été opérée à partir de l'analyse des Cdc et des clustering.

b. Choix du contexte et des paramètres

Plusieurs consommations typiquement quotidiennes sont représentées sur la Fig. 3.11 : C_{max}^{avg} à C_{min}^{avg} et C_{Per}^{avg}. On peut constater que les Cdc journalières de la Zone 2 contiennent deux pics. Le premier se situe entre 8h et 10h et correspond au lavage des légumes et la préparation des déjeuners. Le second a eu lieu de 12h à 14h et correspond à la consommation des lave-vaisselles et au nettoyage du RU. Après avoir analysé les Cdc, nous avons choisi $Start\_h$ = 19h et End_h= 6h pour définir la Zone 1 et la Zone 2.

Nous avons également remarqué que le bâtiment présente des pertes d'eau qui ne dépassent pas 8 l/ jour sans aucune activité. Cela est dû aux pertes dans les installations et équipements qu'on ne peut pas assimiler à des fuites franches. À partir de là, nous avons choisi MNF = 0,33 l/h.

Le débit d'eau a été analysé au cours de la période considérée Per. Le PWNC maximum qui a été extrait est de 3 heures et une minute. Nous avons donc choisi max(PWNC) = 210 minutes afin de prendre en compte un intervalle de confiance de 30 minutes. La Fig. 5.2 représente la courbe de débit F_j avec plusieurs exemples de PWNC. Cette figure présente un PWNC qui a dépassé le seuil prédéfini. Ce seuil a été dépassé d'environ une heure puis la consommation a repris normalement. Dans ce cas, nous pouvons constater qu'une petite fuite d'eau a eu lieu, qui correspond à un robinet ouvert pendant un certain temps.

c. Analyse des résultats de la procédure de détection des fuites

La procédure de détection des fuites d'eau a été configurée avec les valeurs de paramètres précédents et programmée de sorte qu'à chaque minute un test de fuite est répété.

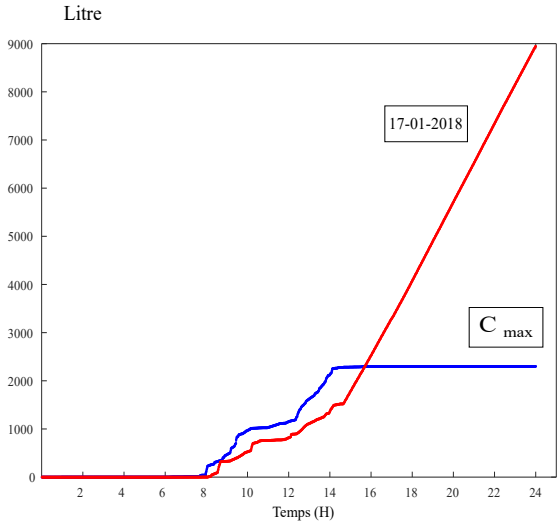

Figure 5.4: Détection d'une grosse fuite avec la Cdc maximale

Le 17-01-2018, l'index $y_j(t_i)$ a atteint le seuil maximum car pendant cette journée, à l'heure t_i avec $i = 942$, à 15h 42mn ($y_j(t_{942}) = 2294$ l) > ($max(y_{Per}(t_{942}) = 2293$ l). Ceci peut être constaté sur la Fig. 5.4, il s'agit là d'une grosse fuite qui a été détectée avec le test de la Cdc maximale.

Le 21-06-2018, une fuite est due à un problème de robinet a été détectée. Il s'agit d'une petite fuite, difficile à détecter par la courbe maximale comme on peut le voir sur la Fig. 5.5 a). Cependant, max(PWNC) a été dépassé car une consommation permanente est apparue entre 1h 43mn et 17h 13mn lorsque l'alerte est déclenchée. Ceci peut être constaté sur la Fig. 5.5 b).

Plusieurs fuites d'eau ont été détectées après l'implantation de la procédure proposée. Afin d'avertir de ces anomalies, des mails et SMS ont été envoyés aux services compétents pour corriger la fuite le plus rapidement possible.

5.3 Modèles utilisés pour la prévision horaire des Cdc de consommation

La prévision de la consommation que ce soit d'eau ou d'électricité ne nécessite pas une connaissance détaillée sur le bâtiment étudié comme par exemple sa surface, le nombre de personnes qui y habitent, leur niveau social, etc. Cependant, la prévision à réaliser dans notre cas d'étude se base essentiellement que sur les données de consommation antérieures et sur des algorithmes ML et approches mathématiques.

Dans ce contexte, nous proposons quatre modèles prédictifs individuels qui correspondent au modèle déterministe [52], SARIMA [117], le réseau des neurones récurrents avec mémoire à court terme (LSTM) [70] et le réseau des neurones MLP [118]. Le choix et l'intérêt d'utiliser ces quatre modèles est basé sur les différences qu'ils ont d'analyser, explorer, modéliser et prédire

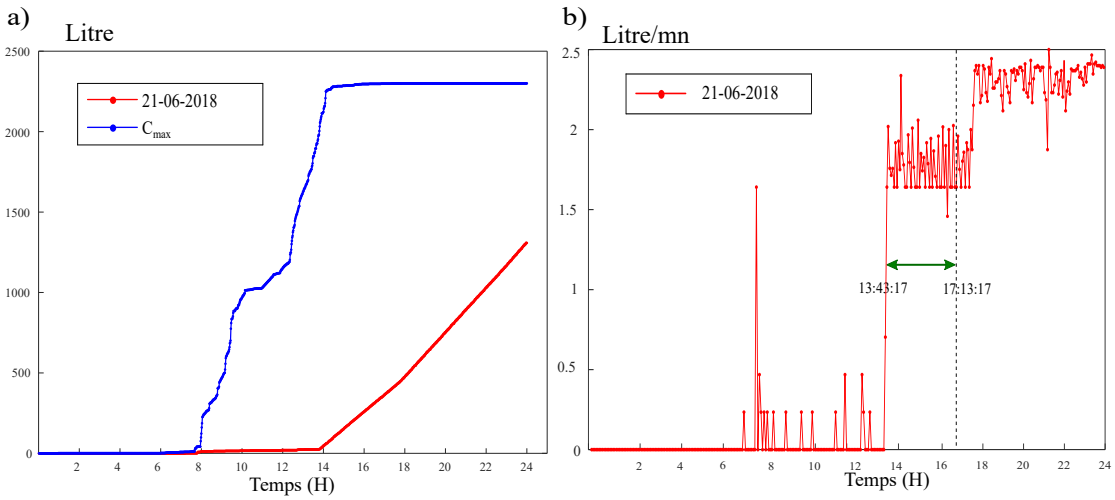

Figure 5.5: Détection de petites fuites: a) comparaison entre la Cdc maximale et une Cdc journalière, b) le débit d'eau journalier associé

les données. Cette différence est due aux aspects déterministe, stochastique, linéaire et non linaire de l'analyse et la modélisation.

Notre objectif, par la suite, consiste à identifier le modèle qui va assurer des prévisions des Cdc de consommations en termes horaires d'une installation tertiaire ainsi que privée avec la plus grande précision. Les modèles prédictifs seront emmenés à prédire le nombre de litres d'eau ou le nombre de Wh de l'énergie électrique à consommer pendant la prochaine heure. À partir des modèles dits "individuels", nous pouvons obtenir 11 différentes combinaisons qui sont représentés dans la Fig. 5.6 et qui correspondent à :

- 6 différentes combinaisons de 2 modèles :
  - **hybride 1** qui est la moyenne du déterministe et le SARIMA.
  - **hybride 2** qui est la moyenne du SARIMA et le LSTM.
  - **hybride 3** qui est la moyenne du SARIMA et le MLP.
  - **hybride 4** qui est la moyenne du MLP et le LSTM.
  - **hybride 5** qui est la moyenne du déterministe et le LSTM.
  - **hybride 6** qui est la moyenne du déterministe et le MLP.
- 4 différentes combinaisons de 3 modèles :
  - **hybride 7** qui est la moyenne du déterministe, le MLP et le LSTM.
  - **hybride 8** qui est la moyenne du SARIMA, le MLP et le LSTM.
  - **hybride 9** qui est la moyenne du déterministe, le SARIMA et le MLP.
  - **hybride 10** qui est la moyenne du déterministe, le SARIMA et le LSTM.

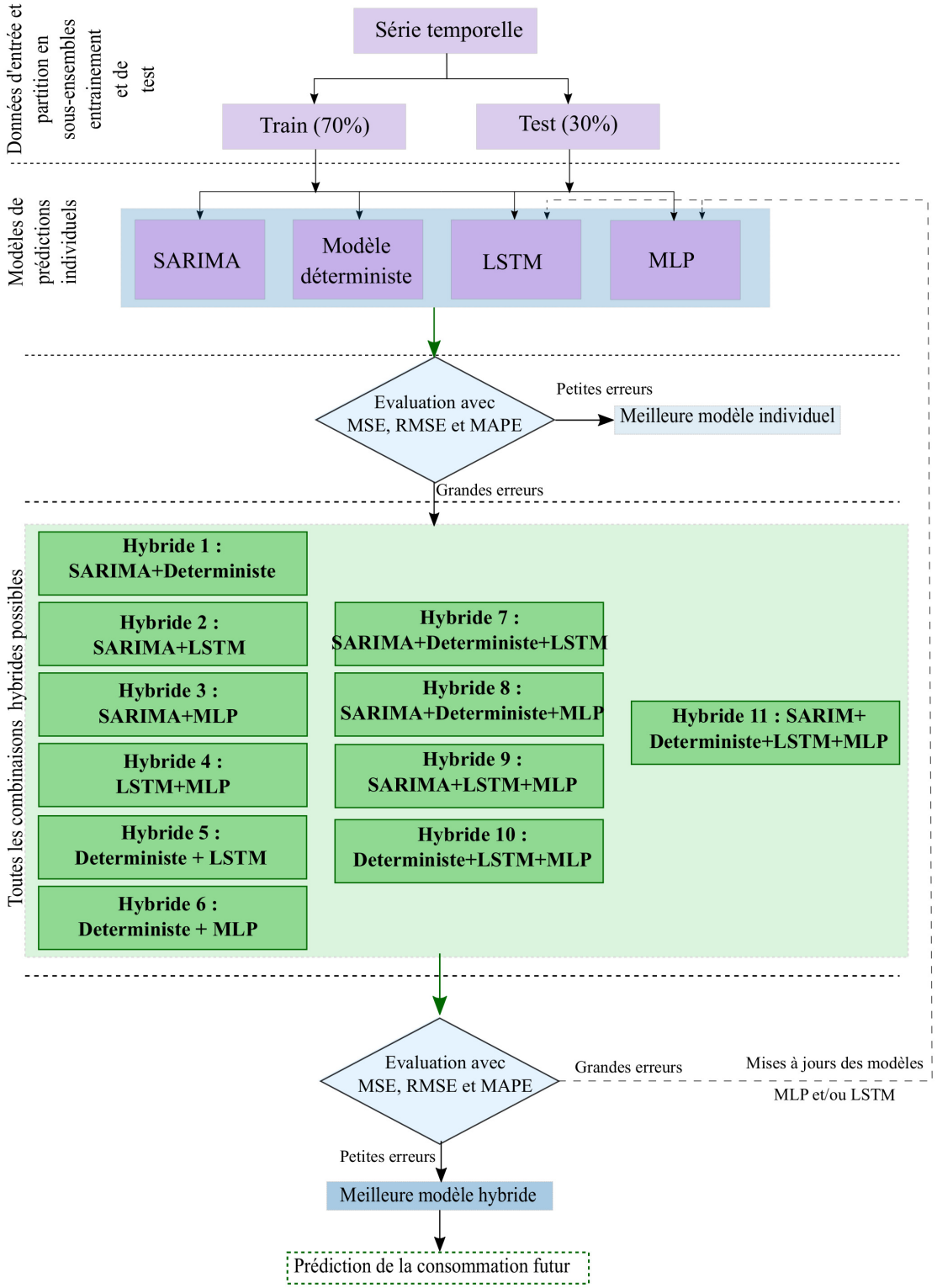

Figure 5.6: Les modèles hybrides proposés à partir des modèles dits "individuels" pour la prévision des séries temporelle

- Une seule combinaison de 4 modèles :
  - **hybride 11** qui est la moyenne de déterministe, le SARIMA, le MLP et le LSTM.

5.4 Prévision des Cdc de la consommation horaire d'eau d'un bâtiment tertiaire

5.4.1 Préparation des données

Nous avons choisi une base de données couvrant environ une année universitaire pour la prévision de la consommation d'eau du 03 Septembre 2018 jusqu'à 30 Juin 2019 avec la base "BaseT1-prévision" dans la Section 3.6.

Afin de construire des Cdc représentant parfaitement la consommation quotidienne d'eau, il est nécessaire d'identifier les consommations anormales (telles que les fuites d'eau, les consommations occasionnelles) qui peuvent influencer l'analyse et la prévision de la consommation future d'eau. Une consommation d'eau anormale [39] se manifeste toujours par un comportement inhabituel et occasionnel par rapport à la majorité des Cdc de la consommation d'eau pendant les autres jours. En pratique, le but de cette tâche est de constituer, à partir d'un ensemble des données, un modèle reflétant au mieux la "normalité" pour détecter comme anormale toute autre consommation s'écartant significativement de la majorité [12] des Cdc de la consommation d'eau. Par conséquent, ce filtrage des Cdc est basé sur la classification des Cdc de la consommation d'eau déjà réalisée dans la Section 4.2.

La détermination des données manquantes, représentée par le bloc (A) dans la Fig. 3.14, dépend de la nature du bâtiment. En ce qui concerne la prévision de la consommation d'eau horaire du bâtiment tertiaire, nous avons choisi une interpolation spline cubique pour les jours travaillés (de lundi au vendredi) et une interpolation linéaire pour les weekends et les vacances [97] en appliquant les équations décrites dans la Section 3.5.3.

Vu la spécificité du RU, la consommation d'eau pendant les weekends, les jours fériés et les vacances universitaires n'est pas très importante et prédire sa consommation ne sera pas intéressant. Par conséquent, nous allons exploiter la spécificité de ce bâtiment avec une prévision des jours travaillés. Cette prévision a été précédée par une classification supervisée des Cdc journalières $\bar{C}_j$, avec le regroupement réalisé avec la HAC et la ED élaborée dans la Section 4.2. Cette classification permet d'identifier les Cdc de la consommation quotidienne représentée par la classe 3. La Fig. 5.7 représente les Cdc de la consommation d'eau qui appartiennent à la classe 3 et qui correspondent aux jours travaillés pendant cette période.

Les données d'entrées pour chaque modèle de prévision sont sous forme des Cdc consécutives. Chaque Cdc est représentée par 24 données correspondant à la quantité d'eau consommée pendant chaque heure. Une normalisation, appelée MIN-MAX Scaling [119] est appliquée afin de simplifier le calcul au niveau des modèles ML et de réduire l'espace de variation de chaque

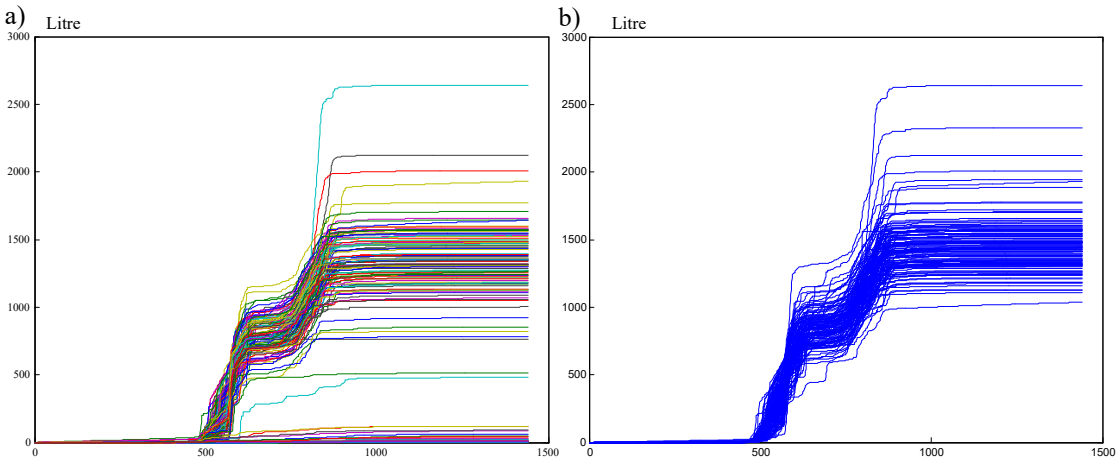

Figure 5.7: a) Cdc de la consommation d'eau journalière des jours d'activité (du lundis au Vendredi) échantillonnées en minute, b) Résultat de la classification supervisée des Cdc de la consommation d'eau journalière des jours d'activité (du lundis au Vendredi) : classe 3

donnée. La dé-normalisation est l'inverse de la normalisation qui consiste à transformer les données normalisées en ses valeurs initiales. La normalisation ainsi que la dé-normalisation sont décrites respectivement par les équations (5.1) et (5.2) comme suit :

$$x_{normaliser} = \frac{x - x_{minimal}}{x_{maximale} - x_{minimale}}, \tag{5.1}$$

$$x = x_{normaliser} \times (x_{maximale} - x_{minimale}) + x_{minimale}, \tag{5.2}$$

où x est un élément de la série, $x_{maximale}$ et $x_{minimale}$ sont respectivement les valeurs maximale et minimale de l'ensemble des données des Cdc.

Après avoir préparer les Cdc d'entrées aux différents modèles, une première étape consiste à diviser les données en trois sous ensembles pour l'entrainement, la validation et le test. Nous avons utilisé 70% des données comme ensemble d'entrainement et 30% des données pour l'ensemble de test. Pour valider notre modèle, nous avons utilisé 30% des données d'entrainement.

5.4.2 Paramètres des modèles prédictifs

a. Modélisation de Cdc de la consommation d'eau avec le modèle déterministe

La méthode de Buys et Ballot, décrite dans la Section 2.5.1, nous a permis de construire la droite des moindres carrés passant par les moyennes et les écarts types de chaque période de Cdc. La période est identifiée avec l'analyse des corrélations qui est 24 comme elle est présentée dans la Fig. 5.8. La pente de cette droite, qui est égale à 2.43, confirme que la nature de la série temporelle, est un modèle multiplicatif. Les figures qui représentent les méthodes de vérification du schéma de décomposition de la Cdc sont présentées dans l'Annexe A.4.

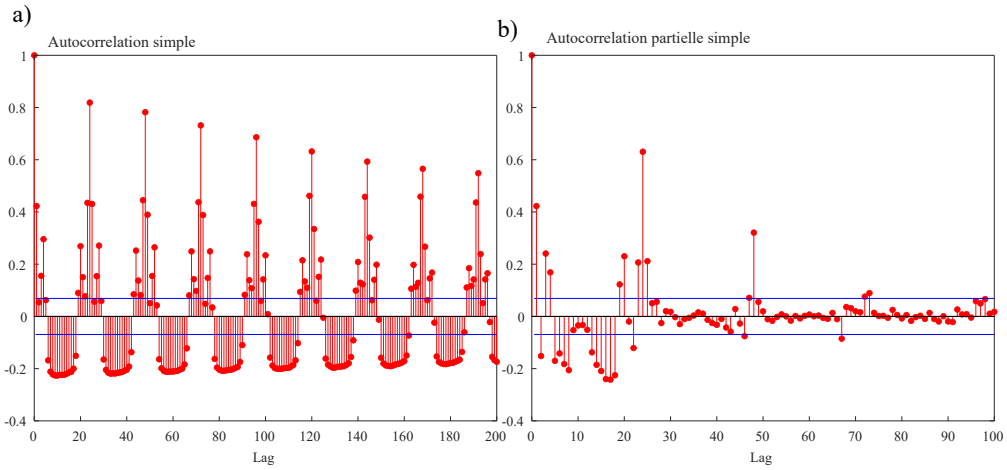

Figure 5.8: a) Auto-corrélation simple et b) partielle de la Cdc de la consommation d'eau

Après avoir déterminé le schéma de décomposition de la série temporelle, qui est multiplicatif, il reste à estimer chacune de ses composantes décrites par l'équation (2.9) qui correspondent à la tendance et la saisonnalité. Les résultats de la fonction de régression sont : $a = -0.002$ et $b = 63.59$. L'estimation du modèle multiplicatif est donné par l'ensemble des équations suivantes (5.3) :

$$X_t = \begin{cases} (63.59 - 0.002t)\hat{c}_1 \ si \ t = 1, 1+24, 1+2\times 24, \ldots \\ (63.59 - 0.002t)\hat{c}_2 \ si \ t = 2, 2+24, 2+2\times 24, \ldots \\ \ldots \\ (63.92 - 0.002t)\hat{c}_{24} \ si \ t = 24, 2+24, 3\times 24, \ldots \end{cases} \tag{5.3}$$

b. Modélisation des Cdc de la consommation d'eau avec le modèle stochastique SARIMA pour la prévision horaire

Les meilleurs paramètres qui minimisent les critères d'informations (AIC et BIC) sont donnés avec le modèle SARIMA $(p = 1,\ d = 0,\ q = 1)(P = 0,\ D = 0,\ Q = 1)_{24}$ et génère l'équation (5.4).

$$(1 - \phi_1 B)(1 - B^{24})X = (1 + \theta_1 B)\epsilon_t. \tag{5.4}$$

c. Prévision des Cdc de la consommation d'eau avec le modèle MLP

Afin de pouvoir prédire avec le modèle MLP, nous devons configurer nos Cdc et transformer la prévision en un problème d'apprentissage supervisé. Cette configuration consiste de transformer les données en plusieurs échantillons des données dont le modèle apprend et se généralise. Chaque échantillon doit avoir à la fois un composant d'entrée et un composant de sortie. Cette décomposition des données est représentée dans le Tableau. 5.1. L'entrée du modèle est un ensemble d'observations antérieures.

Table 5.1: Configuration des données en séquence d'entrée/ sortie pour les méthodes prédictifs ML

Entrée	Sortie
x(t_1), x(t_2), x(t_3), ..., x(t_k)	x(T_{k+1})
x(t_2), x(t_3), x(t_4), ..., x(t_{k+1})	x(T_{k+2})
x(t_3), x(t_4), x(t_5),..., x(t_{k+2})	x(T_{k+3})
...	...

Dans notre cas d'étude, nous avons défini un modèle MLP qui prend 24 données comme entrées. Pour prédire une heure $x(t_{i+1})$, le modèle se base sur les 24 heures précédentes $x([t_{i-24}, .., t_i])$ et pour chaque itération l'ensemble d'observations d'entrée sera décalé d'une composante telle qu'elle est représentée par l'auto-corrélation partielle de la Cdc de la Fig. 5.8. Cette restructuration de Cdc en sous ensemble d'observations est représentée par le Tableau 5.1 avec k =1 à 24.

Le nombre de neurones, le nombre de couches cachées ainsi que les fonctions d'activation (comme Sigmoïde, ReLu, Tanh ...) ont été variés et testées. Le modèle MLP est défini avec une seule couche cachée de 150 neurones et une fonction d'activation Unité Linéaire Rectifiée (ReLu) pour chaque neurone. L'algorithme d'ajustement et d'optimisation est Adam avec un taux d'apprentissage égale 10^{-4}. L'erreur absolue moyenne, MAE, est choisi pour le calcul de la fonction de perte.

d. Modélisation des Cdc de la consommation d'eau avec le LSTM pour la prévision horaire

Le modèle LSTM prend 24 données comme entrées, pour chaque itération l'ensemble d'entrée sera décalé d'une donnée. Le nombre de neurones, le nombre de couches cachées, les fonctions d'activation (comme Sigmoïde, ReLu, Tanh ...) ont été variés et testées. Le modèle LSTM est définit avec trois couches LSTM à 100/150/100 unités consécutives avec une fonction d'activation ReLu et nous avons utilisé l'algorithme Adam pour l'optimisation et la mise à jour du poids du modèle. Nous avons défini comme ajustement du taux d'apprentissage de 10^{-3} de la descente du gradient. Nous avons retenu l'erreur absolue moyenne, MAE, pour le calcul de la fonction de perte. Les résultats ont été obtenu après 150 itérations.

5.4.3 Résultats et discussion

Les résultats des modèles de prévision individuels qui correspondent aux modèle déterministe, SARIMA, MLP et LSTM ont fourni des prévisions avec des erreurs relativement importantes sans avoir bien assuré un bon compromis entre le biais et la variance. Ceci est bien clair à partir des MAPE calculées sur les ensembles de test et représentées par la Fig. 5.10. Les durées

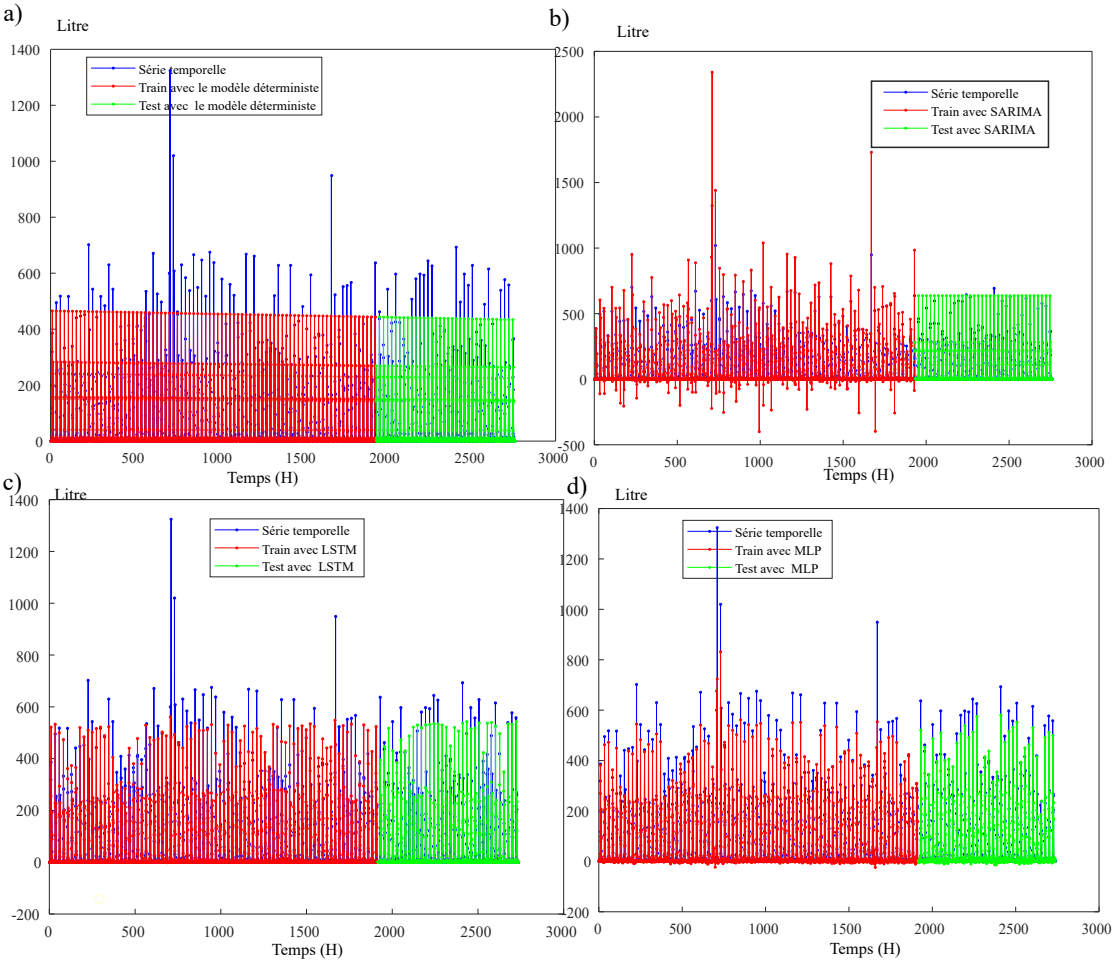

Figure 5.9: Prévision de la consommation d'eau horaire avec le modèle a) déterministe de décomposition en composantes essentielles des séries temporelles, b) SARIMA, c) LSTM et d) MLP d'une période d'une année universitaire

moyennes pour la prévision des Cdc de la consommation d'eau sont respectivement comme suit : 0.18s, 0.36s, 85.24s et 760.34s. Les courbes de prévision de la consommation d'eau horaire avec les quatre modèles prédictifs individuels sont illustrées dans la Fig. 5.9.

Table 5.2: Résultats de la prévision de Cdc de la consommation d'eau horaire d'un bâtiment tertiaire avec les modèles individuels et hybrides

	Entrainement			**Test**			**Modèle**		
	RMSE (l)	**MSE (l^2)**	**MAPE (%)**	**RMSE (l)**	**MSE (l^2)**	**MAPE (%)**	**RMSE (l)**	**MSE (l^2)**	**MAPE (%)**
Modèle déterministe	58.06	3371	0.85	50.23	2523	0.22	55.84	3118	0.58
SARIMA	73.82	5449	0.71	70.48	4967	2.71	72.85	5306	0.39
MLP	35.77	1280	0.68	115.44	13326	3.96	69.98	4897	0.78
LSTM	49.51	2451	0.61	117.26	13750	4.39	76.45	5844	0.69
Hybride 1 : déterministe, SARIMA	28.95	838	0.07	53.9	2905	1.14	38.2	1459	0.09
Hybride 2 : SARIMA, LSTM	30.68	941	0.05	73.92	5464	1.56	47.95	2300	0.15
Hybride 3 : SARIMA, MLP	31.55	996	0.01	72.02	5188	1.09	47.48	2255	0.19
Hybride 4 : MLP, LSTM	44.63	1992	0.77	73.09	5343	0.09	54.76	2999	0.68
Hybride 5 : déterministe, LSTM	51.77	2681	0.73	73.78	5444	0.46	59.25	3511	0.63
Hybride 6 : déterministe, MLP	40.85	1668	0.64	115.64	13374	4.12	72	5183	0.73
Hybride 7 : déterministe, MLP, LSTM	27	729	0.25	61.46	3777	0.55	40.55	1645	0.29
Hybride 8 : SARIMA, MLP, LSTM	24.03	577	0.28	60.47	3657	0.21	38.76	1502	0.32
Hybride 9 : déterministe, SARIMA, MLP	24.14	583	0.19	84.36	7117	1.72	50.45	2545	0.36
Hybride 10 : déterministe, SARIMA , LSTM	45.03	2028	0.71	86.29	7446	1.26	60.46	3655	0.68
Hybride 11 : déterministe, SARIMA, MLP, LSTM	26.63	709	0.36	71.46	5107	0.65	45.05	2030	0.41

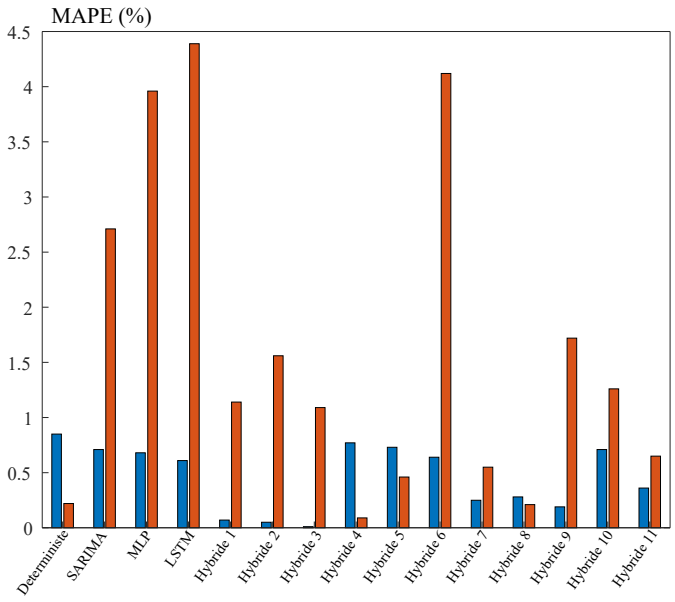

Figure 5.10: L'erreur moyenne en pourcentage absolu MAPE de différents modèles prédictifs sur les données annuelles de la consommation d'eau horaire d'un bâtiment tertiaire éducatif

Les résultats de prévision détaillés sont illustrés dans le Tableau 5.2 en calculant le MSE, RMSE et MAPE. Parmi les quatre modèles individuels, nous pouvons remarquer que le modèle déterministe a fourni la plus petite erreur respectivement sur l'ensemble d'entrainement et de test de 0.85% et 0.22% par rapport à une consommation moyenne de 783.87 l. Cependant plusieurs modèles de combinaison hybride ont pu dépasser les performances des modèles individuels et en particulier le modèle déterministe. À titre d'exemple, le modèle hybride 7 qui est basé sur la combinaison du modèle déterministe, MLP et LSTM correspond à une prévision de 0.25% et 0.55% en termes d'erreur respectivement d'entrainement et de test. Toutefois, le modèle hybride 8 qui est composé par le modèle SARIMA, MLP et LSTM conduit à une prévision de 0.28% et 0.21% en termes d'erreurs respectivement d'entrainement et de test. Par conséquent, nous pouvons confirmer que ces deux modèles hybrides ont pu prédire la consommation et qui ont le pouvoir de généraliser la prédiction au future. D'autres modèles de combinaisons ont permis une légère amélioration par rapport aux modèles individuels comme les modèles hybride 4, hybride 5, hybride 10 et hybride 11.

5.5 Prévision des Cdc de la consommation horaire d'eau d'une installation privée

5.5.1 Analyse et pré-traitement des données

L'analyse et la prévision des données ont été effectuées sur des Cdc, sans données manquantes ni valeurs incohérentes, afin d'obtenir les meilleurs résultats avec l'apprentissage automatique.

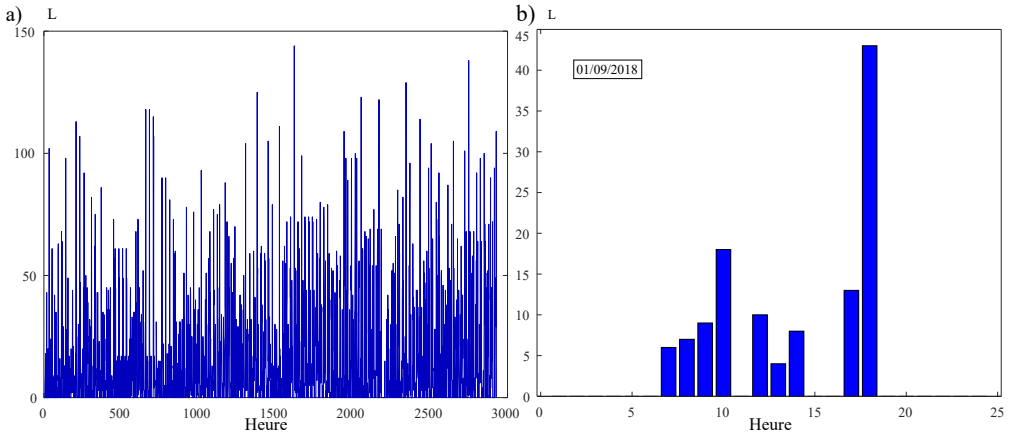

Figure 5.11: Cdc de la consommation d'eau pendant la période du 01 Septembre 2018 au 31 Décembre 2018 et b) Volume d'eau horaire cumulé d'une journée dans une installation privée

Pour corriger et compléter les données manquantes, une recherche d'une courbe de référence standard [12] a été réalisée. Cette étape est représentée par le bloc (A) de la Fig. 3.14.

Afin de procéder à la correction des données manquantes, nous considérons qu'un profil de charge d'un jour *j*, dans une installation privée, est fortement corrélé à celui de la veille *(j-1)*, et le même jour de la semaine précédente *(j-7)* [120].

$$C_j(t_i) = Avg(C_{j-7}(t_i), C_{j-1}(t_i)) \tag{5.5}$$

Notre seule référence est l'ensemble des données récupérées à partir d'un point de mesure unique dans le bâtiment, qui est le compteur d'eau de l'AMI. L'algorithme de détection des données anormales, telles que les anomalies des points définis dans le deuxième chapitre 2.4, qui est basé sur la moyenne (average) et l'écart type (std) des Cdc de l'équation (5.6).

$$An = abs\left[y_j(t) - average(\sum_{i=1}^{n}(y_j(t)))\right] \geq \alpha \times std(\sum_{i=1}^{n}(y_j(t)), \tag{5.6}$$

α est une variable numérique choisie qui est fixé empiriquement et strictement supérieure à 2. Dans notre cas $\alpha = 5$ comme la valeur optimale.

La base des données de consommation d'eau de cette installation privée couvre 3 mois consécutifs, d'Octobre 2018 jusqu'à Décembre 2018 a été nommée "BaseP1-prévision" dans la Section 3.6. Les données sont échantillonnées en heure et sont représentées par la Fig. 5.11.

5.5.2 Paramètres des modèles prédictifs

a. Modélisation des données de la consommation d'eau avec un modèle déterministe pour la prévision horaire

De la même manière que la prévision des Cdc de la consommation d'eau d'un bêtiment tertiaire, le schéma de décomposition de la Cdc est réalisée avec la méthode de Buys et Ballot [60]. La période identifiée avec l'analyse de la corrélation des Cdc est une journée qui correspond à 24h. Le schéma de décomposition de la Cdc de la consommation horaire d'eau d'un bâtiment privé est multiplicatif. Nous estimons maintenant chacune de ses composantes décrites par l'équation (2.9) en se basant sur l'équation (2.12). La courbe de régression est obtenue avec les coefficients suivants : $a = 0.04$ et $b = 8.29$. L'estimation du modèle multiplicatif est donnée par l'ensemble des équations suivantes (5.7) :

$$X_t = \begin{cases} (8.29 + 0.04t)\hat{c}_1 \ si \ t = 1, 1+24, 1+2\times 24, \ldots \\ (8.29 + 0.04t)\hat{c}_2 \ si \ t = 2, 2+24, 2+2\times 24, \ldots \\ \ldots \\ (8.29 + 0.04t)\hat{c}_{24} \ si \ t = 24, 2+24, 3\times 24, \ldots \end{cases} \tag{5.7}$$

b. Modélisation des données de la consommation d'eau avec le modèle stochastique SARIMA pour la prévision horaire

Les meilleurs paramètres du modèle SARIMA qui minimisent les critères d'informations AIC et BIC sont obtenus avec les paramètres suivants $(p = 1, d = 0, q = 2)(P = 0, D = 0, Q = 1)_{24}$ qui génère l'équation (5.8).

$$(1 - \phi_1 B)(1 - B^{24})X = (1 + \theta_1 B + \theta_2 B^2)\epsilon_t. \tag{5.8}$$

c. Modélisation des données de la consommation d'eau avec le modèle MLP pour la prévision horaire

Nous avons défini un modèle MLP qui prend 24 données constituant une journée comme entrées comme la configuration indiquée dans le Tableau 5.1. Pour prédire une heure $x(t_{i+1})$, le modèle se base sur les 24 heures précédentes $x([t_{i-24}, .., t_i])$ et pour chaque itération l'ensemble d'observation d'entrée sera décalé d'une donnée. Le modèle MLP est défini avec cinq couches cachées et chacune à 100 neurones avec une fonction d'activation ReLu pour chaque neurone. L'algorithme Adam est utilisé pour l'optimisation et la mise à jour du poids du modèle. Nous avons pris comme ajustement du taux d'apprentissage10^{-1} de la descente du gradient. Nous avons retenu l'erreur absolue moyenne, MAE, comme critère de performance pour le calcul.

d. Modélisation de données de la consommation d'eau avec le LSTM pour la prévision horaire

Un modèle LSTM est retenu avec les paramètres suivants : il utilise 24 données correspondent à 24 heures de consommation comme entrées et pour chaque itération l'ensemble d'observations d'entrée sera décalé d'une itération à l'autre.

Le modèle LSTM est définit avec une seule couche LSTM à 100 unités consécutives avec une fonction d'activation ReLu. Nous avons utilisé l'algorithme Adam pour l'optimisation et la mise à jour des poids du modèle. Nous avons défini comme taux d'apprentissage 10^{-1} de la descente du gradient. Nous avons retenu l'erreur absolue moyenne, MAE, comme critère de performance pour le calcul.

5.5.3 Résultats et discussion

Les prévisions des Cdc de la consommation d'eau dans un bâtiment privé ont été réalisées avec des modèles individuels qui correspondent au modèle déterministe, SARIMA, MLP et LSTM. Ces modèles ont fourni des prévisions avec des erreurs relativement importantes sans avoir bien assuré un bon compromis entre le biais et la variance. Ceci est bien clair à partir des MAPE sur l'ensemble de test comme indiqué sur la Fig. 5.13. Les durées moyennes pour la prévision des Cdc de la consommation d'eau sont respectivement comme suit : 0.07s, 0.01s, 12.42s et 240.03s. Les Cdc issues de la prévision de la consommation d'eau horaire de ce bâtiment obtenues avec ces quatre modèles prédictifs individuels sont illustrées dans la Fig. 5.12.

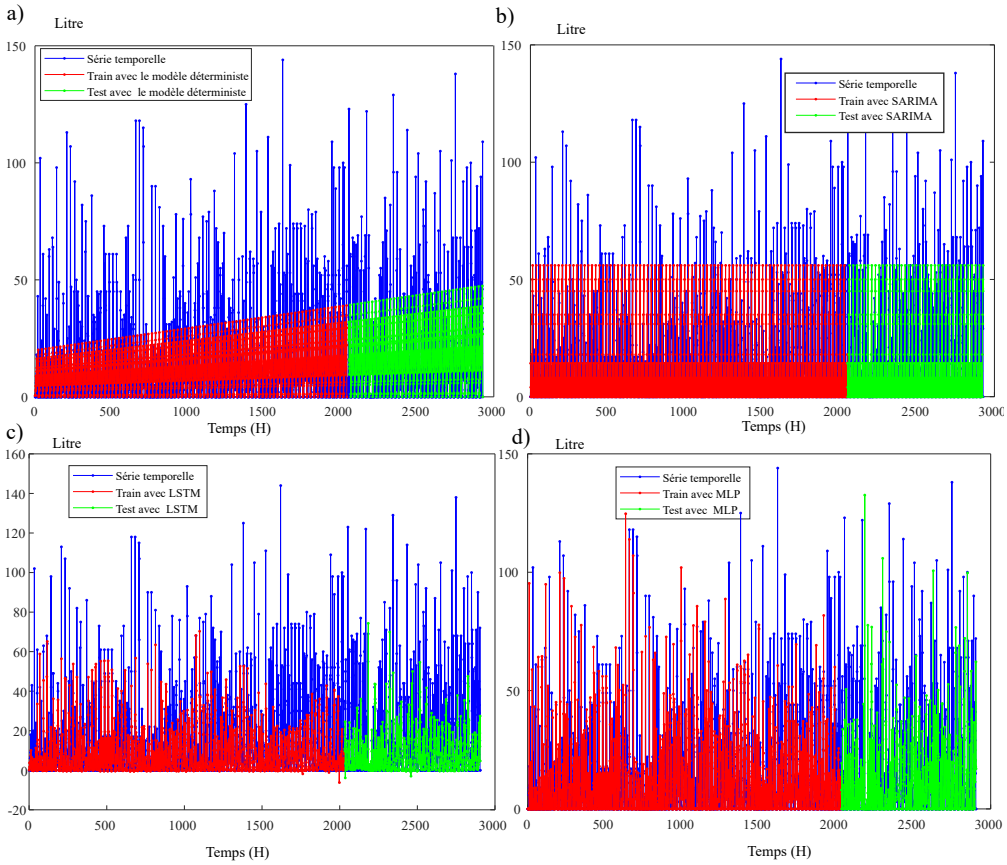

Figure 5.12: Prévision de la consommation d'eau horaire avec le modèle a) déterministe de décomposition en composantes essentielles des séries temporelles, b) SARIMA, c) LSTM et d) MLP

Table 5.3: Résultats de la prévision de la consommation d'eau horaire d'une installation privée avec les modèles individuels et les modèles hybrides

	Entrainement			Test			Modèle		
	RMSE (l)	**MSE (l^2)**	**MAPE (%)**	**RMSE (l)**	**MSE (l^2)**	**MAPE (%)**	**RMSE (l)**	**MSE (l^2)**	**MAPE (%)**
Modèle déterministe	17.322	300	2.775	26.838	720	5.646	20.645	426	2.068
SARIMA	19.113	365	1.72	24.41	596	4.052	20.845	435	0.002
MLP	15.798	250	2.75	23.296	543	0.47	18.373	338	2.171
LSTM	18.69	349	2.427	25.109	630	4.199	20.826	434	1.939
Hybride 1 : déterministe, SARIM	8.563	73	0.527	22.642	513	1.879	14.325	205	1.033
Hybride 2 : SARIMA, LSTM	9.124	83	0.515	22.32	498	0.228	14.416	208	1.085
Hybride 3 : SARIMA, MLP	10.308	106	0.354	22.656	513	1.713	15.115	228	0.969
Hybride 4 : MLP, LSTM	16.404	269	2.601	23.384	547	2.196	18.774	352	2.004
Hybride 5 : déterministe, LSTM	15.685	246	2.762	22.956	527	0.539	18.176	330	2.119
Hybride 6 : déterministe, MLP	16.305	266	2.589	23.481	551	1.366	18.75	352	2.055
Hybride 7 : déterministe, MLP, LSTM	8.79	77	1.268	21.839	477	0.681	14.045	197	1.412
Hybride 8 : SARIMA, MLP, LSTM	9.31	87	1.161	21.971	483	0.421	14.338	206	1.335
Hybride 9 : déterministe, SARIMA, MLP	9.76	95	1.153	22.317	498	0.145	14.703	216	1.369
Hybride 10 : déterministe, SARIMA, LSTM	15.723	247	2.651	22.626	512	0.074	18.074	327	2.059
Hybride 11 : déterministe, SARIMA, MLP, LSTM	10.145	103	1.558	21.821	476	0.843	14.662	215	1.544

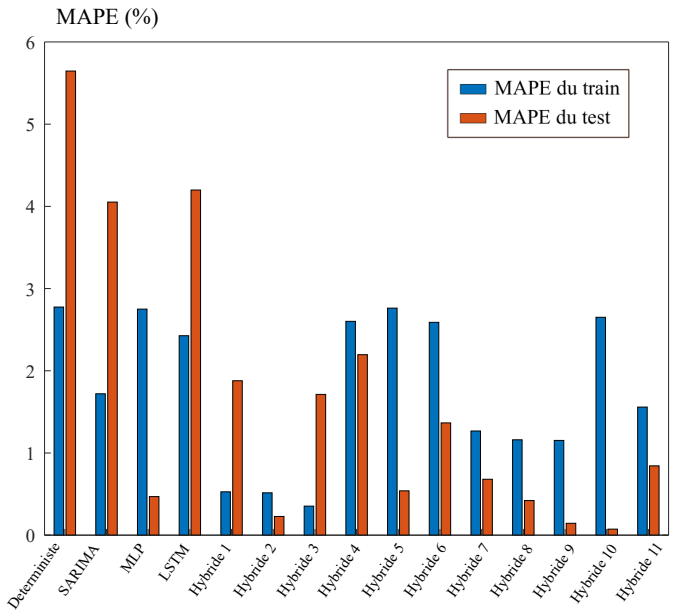

Figure 5.13: L'erreur moyenne en pourcentage absolu MAPE de différents modèles prédictifs de la consommation d'eau horaire d'un bâtiment résidentiel privé

Les résultats détaillés de la prévision sont illustrés dans le Tableau 5.3 en calculant le MSE, RMSE et MAPE. Parmi les quatre modèles individuels, nous pouvons remarquer que le modèle SARIMA a fourni la plus petite erreur de 1.72% sur l'ensemble d'entrainement. Cependant, le modèle MLP a fourni la plus petite erreur de 0.47% sur l'ensemble de test par rapport à une consommation moyenne de 187.1 l. Cependant, le modèle hybride 2 qui se compose du modèle SARIMA et LSTM a fourni les plus petites erreurs de prévision respectivement de 0.515% et 0.228% sur l'ensemble d'entrainement et l'ensemble de test.

5.6 Prévision de la Cdc de la consommation horaire d'électricité

5.6.1 Description et préparation des données

La consommation électrique représente la puissance électrique en Wh consommée pendant 7 semaines dans un bâtiment tertiaire. Nous avons choisi ces semaines de la base de données "BaseT2-prévision" qui a été définie dans la Section 3.6. La sélection des semaines est basée sur la complétude des données. Le pré-traitement des données de consommation d'électricité est nécessaire pour éliminer le bruit du processus de collecte des données, de déclenchement ou de désactivation des appareils électriques. Nous avons pré-traité les données brutes de la puissance avec un lissage de moyenne mobile calculée sur chaque sous-ensemble de F valeurs consécutives de la Cdc constituées d'une fenêtre glissante qui correspond à 60 données. Le

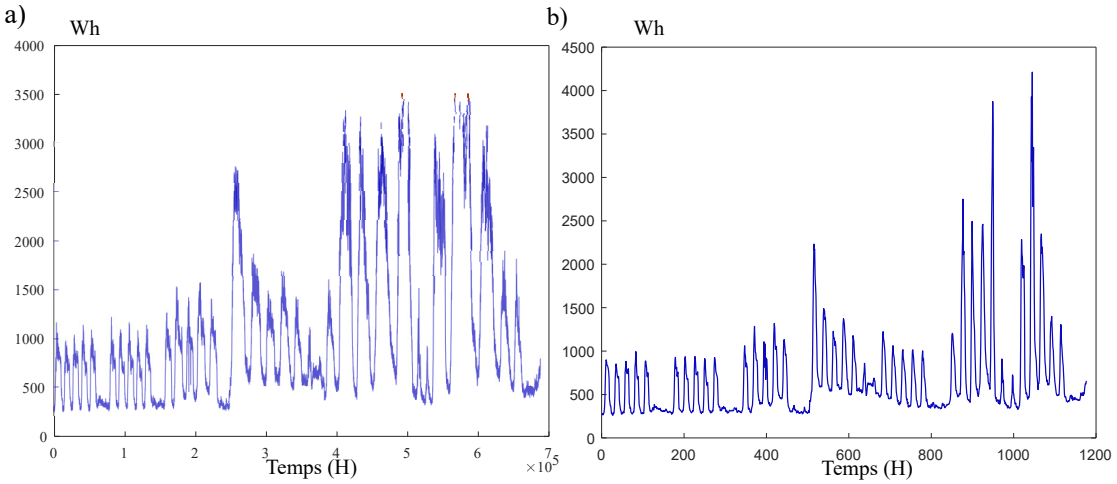

Figure 5.14: a) La Cdc de charge de la puissance électrique brute , b) Cdc de charge de la puissance électrique échantillonnée par heure pendant 7 semaines d'une installation publique

lissage est obtenu par l'équation (5.9) :

$$\bar{x}_n = \frac{1}{F}\sum_{k=0}^{F-1} x_n - k \tag{5.9}$$

avec $F = 60$.

La Fig. 5.14 représente la Cdc de la puissance électrique brute constituée de plus de 600 000 données associée à sa Cdc de la puissance lissée et échantillonnée en heure.

5.6.2 Paramètres des modèles prédictifs

a. Modélisation de données de la puissance électrique avec le modèle déterministe pour la prévision horaire

La modélisation des données de la consommation électrique avec le modèle déterministe est réalisée sur un ensemble d'entrainement de 70% des données. Afin de pouvoir former les modèles prédictifs une analyse de la corrélation des données de la Cdc a été opérée. Nous avons constaté que la série représente une composante saisonnière d'une semaine (24×7 jours) comme représentée dans la Fig. 5.15.

À l'instar de la prédiction de la consommation d'eau, nous avons suivi les mêmes étapes afin de déterminer le schéma de décomposition mathématique des Cdc de la consommation électrique. Nous avons obtenu un schéma de décomposition multiplicatif et nous avons estimé chacune de ses composantes. La courbe de régression a comme coefficients : $a = 0.26$ et $b = 460.7$.

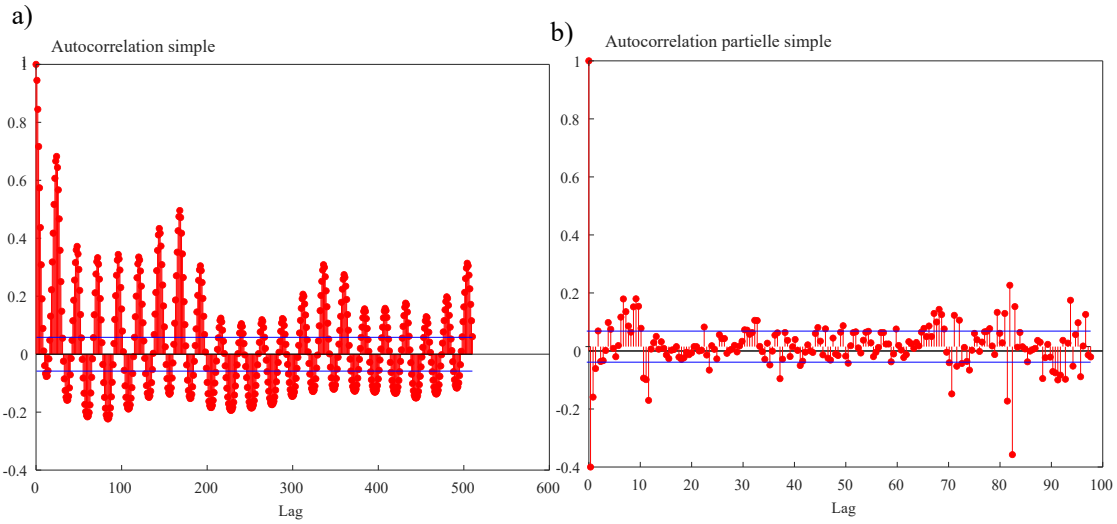

Figure 5.15: a) Auto corrélation simple et b) partielle des Cdc de la puissance électrique

L'estimation du modèle multiplicatif est donnée par l'ensemble des équations suivantes (5.10) :

$$X_t = \begin{cases} (460.7 + 0.26t)\hat{c}_1 \ si \ t = 1, 1+168, 1+2\times 168, \ldots \\ (460.7 + 0.26t)\hat{c}_2 \ si \ t = 2, 2+168, 2+2\times 168, \ldots \\ \ldots \\ (460.7 + 0.26t)\hat{c}_{168} \ si \ t = 168, 2+168, 3\times 168, \ldots \end{cases} \tag{5.10}$$

Afin de prédire les valeurs futures de la Cdc de la consommation horaire de la puissance électrique, nous allons utiliser l'estimation de la tendance et celle de la composante saisonnière.

b. Modélisation des données de la puissance électrique avec le modèle stochastique SARIMA pour la prévision horaire

Les meilleurs paramètres qui minimisent les critères d'informations AIC et BIC sont : $p = 5$, $q = d = 0$, $D = 1$ et $P = Q = 0$. De ce fait, notre modèle est obtenu avec SARIMA $(5,0,0)(0,1,0)_{168}$ et génère l'équation (5.11) :

$$(1 - \phi_1 B - \ldots - \phi_5 B^5)(1 - B^{168})X = \epsilon_t. \tag{5.11}$$

c. Modélisation des données de la puissance électrique avec le modèle MLP pour la prévision horaire

À l'instar du modèle MLP pour la prédiction de la consommation horaire d'eau, nous avons validé un MLP pour l'électricité qui prend 24 données comme entrées pour chaque itération. Il est définit avec deux couches cachées avec 100 neurones dans chacune et avec une fonction d'activation ReLu pour chaque neurone. L'algorithme Adam est utilisé pour l'optimisation et

nous avons pris comme valeur du taux d'apprentissage10^{-4} de la descente du gradient pour la mise à jour des poids du modèle MLP. Nous avons retenu l'erreur absolue moyenne, MAE, comme critère de performance.

d. Modélisation des données de la consommation de la puissance électrique avec le LSTM pour la prévision horaire

Un modèle LSTM est retenu avec les paramètres suivants. Il prend 24 données comme entrées et pour chaque itération l'ensemble d'observation d'entrée sera décalé d'une donnée. Le modèle LSTM est défini avec trois couches à 100,150 et 150 unités respectivement dans chacune et avec une fonction d'activation ReLu. L'algorithme d'ajustement et d'optimisation est Adam et un taux d'apprentissage égal 10^{-3} de la descente du gradient pour la mise à jour des poids du modèle LSTM. L'erreur absolue moyenne, MAE, est employée comme critère de performance.

Les prévisions de la puissance électrique horaire avec les quatre modèles prédictifs sont illustrées dans la Fig. 5.16. Les résultats des modèles de prévision déterministe, SARIMA, MLP et LSTM ont fourni des prévisions avec des erreurs pratiquement importantes en les comparant avec celles calculées sur l'ensemble d'entrainement et de test. De ce fait, nous proposons de combiner les modèles individuels dont le but est de minimiser l'erreur et de trouver un modèle optimal pour la prévision de la consommation électrique du bâtiment tertiaire éducatif sur une base de données couvrant quelques semaines.

5.6.3 Résultats et discussion

Les résultats des modèles de prévision des Cdc de la puissance électrique horaire sont obtenus avec des modèles individuels qui correspondent au modèle déterministe, SARIMA, MLP et LSTM. Ainsi, les durées moyennes pour la prévision des Cdc de la puissance moyenne électrique sont respectivement 0.11s, 0.12s, 18.43s et 99.74s. Ces modèles ont fourni des prévisions avec des erreurs relativement importantes sans avoir bien assurer un judicieux compromis entre le biais et la variance. Ceci est bien clair à partir des MAPE obtenus qui sont représentées sur la Fig. 5.17. Les Cdc de la consommation d'électricité issues de la prévision horaire de ce bâtiment avec ces quatre modèles prédictifs individuels et de combinaison sont représentés dans la Fig. 5.16.

Les résultats de prévision détaillés sont illustrés dans le Tableau 5.4. Des modèles de combinaison hybride ont pu dépassé les performances des modèles individuels. À titre d'exemple, le modèle hybride 9 qui se compose du modèle déterministe, MLP et SARIMA fournit des erreurs de 0.56% et 1.55% en terme d'erreurs respectivement d'entrainement et de test. Toutefois, le modèle hybride 10 qui est composé du modèle déterministe, LSTM et SARIMA correspond à une erreur de prévision de 0.74% et 1.33% en termes d'erreur respectivement en entrainement et en test. Ces deux modèles hybrides ont pu dépassé les performances des modèles individuels tels que le MLP avec des taux d'erreurs respectives de 0.94% et 2.62% en entrainement et en test. Toutefois,

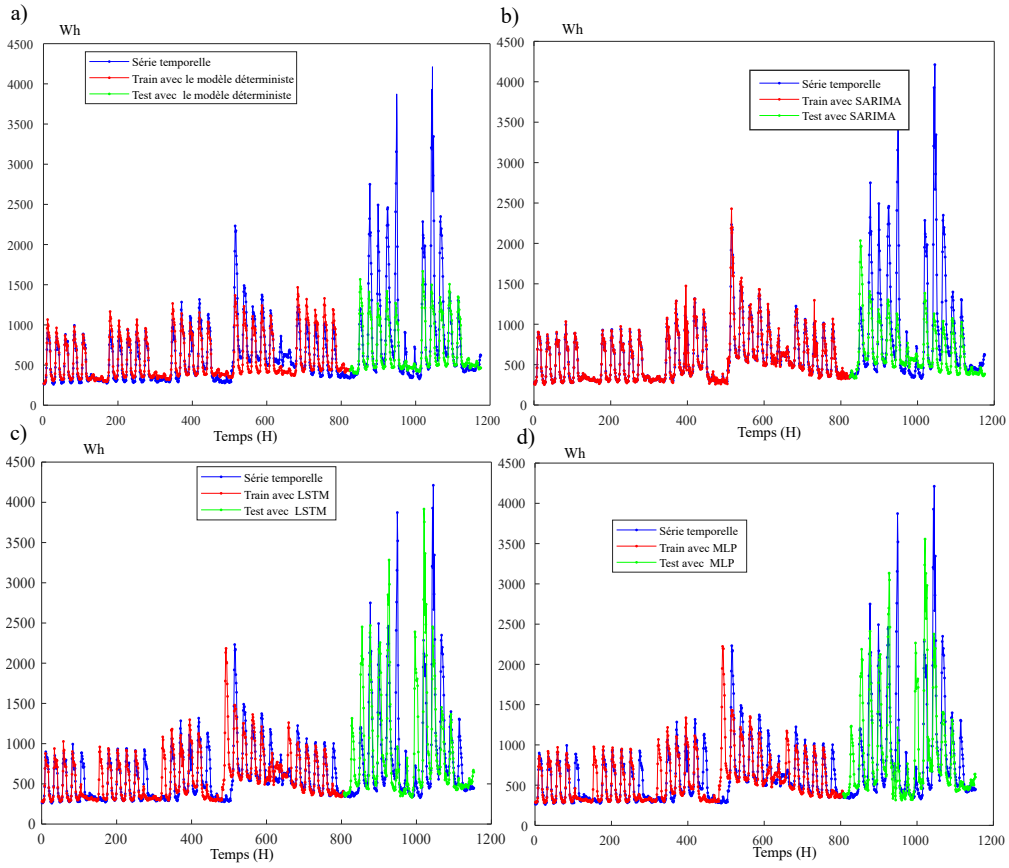

Figure 5.16: Prévision de la puissance électrique horaire avec le modèle a) déterministe de décomposition en composantes essentielles des séries temporelles, b) SARIMA, c) LSTM et d) MLP

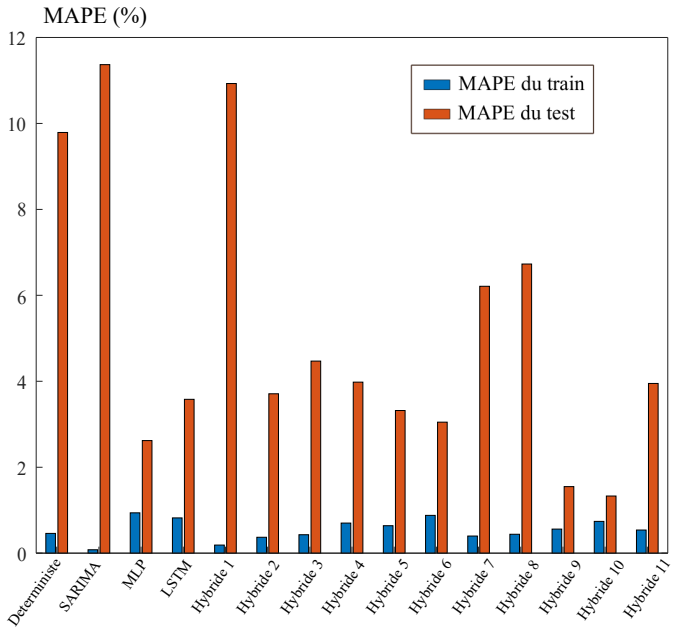

Figure 5.17: L'erreur moyenne en pourcentage absolu MAPE de différents modèles prédictifs de la puissance électrique horaire

ces modèles hybrides ont dépassé en terme de performances de prévision le modèle LSTM avec des taux d'erreurs respectives de 0.82% et 3.58% en entrainement et en test.

Table 5.4: Résultats de la prévision de la puissance électrique avec les modèles individuels et les modèles hybrides

	Entrainement			Test			Modèle		
	RMSE (Wh)	**MSE (Wh^2)**	**MAPE (%)**	**RMSE (Wh)**	**MSE (Wh^2)**	**MAPE (%)**	**RMSE (Wh)**	**MSE (Wh^2)**	**MAPE (%)**
Modèle déterministe	135.29	18302	0.46	501.88	251886	9.79	297.39	88442	1.89
SARIMA	47.38	2244	0.08	579.61	335950	11.37	320.13	102480	2.05
MLP	244.86	59955	0.94	559.19	312689	2.62	368.51	135803	1.05
LSTM	238.87	57060	0.82	582.23	338987	3.58	376.4	141678	0.93
Hybride 1 : déterministe, SARIM	67.18	4513	0.19	536.73	288080	10.93	299.46	89677	1.97
Hybride 2 : SARIMA, LSTM	120.31	14474	0.37	506.65	256692	3.71	295.29	87194	1.49
Hybride 3 : SARIMA, MLP	123.4	15227	0.43	501.01	251013	4.47	293.28	86014	1.55
Hybride 4 : MLP, LSTM	152.96	23396	0.7	464.39	215659	3.98	284.79	81105	1.47
Hybride 5 : déterministe, LSTM	148.57	22074	0.64	469.55	220480	3.32	285.71	81631	1.41
Hybride 6 : déterministe, MLP	241.03	58095	0.88	567.87	322472	3.05	370.73	137442	0.99
Hybride 7 : déterministe, MLP, LSTM	98.11	9625	0.4	486.96	237129	6.21	279.16	77932	1.63
Hybride 8 : SARIMA, MLP, LSTM	101.13	10228	0.44	485.17	235388	6.73	278.98	77830	1.67
Hybride 9 : déterministe,SARIMA, MLP	160.52	25765	0.56	509.17	259258	1.55	309.6	95854	1.35
Hybride 10 : déterministe,SARIMA, LSTM	176.04	30992	0.74	486.57	236752	1.33	304.55	92749	1.29
Hybride 11 : déterministe, SARIMA, MLP, LSTM	131.32	17245	0.54	483.4	233678	3.95	286.74	82218	1.48

5.7 Conclusion

Dans ce chapitre, un algorithme empirique a été développé pour la détection des fuites tout en faisant la distinction entre les petites et les grandes fuites d'eau. La mise en œuvre de cet algorithme au sein des infrastructures de comptage intelligent a été réalisé sans aucun investissement matériel. Il a permis de générer un signal d'alarme pour la notification d'une éventuelle fuite d'eau et éviter au plus tôt des dégâts qui peuvent être lourd de conséquence. En effet, l'algorithme de détection des fuites est basé sur le débit minimum nocturne (MNF) et la période maximale sans consommation nulle (PWNC). Les paramètres de la MNF et de la PWNC ont été adaptés et calculés à partir de la consommation relevée dans un bâtiment. Une Cdc quotidienne maximale cumulée a également été définie comme référence pour chaque instant et permet de détecter les fuites importantes en cas de dépassement.

Pour la prévision des Cdc de la consommation d'eau et d'électricité, plusieurs modèles prédictifs ont été proposé : le modèle déterministe, le modèle SARIMA, le réseau de neurones multi-couches MLP et le réseau de neurones récurrent avec mémoire LSTM. À partir de ces modèles individuels, différentes combinaisons ont été proposés pour former des stacking permettant de réduire les erreurs de prévision des modèles individuels.

La prévision horaire des Cdc constitue une prévision classique des séries temporelles échantillonnées au cours du temps. Elle consiste à prévoir le comportement de la consommation futur. Cela nécessite des Cdc parfaitement échantillonnées par rapport au temps selon une précision imposée (horaire, journalière, hebdomadaire ...). Nous avons constaté que pour chaque base de données, des modèles de combinaisons hybrides ont été adaptés. Nous pouvons toutefois retenir que la prévision des Cdc de la consommation d'eau horaire de deux bâtiments a été bien améliorée avec les modèles de combinaisons hybrides 7 et 8 pour la consommation du bâtiment tertiaire. Ces combinaisons sont basées sur la moyenne de prévision de deux modèles non linéaires avec un modèle stochastique ou déterministe. Toutefois, le modèle de combinaison hybride 2 a pu prédire parfaitement la consommation d'eau horaire du bâtiment privé. Ce modèle est basé sur la moyenne de prévision de deux modèles qui sont le SARIMA et le LSTM. Cependant, nous avons pu constater que la prévision de la puissance moyenne de la consommation électrique a été améliorée avec les deux modèles hybrides 9 et 10 qui constituent une combinaison de trois modèles différents : linéaire, non linéaire et stochastique.
Á partir de la plateforme de collecte des données de consommation proposée, nous avons à notre disposition des données de très grande précision. De ce fait, une exploitation de ces données dans le chapitre qui suit sera opérée pour une prévision des séries temporelles en termes d'événements datés d'eau et d'électricité.

6 Prévision des séries temporelles de consommation

6.1 Introduction

La prévision de Cdc de consommations était élaborée dans le chapitre précédent avec une précision horaire. Une prévision plus accrue peut générer des résultats aberrants et un blocage au niveau des modèles de prévision à cause de l'importance de la masse des données. En prenant par exemple des Cdc de consommations avec un échantillonnage en minutes et une saisonnalité d'une semaine, le modèle déterministe et le modèle SARIMA comportent 10 080 données pour former une saisonnalité. Ensuite, l'échantillonnage et le pré-traitement des Cdc peuvent modifier les valeurs des données récupérées à partir du serveur par le fait d'ajouter ou de supprimer des données. Pour remédier à ces problèmes, les séries temporelles des évènements de consommations représentent la solution pour une prévision en terme de seconde en utilisant sa forme brute qui est la plus proche du comportement réel de la consommation.

Ce chapitre traite la prévision des séries temporelles de consommations d'eau et d'électricité en utilisant la représentation des Cdc sous formes de séries temporelles des évènements. Ces derniers ne nécessitent pas un pré-traitement ou un échantillonnage au préalable avant sa modélisation pour une prévision future. De ce fait, nous proposons une approche de prévision des séries temporelles de consommations. Il s'agit là d'une prévision des dates possibles d'occurrence de chaque événement de consommation. Les modèles de combinaisons hybrides seront également testés afin d'identifier les modèles des prédictions les plus performants tout en assurant le compromis biais/variance. Une comparaison sera proposée entre la prévision des Cdc qui a été élaborée dans le chapitre précédent et les séries temporelles de consommation.

6.2 Structure des séries temporelles d'évènements datés

Une série temporelle X_t, dite aussi chronologique est constituée d'une succession de valeurs numériques. Chaque valeur est une observation étiquetée par un instant temporel t, avec $t = 1, 2, 3, ...n$. Le nombre n est la longueur de la série et l'étiquetage temporel peut être en

minute, heure, jour, ou en ordre naturel d'apparition. Prédire une valeur future revient à chercher $X_t + h$ avec $h = 1, 2, 3,$ etc.

La série temporelle est un ensemble d'évènements successivement ordonnés dans le temps. Elle est implicitement échantillonnée dans le temps selon l'ordre d'apparition des consommations [89]. Une série temporelle échantillonnée en fonction du temps peut correspondre à une Cdc. Dans ce cas la prévision peut être élaborée directement sans procéder à un pré-traitement ou un échantillonnage.

Dans ce qui suit, les séries temporelles de consommations d'eau et d'électricité à prédire se composent de données naturellement échantillonnées directement collectées des compteurs respectifs. Chaque donnée est une durée en milliseconde (ou en seconde) qui correspond à l'écart temporel δ, qui est représenté dans la Fig 3.4, entre deux événements de consommation successifs. Dans ce cadre, une prévision d'un évènement revient à prédire $\delta(t + h)$.

Pour faciliter la tâche de la modélisation mathématique, nous considérons que la série temporelle des écarts des évènements datés est un phénomène qui évolue au cours du temps avec une saisonnalité graphiquement visible dans le cas des Cdc échantillonnées. Cette saisonnalité est généralement identifiée à partir de la fonction d'auto-corrélation et celle d'auto-corrélation partielle. Dans le cas où cette saisonnalité est difficilement déterminée à partir des séries temporelles avec les fonctions d'auto-corrélations, elle sera choisie de sorte qu'elle couvre un jour de consommation complet. Cela veut dire que la saisonnalité reproduite par cette série temporelle représente le volume d'eau ou la quantité d'énergie journalière consommée.

6.3 Prévision de la série temporelle de la consommation d'eau d'une installation privée

6.3.1 Données d'entrée pour les modèles prédictifs

Pour une prévision événementielle, nous avons sélectionné une base de données composée de 2000 données représentant chacune l'écart temporel de 2001 litres consommés dans une installation privée. Cette base de données est nommée "BaseP1-évènement-prévision" dans la Section 3.6. Notre base d'évènements datés comporte les 2000 premières données de la Fig. 6.1. Cette figure correspond à la série temporelle des évènements de consommation de chaque litre d'eau consommé en fonction des écarts temporels. Les données sont divisées en trois ensembles, d'entrainement, de validation et de test. Ces ensembles représentent un pourcentage par rapport au nombre total des données : 70% pour l'entrainement y compris 30% pour la validation et 30% pour le test des modèles.

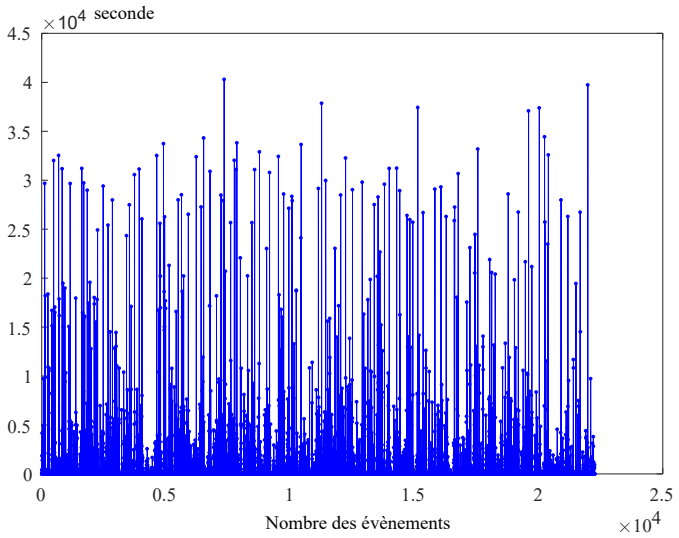

Figure 6.1: Série temporelle des évènements de consommation des litres d'eau consommés du 01/09/2018 07:44:58.250 jusqu'à 31/12/2018 23:54:01.438

6.3.2 Paramètres des modèles prédictifs

Avec la série temporelle de la consommation d'eau, il parait difficile d'identifier la saisonnalité avec la fonction d'auto-corrélation et la fonction d'auto-corrélation partielle. Cette difficulté est due aux données de natures indépendantes les unes des autres. Pour cette raison, nous avons sélectionné le volume journalier maximal sur l'ensemble des jours $max(y_j(n_j))$ comme saisonnalité. Avec cette hypothèse, une décomposition de la série temporelle en composantes essentielles et une modélisation avec le modèle SARIMA sont devenues possibles.

a. Modélisation de la série temporelle avec le modèle déterministe

La méthode de Buys et Ballot [60] nous a permis de construire la droite des moindres carrés passant par les moyennes et les écarts types de chaque période de notre série. La pente de cette droite, qui est égale à 5.84, peut confirmer la nature de la série temporelle qui suit un modèle multiplicatif. Après avoir déterminé le schéma de décomposition de la série temporelle, il reste à estimer chacune de ses composantes décrites par l'équation (2.9). Nous avons obtenu les coefficients suivants pour représenter la fonction de régression : $a = -0.02$ et $b = 520.5$.

L'estimation du modèle multiplicatif est données par l'ensemble des équations suivantes (6.3) :

$$X_t = \begin{cases} (520.5 - 0.02t)\hat{c}_1 \ si \ t = 1, 1 + max(y_j(n_j)), 1 + 2 \times max(y_j(n_j)), ... \\ (520.5 - 0.02t)\hat{c}_2 \ si \ t = 2, 2 + max(y_j(n_j)), 2 + 2 \times max(y_j(n_j)), ... \\ ... \\ (520.5 - 0.02t)\hat{c}_{max(y_j(n_j))} \ si \ t = max(y_j(n_j)), 2 + max(y_j(n_j)), 3 \times max(y_j(n_j)), ... \end{cases}, \tag{6.1}$$

pour notre cas d'étude, nous avons $max(y_j(n_j)) = 489$ litres.

b. Modélisation de la série temporelle des écarts temporels des évènements datés avec SARIMA

Les meilleurs paramètres qui minimisent les critères d'informations AIC et BIC ont permis d'identifier le modèle SARIMA avec les paramètres suivants : $(p = 0, d = 0, q = 1)(P = 0, D = 0, Q = 1)_{max(y_j(n_j))}$. Ce modèle génère l'équation (6.2).

$$(1 - B^{max(y_j(n_j))})X = (1 + \phi_1 B)\epsilon_t. \tag{6.2}$$

c. Prévision de la série temporelle des écarts des évènements datés d'eau avec le modèle MLP

Le modèle MLP est définit avec deux couches cachées à 150/150 unité et une fonction d'activation ReLu pour chaque neurone. L'algorithme Adam est utilisé pour l'optimisation. L'ajustement [68] de la descente du gradient ainsi que le MAE pour le calcul de la fonction de perte avec un taux d'apprentissage très petit et égal à 10^{-3}.

d. Prévision de la série temporelle des écarts des évènements datés d'eau avec le modèle LSTM

Le modèle LSTM prend une seule donnée comme entrée. Il est défini avec trois couches LSTM avec respectivement 100,150 et 100 unités consécutives et avec une fonction d'activation ReLu. L'algorithme d'ajustement et d'optimisation est Adam et un taux d'apprentissage égale à 10^{-3} de la descente du gradient pour la mise à jour des poids du modèle. L'erreur absolue moyenne, MAE, est employée pour le calcul de la fonction de perte.

6.3.3 Résultats et discussion

La prévision des instants de consommation des litres d'eau avec les quatre modèles prédictifs est illustrée dans la Fig. 6.2.

Les différents modèles de prédiction qui correspondent au modèle déterministe, SARIMA, MLP et LSTM, ont la capacité de prédire les instants pendant lesquels un litre d'eau sera consommé. Les durées moyennes pour la prévision des Cdc de la consommation d'eau sont respectivement comme suit : 0.11s pour le modèle déterministe, 12.3s pour le modèle SARIMA, 23.22s pour le MLP et 111.96s pour le LSTM.

Les résultats détaillés de la prévision sont illustrés dans le Tableau 6.1 en calculant le MSE, RMSE et MAPE. Parmi les quatre modèles individuels, nous pouvons remarquer que le modèle

Table 6.1: Résultats de la prévision de la série temporelle de la consommation d'eau d'une installation privée avec les modèles individuels et les modèles hybrides

	Entrainement		Test		Modèle	
	RMSE (s)	MAPE (%)	RMSE (s)	MAPE (%)	RMSE (s)	MAPE (%)
Modèle déterministe	2276.493	4.296	3490.725	16.125	2699.877	3.67
SARIMA	3061.699	7.003	3232.09	13.952	3114.634	4.861
MLP	2973.976	6.835	3154.283	13.48	3030.001	4.792
LSTM	3168.976	6.823	3405.561	14.454	3242.636	4.757
Hybride 1 : déterministe, SARIM	2418.093	5.649	3139.983	13.935	2656.275	4.265
Hybride 2 : SARIMA, LSTM	2831.377	6.919	3046.725	11.513	2898.459	4.826
Hybride 3 : SARIMA, MLP	2882.404	6.913	3108.596	12.372	2952.894	4.809
Hybride 4 : MLP, LSTM	2435.999	5.56	3189.545	14.062	2685.307	4.213
Hybride 5 : déterministe, LSTM	2374.612	5.565	3139.457	14.091	2628.48	4.231
Hybride 6 : déterministe, MLP	3023.87	6.829	3223.445	13.738	3085.921	4.775
Hybride 7 : déterministe, MLP, LSTM	2463.278	6.045	3044.335	11.979	2651.874	4.441
Hybride 8 : SARIMA, MLP, LSTM	2489.07	6.041	3065.123	12.164	2675.818	4.429
Hybride 9 : déterministe, SARIMA, MLP	2859.953	6.887	3079.543	12.028	2928.361	4.803
Hybride 10 : déterministe, SARIMA, LSTM	2555.98	5.985	3126.01	13.304	2740.353	4.406
Hybride 11 : déterministe, SARIMA, MLP, LSTM	2557.689	6.239	3046.962	11.698	2714.599	4.52

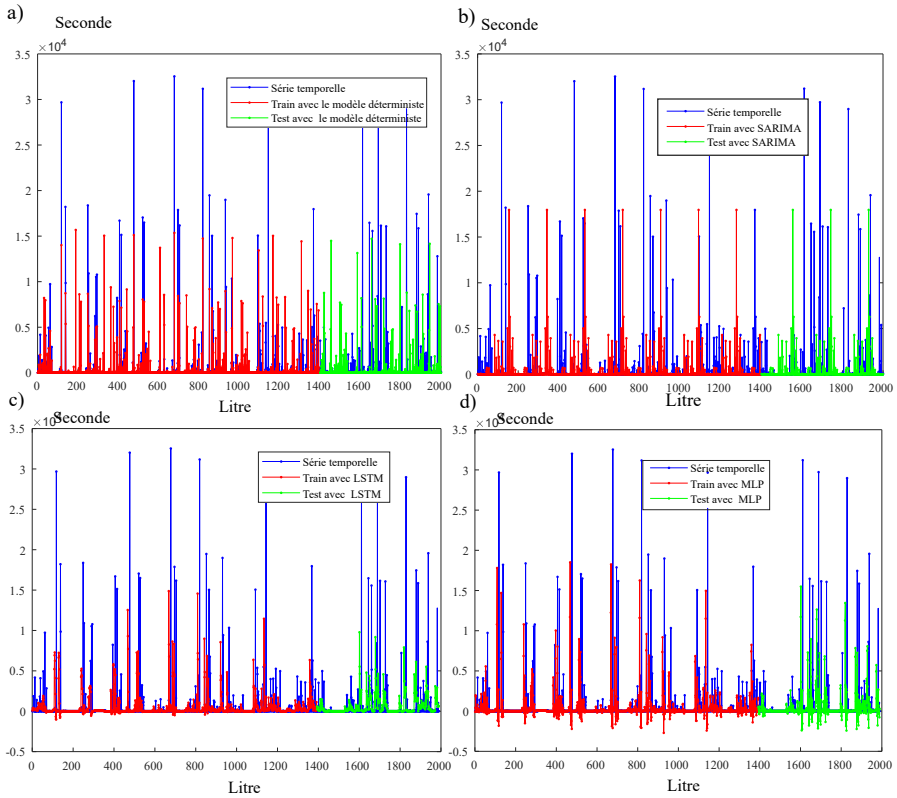

Figure 6.2: Prévision des écarts temporels des consommations des litres d'eau avec le modèle a) déterministe de décomposition en composantes essentielles des séries temporelles, b) SARIMA, c) LSTM et d) MLP

déterministe a fourni la plus petite erreur en entrainement d'environ 4.29% par contre le modèle SARIMA a généré une erreur de 7% en sachant que la moyenne des écarts est 438s, avec un minimum de 3s et un maximum de 40302s. Ces petites erreurs peuvent valider notre proposition pour choisir une saisonnalité non apparente au sein de la série temporelle de consommation. Toutefois, les modèles de combinaison hybride ont pu légèrement dépasser les performances des modèles individuels. À titre d'exemple, le modèle hybride 2 et 3 ont fourni des erreurs de prévision de 6.919% et 6.913% pour le premier et 11.513% et 12.372% pour le deuxième en terme d'erreurs respectivement d'entrainement et de test.

6.4 Prévision de série temporelle de la consommation électrique d'une installation tertiaire : cas d'un bureau

6.4.1 Données d'entrée pour les modèles prédictifs

Pour une prévision événementielle, nous avons choisi une base de données composée de quelques jours de consommation consécutifs de plus de 2000 données stockées dans la base "BaseT2-

évènement-prévision" de la Section 3.6. Cette base de donnée est échantillonnée naturellement selon l'ordre de la consommation de chaque Wh. Les données sont divisées en trois ensembles d'entrainement, de validation et de test. Ces ensembles représentent un pourcentage par rapport au nombre total des données : 70% pour l'entrainement y compris 30% pour la validation et 30% pour le test du modèle.

De même que la consommation d'eau en terme évènementiel, il est difficile d'identifier la saisonnalité avec la fonction d'auto-corrélation et la fonction d'auto-corrélation partielle. Pour cette raison, nous avons sélectionné l'énergie maximale consommée en une journée sur l'ensemble des jours $max(y_j(n_j))$. Avec cette hypothèse, une décomposition de la série temporelle en composantes essentielles et une modélisation avec le modèle SARIMA sont devenues possibles.

6.4.2 Paramètres des modèles prédictifs

a. Modélisation des évènements datés de la consommation électrique avec le modèle déterministe

La méthode de Buys et Ballott [60] nous a permis de construire la droite des moindres carrés passant par les moyennes et les écarts types de chaque période de notre série. La pente de cette droite, qui est égale à 3.6, peut confirmer la nature de la série temporelle qui suit un modèle multiplicatif. Ces méthodes sont représentées par la Fig. A.4 donné en annexe. Après avoir déterminé le schéma de décomposition de la série temporelle, qui est multiplicatif, il reste à estimer chacune de ses composantes décrites par l'équation (2.9) tel qu'il est défini précédemment. Avec une régression automatique, nous avons obtenu les coefficients suivants : $a = -3.76e-05$ et $b = 6.107$.

L'estimation du modèle multiplicatif est donnée par l'ensemble des équations suivantes (6.3) :

$$X_t = \begin{cases} (6.107 - 3.76e-05t)\hat{c}_1 \ si \ t = 1, 1 + max(y_j(n_j)), 1 + 2 \times max(y_j(n_j)), ... \\ (6.107 - 3.76e-05t)\hat{c}_2 \ si \ t = 2, 2 + max(y_j(n_j)), 2 + 2 \times max(y_j(n_j)), ... \\ ... \\ (6.107 - 3.76e-05t)\hat{c}_{max(y_j(n_j))} \ si \ t = max(y_j(n_j)), 2 + max(y_j(n_j)), ... \end{cases}, \quad (6.3)$$

pour notre cas d'étude, nous avons $max(y_j(n_j)) = 688$ Wh.

b. Modélisation de la série temporelle des écarts des évènements datés de la consommation électrique avec le modèle stochastique SARIMA

Les meilleurs paramètres qui minimisent les critères d'informations (AIC et BIC) sont donnés tels que le modèle est défini avec SARIMA $(p = 1, d = 0, q = 1)(P = 0, D = 0, Q = 1)_{max(y_j(n_j))}$ et

génère l'équation (6.4).

$$(1-\phi_1)(1-B^{max(y_j(n_j))})X = (1+\phi_1 B)\epsilon_t. \tag{6.4}$$

c. Prévision de la série temporelle des écarts des évènements datés de la consommation électrique avec le modèle MLP

Le modèle MLP est défini avec deux couches cachées avec 150 et 150 neurones dans chacune et ayant une fonction d'activation ReLu. Ce modèle est initialisé avec une seule entrée. L'algorithme Adam est utilisé pour l'optimisation et l'ajustement [68] de la descente du gradient. Ainsi l'erreur absolue moyenne pour le calcul de la fonction de perte avec un taux d'apprentissage très petit égal à 10^{-4}.

d. Prévision de la série temporelle des écarts des évènements datés de la consommation électrique avec le modèle LSTM

Le modèle LSTM prend une seule donnée comme entrée. Le nombre de neurones, le nombre de couches cachées, les fonctions d'activation (comme Sigmoïde, ReLu, Tanh ...) ont été variés et testés. Le modèle LSTM est défini avec une couche LSTM à 150 unités consécutives et avec une fonction d'activation tangente Tanh. L'algorithme d'ajustement et d'optimisation est Adam et un taux d'apprentissage égal 0.005 de la descente du gradient. Le RMSE est employé pour le calcul de la fonction de perte.

6.4.3 Résultats et discussion

La prévision des instants de consommation des Wh de la puissance électrique avec les quatre modèles prédictifs sont illustrées dans la Fig. 6.3. Les résultats des modèles de prévision déterministe, SARIMA, MLP et LSTM sont illustrés dans le Tableau 6.2.

Parmi les quatre modèles individuels, nous pouvons remarquer que le modèle LSTM a fourni la plus petite erreur en entrainement qui est proche de 0 soit 0,00001% et en test de 0,0001% telle que la moyenne des écarts est 125s. Toutefois, les modèles de combinaison hybride n'ont pas pu cette fois-ci améliorer les performances des modèles individuels. Selon la Fig. 6.3, nous pouvons remarquer que les modèles neuronales MLP et LSTM ont bien modélisé la dynamique et le comportement aléatoire de la consommation. Par contre, les modèles mathématiques tels que SARIMA a bien modélisé l'ensemble d'entrainement qui peut confirmer une autre fois notre hypothèse et notre choix de la saisonnalité non apparente au sein de la série temporelle de consommation.

Table 6.2: Résultats de la prévision des $\delta(t_i, t_{i+1})$ des évènements datés de la consommation électrique avec les modèles individuels et les modèles hybrides

	Ensemble de train		Ensemble de test		Modèle	
	RMSE (s)	MAPE (%)	RMSE (s)	MAPE (%)	RMSE (s)	MAPE (%)
Modèle déterministe	0.09	0	0.44	0.05	0.25	0.01
SARIMA	0	0	0.22	0.02	0.12	0
MLP	0.19	0.01	0.19	0.02	0.19	0.01
LSTM	0.01	0	0	0	0.01	0
Hybride 1 : déterministe, SARIM	0.05	0	0.32	0.04	0.18	0
Hybride 2 : SARIMA, LSTM	0.01	0	0.11	0.01	0.06	0
Hybride 3 : SARIMA, MLP	0.1	0.01	0.19	0.02	0.13	0
Hybride 4 : MLP, LSTM	0.11	0.01	0.31	0.04	0.19	0.01
Hybride 5 : déterministe, LSTM	0.05	0	0.22	0.03	0.13	0
Hybride 6 : déterministe, MLP	0.1	0.01	0.1	0.01	0.1	0
Hybride 7 : déterministe, MLP, LSTM	0.03	0	0.21	0.03	0.12	0
Hybride 8 : SARIMA, MLP, LSTM	0.07	0	0.27	0.03	0.16	0
Hybride 9 : déterministe, SARIMA, MLP	0.06	0	0.13	0.02	0.09	0
Hybride 10 : déterministe, SARIMA, LSTM	0.07	0	0.21	0.03	0.13	0
Hybride 11 : déterministe, SARIMA, MLP, LSTM	0.06	0	0.21	0.03	0.12	0

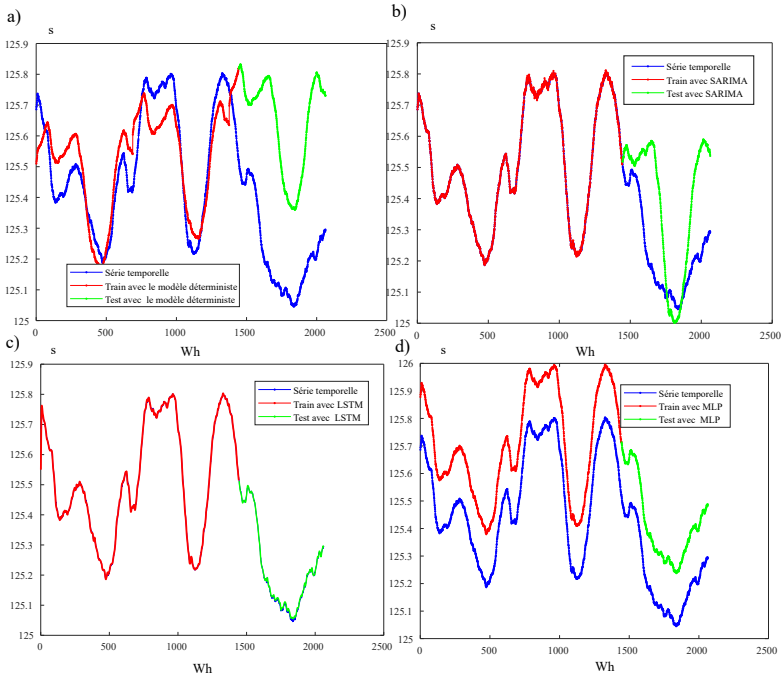

Figure 6.3: Prévision des écarts temporels des consommations des Wh de la consommation électrique avec le modèle a) déterministe de décomposition en composantes essentielles des séries temporelles, b) SARIMA, c) LSTM et d) MLP

6.5 Prévision de série temporelle de la consommation électrique d'une installation tertiaire

6.5.1 Données d'entrée et choix du modèle prédictif

Pour une prévision événementielle, nous avons choisi une base de données composée de quelques jours de consommation consécutifs de plus de 36 300 données de la base "BaseT3-évènement-prévision" décrite dans la Section 3.6. Cette base de donnée est échantillonnée naturellement selon l'ordre de la consommation de chaque Wh. Les données sont divisées en trois ensembles d'entrainement, de validation et de test. Ces ensembles représentent un pourcentage par rapport au nombre total des données : 70% pour l'entrainement y compris 30% pour la validation et 30% pour le test du modèle.

Le choix de ce modèle est basé sur sa capacité d'apprendre et de prédire la série temporelle de la consommation électrique mieux que les autres modèles individuels qui correspondent au MLP, SARIMA et le modèle déterministe. Le serveur sur lequel notre base des données est sauvegardée réalise une mise à jour pendant des périodes en début de la journée. Par conséquent, la consommation est retardée pour quelques minutes vers 1h du matin ce qui provoque des écarts temporels plus longs. Pour cela, nous avons défini deux modèles prédictifs représentés par deux LSTM.

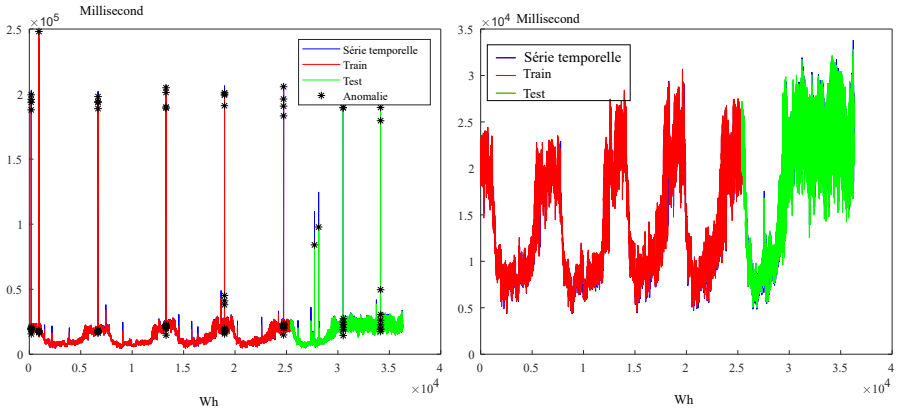

Figure 6.4: a) Prévision de la série temporelle des écarts des Wh électriques sur la base des données brutes, b) prévision de la série temporelle des écarts des Wh électriques après la suppression des pics

Table 6.3: Résultats de la prévision de la série temporelle de la consommation d'électricité avec et sans les pics

	Modèle 1			**Modèle 2**		
	Entrainement	**Test**	**Modèle**	**Entrainement**	**Test**	**Modèle**
RMSE (s)	8.29	7.79	8.15	2.38	2.78	2.87
MAPE (%)	0.06	0.044	0.035	0.002	0.02	0.003
Temps (s)	3469.87			3075.32		

6.5.2 Paramètres des modèles LSTM

Le premier modèle prédit la série temporelle des écarts de la consommation électrique qui comporte tous les instants y compris les pics du début des journées. Le second modèle prédit la série temporelle des écarts de la consommation électrique en supprimant les instants qui représentent les pics du début des journées. Le premier modèle prend six données comme entrées. Cette valeur est déterminée à partir de la fonction d'auto-corrélation partielle de la série. Le modèle est défini avec deux couches LSTM à 150/150 unités consécutives avec une fonction d'activation ReLu. L'algorithme d'ajustement et d'optimisation est Adam et un taux d'apprentissage égal à 10^{-3} de la descente du gradient pour la mise à jour des poids du modèle. L'erreur absolue moyenne, MAE, est employée pour le calcul de la fonction de perte. Le deuxième modèle LSTM est défini avec la même configuration que celui du premier modèle. Cependant, il prend 13 données comme entrées.

6.5.3 Résultats et discussion

La série temporelle est représentée avec un écart minimum de 4s, un maximum de 243s, une moyenne de 16s et une écart type de 9s. Les résultats de la prévision sont donnés dans le tableau 6.3. Nous pouvons remarquer, en premier lieu, que les deux modèles LSTM ont bien

modélisé notre série temporelle. Le modèle 2 qui a été formé sur la série temporelle sans les pics a optimisé la prévision en terme d'erreur avec un gain de 0.032% et en temps de calcul avec un gain de 400s.

À partir de deux modèles LSTM, nous pouvons toutefois remarquer que le modèle 1 n'a pas bien modélisé et prédit les pics qui représentent des périodes pendant lesquelles il n y a pas de transmission de la consommation. En se basant sur l'écart type σ et la fonction de moyenne mobile selon l'équation (5.6), nous pouvons sélectionner les valeurs difficiles à apprendre et qui peuvent également refléter un comportement différent de celui de la consommation habituelle. Pour cela, le paramètre α dans cette équation est une variable numérique choisit qui est identifiée empiriquement et strictement supérieure à 2. Dans notre cas $\alpha = 4$ est la valeur optimale. Cette valeur est basée sur la règle empirique de la distribution normale selon laquelle toutes les valeurs qui se situent à environ 3σ de la moyenne μ sont considérées comme valeurs normales modélisant un comportement habituel de la consommation électrique. Par conséquent, une valeur en dehors de 3σ peut être considérée automatiquement comme anormale et par la suite elle reflète une anomalie de consommation. Ces valeurs sont représentées dans la Fig.6.4.

6.6 Comparaison entre la prévision de série temporelle et la prévision des Cdc de consommation

La prévision des Cdc échantillonnée en heure offre la possibilité d'explorer la quantité, en litre ou en Wh, de la consommation horaire future. Par contre, la prévision des séries temporelles offre la possibilité d'explorer l'instant précis de la consommation future. À partir de la prévision de la consommation basée sur les Cdc journalières échantillonnées en heures abordée dans le chapitre 5, nous pouvons constater que les modèles prédictifs ont pu modéliser le processus de consommation et de le prédire correctement. Cependant, la prévision de la consommation d'eau basée sur les séries temporelles sans échantillonnage constituent un processus à évènement de consommation discret selon un ordre naturel. La prévision selon les deux principes est décrite dans la Fig. 6.5. Nous avons constaté que les modèles prédictifs n'ont pas pu modéliser ces séries temporelles d'une manière plus fluide que les Cdc échantillonnées. Néanmoins, les séries temporelles représentent une alternative intéressante pour la prévision des processus de consommation. Cela est due au fait qu'elles n'ont pas subi des modifications ou de traitement et correspondent au processus réel de consommation.

Nous allons comparer la prévision de Cdc de la consommation par rapport à la prévision de la série temporelle avec le modèle prédictif qui a été choisi. Cette comparaison est réalisée sur la base du modèle hybride 2 qui est à la base une combinaison du modèle SARIMA et le LSTM. Prenons comme exemple les résultats du modèle prédictif de la consommation d'eau pour sept heures pendant la journée du 01/09/2018. Selon les résultats qui sont détaillés dans le tableau 6.4, nous pouvons retenir deux importantes constatations :

- Il existe des heures pendants lesquelles la consommation peut être nulle par exemple

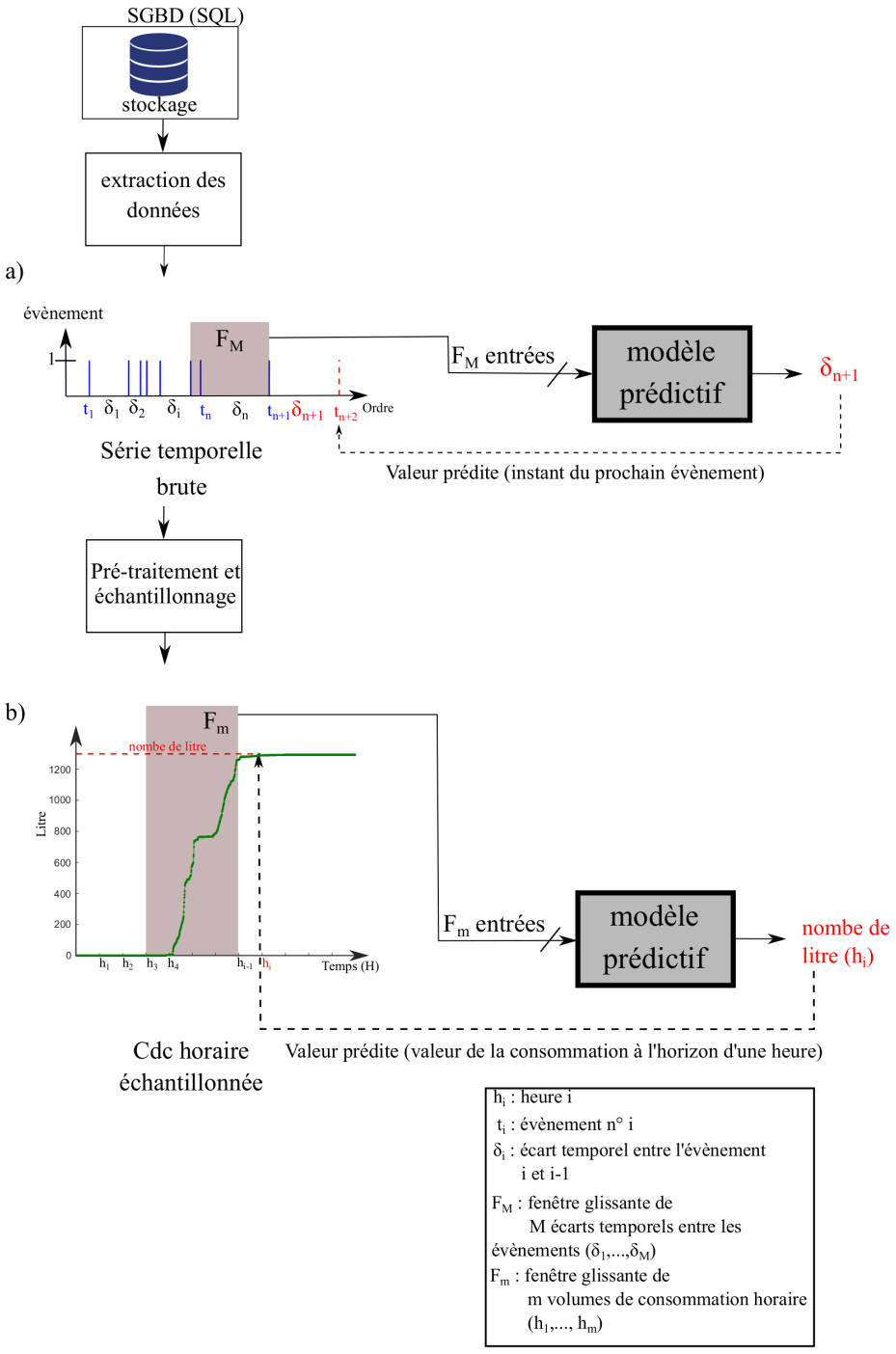

Figure 6.5: Comparaison entre a) la prévision de série temporelle de la consommation d'eau avec des données brutes et b) la prévision de Cdc de la consommation d'eau avec des données échantillonnées

Table 6.4: Comparaison de la prévision de série temporelle par rapport à la prévision avec les Cdc de consommation d'eau de deux heures

Heure	Instants des évènements réels	Instants des évènements prédits	Volume d'eau réel	Volume d'eau prédit
01			0	0
02			0	0
03			0	0
04			0	0
05			0	0
06			0	0
07	07:44:58.250	07:44:58.250	6	6
	07:45:06.500	07:45:06.028		
	07:45:14.750	07:45:13.836		
	07:52:19.375	07:48:38.147		
	07:53:16.188	07:49:08.874		
	07:54:29.125	07:49:47.150		

à 1h, 2h ... pendant des périodes données dans la journée. C'est un comportement de consommation en lien avec le type du bâtiment étudié, la précision de l'échantillonnage, etc. Dans notre cas étude, la consommation d'eau n'a eu lieu pendant ce jour qu'à partir du 7h du matin. Pour prédire la consommation horaire de ce jour, il est essentiel de prédire au début la consommation de 1h, 2h, ..., 6h du matin et après prédire 7h. Tandis que la consommation avant 7h et égale à 0 litre. De ce fait, prédire à l'horizon d'une heure, deux heures voir 6 heures n'a pas d'intérêt.

- La prévision de série temporelle de la consommation d'eau consiste à prédire que les instants pendant lesquels un évènement a eu lieu. Autrement dit, prédire à l'horizon d'un seul évènement, deux évènements voir même 10 évènements qui peuvent avoir lieu sur plusieurs heures rend la prédiction synchrone avec la consommation et optimise les tâches de l'algorithme. Nous pouvons ainsi déduire l'intérêt de prédire en terme d'évènements que de prédire des Cdc de consommation échantillonnées.

6.7 Conclusion

Le pré-traitement a pour objectif de corriger toutes anomalies qui peuvent apparaitre lors de l'élaboration des Cdc de consommation. L'utilisation des données brutes de consommation sous forme de séries temporelles ne nécessite aucun pré-traitement. Elles correspondent à des consommations réelles avec une précision en milliseconde selon l'ordre naturel d'apparition.

Les données constituant les Cdc de la consommation électrique ont été lissées puis échantillonnées en heure. Par conséquent, ces modifications peuvent influencer la qualité des prévisions futures. Les séries temporelles qui correspondent aux instants des évènements de consommation

sont une solution permettant de constituer une base des données qui reflète le comportement de consommation réel.

Les données évènementielles ont offert une possibilité d'analyser et prédire les consommations sans pré-traitement préalable des données avec une prévision synchrone à la consommation réelle. Elles ont permis également de fournir une précision de prévision très importante en seconde voir même en milliseconde. Les modèles hybrides sont moins adaptés aux séries temporelles que les Cdc de consommation. Cela peut être expliqué par la complexité des données ainsi la grande masse qu'elle les constitue.

Toutefois, les modèles individuels ont montré leurs performances pour les séries temporelles de la consommation d'eau et d'électricité. Nous avons ainsi remarqué que les modèles mathématiques tels que SARIMA et le modèle déterministe ont bien modélisé le processus de consommation. Cela confirme notre hypothèse de départ pour la recherche de la saisonnalité non apparente au sein de la série temporelle de consommation d'eau et ça n'était pas le cas avec les séries temporelles de la consommation électrique. Avec ces séries temporelles de consommation, la fonction d'auto-corrélation et la fonction d'auto-corrélation partielle ne peuvent être toujours la solution pour déterminer la saisonnalité et analyser la relation entre ses données. Malgré qu'il existe d'autres modèles d'apprentissage destinés à la modélisation et la prévision des séries temporelles comme le ARMA, le ARIMA, ... qui ne tiennent pas compte de la notion de la saisonnalité, ces modèles n'ont pas fourni des résultats satisfaisants. Par conséquent, sur la base des informations sur les Cdc, nous avons choisi d'appliquer les modèles prédictifs représentés par le modèle déterministe, SARIMA, MLP et LSTM ainsi que toutes ses combinaisons. Nous avons supposé qu'une saisonnalité non apparente peut être définie à partir du volume d'eau ou la quantité d'énergie journalière consommée. Cette hypothèse était validée à travers les performances de la prévision de la série temporelle de la consommation d'eau obtenue avec le modèle déterministe. Cependant, la prévision de la série temporelle de la consommation d'électricité était performante avec le modèle neuronal LSTM.

La saisonnalité comme caractéristique peut être introduite comme par exemple un paramètre au niveau des modèles neuronales en tant que taille de la fenêtre glissante à partir de laquelle une prévision de la valeur suivante peut être déterminée.

7 Conclusion générale

7.1 Conclusion

La gestion des ressources énergétiques et hydrauliques est un sujet contemporain important. L'eau est devenue un bien de plus en plus précieux et nous devrons surveiller son usage. Toutefois, la production d'électricité est considérée comme une source de pollution environnementale qui a connu une augmentation en lien avec la croissance démographique et l'urbanisation. Par conséquent, la gestion durable de ces ressources avec les défis environnementaux constituent la motivation principale de cette thèse.

Ainsi notre travail se focalise sur la collecte de masse des données des consommations via des compteurs intelligents dont le but de modéliser, d'analyser des profils de consommation afin d'alerter sur les anomalies et de prédire les consommations futures.

Une plateforme d'acquisition des données de consommation d'eau et d'électricité a été mise en place. La première étape dans l'analyse du comportement des utilisateurs était la classification des Cdc journalières. La classification non supervisée et supervisée des Cdc de la consommation d'eau a permis de les regrouper en des classes caractérisant des comportements de consommation différents. À son tour, cette classification a été considérée en tant que pré-traitement des données qui a permis d'avoir une base de Cdc journalière de référence. Cette base a permis de détecter des changements d'habitudes de consommation à l'échelle d'un consommateur, également à des fuites d'eau ou à des phénomènes exceptionnels (tels que : fêtes, canicules ...). Différents modèles prédictifs ont été également étudiés sur la consommation d'eau et de l'électricité de différents types des bâtiments : tertiaire et privé. Dans ce cadre, nous avons procédé une approche horaire et une évènementielle. La première a été réalisée avec une prévision des Cdc échantillonnées en heure. Cependant, la deuxième a été réalisée avec une prévision des séries temporelles naturellement échantillonnée selon l'ordre d'apparition en l'occurrence les évènements datés.

Les contributions ont été élaborés dans les différents chapitres comme suit :

Dans le **chapitre 2**, nous avons décrit les méthodes et les modèles machine learning utilisés dans le cadre d'un processus de classification, de prédiction ou de détection des anomalies des consommations. Chaque méthode est étudiée par rapport aux recherches et aux travaux existants selon le type de consommation (eau, électricité, gaz ...) et la précision des données (horaire, journalière, saisonnière, ...). Nous avons toutefois introduit les contraintes et les problèmes qui peuvent être recentrées en appliquant les méthodes ML qui correspondent à un sur-apprentissage ou un sous apprentissage à travers la vérification du compromis biais/variance.

Dans le **chapitre 3**, nous avons présenté notre contribution sur la collecte des données de consommation. Nous avons mis en œuvre une stratégie de collecte automatisée à partir des compteurs d'eau et d'électricité intelligents. Cette collecte en temps réel nous a permis de collecter des données de consommation dans différentes installations : tertiaire et privé.

La collecte continue nous a permis d'avoir différentes bases des données avec une taille et une précision différentes. La taille de la base a dépendu de la période pendant laquelle nous avons collecté et traité ces données. La précision était en fonction de sa disponibilité. En premier, nous avons validé la technologie du compteur d'eau proposé par [1] à partir des index collectés. Ces derniers ont offert la possibilité de monitorer et surveiller le bon fonctionnement de la plateforme du comptage intelligente. Cette surveillance était basée sur la vérification du manque des données à partir des watchdog fonctionnent en temps réel. Les données collectées sont inégalement espacées dans le temps. Afin de comparer les consommations journalières, un pré-traitement a été proposé. Il est basé sur la détermination du manque des données et la transformation en des données également espacées dans le temps avec un échantillonnage temporel.

Dans le **chapitre 4**, Nous avons proposé une première contribution sur le clustering des Cdc de la consommation d'eau. Ce clustering a permis d'attribuer à chaque Cdc une classe caractérisée par un profil journalier particulier (normale/anormale, jours d'activités/weekends, vacances, fériés). Dans ce cadre, une approche probabiliste PDF a été proposée qui explore les profils de consommation d'eau journalière des utilisateurs afin de rechercher une partie de cette Cdc comportant le plus de consommation. Ainsi, nous avons été en mesure de les classifier avec une erreur très faible et un nombre de données plus restreint. Avec les portions des Cdc ayant la part la plus importante de consommation, nous avons pu procéder à une classification supervisée de la consommation journalière en temps réel dans la journée et d'alerter le plus rapidement en cas d'une consommation anormale. Nous avons mené une analyse plus détaillée de cette classe de consommation anormale avec un cas d'étude de la consommation d'eau d'un RU.

Ensuite, notre deuxième contribution dans ce chapitre est la classification des appareils électriques représentés par des caractéristiques harmoniques déterminées à partir du courant de charge agrégé. Plusieurs méthodes et algorithmes ML ont été implémentées et testées pour classifier ces appareils électriques : un MLP, un arbre de décision DT de type CART été mis en place ainsi que la méthode des k-voisins les plus proches KNN. Ces méthodes consistent à classifier 8 appareils électriques en se basant sur leurs signatures harmoniques représentées par

16 caractéristiques.

Les résultats montrent que la DT a donné les meilleures performances. Cependant, un problème de sur-apprentissage peut apparaitre si l'arbre est trop profond et les données plus nombreuses. Pour palier à ce problème, nous avons proposé deux approches ensemblistes qui ont consisté à utiliser un ensemble de classificateurs optimaux basé sur une combinaison qui regroupe les résultats issus des MLP et les principes de ECOC et OAA. L'étude et la comparaison des méthodes classiques telles que la DT, KNN, MLP, a permis de comprendre les qualités respectives des modèles ensemblistes de type ECOC et OAA. Dans le contexte de notre application, l'approche ensembliste ECOC était plus performante par rapport aux modèles individuels MLP, le DT et le KNN. Nous avons ainsi comparé les résultats obtenus avec l'approche ECOC avec les resultats obtenus avec l'approhe OAA. Nous pouvons confirmer que cette dernière a dépassé toutes les autres classifications en terme de performance.

Une troisième contribution consiste à simplifier les calculs à travers la diminution des caractéristiques des appareils électriques. Ceci a consisté en une analyse et une extraction des caractéristiques. À partir de 16 initialisées pour chaque modèle de classification, une analyse avec un algorithme de gradient boosting basé sur un ensemble des arbres de décision a été proposé pour attribuer des scores d'importance à chaque caractéristique. Cette méthode a été testée sur quatre ensembles des données dont chacun se caractérise par un aspect différent : bruitées, aléatoires, normales et trafiquées. Ainsi, la classification avec un nombre plus réduit a amélioré les performances de la classification en taux et en durée.

Le **chapitre 5** présente notre contribution pour la détection des anomalies de consommation, en temps réel, en l'occurrence les fuites d'eau. En effet, un algorithme de détection de fuite a été proposé qui se base sur le débit minimum nocturne (MNF), la période maximale sans consommation nulle (PWNC) et la Cdc maximale qui représente le plafond supérieur qu'une consommation ne devrait pas le dépasser. Cet algorithme a permis de détecter les petites ainsi que les grandes fuites. Le paramétrage de cet algorithme est lié au type du bâtiment.

Pour la prédiction de la consommation d'eau et d'électricité en terme horaire, nous avons testé le modèle déterministe, SARIMA, MLP et LSTM. Le choix des modèles prédictifs a été basé sur la nature de la série temporelle de la consommation.

- Le modèle déterministe traite la série selon les trois composantes essentielles : la saisonnalité, la tendance et le bruit,
- Le modèle SARIMA, qui est à son tour, traite la série temporelle selon des variations périodiques avec un caractère aléatoire,
- Le modèle MLP, qui est un réseau de neurones constitué de plusieurs couches, modélise la série temporelle avec des fonctions non linéaires,
- Le modèle LSTM, qui est un réseau de neurones récurrent, modélise la série temporelle avec des fonctions non linéaires en introduisant la mémoire des entrées précédentes.

À partir de ces modèles individuels, différentes combinaisons ont été proposées pour former des stacking permettant de réduire les erreurs de prévisions des modèles individuels. 11 différentes combinaisons ont été proposées à partir de ces modèles individuels.

La prévision a été réalisée avec des Cdc de consommation parfaitement échantillonnées par rapport au temps selon une précision horaire. Dans ce cas, nous avons remarqué que les modèles combinés ont permis de réduire les erreurs et de trouver un compromis biais/variance pour la prévision avec l'ensemble d'entrainement et l'ensemble de test.

Le **chapitre 6** présente notre contribution pour la prévision des séries temporelles de consommation qui correspondent aux données évènementielles. Ces séries ont permis une précision très importante en seconde voir même en milliseconde et ont évité l'échantillonnage. Ces données présentent toutefois une distribution très aléatoire et non linéaire difficile de les modéliser par la méthode mathématique déterministe.

Par rapport aux Cdc des consommations, les données ont été pré-traitées puis échantillonnées en heures. Cela peut engendrer une modification des données. Par conséquent, ces modifications peuvent influencer la qualité des prévisions futures. Les séries temporelles qui correspondent aux instants des évènements électriques ont été une solution permettant de constituer une base des données qui reflète le comportement réel de consommation. De ce fait, ces données représentent des séries temporelles des évènements datés ont offert une possibilité d'analyser et prédire les consommations sans pré-traitement préalable des données. Elles ont permis également de fournir une précision de prévision très importante en seconde voir même en milliseconde.

La prévision avec les séries temporelles des évènements datés est synchrone à la consommation réelle. Cela permet d'optimiser le processus de prévision dans le cas où la série temporelle présente des intervalles où la consommation est absente.

Les modèles hybrides sont moins adaptés aux séries temporelles que les Cdc de consommation. Cela peut être expliqué par la complexité et le nombre important des données. Les modèles de prédiction individuels ont montré leurs performances pour les séries temporelles de la consommation d'eau et d'électricité. Nous avons ainsi remarqué que les modèles mathématiques tels que SARIMA et le modèle déterministe ont bien modélisé le processus de consommation. Cela confirme notre hypothèse de départ pour la recherche de la saisonnalité non apparente au sein de la série temporelle de consommation d'eau et qui n'était pas le cas avec les séries temporelles de la consommation électrique.

Les différentes approches et algorithmes ont été appliqués et testés sur plusieurs jeux des données couvrant différentes périodes de consommation. Nous avons pu valider et confirmer tous les résultats trouvés sur une masse de données plus grande. À titre d'exemple, la détection des fuites proposée dans 5.2 est à jour et fonctionne depuis plus de deux ans. Une seule intervention a été réalisée lors de l'installation d'un nouvel appareil qui consomme de l'eau. Par conséquent, une mise à jour de MNF a été opérée.

7.2 Perspectives

Nous proposons comme première perspective de généraliser les analyses proposées dans cette thèse sur d'autres bâtiments pour envisager l'études d'autres comportements individuels ou industriels. Les mêmes techniques peuvent être employées sur d'autres usages, avec d'autres comportements, et donc d'autres classes. Cette étude permettrait de contribuer et de définir de nouveaux services et de nouveaux outils.

La deuxième perspective que nous proposons serait de calculer les pondérations associées aux prédictions issues de chaque modèle individuel de manière différente. Une piste serait par exemple, d'utiliser la méthode des arbres à gradient boostés pour déterminer un score ou une importance qui pourrait servir à calculer les pondérations.

Détecter une fuite d'eau est un exercice difficile surtout lorsqu'elle se produit et il y a d'autres consommations. Une piste pour détecter les fuites quel que soit le contexte serait d'utiliser une information plus précise que la consommation en lui en associant par exemple des données issues d'un débitmètre. Ces informations permettraient de définir des signatures de la consommation d'eau de chaque appareil.

Une perspective supplémentaire serait d'analyser la corrélation qui existe entre la consommation d'eau et la consommation d'électricité. Cette démarche est à nuancer parce qu'il n'y a pas nécessairement de lien direct entre les deux grandeurs. Il serait donc judicieux de détecter les périodes ou les phases pendant lesquelles les deux grandeurs sont liées. Ainsi les données relatives à la consommation de la première grandeur permettrait de renforcer la prédiction de la deuxième.

Enfin, il serait intéressant de modéliser les consommations d'eau, d'électricité ou de gaz avec des systèmes multi-agents. Dans cette modélisation, les agents représenteraient un usager, un appareil, un consommateur, un regroupement d'appareils, un fournisseur, un distributeur, etc. L'utilisation des systèmes multi-agents dans ce contexte offriraient la possibilité de combiner implicitement les données de consommation des différents compteurs pour modéliser au mieux les contraintes, les objectifs, les habitudes de tous les acteurs entre la production et la consommation des ressources en eau, électricité et gaz.

A Annexes

A.1 Liste des acronymes

AMI	Advanced Metering Infrastruction
AMR	Automatic Meter Reading
ANN	Artificiel Neural Network
ARIMA	Auto Regressive Integrated Moving Average
HAC	Hierarchical Agglomerative Clustering
CART	Classification and Regression Trees
Cdc	Courbe de charge
DT	Decision Tree
CH	Classification Hiérarchique
DTW	Dynamic Time Warping
ECOC	Error Correcting Output Code
EM	Expectation Maximisation
IoT	Internet of Things
KDE	Kernel Density Estimator
KNN	k-Nearest Neighbors
LSTM	Long Short Term Memory
ML	Machine Learning
MLP	Multi Layer Perceptron
MNF	Minimum Night Flow
MSE	Mean Squared Error
NILM	Non-Intrusive Load Monitoring
OAA	One Against All
PDF	Probability Density Function
PWNC	Period Without Null Consumption
ReLu	Rectified Linear Unit
RMSE	Root Mean Squared Error
RNN	Reccurrent Neural Network
RU	Restaurant Universitaire
SARIMA	Seasonal ARIMA
SOM	Self Organisation Maps
SQL	Structured Query Language
WDN	Water Distribution Network

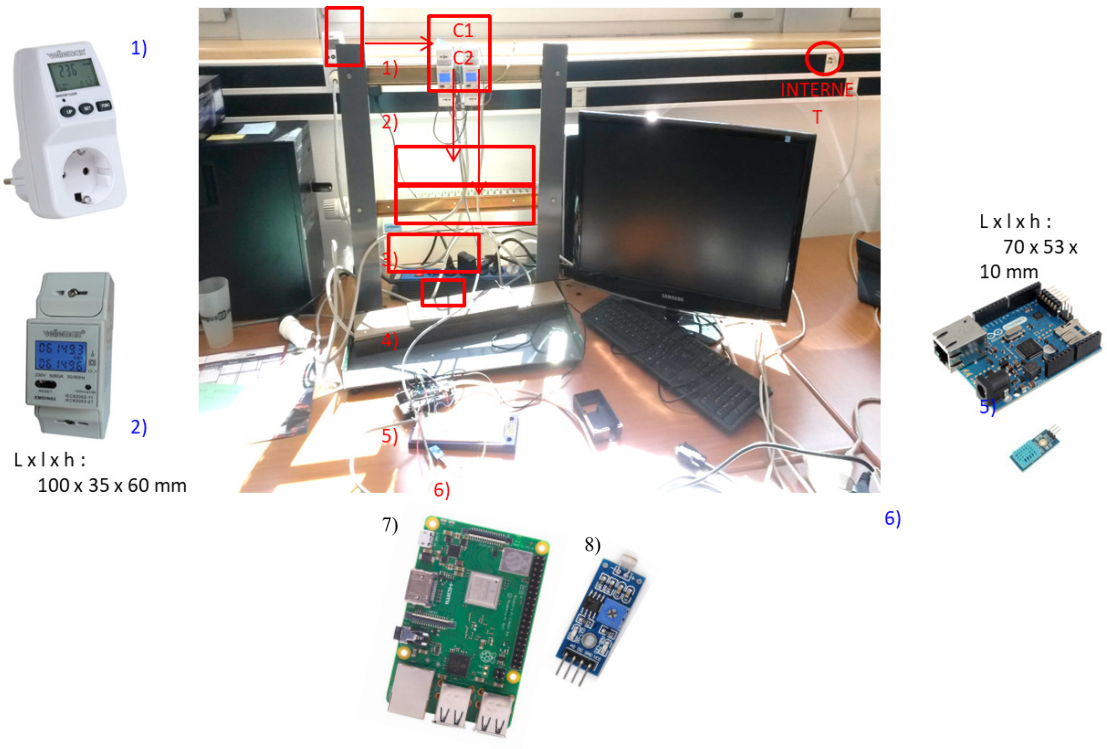

Figure A.1: Plateforme expérimentale IoT de l'acquisition de données de la consommation électrique

A.2 Compteurs intelligents utilisés

A.2.1 Compteur intelligent d'électricité

Des expérimentations de recueil de données d'électricité sont proposées. Une première plateforme a été réalisée par [121] afin de récolter la consommation d'électricité instantanée. La consommation électrique est collectée à partir d'un point de mesure unique via un compteur général. Le dispositif matériel est représenté dans la Fig. A.1. Dans cette figure représente les équipements installés dans le laboratoire IRIMAS (à l'IUT de Mulhouse).

Plusieurs d'autres compteurs ont été installés au fur et à mesure [122]. La plateforme de la collecte de données de la consommation électrique est composé d'un wattmètre classique (Velleman EMDIN02 avec impulsions de sortie (1 impulsion pour 1 Wh)) et d'un capteur de lumière à photo-résistance (LDR5516) comme le montre la Fig. A.1. Les données de consommation envoyées par les compteurs intelligents et les transmettent à une plateforme cloud où elles sont traitées afin de renvoyer des informations utiles aux utilisateurs en temps réel. Les données récoltées sont stockées sur le serveur de l'université.

Les technologies et les équipements techniques de la plateforme de la collecte de données de la consommation électrique sont les suivants :

Table A.1: Compteurs installés et utilisés pour la collecte de données d'eau et d'électricité

identifiant	type de compteur	Précision de mesure	type de bâtiment
6	compteur d'eau	1 litre	tertiaire
10	compteur d'eau	10 litres	tertiaire
2004	compteur d'eau	10 litres	tertiaire
2005	compteur d'eau	1 litre	tertiaire
2006	compteur d'eau	10 litres	tertiaire
2007	compteur d'eau	1 litre	privé
5	compteur d'électricité	800 imp= 1 kWh	tertiaire
11	compteur d'électricité	1 imp=1 Wh	tertiaire
206	compteur d'électricité	1 imp=0,1 Wh	tertiaire

1. Capteur de puissance Velleman PM230 / E avec affichage instantané de la consommation
2. 2 wattmètres (C1 et C2) de type Velleman EMDIN02 avec impulsions de sortie (1 impulsion pour 1 Wh)
3. Barrette électrique connectée à C1 (sur laquelle sont connectés différents appareils)
4. Barrette électrique connectée à C2
5. Carte microcontrôleur
6. Capteur de température et d'humidité DHT11
7. Carte Raspberry Pi 3 B+
8. Capteur de lumière à photo-résistance (LDR5516)

A.2.2 Compteur intelligent d'eau

La consommation d'eau est collectée à partir d'un point de mesure unique via un compteur général. Cette plateforme est proposée par [1]. Différents compteurs d'eau intelligents installés au sein du bâtiment tertiaire éducatif (IUT de Mulhouse) sont représentés dans la Fig. A.2.

La transmission des messages est assurée par une fenêtre glissante composée de six émissions redondantes, comme indiqué sur la Fig. A.3.

Afin de garantir la réception des informations, les transmissions sont numérotées de un à six et peuvent être considérées comme des paquets indépendants dans la trame transmise. La longueur du code lp de chaque paquet est limitée afin de définir une longueur de trame maximale Lf en fonction de la technologie radio employée. Dans le cas de cette plateforme, le choix est basé sur la longueur de Lf qui est limité à 120 octets maximum par trame transmise.

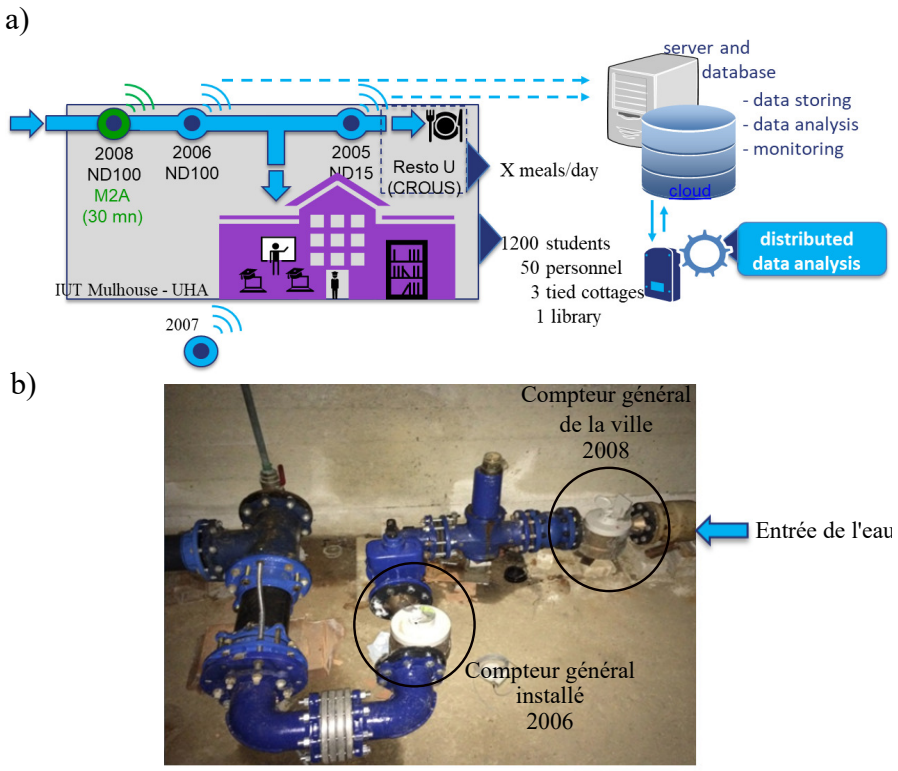

Figure A.2: a) Plateforme expérimentale de la collecte de données de la consommation d'eau, b) compteur d'eau générale à l'IUT de Mulhouse

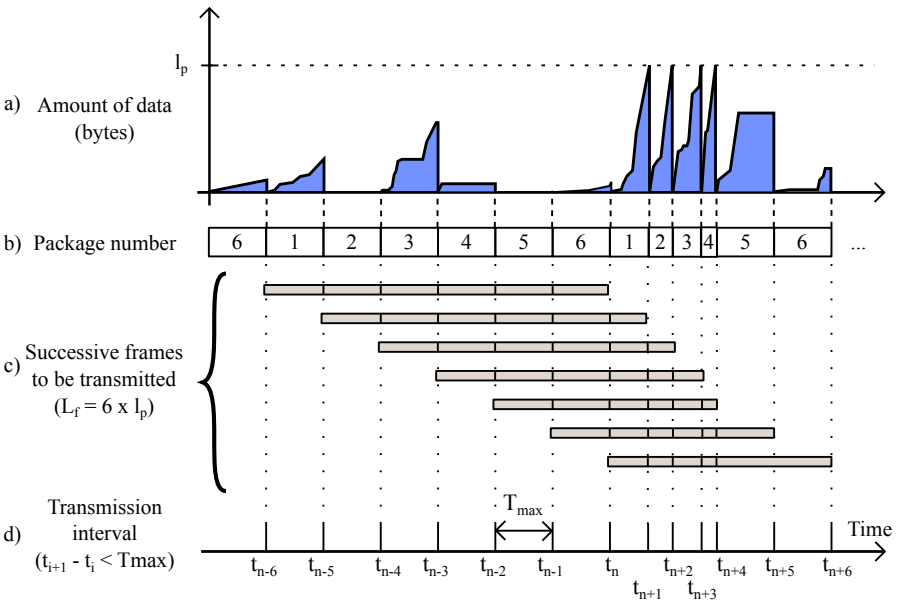

Figure A.3: Principe de fonctionnement de la fenêtre coulissante pour assurer la redondance des données transmises des compteurs d'eau intelligents à travers des trames successives [1]

Les compteurs d'eau et d'électricité installés dans les différents bâtiments correspondent aux identifiants comme il est figuré dans le Tableau A.1.

A.3 Descriptifs des données de consommation

A.3.1 Données de consommation électrique

La transmission des données (requête) repose sur le protocole HTTP du Web. Dans la spécification du protocole HTTP, GET permet de demander une page Web en lui transmettant des paramètres.

Les données récoltées sont stockées sur le serveur de l'université que nous pouvons les consultés via internet sous forme d'un tableau via l'adresse : http://www.mmi.iutmulhouse.uha.fr/eread.php?idd=11&mm=12&yyyy=2020&bf=actualiser&iddd=&ppi=) . Cette adresse permet de visualiser la consommation instantanée du compteur électrique avec l'identifiant id=11 pendant le mois mm=12 de l'année yyyy=2020.

La fonction **"Load_Pulses(date)"** permet la collecte des données de la consommation d'électricité du côté client qui correspondent aux écarts temporels δ_i de chaque Wh consommé.

```
function [mD]= Load_Pulses(date)
format long;
meter = '11';
webfile = strcat('http://www.mmi.iutmulhouse.uha.fr/etseries.php?idd=', meter,'&ddate='
  ,date);
Pulse=['Power_Pulse_' date];
urlwrite(webfile, [Pulse '.txt']);
fname=[Pulse '.txt'];
fileID = fopen(fname,'r');
Dir = fscanf(fileID, '%s,',1);
mD = fscanf(fileID, '%d, ');
fclose(fileID);
dd=sprintf('%d-%d-%d %d:%d:%d.%d', mD(1), mD(2), mD(3), mD(4), mD(5), mD(6), mD(7));
trs=datestr(dd,'dd-mmm-yyyy HH:MM:SS.FFF');
trv=datevec(dd,'dd-mm-yyyy HH:MM:SS.FFF');
trn=datenum(dd,'dd-mm-yyyy HH:MM:SS.FFF');
```

Cette fonction prend comme entrée une date par exemple "2018-09-16" qui permet de collecter les écarts temporels des Wh consommés pendant ce jour et les sauvegarder dans un vecteur mD.

Les sept premières données collectées représentent la date de référence journalière qui est l'écart temporel entre le premier et la deuxième Wh consommé. La huitième donnée représente l'identifiant du compteur qui est "11" dans cet exemple. Deux champs de séparation entre ces données ainsi que les écarts temporels représentés par 0, 0. Les données collectées correspondent à la liste suivante :

```
[2020, 01, 01, 00, 00, 05, 152, 206, 0, 0, 7847, 8027, 10376, 9373, 10698, 9384,
10920,11152, 11137, 8500, 10734, 7115, 8223, 8518, 9928, 11203, 11410, 8733, 11238,
11456,11253, 12714, 9783, 7299, 9840, 11416, 11374, 9734, 8563, 10019, 9957, 11305,
8626, 10069, 10533, 11323, 12776, 11391, 12574, 7105, 8521, 9854, 8823, 11067, 9674,
9440, 8823, 6911, 10684, 8549, 8718, 9568, 9907, 10134, 7176, 8415, 9706, 9652, ...]
```

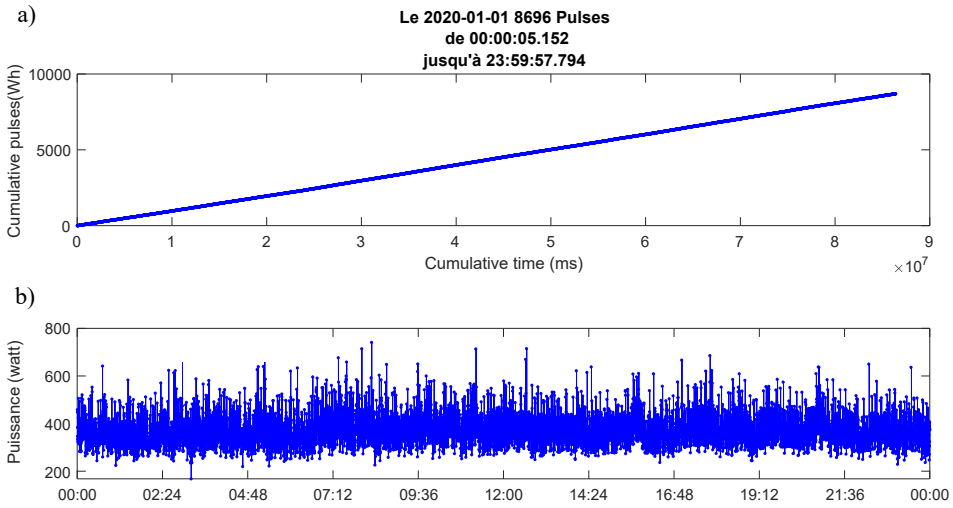

Figure A.4: a) Cdc de la consommation d'énergie électrique et b) la Cdc de la puissance pendant le 01/01/2020

A partir de ces données, plusieurs courbes pouvant être par la suite représenté telle qui correspondent à la Cdc de l'énergie consommée et la courbe de la puissance moyenne qui sont figurées dans la Fig. A.4. L'énergie consommé pendant ce jour est le nombre des écarts temporels collectés +1.

A.3.2 Données de consommation d'eau

De même que la collecte des données électrique, la transmission des données et sa collecte (requête) repose sur le protocole HTTP du Web. Dans la spécification du protocole HTTP, GET permet de demander une page Web en lui transmettant des paramètres. La récupération de ces données nécessite une connaissance d'un identifiant et un mot de passe définit au préalable pour de chaque compteur.
La fonction **WaterData** ci-dessous permet la collecte de données de la consommation d'eau côté client en terme d'index.

```
function [] = WaterData(meter, pwd)
webfile = strcat('http://www.mmi.iutmulhouse.uha.fr/waterread.php?id=', meter,'&pwd=',
  pwd);
localcsvfile2005='localcsvfile2005';
urlwrite(webfile, [localcsvfile2005 '.csv']);
webfile = 'http://www.mmi.iutmulhouse.uha.fr/water_data2005.csv'; localcsvfile2005='
  localcsvfile2005';
urlwrite(webfile, [localcsvfile2005 '.csv']);
fileID = fopen([localcsvfile2005 '.csv']);
cdata = textscan(fileID,'%s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s
   %s %s %s %s %s %s %s %s');
fclose(fileID);  %celldisp(cdata);
climit=';';
sdata(1)=size(cdata{1,1},1);  % nb rows
sdata(2)=size(cdata,2);       % nb columns
%removing the delimiters and data format conversion
```

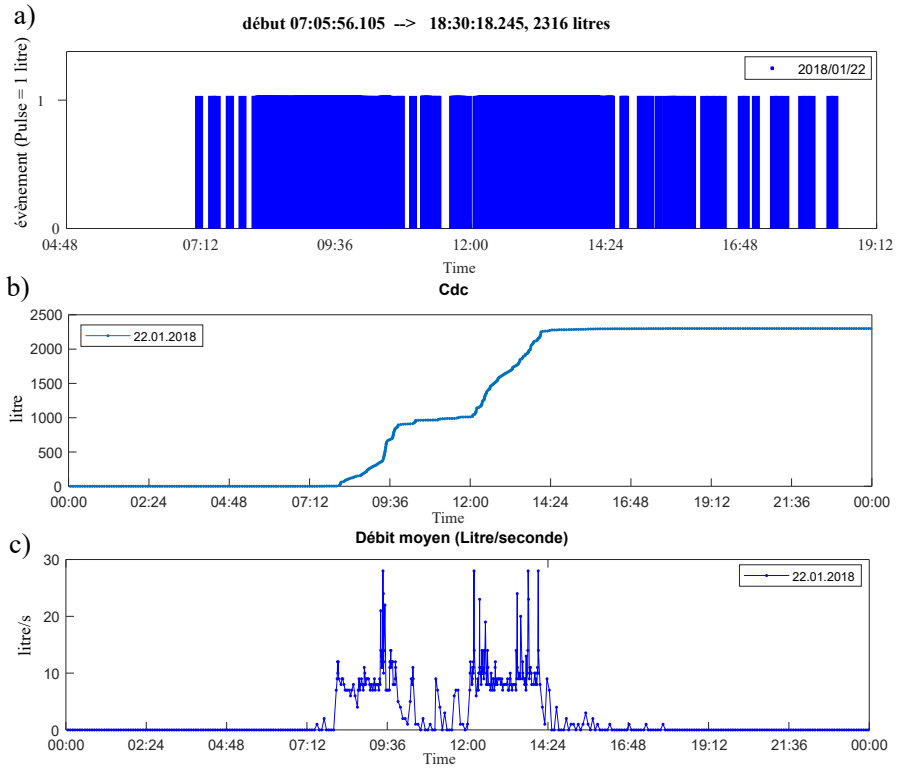

Figure A.5: a) Les évènements générés par le compteur d'eau, b) la Cdc de la consommation d'eau et b) sa Cdc représentant le débit moyen d'eau consommé pendant le 22/01/2018

```
cdata{1,1}=strrep(cdata{1,1}, climit,'');
ndata=cellfun(@str2double,cdata{1,1}(:));
cdata{1,3}=strrep(cdata{1,3}, climit,'');
for ii=5:sdata(2)-1
cdata{1,ii}=strrep(cdata{1,ii}, ';','');
ndata(:,ii-2)=cellfun(@str2double,cdata{1,ii}(:));
end
strame=cdata{1,29};
%% date format conversion
for ii=1:sdata(1)
ndata(ii,2)=datenum(sprintf('%s %s', char(cdata{1,2}(ii)), char(cdata{1,3}(ii))));
ndata(ii,3)=datenum(sprintf('%s %s', char(cdata{1,4}(ii)), char(cdata{1,5}(ii))));
end
clear cdata
%% save into a mat file
save([localcsvfile2005 '.mat'], 'ndata', 'sdata');
%example to plot mes_count in time axis
figure;
plot(ndata(:,3), ndata(:,4)/1000, '.')
datetick('x', 'dd.mmm.yy','keeplimits', 'keepticks')
xlabel('Time');
ylabel('m^3');
grid on;
```

Le Tableau A.2 correspond aux 10 premiers index d'eau collectés à partir de la plateforme

proposée. Le dernier champs dans le tableau consiste aux données de dataframe représentant toutes les informations concernant la consommation d'eau en chaque instant, transmises vers le serveur et sauvegardées dans la base de données. Il représente les 7 premiers champs du tableau cryptés et compressé selon la stratégie proposée dans la thèse de Spiegel [1].

Il est possible de collecter les évènements de la consommation d'eau en terme des écarts temporels δ_i des litres d'eau consommés. A partir de toutes ces données, plusieurs courbes peuvent être par la suite représenté comme dans la Fig A.5. Ces courbes correspondent à la série temporelle des évènements, qui sont sous forme des impulsions de 1 litre dans a), la Cdc cumulée dans b), et la courbe de débit moyen représentées dans c).

Table A.2: Les données brutes des index d'eau collectés à partir de la plateforme proposée

Id mesure	instant du compteur	instant du serveur	valeur index	status	Id compteur	var 1	var 8 ... var 20	dataframe crypté
8293	2017-11-30 23:01:12	2017-12-01 00:01:17	100009	1	2005	846263	0..0	01600f2997fc1244304c6a1a 95030000b1a9860100a2f27032
8295	2017-11-30 23:05:28	2017-12-01 00:05:33	100009	1	2005	846279	0..0	01600f2d7ffc1244304c6a1a9503 0000b1a9860100a2327132
8297	2017-11-30 23:10:16	2017-12-01 00:10:21	100009	1	2005	846298	0..0	01600f31e4f81244304c6a1a95030000 b1a9860100a27a7132
8299	2017-11-30 23:14:32	2017-12-01 00:14:37	100009	1	2005	846314	0..0	01600f35ccfb1244304c6a1a95030000 b1a9860100a2ba7132
8301	2017-11-30 23:18:48	2017-12-01 00:18:53	100009	1	2005	846331	0..0	01600f39b4fb1244304c6a1a95030000b 1a9860100a2fa7132
8303	2017-11-30 23:23:04	2017-12-01 00:23:09	100009	1	2005	846347	0..0	01600f3d9cfa1244304c6a1a95030000 b1a9860100a23a7232
8305	2017-11-30 23:27:20	2017-12-01 00:27:25	100009	1	2005	846363	0..0	01600f4184fa1244304c6a1a950300 00b1a9860100a27a7232
8307	2017-11-30 23:31:36	2017-12-01 00:31:41	100009	1	2005	846380	0..0	01600f456cf91244304c6a1a9503000 0b1a9860100a2ba7232
8309	2017-11-30 23:35:52	2017-12-01 00:35:57	100009	1	2005	846396	0..0	01600f4954f51244304c6a1a950300 00b1a9860100a2fa7232
8311	2017-11-30 23:40:08	2017-12-01 00:40:13	100009	1	2005	846412	0..0	01600f4d3cf81244304c6a1a950300 00b1a9860100a23a7332

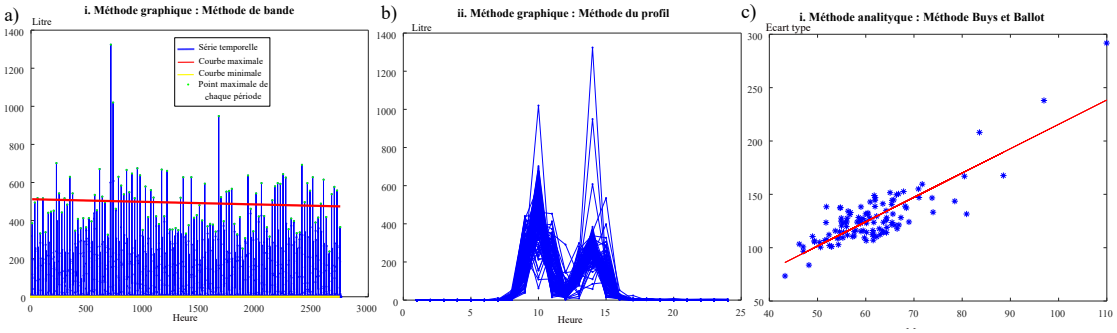

Figure A.6: Application de la méthode graphique avec a) la méthode de bande, b) la méthode de profil ainsi que c) la méthode de Buys et Ballot pour identifier le schéma de décomposition des données

A.4 Schémas de décomposition de Cdc ainsi que les séries temporelles avec le modèle déterministe

L'application de la méthode de bande, définie dans 2.5.1, sur les différentes données de consommation est présentée dans cette section. En effet, l'identification du schémas de décomposition de Cdc ainsi que les séries temporelles avec le modèle déterministe est validée par la méthode graphique et analytique comme suit. Nous avons remarqué que l'identification du schémas de décomposition de toutes nos données est réalisée de la même manière.
Les graphiques montre plus au moins que les droites passant par les minima des données et les autres droites passant par ses maxima ne sont pas parfaitement parallèles. En outre, l'application de la méthode du profil a montré que les courbes en superpositions sur les données sur les données qui représentent la saisonnalité. De ce fait, nous pouvons déduire que les schéma de décomposition de toutes nos données est de type multiplicatif.
La Fig. A.6 est un exemple de représentation graphique permettant à identifier le schéma de décomposition des données pour la modélisation déterministe.

Bibliographie

[1] Julien Spiegel. *Nouvelle stratégie de collecte de données pour les compteurs d'eau communicants*. PhD thesis, Université de Haute Alsace, 2019.

[2] William Hurst and Chelsea Dobbins. Guest editorial special issue on: Big data analytics in intelligent systems. *Journal of Computer Sciences and Applications*, 3(3):1–9, 2015.

[3] Ioan Szilagyi and Patrice Wira. Ontologies and semantic web for the internet of things. In *42nd IEEE Industrial Electronics Conference (IECON 2016)*, pages 6949–6954, 2016.

[4] Roberto Fernandez Molanes, Kasun Amarasinghe, Juan Rodriguez-Andina, and Milos Manic. Deep learning and reconfigurable platforms in the internet of things: Challenges and opportunities in algorithms and hardware. *IEEE Industrial Electronics Magazine*, 12(2):36–49, 2018.

[5] Sean Dieter Tebje Kelly, Nagender Kumar Suryadevara, and Subhas Chandra Mukhopadhyay. Towards the implementation of IoT for environmental condition monitoring in homes. *IEEE Sensors Journal*, 13(10):3846–3853, 2013.

[6] Mabrouka El Guedri. *Caractérisation aveugle de la courbe de charge électrique : Détection, classification et estimation des usages dans les secteurs résidentiel et tertiaire*. Phd thesis, Université Paris Sud - Paris XI, 2009.

[7] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[8] Mark P. McHenry. Technical and governance considerations for advanced metering infrastructure/smart meters: Technology, security, uncertainty, costs, benefits, and risks. *Energy Policy*, 59:834–842, 2013.

[9] Aida Boudhaouia and Patrice Wira. Collecting and analyzing data from autonomous iot sensors with application to smart buildings. In *Upper Rhine Cluster for Sustainability Research - International Conference (URCforSR 2018)*, 2018.

[10] Aida Boudhaouia and Patrice Wira. Detection et classification de comportements utilisateurs par apprentissage machine des consommations. In *Journée doctorale des Sciences Exactes*, 12 Juin 2020.

[11] Aida Boudhaouia, Julien Spiegel, and Patrice Wira. Recueil et exploitation de données issues de capteurs connectés dans un contexte de télémonitoring pour bâtiments intelligents. In *Congrès National de la Recherche en IUT (CNRIUT 2018), Aix-en-Provence*, page 3, 2018.

[12] Aida Boudhaouia and Patrice 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, IEEE)*, pages 1–6, 2018.

[13] Anissa Ticherahine, Aida Boudhaouia, Patrice Wira, and Abdenacer Makhlouf. Time series forecasting of hourly water consumption with combinations of deterministic and learning models in the context of a tertiary building. In *International Conference on Decision Aid Sciences and Application (DASA 2020)*.

[14] Aida Boudhaouia and Patrice Wira. Comparison of machine learning algorithms to predict daily water consumptions. In *2021 International Conference on Design & Test of integrated micro & nano-Systems (DTS, IEEE)*, pages 1–6, 2021.

[15] Aida Boudhaouia and Patrice Wira. Power and water consumption monitoring with IoT devices and machine learning methods in a smart building. volume 346. Presses Universitaires de Strasbourg, Philippe Hamman, Stéphane Vuilleumier edition, 2019.

[16] Aida Boudhaouia and Patrice Wira. A real-time data analysis platform for short-term water consumption forecasting with machine learning. volume 3, pages 682–694, 2021.

[17] Anissa Ticherahine, Aida Boudhaouia, Patrice Wira, and Abdenacer Makhlouf. Hourly prediction of time series with deterministic and learning models-application to water and electricity consumption in a tertiary building. *Applied Mathematics and Computation*, 2021.

[18] Aida Boudhaouia and Patrice Wira. Probabilistic and learning approaches for real-time classification of time series in the context of water consumption. *Applied Artificial Intelligence*, 2021.

[19] Tiago M. Fernandez-Carames and Paula Fraga-Lamas. A review on human-centered IoT-connected smart labels for the industry 4.0. 6:25939–25957, 2018.

[20] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006.

[21] Stephen Marsland. *Machine Learning. An algorithmic perspective*. Chapman and Hall/CRC Press, 2 edition, 2015.

[22] Gérard Dreyfus, Jean-Marc Martinez, Manuel Samuelides, Mirta B. Gordon, Fouad Badran, and Sylvie Thiria. *Apprentissage statistique: Réseaux de neurones - Cartes topologiques - Machines à vecteurs supports*. Editions Eyrolles, 2011.

[23] Mathieu Bourdeau, Xiao qiang Zhai, Elyes Nefzaoui, Xiaofeng Guo, and Patrice Chatellier. Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society*, 48:101533, 2019.

[24] Félix Biscarri, Iñigo Monedero, Antonio García, Juan Ignacio Guerrero, and Carlos León. Electricity clustering framework for automatic classification of customer loads. *Expert Systems with Applications*, 86:54–63, 2017.

[25] Selina. Chu, Eamonn. Keogh, David. Hart, and Michael. Pazzani. Iterative deepening dynamic time warping for time series. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, Proceedings, pages 195–212. Society for Industrial and Applied Mathematics, 2002.

[26] Ruizhe Ma and Rafal Angryk. Distance and density clustering for time series data. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 25–32, 2017.

[27] Xiaozhe Wang, Kate A Smith, Rob Hyndman, and Damminda Alahakoon. A scalable method for time series clustering. *Proc. Unrefereed Res. Papers*, 2004.

[28] Chrysi Laspidou, Elpiniki Papageorgiou, Konstantinos Kokkinos, Sambit Sahu, Arpit Gupta, and Leandros Tassiulas. Exploring patterns in water consumption by clustering. *Procedia Engineering*, 119:1439–1446, 01 2015.

[29] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[30] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. *Information Systems*, 53:16–38, 2015.

[31] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[32] Germain Forestier, Florent Lalys, Laurent Riffaud, Brivael Trelhu, and Pierre Jannin. Clustering de séquences d'activités pour l'étude de procédures neurochirurgicales. In *Journées Francophones Extraction et Gestion des Connaissances (EGC)*, pages 489–494, 2012.

[33] Simon S. Haykin and Simon S. Haykin. *Neural networks and learning machines*. Prentice Hall, 3rd ed edition, 2009.

[34] D. Praveen Kumar, Tarachand Amgoth, and Chandra Sekhara Rao Annavarapu. Machine learning algorithms for wireless sensor networks: A survey. *Information Fusion*, 49:1–25, 2019.

[35] Tom FAWCETT and Foster PROVOST. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1997.

[36] William DuMouchel and Matthias Schonlau. A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities. *in Proc. KDD*, page 5, 1998.

[37] Megha Gaur, Stephen Makonin, Ivan V. Bajic, and Angshul Majumdar. Performance evaluation of techniques for identifying abnormal energy consumption in buildings. *IEEE Access*, 7:62721–62733, 2019.

[38] Y. Zhang, W. Chen, and J. Black. Anomaly detection in premise energy consumption data. *2011 IEEE Power and Energy Society General Meeting*, pages 1–8, 2011.

[39] Seif-Eddine Benkabou. *Détection d'anomalies dans les séries temporelles: application aux masses de données sur les pneumatiques.* PhD thesis, Université Claude Bernard, 2018.

[40] Lili Yang and Shuang-Hua Yang. Domestic water consumption monitoring and behaviour intervention by employing the internet of things technologies. *Procedia Computer Science*, pages 367–375, 2017.

[41] R. Puust, Z. Kapelan, D. A. Savic, and T. Koppel. A review of methods for leakage management in pipe networks. *Urban Water Journal*, 7(1):25–45, 2010.

[42] Chiara Luciani, Francesco Casellato, Stefano Alvisi, and Marco Franchini. From water consumption smart metering to leakage characterization at district and user level: The GST4 water project. 2(11):675, 2018.

[43] Chiara Luciani, Francesco Casellato, Stefano Alvisi, and Marco Franchini. Green smart technology for water (GST 4water): Water loss identification at user level by using smart metering systems. *Water*, 11(3):405, 2019.

[44] Gustavo de Souza Groppo, Marcelo Azevedo Costa, and Marcelo Libânio. Predicting water demand: a review of the methods employed and future possibilities. 19(8):2179–2198, 2019.

[45] Kazeem B Adedeji, Yskandar Hamam, Bolanle T Abe, and Adnan M Abu-Mahfouz. Leakage detection and estimation algorithm for loss reduction in water piping networks. *Water*, 9(10):773, 2017.

[46] Rustom Mamlook and Odeh Al-Jayyousi. Fuzzy sets analysis for leak detection in infrastructure systems: a proposed methodology. *Clean Technologies and Environmental Policy*, 6(1):26–31, 2003.

[47] Jui-Sheng Chou and Abdi Suryadinata Telaga. Real-time detection of anomalous power consumption. *Renewable and Sustainable Energy Reviews*, 33:400–411, 2014.

[48] Weihui Deng, Guoyin Wang, Xuerui Zhang, Yishuai Guo, and Guangdi Li. Water quality prediction based on a novel hybrid model of ARIMA and RBF neural network. In *IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, pages 33–40. IEEE, 2014.

[49] J Chen and D Boccelli. Demand forecasting for water distribution systems. *Procedia Engineering*, 70:339–342, 2014.

[50] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017-07-01.

[51] Siddharth Arora and James W. Taylor. Forecasting electricity smart meter data using conditional kernel density estimation. *Omega*, 59:47–59, 2016.

[52] Laouafi Abderrezak. *Contribution à la modélisation de la courbe de charge électrique par des techniques intelligentes.* PhD thesis, Université 20 Aout 1955, Skikda, 2017.

[53] Jean-Marc Mercantini, Leïla Sakli, and Jean-Claude Hennet. Study of supply chain vulnerabilities based on cognitive engineering and arima formal models. In *11th International Conference on Integrated Modeling and Analysis in Applied Control and Automation (IMAACA)*, Budapest, Hungary, 2018.

[54] Jolanta Szoplik. Forecasting of natural gas consumption with artificial neural networks. *Energy*, 85:208–220, 2015.

[55] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang. Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2019.

[56] Muhammad A. Al-Zahrani and Amin Abo-Monasar. Urban residential water demand prediction based on artificial neural networks and time series models. 29(10):3651–3662, 2015.

[57] Milad Leyli Abadi, Allou Same, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cedric Feliers, and Olivier Chesneau. Predictive classification of water consumption time series using non-homogeneous markov models. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 323–331, 2017.

[58] Stefano Alvisi, Marco Franchini, and Alberto Marinelli. A short-term, pattern-based model for water-demand forecasting. *Journal of Hydroinformatics*, 9:39–50, 2007.

[59] J. Durbin and P. B. Kenny. *Seasonal Adjustment When the Seasonal Component Behaves Neither Purely Multiplicatively nor Purely Additively.* NBER, 1978.

[60] Agnes Lagnoux. *Séries chronologiques.* ISMAG, Master1-MI00141X, Université de Toulouse Le Mirail, 53p, 1996.

[61] Jacques J. F. Commandeur and S. J. Koopman. *An introduction to state space time series analysis.* Oxford University Press, 2007.

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

[63] Yves Aragon. *Séries temporelles avec R Méthodes et cas*. New York : Springer, 2011.

[64] Sana Buhamra, Nejib Smaoui, and Mahmoud Gabr. The box–jenkins analysis and neural networks: prediction and time series modelling. *Applied Mathematical Modelling*, 27(10):805–815, 2003.

[65] Sylvain Rubenthaler. Séries chronologiques (avec R) (Cours et exercices). In *JAD - Laboratoire Jean Alexandre Dieudonné*, France, 2019.

[66] Dehua Zhang and Sha Lou. The application research of neural network and bp algorithm in stock price pattern classification and prediction. *Future Generation Computer Systems*, 115:872 – 879, 2021.

[67] Jeanny Hérault and Christian Jutten. *Réseaux neuronaux et traitement du signal*. Hermès science publications, 1994.

[68] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[69] Xuan-Hien Le, Hung Viet Ho, Giha Lee, and Sungho Jung. Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, 11(7):1387, 2019.

[70] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang. Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2017.

[71] Xiangyun Qing and Yugang Niu. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, 148:461 – 468, 2018.

[72] Mehdi Khashei and Mehdi Bijari. An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications*, 37(1):479–489, 2010.

[73] Mahmut Firat, Mustafa Erkan Turan, and Mehmet Ali Yurdusev. Comparative analysis of neural network techniques for predicting water consumption time series. *Journal of hydrology*, 384(1-2):46–51, 2010.

[74] Mohammad Ebrahim Banihabib and Pezhman 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*, 48:101585, 2019.

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

[76] Yukio Maruyama and Hisashi Yamamoto. A study of statistical forecasting method concerning water demand. *Procedia Manufacturing*, 39:1801–1808, 2019.

[77] S. Saigal and D. Mehrotra. Performance comparison of time series data using predictive data mining techniques. *Advances in Information Mining*, pages 57–66, 2012.

[78] Moataz Soliman, Tobi Abiodun, Tarek Hamouda, Jiehan Zhou, and Chung-Horng Lung. Smart home: Integrating internet of things with web services and cloud computing. In *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, pages 317–320. IEEE, 2013.

[79] Wei Li, Guogang Liu, and Junho Choi. Environmental monitoring system for intelligent stations. *Concurrency and Computation: Practice and Experience*, page e5131, 2019.

[80] Chuan Zhang, Alessandro Romagnoli, Li Zhou, and Markus Kraft. From numerical model to computational intelligence: The digital transition of urban energy system. *Energy Procedia*, 143:884–890, 2017.

[81] Simon Marvin, H. Chappells, and S. Guy. Pathways of smart metering development: shaping environmental innovation. *Computers, Environment and Urban Systems*, 23(2):109–126, 1999.

[82] Timo Koskela, Mikko Lehtokangas, Jukka Saarinen, and Kimmo Kaski. Time series prediction with multilayer perceptron, fir and elman neural networks. *Proceedings of the World Congress on Neural Networks*, 1996.

[83] Rodney A. Stewart, Khoi Nguyen, Cara Beal, Hong Zhang, Oz Sahin, Edoardo Bertone, Abel Silva Vieira, Andrea Castelletti, Andrea Cominola, Matteo Giuliani, Damien Giurco, Michael Blumenstein, Andrea Turner, Ariane Liu, Steven Kenway, Dragan A. Savić, Christos Makropoulos, and Panagiotis Kossieris. Integrated intelligent water-energy metering systems and informatics: Visioning a digital multi-utility service provider. *Environmental Modelling & Software*, 105:94–117, 2018.

[84] Kathryn B. Janda. Buildings don't use energy: people do. *Architectural Science Review*, 54(1):15–22, 2011.

[85] Soma Shekara Sreenadh Reddy Depuru, Lingfeng Wang, Vijay Devabhaktuni, and Nikhil Gudi. Smart meters for power grid—challenges, issues, advantages and status. In *2011 IEEE/PES Power Systems Conference and Exposition*, pages 1–7. IEEE, 2011.

[86] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, and A.E. Rizzoli. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environmental Modelling & Software*, 72:198–214, 2015.

[87] Julien Spiegel, Gilles Hermann, and Patrice Wira. A comparative experimental study of compression algorithms for enhancing energy efficiency in smart meters. In *IEEE 16TH International Conference of Industrial Informatics (INDIN 2018)*, 2018.

[88] GEORGE W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[89] Kira Rehfeld, Norbert Marwan, Jobst Heitzig, and Juergen Kurths. Comparison of correlation analysis techniques for irregularly sampled time series. 18(3):389–404, 2011.

[90] Andreas Eckner. Algorithms for unevenly spaced time series : Moving averages and other rolling operators. In *Working Paper*, 2015.

[91] D. Kang. Real-time optimal control of water distribution systems. *Procedia Engineering*, 70:917–923, 2014.

[92] Julien Spiegel, Patrice Wira, and Gilles Hermann. Energy efficiency optimization in fluid flow metering. In *IEEE 19th International Conference on Industrial Technology (ICIT)*, Lyon, France.

[93] David Walker, Enrico Creaco, Lydia Vamvakeridou-Lyroudia, Raziyeh Farmani, Zoran Kapelan, and Dragan Savić. Forecasting domestic water consumption from smart meter readings using statistical methods and artificial neural networks. *Procedia Engineering*, 119:1419–1428, 2015.

[94] Kadir Amasyali and Nora M. El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018.

[95] Mathieu Fauvernier. *Splines multidimensionnelles pénalisées pour modéliser le taux de survenue d'un événement : application au taux de mortalité en excès et à la survie nette en épidémiologie des maladies chroniques*. PhD thesis, Université Claude Bernard Lyon 1.

[96] Thien-Minh Nguyen. *Contribution to the analysis and understanding of electrical-grid signals with signal processing and machine learning techniques*. PhD thesis, Université de Haute Alsace, 2017.

[97] Jose Maria Giron-Sierra. *Digital signal processing with Matlab examples*. Springer Berlin Heidelberg, 2017.

[98] Albrecht Zimmermann. Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. *WIREs Data Mining and Knowledge Discovery*, 10(2), 2020.

[99] Robert R. Sokal and F. James Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40, 1962.

[100] Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

[101] Paul Mangiameli, Shaw K. Chen, and David West. A comparison of som neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93(2): 402–417, 1996.

[102] Matias Callara and Patrice Wira. A probabilistic learning approach for predicting application launches in cloud computing architectures. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 584–589, 2019.

[103] Shaomin Wang, Shouxiang Wang, and Dan Wang. Combined probability density model for medium term load forecasting based on quantile regression and kernel density estimation. *Energy Procedia*, pages 6446–6451, 2019.

[104] Daniela G. Calò, Angela Montanari, and Cinzia Viroli. A hierarchical modeling approach for clustering probability density functions. *Computational Statistics & Data Analysis*, 71:79–91, 2014.

[105] Ricardo Cao, Antonio Cuevas, and Wensceslao González Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, 17(2):153–176, 1994.

[106] Matias Callara. Suivi et prédiction des comportements par auto-apprentissage pour l'optimisation y compris énergétique des ressources et des services informatiques, 2019.

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

[108] Stephen Makonin. *Real-time embedded low-frequency load disaggregation.* PhD thesis, Simon Fraser University, School of Computing Science, 2014.

[109] D. Srinivasan, W. S. Ng, and A. C. Liew. Neural-network-based signature recognition for harmonic source identification. *IEEE Transactions on Power Delivery*, 21(1):398–405, 2006.

[110] Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, and Piotr Duda. The CART decision tree for mining data streams. 266:1–15.

[111] Mohamed Chiheb Ben Nasr, Sofia Ben Jebara, Samuel Otis, Bessam Abdulrazak, and Neila Mezghani. Respiratory activity classification based on ballistocardiogram analysis. In Mohamed Jmaiel, Mounir Mokhtari, Bessam Abdulrazak, Hamdi Aloulou, and Slim Kallel, editors, *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*, volume 12157, pages 79–88. Springer International Publishing, 2020.

[112] S. Escalera and O. Pujol. ECOC-ONE: A novel coding and decoding strategy. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 578–581. IEEE, 2006.

[113] Raphael Puget. *Étude de la classification dans un très grand nombre de catégories.* phdthesis, Université Pierre et Marie Curie, 2016.

[114] Nima Hatami, Reza Ebrahimpour, and Reza Ghaderi. ECOC-based training of neural networks for face recognition. In *2008 IEEE Conference on Cybernetics and Intelligent Systems*, pages 450–454. IEEE, 2008.

[115] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016.

[116] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

[117] H. Matsila and P. Bokoro. Load forecasting using statistical time series model in a medium voltage distribution network. In *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, pages 4974–4979, 2018.

[118] Zulifqar Ali, Ijaz Hussain, Muhammad Faisal, Hafiza Mamona Nazir, Tajammal Hussain, Muhammad Yousaf Shad, Alaa Mohamd Shoukry, and Showkat Hussain Gani. Forecasting drought using multilayer perceptron artificial neural network model. *Advances in Meteorology*, 2017:1–9, 2017.

[119] Weijun li and Zhenyu Liu. A method of SVM with normalization in intrusion detection. *Procedia Environmental Sciences*, 11:256–262, 2011.

[120] Jaehyun Lee, Jinho Kim, and Woong Ko. Day-ahead electric load forecasting for the residential building with a small-size dataset based on a self-organizing map and a stacking ensemble learning method. *Applied Sciences*, 9(6):1231, 2019.

[121] Wided Hammedi. *Analyse de la courbe de charge électrique au sein d'un bâtiment intelligent.* PFE, IUT Mulhouse, 2017.

[122] Derdah Younes. *Collecte, transmission, traitement et analyse de signaux et de données massives issus d'objets connectés.* Rapport technique master 2 EEA, IUT Mulhouse, 2021.