UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

CARACTÉRISATION ÉLECTRIQUE, THERMIQUE ET COMPORTEMENTALE HIÉRARCHISÉE D'UN AGRÉGAT DE RÉSIDENCES EN VUE DE LA PRÉVISION ÉNERGÉTIQUE À COURT TERME

THÈSE PRÉSENTÉE
COMME EXIGENCE PARTIELLE DU
DOCTORAT EN GÉNIE ÉLECTRIQUE

PAR
KHANSA DAB

MAI 2024

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

Doctorat en génie électrique (Ph.D.)

Direction de recherche :

Kodjo Agbossou	Directeur de recherche
Yves Dubé	Codirecteur de recherche

Jury d'évaluation

Ahmed Ouameur Messoud	Président du jury
Nassim Noura	Évaluateur
José Agustín Candanedo	Évaluateur externe
Kodjo Agbousso	Directeur de recherche
Yves Dubé	Codirecteur de recherche

Thèse soutenue le 19 avril 2024

Résumé

L'efficacité de la gestion des réseaux électriques dépend grandement de la capacité à anticiper de manière fiable et précise la consommation énergétique future. Cette tâche se révèle particulièrement complexe en raison de facteurs stochastiques variés, tels que les comportements imprévisibles des consommateurs et l'influence de variables externes comme les conditions météorologiques. Ces éléments sont particulièrement pertinents dans des contextes spécifiques, comme celui du Québec, où les variations de température peuvent grandement affecter la consommation énergétique, surtout pendant les pics hivernaux. La prévision à court terme est un élément essentiel pour la gestion et l'opération efficaces des réseaux électriques. Elle contribue à assurer l'équilibre entre l'offre et la demande et réduisant ainsi le risque de surcharges ou de déficits de puissance. Le processus de prévision réalisé à partir de systèmes automatisés présente des défis importants. Ceux-ci sont principalement liés à la gestion des incertitudes et à la complexité de la manipulation de données avec une granularité spatiale et temporelle poussée. La variabilité significative introduite par des facteurs imprévisibles tels que les aléas météorologiques et les fluctuations brusques des habitudes de consommation complique grandement l'élaboration de modèles prédictifs fiables. En outre, l'analyse de larges ensembles de données, englobant des informations sur la consommation historique et les comportements des consommateurs à différentes échelles, requiert des algorithmes avancés et des techniques d'analyse sophistiqués. Pour relever ces défis, la littérature scientifique propose plusieurs méthodes de prévision à court terme de la consommation énergétique, qui se basent sur des approches statistiques, des analyses de séries temporelles ou l'apprentissage automatique. Ces modèles ont prouvé leur efficacité dans certaines situations, mais ils présentent aussi des limites significatives. Parmi celles-ci, on note une dépendance forte aux données historiques et une tendance à sous-estimer des facteurs

exogènes, comme les basses températures caractéristiques des hivers québécois. De plus, ces méthodes peinent souvent à quantifier l'incertitude inhérente aux prévisions, ce qui soulève des questions quant à la fiabilité de ces dernières. En résumé, bien que ces approches apportent des contributions importantes à la prévision de la demande énergétique, elles ne fournissent pas toujours une mesure explicite de la fiabilité de leurs prédictions, ce qui est crucial pour la gestion optimale des réseaux électriques.

Cette thèse adresse ces défis en examinant de nouvelles méthodes pour anticiper la demande d'énergie, en mettant particulièrement l'accent sur les modèles probabilistes non paramétriques, tels que le Processus gaussien additif (AGP) bayésien multivarié. Cette approche innovante fusionne l'inférence bayésienne avec des techniques de classification afin de traiter la complexité des données multivariées. Cette méthode novatrice combine l'inférence bayésienne et des techniques de classification pour aborder la complexité des données multivariées, tout en répondant aux fluctuations des profils de consommation agrégée influencés par des variables climatiques et comportementales. Le travail de recherche se divise en plusieurs phases. Dans un premier temps, un modèle de prévision est appliqué aux composantes sensibles au climat et aux aspects calendaires des profils de charge agrégés par un système additif. Ensuite une procédure de classification est appliquée aux profils de charges dans le but d'améliorer ces prévisions. La dernière analyse de cette thèse réside dans l'analyse de l'incertitude des prévisions, permettant de déterminer la flexibilité offerte par un groupe de résidences dans leur participation à la gestion locale de la demande d'énergie. Les performances du modèle proposé sont évaluées à l'aide de données réelles de la consommation électrique agrégée d'un ensemble de maisons au Québec. Les résultats démontrent que cette approche non paramétrique surpasse les méthodes existantes, offrant des prévisions plus précises et fiables. Cette contribution significative à la littérature scientifique apporte des perspectives prometteuses

à l'étude de la gestion et de la transition énergétique, offrant aux gestionnaires de réseau, tels qu'Hydro-Québec, des outils plus efficaces permettant une participation active des consommateurs finaux dans la gestion globale du réseau électrique.

Dédicaces

Cette thèse est dédiée à tout mon monde : ma famille, mes amis, ceux et celles qui sont là aujourd'hui, qui l'ont été avant et qui le seront plus tard; sans vous autres, toute cette folie-là ne vaudrait pas la peine.

Remerciements

Je tiens à remercier sincèrement mon directeur de recherche Monsieur **Kodjo Agbossou,** professeur du département de génie électrique et génie informatique de l'UQTR et directeur du Laboratoire d'Innovation et de Recherche en Énergie Intelligente (LIREI), pour m' avoir confié ce projet, pour la qualité de son encadrement, pour ses conseils, son soutien et ses encouragements. J'aimerais remercier Monsieur **Nilson Henao** agent de recherche du LIREI pour son aide et tous les encouragements et le soutien durant ma thèse, son expérience, ses connaissances et son appétence pour le domaine ont très fortement contribué à la réussite de ma thèse, aussi je remercie mon codirecteur de recherche Monsieur **Yves Dubé**, professeur du département de génie mécanique de l'UQTR, pour son aide, son appui scientifique, pour ses précieuses contributions et sa disponibilité tout au long de ce travail de thèse. Je tiens a remercier Monsieur **Shaival Nagarsheth** post doctorant dans notre équipe de recherche de m'avoir orienté, aidé beaucoup et conseillé durant ma thèse. Je remercie également Messieurs **Sayed Saeed Hosseini**, **Michael Fournier** et **Simon Sansregret**, chercheurs du Laboratoire de Technologies de l' Énergie (LTE) d'Hydro-Québec, pour leurs participations scientifiques, ainsi que pour le temps qu' ils ont consacré à nos échanges, réunions et révisions de documents relatifs au travail de recherche. Ce travail n' aurait pas été possible sans l'aide de différentes instituts et organismes qui, au travers de leur soutien financier ou technique, ont reconnu mon travail et m'ont fait confiance: le LTE d'Hydro-Québec, le Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG), l' Institut de Recherche sur l' Hydrogène (IRH) et la Fondation de l'UQTR.

Table des matières

Liste des figures

Liste des abréviations

AGP	Processus Gaussien Additive
CBAF	Cluster based Aggregate Forecasting.
CC	Consensus Clustering.
CDF	Cumulative Distribution Function
DER	Ressources Énergétiques Distribuée
DIRH	Direct Horizontal Irradiation
DIRV	Direct Vertical Irradiation
DR	Décision Making
DR	Réponse à la Demande.
DSO	Distribution System Operators
ET	Energie Transactionnelle (Transactive Energy)
FB	Facebook Prophet
GBR	Gradient Boosting Regressor
HVAC	Chauffage Ventilation Climatisation
IA	Intelligence Artificielle.
KDE	Kernel Density Estimation
LSTM	Long Short-Term Memory
MAP	Maximum a Posteriori
PDF	Fonction de Densité de Probabilité
RFR	Ramdom Forest Reggression
REI	Réseau Énergétique Intelligent
SoD	Subset Of Data.
STLF	Short Term Load Forecasting.
SVR	Support Vector Reggression

Chapitre 1 - Introduction

1.1 Contexte général de recherche

La décentralisation de la gestion dans les réseaux de distribution joue un rôle clé pour simplifier le contrôle d'un grand nombre d'éléments, tels que les sources d'énergie renouvelables, les systèmes de stockage et les charges des consommateurs. En déplaçant la gestion au niveau local, on réduit la complexité inhérente au contrôle centralisé, rendant la surveillance et l'optimisation du réseau plus pratiques [1]. L'approche décentralisée permet également une réactivité plus rapide face aux variations locales de consommation ou de production d'énergie, ce qui est particulièrement bénéfique dans des scénarios où les conditions énergétiques peuvent varier significativement d'une région à l'autre. Afin de maintenir une coordination efficace dans un tel réseau décentralisé, il est essentiel d'avoir des systèmes de gestion avancés, capables d'estimer l'état du réseau et de fournir des prévisions précises, pour s'adapter rapidement aux changements [2].

D'ailleurs, la création des groupes de consommateurs au sein du réseau de distribution présente des défis en ce qui concerne la coordination et le contrôle. Chaque groupe peut avoir des caractéristiques et des besoins uniques, nécessitant une gestion décentralisée et une prise de décision agile au niveau local [3]. Cependant, en anticipant les modèles de demande, la gestion décentralisée peut ajuster de manière proactive la production et la distribution d'énergie optimisant ainsi la réponse à la demande. Par conséquent, la prévision du comportement des agrégats permet une meilleure compréhension des tendances de consommation d'énergie au sein de différents groupes. De plus, avec des prévisions précises, les gestionnaires de réseau décentralisés peuvent mieux équilibrer l'offre et la demande au niveau local. Ceci revêt une importance particulière pour intégrer efficacement les sources d'énergie renouvelable, qui peuvent être intermittentes et

dépendant des conditions locales. Dans ce cas, les opérateurs peuvent minimiser les pertes d'énergie en ajustant les flux d'énergie pour répondre plus précisément à la demande locale par la bonne compréhension du comportement des agrégats [4]. D'une part, en anticipant les variations de la demande, la gestion décentralisée peut maintenir plus efficacement la stabilité et la fiabilité du réseau, en s'adaptant rapidement aux changements et en assurant une fourniture d'énergie constante. Également, en prévoyant les modèles de charge liés à la recharge, par exemple les véhicules électriques et l'augmentation des pointes dû à une électrification complète des bâtiments, les gestionnaires de réseau peuvent ajuster la distribution d'énergie pour éviter les surcharges et optimiser l'utilisation des ressources énergétiques [5]. En outre, le développement des réseaux électriques intelligents au Canada et au Québec ouvre d'importantes opportunités pour améliorer l'efficacité énergétique, renforcer la fiabilité et la résilience du réseau, et faciliter l'intégration des énergies renouvelables. Ces réseaux avancés permettent une meilleure gestion de la demande et offrent aux consommateurs un plus grand contrôle sur leur consommation énergétique grâce à des technologies comme les compteurs intelligents et les systèmes de gestion de l'énergie domestique. Ces réseaux envisagent une amélioration des prévisions en fournissant des données avec une granularité spatiale et temporelle plus élevées [6].

Au Québec, le nombre de nouveaux bâtiments résidentiels a considérablement augmenté entre 2011 et 2023 [7]. Ces bâtiments utilisent l'électricité en tant que source principale d'énergie (35%) [8]. Ce phénomène a eu pour conséquence une augmentation substantielle de la consommation d'énergie qui est attribuable à l'élargissement de la superficie à chauffer. La consommation de l'énergie électrique a également augmenté dans les anciennes résidences due à l'électrification des usages tels que le chauffage, chauffe-eau, ainsi l'achat de voitures électriques rechargeables comme illustré sur la

Figure 1.4. Selon les statistiques d'Hydro-Québec, en 2019, un pourcentage de (80%) [1] d'énergie consommée des résidences au Québec est utilisé pour le chauffage des espaces et de l'eau des bâtiments en raison de la baisse des températures extérieures [9]. Le profil de la consommation d'énergie sera donc modifié selon les périodes de pointe qui se produisent pendant les basses températures. De ce fait, la consommation énergétique résidentielle est désormais une préoccupation majeure des gestionnaires du réseau électrique. Par conséquent, pour analyser la consommation résidentielle, des modèles de prévision ont été proposés dont l'élaboration a été principalement assurée par deux groupes de chercheurs: les ingénieurs et les économistes [5]. Les modèles des premiers se basent sur les fondements techniques de la demande, décrivant les prédictions à partir des technologies variées, alors que ceux des seconds décrivent la demande comme une résultante des interactions humaines en rapport avec le contexte économique, réagissant ainsi aux prix de l'énergie. Par ailleurs, un autre aspect à considérer est la périodicité des données dans le processus de prévision. Les prévisions de la charge peuvent être établies à des intervalles horaires, quotidiens, mensuels, saisonniers et annuels. Cette classification prend également en compte les besoins énergétique en électricité, qu'il s'agisse de prévisions de charge opérationnelles ou de planification à plus long terme. Ainsi, l'approche de modélisation varie en fonction des caractéristiques techniques ou économiques prises en compte, ainsi que de la période temporelle sur laquelle les prévisions sont effectuées.

Les gestionnaires de réseaux électriques de l'autre côté ont réfléchi à un nouveau concept de réseaux électriques intelligents (RÉI) pour améliorer l'efficacité énergétique, la réduction des pertes d'énergie et la gestion de la demande et de la charge. Selon ce paradigme, des solutions ont été proposées qui permettent de mettre en place les

1. https://www.hydroquebec.com/residential/customer-space/electricity-use/electricity-consumption-by-use.html

comportements des occupants et réduire les consommations d'énergie, tout en gardant comme objectif l'amélioration des performances des systèmes et réseaux d'énergie électrique à tous les niveaux pour les consommateurs ainsi que pour le producteur. En 2017, la consommation totale d'énergie au Québec, tous secteurs confondus, était de 1749 PJ. Ce niveau de consommation est très élevé à l'échelle mondiale, comme le montre la Figure 1.1. Cette grande consommation s'explique en partie par la consommation industrielle liée à l'hydroélectricité, mais aussi par une consommation énergétique dans les transports et les bâtiments (résidentiels et commerciaux) supérieurs à celle des pays européens dont le niveau de vie est comparable ou supérieur comme le montre la Figure 1.2. [2]

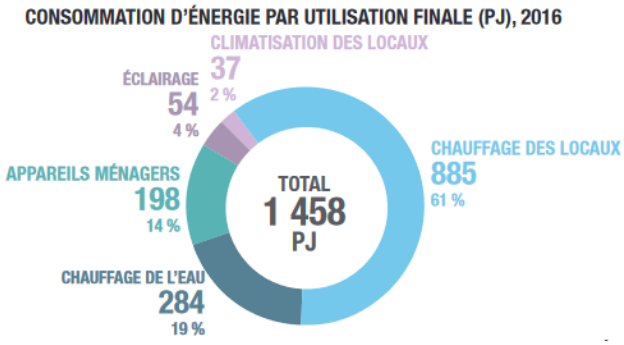

Figure 1.1 Consommation résidentielle d'énergie pendant la période hivernale : cas du Canada.

Défis liés à la prévision : Pour soutenir le développement du réseau électrique au Canada et au Québec, l'installation de compteurs intelligents se généralise à différents niveaux du réseau, allant des consommateurs individuels aux alimentateurs basse tension et aux postes de transformation. Cette prolifération des compteurs intelligents nécessite le développement de techniques de prévision évolutives capables de gérer un grand nombre de points de mesure avec des caractéristiques diverses. Ces efforts sont essentiels pour

2. Source: État de l'énergie au Québec édition 2020.

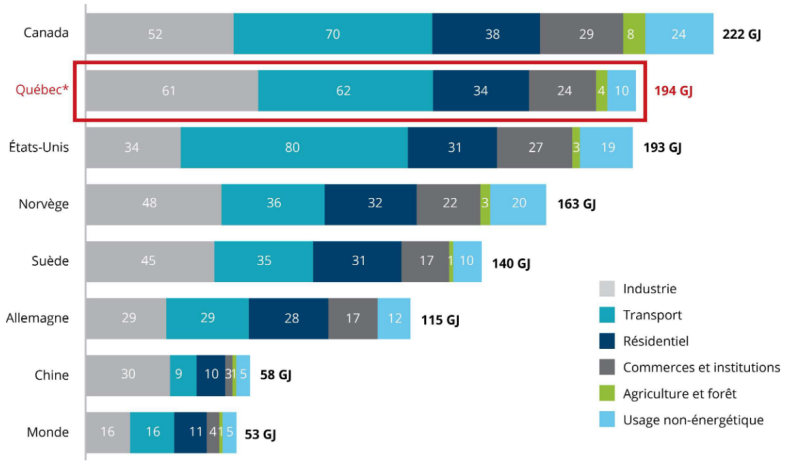

Figure 1.2 Comparaison de la consommation énergétique par habitant pour l'année 2017.

garantir des prédictions précises et efficaces, permettant ainsi une gestion optimale du réseau électrique dans un contexte canadien et québécois, où la durabilité et l'efficacité énergétique sont de plus en plus prioritaires.

Les smart grids génèrent une grande quantité de données qui doivent être analysées et traitées efficacement. Cependant, transformer ces données en prévision de charge précises et en informations exploitables pour la prise de décisions opérationnelles exige des capacités analytiques et des technologies de traitement de données sophistiquées. L'agrégation des données de multiples sources sur le réseau de distribution, telles que les compteurs intelligents, les capteurs, et les dispositifs de contrôle, rend la gestion de ces données plus complexe. Il faut non seulement collecter, mais aussi analyser efficacement ces données agrégées pour obtenir des insights pertinents, ce qui requiert des systèmes d'analyse avancés et une capacité à gérer des volumes de données considérables. Ces réseaux offrent également la possibilité d'une gestion plus dynamique et efficace du réseau électrique grâce à des programmes de réponse à la demande bien structurés. En

outre, l'intégration des énergies renouvelables dans le réseau est facilitée, favorisant ainsi une transition vers des sources d'énergie plus durables. Enfin, les smart grids stimulent l'innovation technologique, notamment par l'application de l'intelligence artificielle et de l'apprentissage automatique pour l'analyse des données, ouvrant la voie à des avancées significatives dans la gestion énergétique et la précision des prévisions de charge.[3] Cette analyse fine des données est d'autant plus essentielle dans des contextes comme celui du Canada, où le climat nordique assez froid entraîne une utilisation énergétique importante pour le chauffage d'espace, représentant jusqu'à 61% de la consommation totale d'énergie. Par exemple, en 2016, la consommation totale d'électricité au Canada était de 222 gigajoules (Gj), et parmi les différentes provinces, le Québec se démarquait avec une consommation de 194 gigajoules (Gj) comme montré sur la Figure 1.2[4].

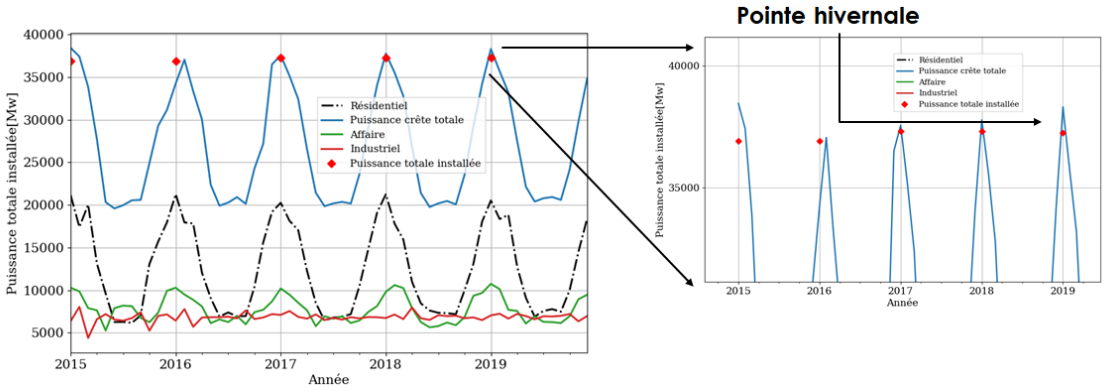

Figure 1.3 Comparaison de la consommation résidentielle par habitant en 2017

Au Québec, le secteur industriel a consommé le plus d'électricité en 2019, totalisant 93 TWh, suivi par les secteurs résidentiel et commercial avec respectivement 71 TWh et 40 TWh. La consommation d'électricité dans le secteur résidentiel est significative, en partie en raison de l'augmentation des nouvelles constructions chauffées à l'électricité,

3. Source: Guide de données sur la consommation d'énergie, L'Office de l'efficacité énergétique Ressources naturelles Canada.

4. Source: https://eneroutlook.enerdata.net/canada-energy-forecast.html

représentant 80% entre 2011 et 2020. Afin de réduire la consommation d'électricité, il est important de prendre des mesures telles que l'amélioration de l'efficacité énergétique et l'adoption de comportements de consommation responsables. L'Agence internationale de l'énergie (AIE) a également émis 25 recommandations visant à améliorer l'efficacité énergétique, couvrant des domaines tels que les bâtiments, les appareils électriques, l'éclairage, le transport, les industries, ainsi que les producteurs et distributeurs d'énergie. L'Agence effectue également l'évaluation et le suivi des recommandations. On peut voir que dans la province de Québec, la puissance crête de l'année 2015 jusqu'à 2019 se situe autour de 38 GW, dépassant momentanément la puissance totale installée. Le but est de diminuer les besoins de puissance à la pointe en essayant d'aplatir la courbe de la puissance résidentielle comme l'illustre la Figure 1.3 [5].

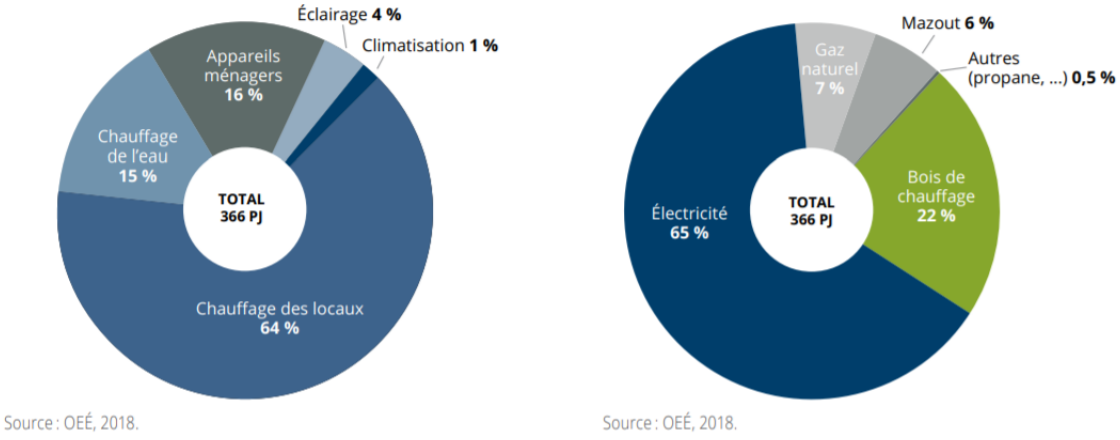

Figure 1.4 Consommation d'énergie par type d'utilisation à gauche et par source d'énergie à droite dans le secteur résidentiel au Québec, 2017.

5. Source: Rapport annuel d'Hydro-Québec 2015 jusqu'à 2020

1.2 Problématique de la thèse

La littérature scientifique sur la prévision à court terme de la charge résidentielle a identifié quelques difficultés majeures associées à l'exploitation de données. Ainsi, le défi qui se pose consiste à explorer ces méthodes d'apprentissage en utilisant des modèles non paramétriques, tout en prenant en compte les facteurs exogènes qui influencent la consommation pour un agrégat de résidences. La prévision joue un rôle essentiel dans le fonctionnement efficace et fiable des réseaux électriques intelligents. Ces prévisions sont effectuées pour une période allant de quelques heures à quelques jours à l'avance en très court terme court terme moyen terme ou long terme. La problématique à laquelle nous sommes confrontés concerne la réalisation de prévisions à court terme pour la consommation d'électricité, avec une approche quotidienne. Bien que la période à court terme puisse varier de quelques heures à deux semaines, notre attention se concentre sur le marché d'une seule journée, avec un horizon temporel de 24 heures. Dans ce contexte, la question demeure : comment pouvons-nous réduire l'intervention humaine dans la conception de l'algorithme de prédiction ? De plus, afin d'améliorer la précision des prévisions pour un ensemble de résidences, nous envisageons d'exploiter les données fournies par les capteurs intelligents et les stations météorologiques.

Certains facteurs comme les données climatiques, les compteurs intelligents et la nature stochastique des occupants exigent une étude plus rigoureuse à l'aide des modèles qui éliminent la redondance et l'incertitude [10]. Par ailleurs, le profil de la consommation d'électricité dépend de l'heure de la journée, le jour de la semaine, les fins de semaine, les jours fériés, les jours spéciaux/les festivals, les saisons et des habitudes dynamiques des clients résidentiels. Il peut y avoir des différences importantes entre les valeurs de la demande de pointe et de la demande hors pointe, pour une journée donnée dans les

profils de charge. Cependant, le problème de prévision se complexifie à cause de la quantité de données disponibles dans le réseau intelligent [11]. Cette situation implique une complexité algorithmique au niveau des simulations de ces modèles de prévision qui ne nécessitent pas l'intervention humaine. D'un côté, l'application des approches existantes devient complexe en raison du nombre important de paramètres dans les approches paramétriques. De l'autre côté, les approches non paramétriques présentent une complexité considérable, notamment en termes d'utilisation de l'espace mémoire et de temps de calcul élevé dans ces scénarios. Nous sommes confrontés au défi de la prévision en tenant compte des contraintes de ces approches, telles que la réduction de dimension des entrées et la minimisation de la propagation des erreurs liées aux prévisions météorologiques, tout en prenant en considération le temps calendaire. Les modèles de prévision actuels rencontrent des défis pour aborder simultanément toutes ces difficultés.

Alors cette problématique mène à une deuxième qui est la caractérisation comportementale de l'agrégat des résidences avec des stratégies de classification différentes. Le défi qui se pose ici réside dans cette caractérisation, nécessitant l'application de différentes techniques de classification. La prévision de la charge devient complexe en raison de l'utilisation de flux de données temporelles évolutives. Cette complexité est accentuée par la diversité des périodes temporelles dans l'ensemble de données, dont le besoin de robustesse face aux séries temporelles incomplètes. La contrainte principale est la limitation de la capacité de regrouper des utilisateurs en fonction de séries temporelles synchronisées, accentuant le défi de distinction des groupes. Par ailleurs, les données de séries temporelles, souvent bruyantes ou stochastiques, compliquent davantage la caractérisation des groupes. Les méthodes de regroupement classiques se trouvent confrontées à une inefficacité prévisionnelle, résultant de l'accumulation d'erreurs induite par le comportement stochastique des

utilisateurs.

Cependant, la compréhension indispensable des caractéristiques des modèles de charge est souvent perdue au cours des processus d'agrégation. Les fluctuations des données révèlent des changements qui peuvent détériorer l'efficacité du processus de prévision [12]. Des recherches connexes montrent que les techniques de classification des charges peuvent accroître la précision des prévisions pour les charges agrégées en réduisant les caractéristiques stochastiques des profils de charge, ce qui permet à l'algorithme de prévision d'être formé sur des données fortement corrélées [13]. Les prévisions de séries temporelles et la classification sont des méthodes standard utilisées pour faciliter la prise de décision; il est donc avantageux de diviser les consommateurs existants en groupes plus petits en fonction de leurs caractéristiques communes. Les stratégies de classification appliquées à la procédure de prévision pourraient permettre une amélioration statistiquement significative de la précision par rapport aux prévisions agrégées traditionnelles en fonction du nombre de groupes et de la taille de la base de données. [14]. Ainsi, la CBAF peut offrir un aperçu supplémentaire aux praticiens qui souhaitent mettre en œuvre la stratégie dans le monde réel et améliorer la prévision de la charge pour un ensemble de données classifié. En outre, l'algorithme de consensus joue un rôle essentiel dans garantir la cohérence des données en coordonnant les différentes partitions. [15] [16]. Il aboutit à des partitions stables et gère la diversité en générant des partitions avec différents sous-ensembles d'attributs cohérents [17].

Les marchés de flexibilité représentent une composante essentielle dans la gestion des défis dynamiques des réseaux électriques, notamment en matière de gestion de la congestion et de lissage des pointes de demande. Au cœur de ces marchés, les agrégateurs jouent un rôle essentiel en tant qu'intermédiaires qui rassemblent et commercialisent la

flexibilité offerte par les consommateurs aux opérateurs de réseau ou aux gestionnaires de réseau de distribution (GRD). Par exemple, une intégration réussie des énergies renouvelables dépend en grande partie de la disponibilité de prévisions précises de la demande, ce qui renforce la fiabilité du réseau électrique. De même, une prévision précise est indispensable pour gérer la flexibilité des systèmes de stockage, permettant ainsi de déterminer les moments optimaux de recharge ou de décharge, et d'ajuster les signaux envoyés aux consommateurs en fonction du niveau de saturation prévu du réseau. Les agrégateurs proposent des réductions de charge basées sur des références, principalement des estimations de la charge anticipée par les consommateurs, qui servent de point de référence pour évaluer l'efficacité des initiatives de réponse à la demande. Toutefois, une difficulté pour les agrégateurs réside dans l'incertitude inhérente au comportement des consommateurs.

En effet, les agrégateurs ne peuvent pas exercer un contrôle direct sur les actions individuelles des consommateurs. Les incertitudes liées à des facteurs peu prévisibles tels que les habitudes des consommateurs, à l'efficacité des équipements, peuvent entraîner des écarts entre les réductions de charge projetées et les résultats réels. Pendant la phase d'enchères des marchés de flexibilité, il est essentiel que les agrégateurs tiennent compte de ces écarts potentiels. Une sous-estimation pourrait les conduire à s'engager sur un objectif de réduction qu'ils ne pourront pas atteindre, compte tenu des incertitudes liées au comportement des consommateurs. Une surestimation, en revanche, pourrait leur faire manquer des opportunités de revenus, car ils auraient pu se porter garants de réductions plus importantes. Ce défi est aggravé par la nature volontaire de la participation des consommateurs. Les agrégateurs impliquent généralement les consommateurs sur une base volontaire, où les participants offrent de la flexibilité sans engagement contraignante. Ainsi, si un consommateur ne parvient pas à atteindre

la réduction de charge anticipée, il n'est généralement pas pénalisé. Cela place la responsabilité sur l'agrégateur de fournir des incitations qui motivent une participation cohérente et optimale des consommateurs. Étant donné que la responsabilité d'évaluer la réduction potentielle incombe principalement aux agrégateurs, ils doivent trouver un équilibre entre l'incitation à la participation des consommateurs et l'estimation précise des réductions de charge pour s'assurer que leurs offres sont à la fois compétitives et réalisables. Imaginons qu'un agrégateur énergétique, par exemple, planifie une réduction de charge de 100 mégawatts (MW) lors d'une phase d'enchères sur les marchés de flexibilité. Cependant, en raison des facteurs imprévisibles tels que les habitudes variables des consommateurs et l'efficacité des équipements, il existe un écart entre les réductions de charge projetées et les résultats réels. Si l'agrégateur sous-estime cet écart, il pourrait s'engager sur un objectif de réduction qu'il ne pourra pas atteindre, risquant ainsi de ne pas respecter ses engagements envers les opérateurs de réseau électrique. À l'inverse, une surestimation pourrait lui faire manquer des opportunités de revenus, car il aurait pu garantir des réductions plus importantes et potentiellement bénéficier de prix plus élevés sur le marché de la flexibilité.

1.3 Objectifs et contributions

L'objectif principal de cette thèse est de développer une méthodologie pour la prévision agrégée de la consommation résidentielle en utilisant des approches non paramétriques. Cela implique de prendre en compte la composante liée aux variables climatiques influençant la consommation et la composante calendaire liée aux charges et aux comportements des occupants. Ensuite, la validation de cette approche sera effectuée dans le contexte d'une application de gestion transactionnelle en analysant les incertitudes associées aux prévisions résultantes. Ainsi, les objectifs spécifiques sont proposées:

Objectifs spécifiques :

1. Développer un modèle de prévision à court terme pour la demande électrique résidentielle agrégée. Ce modèle repose sur l'application d'une approche non paramétrique à l'agrégat, permettant de combiner la sensibilité au climat et la saisonnalité de la demande d'électricité dans un modèle bayésien, tout en tenant compte des incertitudes associées
2. Caractériser le comportement de l'agrégat des résidences avec des stratégies de classification différentes. Ceci par le développement d'une métrique pour la flexibilité énergétique permettant de mettre en place des stratégies transactionnelles de la puissance agrégée.
3. Proposer un mécanisme pour renforcer la capacité des agrégateurs à anticiper et à intégrer efficacement les écarts potentiels dans leurs stratégies de prévisions. Cette amélioration vise à assurer la viabilité économique des agrégateurs tout en contribuant de manière significative à l'efficacité et à la stabilité globales des marchés de flexibilité en quantifiant les incertitudes liées aux prévisions.

La réalisation de ces objectifs a donné lieu à trois contributions principales

Contributions réalisées : Les approches proposées dans le cadre de cette étude ont abouti à l'amélioration des connaissances dans le domaine de la prévision des consommations électriques résidentielles agrégé à court terme. Nous énumérons les contributions suivantes :

1. Un modèle multivarié non paramétrique qui divise la consommation d'énergie en composantes climat sensible et calendaire. En particulier, un processus gaussien additif est utilisé comme structure probabiliste pour permettre la modélisation des variables d'entrée en utilisant des facteurs de la composante sensible au climat et du

facteur calendaire ainsi que la composante de la meilleure combinaison de climat et du temps.

2. Une structure nettement réduite en termes du temps et d'espace mémoire qui déploie deux stratégies utiles. Cette méthode fait appel aux techniques d'approximation et de classification pour sélectionner un sous-ensemble fini de points d'induction afin de gérer les défis du traitement de données volumineuses. Ces points d'induction sont les données sélectionnés de manière stratégique à partir d'un ensemble plus vaste pour représenter efficacement l'ensemble des données. L'objectif est de réduire le temps de calcul et l'utilisation de la mémoire en travaillant avec un ensemble de données plus petit mais représentatif.
3. Un modèle de classification basée sur la classification CBAF des profils de charges des maisons afin d'améliorer la prévision globale de la charge. Ce modèle utilise l'algorithme de clustering basé sur les k-médoides et l'AGP et également l'utilisation de l'indice de similarité de Jaccard pour évaluer la ressemblance entre les clusters.
4. Une méthodologie pour quantifier les incertitudes de prévision au niveau de l'agrégateur. Cette dernière est basée sur l'utilisation de distributions complètes pour une représentation plus précise de l'incertitude et pour faciliter l'identification des besoins en flexibilité. Ensuite l'évaluation de l'efficacité de la stratégie sur un ensemble de données synthétiques et une étude de cas spécifique en comparant a la fin l'AGP et le modèle Prophet, soulignant la supériorité de l'AGP.

Enfin, des simulations numériques sur des ensembles de données synthétiques et réelles, en insistant sur l'amélioration de la précision des prévisions. Ensuite l'implication pratique de différentes études proposées pour la gestion de l'énergie et de la planification du réseau électrique. Le développement de ces contributions

est décrit dans le chapitre3 dans les trois articles publiés.

1.4 Infrastructure de recherche

Ce travail de recherche fait partie du projet d'Hydro Québec Chaire de recherche Hydro Québec sur la gestion transactionnelle de la demande résidentielle en puissance et en énergie. Dans cette chaire, nous nous pencherons sur l'importance de la gestion transactionnelle de la demande résidentielle en puissance et en énergie, et surtout ses implications potentielles pour l'efficacité énergétique. En intégrant les données de la Figure 1.5, nous pourrons voir un aperçu complet des opportunités liés à ce projet.

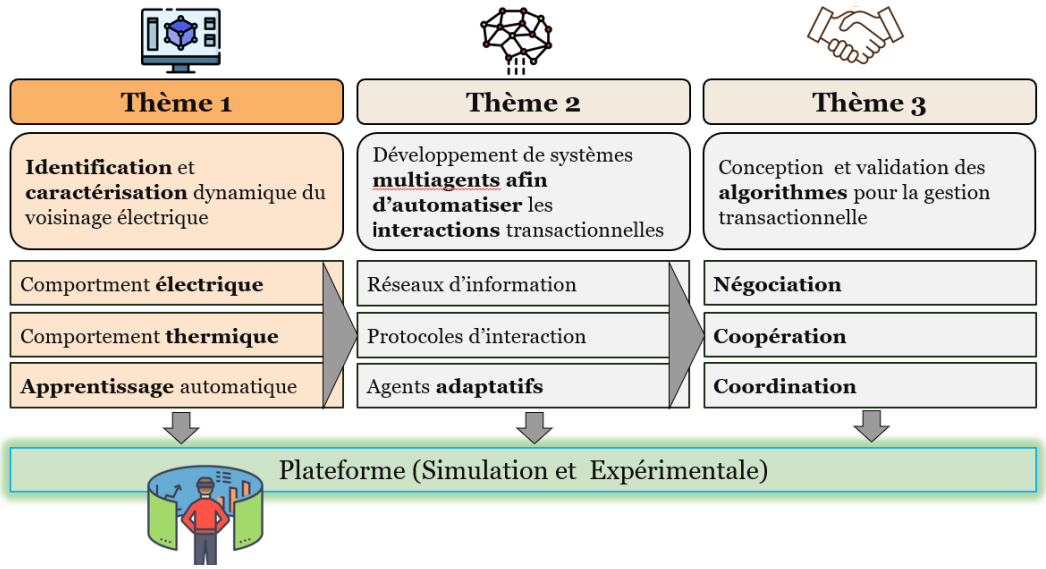

Figure 1.5 Projet de Chaire Hydro Quebec

Il comprend une partie du volet concernant le développement d'une connaissance approfondie du système énergétique résidentiel québécois afin de réaliser une gestion adaptative de la demande locale. Dans la chaire de HQ, il y a trois composantes, dont l'une concerne l'étude du voisinage d'un groupe de résidences. Dans ce projet, nous nous concentrons pas sur le comportement à l'intérieur des résidences, mais plutôt sur la

prévision en prenant en compte uniquement l'extérieur des résidences. Dans les échanges avec Hydro-Québec, il existe une plateforme de simulation expérimentale, et mon modèle sera ainsi inclus dans cette plateforme.

Des mesures ont été réalisées dans des bâtiments résidentiels québécois à Trois-Rivières afin d'obtenir certaines informations utiles à la validation du modèle proposé. Les données utilisées dans cette étude comprennent des observations horaires de la consommation d'électricité domestique. Ces données proviennent de 20 maisons réelles et de 1000 maisons simulées à partir du parc virtuel. Elles incluent des mesures telles que la température, les radiations solaires et l'humidité relative de la région de Trois-Rivières, dans la province du Québec. Toutes ces données ont été enregistrées chaque heure tout au long de la période 2018-2019 par Hydro Québec, la plus grande compagnie de production et de distribution d'électricité au Québec. La température est exprimée en degrés Celsius (°C), l'humidité relative en pourcentage (%), et les radiations solaires en watt par mètre carré (W/m^2). Nous avons spécifiquement retenu les variables climatiques mesurées à Trois-Rivières.

1.5 Méthodologie

Une planification de la gestion de la demande ou de la gestion transactionnelle avancée de la consommation résidentielle nécessite une connaissance approfondie des modèles de prévision. Bien que la structure générale des modèles de prédiction de la consommation résidentielle est construite à partir de certains modules, nous développons dans le cadre de ce projet des techniques proposées dans la littérature en validant les résultats par des simulations. De ce fait, la première étape de cette thèse consiste à définir le problème de prédiction d'un agrégat de résidences. Cette première phase concerne l'exploration du problème de prévision posé à travers la recherche bibliographique reliée à ce domaine

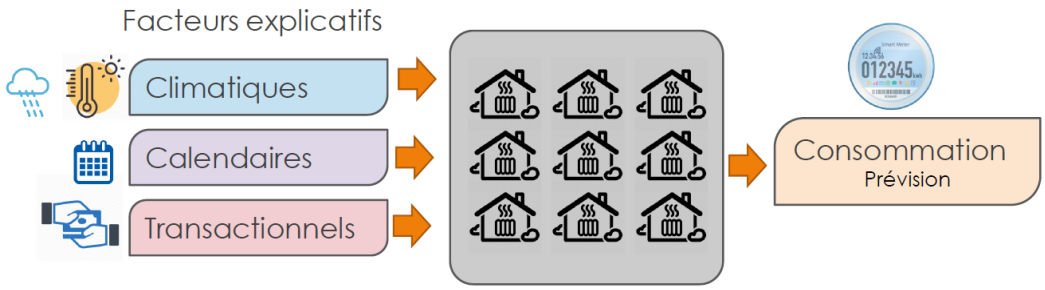

Figure 1.6 Schéma de la méthodologie de recherche.

d'étude. Ensuite, il sera pertinent d'étudier l'état de l'art sur les stratégies de prévisions utilisées pour développer des modèles efficaces, capables d'anticiper la demande en énergie des charges se basant sur les travaux déjà réalisés dans la littérature. La deuxième étape consiste à proposer et implémenter des algorithmes de prévision à court terme pour les composantes dépendantes du climat et de la saisonnalité de la demande. Une autre phase consiste à choisir les données d'entrées et déterminer les facteurs significatifs dans la consommation par une étude de sensibilité sur les données météorologiques. Cette étude nous donne un aperçu sur l'environnement climatique des bâtiments (température, humidité, exposition au vent et au soleil) ; qui contribue avec un important pourcentage dans la variation du profil de puissance agrégé. Justement, ce développement entraîne des améliorations des prévisions à court terme plus efficace. Procéder également à une analyse d'incertitude et de sensibilité des prévisions des données météorologiques et réduire les dimensions des données nous amène à réduire le problème de la complexité algorithmique et du temps de calcul trop élevé. Finalement, on appliquera ces prévisions dans le cadre d'une application de gestion transactionnelle en incluant le prix de consommation comme variable exogène. Un schéma résumant la méthodologie de recherche suivie dans ce projet est montré à la Figure 1.6 Une dernière phase correspond à comparer la précision et la performance de prévision de l'approche proposée avec d'autres méthodes proposées dans

la littérature et valider les résultats par des simulations.

1.6 Structure de la thèse

Le travail entrepris dans cette thèse se concentre sur le développement d'une méthodologie robuste pour la modélisation et la prévision à court terme de la demande d'électricité. Chaque chapitre contribue de manière significative à la compréhension et à l'amélioration de la modélisation et de la prévision à court terme de la demande d'électricité. Voici une brève présentation de chaque chapitre :

1. Introduction (Chapitre 1) : Le premier chapitre introduit de manière générale le sujet de la thèse, en exposant la problématique, les objectifs et la méthodologie de recherche. Ce chapitre établit les bases nécessaires pour comprendre le contexte global de la thèse.
2. État de l'art (Chapitre 2) : Le deuxième chapitre propose une revue de littérature sur les modèles de prévision de la demande énergétique à court terme. Il explore les stratégies d'acquisition des données sur le comportement humain, mettant en évidence les différentes approches utilisées pour la classification comportementale dans la gestion locale de la demande. En plus de décrire les différentes catégories d'horizon de prévision, ce chapitre aborde les problèmes généraux liés à la prévision énergétique, mettant particulièrement l'accent sur l'application de l'apprentissage statistique dans la modélisation énergétique des bâtiments. L'approche choisie pour cette étude est justifiée, et la structure complète des modèles proposés pour la prévision de la demande est présentée.
3. Présentation des résultats par articles (Chapitre 3) : Le troisième chapitre détaille les approches novatrices développées pour atteindre les objectifs fixés. La première section expose la description statistique du modèle, illustrée par l'article "A

compositional kernel based Gaussian process approach to day-ahead residential load forecasting". L'analyse s'appuie sur des données de demande électrique de résidences réelles à Québec, évaluant la relation entre la demande et la température extérieure. Une évaluation de la performance est réalisée, avec une attention particulière portée à la sensibilité du modèle aux données d'entrée. La deuxième section traite de la classification pour améliorer le développement comportemental des résidences, comme présenté dans l'article "Consensus-Based Time-Series Clustering Approach to Short-Term Load Forecasting for Residential Electricity Demand". La troisième section explore les incertitudes liées au modèle de prévisions, exposées dans l'article "Uncertainty Quantification in Load Forecasting for Smart Grids Using Non-parametric Statistics". Cette partie se penche sur la quantification des incertitudes de prévision des profils de puissance au niveau de l'agrégateur. La méthodologie proposée introduit une approche basée sur un modèle pour fournir une représentation plus complète de l'incertitude et l'étude des variations de charge.

4. Discussion et perspectives (Chapitre 4) : Le quatrième chapitre est dédié à la discussion des résultats obtenus à travers les méthodes proposées. Il évalue les contributions dans divers scénarios, synthétise les travaux réalisés, expose quelques limitations et défis rencontrés. Cette section ouvre également la voie aux opportunités futures de recherche et d'amélioration des méthodes développées. Ensuite ce chapitre synthétise les principales contributions, les limites de la recherche et propose des perspectives futures. Ce dernier chapitre clôture la thèse en présentant des conclusions solides tirées des résultats obtenus. Des recommandations sont formulées pour orienter les futurs développements et améliorations des deux méthodes proposées pour la prévision agrégée.

Chapitre 2 - État de l'art

2.1 Généralités sur les modèles de prévisions à court terme

La prévision de la consommation à court terme (STLF en anglais) anticipe la demande des utilisateurs dans un futur proche, ce qui donne les informations clés pour prendre des décisions dans la gestion de la demande. Les prévisions peuvent être appliquées à une seule ou à un ensemble de résidences. Toutefois, les charges à différents niveaux d'agrégation peuvent être obtenues en agrégeant différents nombres de compteurs intelligents et une grande présence d'équipements dans les résidences. Comme le caractère aléatoire est atténué après l'agrégation, la prévision de la charge devient dans ce cas plus facile. Par exemple, si les résidents partent en vacances demain, cela entraînera des changements dans le profil de charge considérables qui ne peuvent pas être déduits des données historiques de puissance sans nouvelles informations. Quelles que soient les techniques de prévision utilisées, les erreurs de prévision des charges d'une seule résidence sont toujours importantes. D'une part, la prévision à court terme pour un agrégat de résidences peut contribuer à la satisfaction de la demande d'électricité des consommateurs et à la réduction des risques de défaillances ou pannes. Si par exemple le résultat de la (STLF) indique que la demande d'électricité des utilisateurs excédera la capacité du système dans une zone résidentielle, la compagnie d'électricité peut inciter les utilisateurs résidentiels à changer leur consommation d'électricité en augmentant le prix de l'électricité. D'autre part, la (STLF) peut contribuer à faire profiter l'utilité en calculant des stratégies de tarifs optimaux pour la consommation d'électricité résidentielle pour donner suite aux résultats de (STLF). Dans la littérature, les techniques basées sur les données ont été largement utilisées pour modéliser et prévoir les profils de charge électrique agrégée résidentiels. La prévision de la charge est essentielle pour améliorer les systèmes de gestion de la demande

(DSM). Elle augmente la fiabilité des systèmes électriques et a de multiples applications dans les réseaux intelligents, comme la planification de la capacité pour réduire les pics de charge dans les environnements de réseau intelligent. Étant donné que la consommation quotidienne est considérablement influencée par les conditions météorologiques, il est nécessaire d'étudier la corrélation entre la consommation d'électricité et les variables climatiques. Plusieurs méthodes ont été proposées dans la littérature pour traiter les questions ci-dessus. Une première approche analyse la prévision de la consommation d'énergie résidentielle en utilisant uniquement la température. De nombreuses études considèrent d'autres facteurs climatiques qui influencent le plus la demande d'électricité. Une autre approche analyse la prévision de la consommation d'énergie résidentielle en utilisant uniquement une seule maison d'autres de plusieurs maisons. Notre choix se porte donc sur le modèle de prévision basée sur PG. Le PG est choisi pour la prévision des séries temporelles en raison de sa capacité exceptionnelle à capturer automatiquement des structures complexes au sein des données.

2.2 Méthodes de prévision basées sur l'apprentissage automatique

L'analyse de prévision débute avec un modèle, lequel peut être déterministe ou probabiliste. Il peut aussi prendre la forme d'un modèle unique ou résulter de la combinaison de deux ou plusieurs modèles, communément appelés modèles hybrides. Par ailleurs, l'approche bayésienne intègre le modèle probabiliste en tant qu'hypothèse sur le processus de génération à l'origine des données observées. Ce modèle constitue une structure décrivant les interdépendances entre ses composants. En fonction du nombre de paramètres, deux stratégies couramment employées pour la modélisation de la prévision résidentielle incluent des modèles paramétriques et d'autres, non paramétriques.

2.2.1 Apprentissage supervisé

L'apprentissage automatique supervisé est un domaine dont l'intérêt majeur est le développement des algorithmes permettant à une machine d'apprendre à partir d'un ensemble de données étiquetées.

- **Modèles paramétriques:** Les modèles paramétriques peuvent être caractérisés par une ensemble fini de paramètres. La manière dont ces derniers sont définis et utilisés peut donner lieu à des modèles statistiques et physiques. Les modèles physiques peuvent être plus ancrés dans les principes fondamentaux des sciences physiques, tandis que les modèles statistiques s'appuient souvent sur des distributions probabilistes pour décrire les relations entre les variables. Cette stratégie englobe les modèles statistiques qui déduisent les informations sous-jacentes d'un ensemble de données en utilisant un nombre fini de paramètres. Les modèles statistiques paramétriques supposent une distribution spécifique pour les données et incluent des paramètres numériques qui définissent cette distribution. Par exemple, dans un modèle linéaire simple, les paramètres sont les coefficients de la régression. Plusieurs algorithmes de cette catégorie ont été utilisés pour la prévision de la charge. La moyenne mobile intégrée autorégressive et le modèle saisonnier (ARIMA) et (SARIMA) ont été exploités pour examiner la corrélation entre la charge et les variables météorologiques dans [18] et [19] respectivement. En outre, des régressions linéaires multiples et simples, polynomiales et de moindres carrés ordinaires (OLS) ont été envisagées dans le but prédire la charge à court terme comme l'explique la référence [20]. Il y a aussi les modèles physiques paramétriques qui reposent sur des équations mathématiques qui décrivent le comportement d'un système physique. Ces équations peuvent contenir des paramètres qui représentent des propriétés physiques du système,

comme la masse, la résistance, etc. Par exemple, les modèles thermodynamiques et les modèles HVAC [21].

Les modèles hybrides pour la prévision de la consommation d'électricité combinent deux modèles de prévision de natures différentes. Afin d'améliorer la précision de la prévision de la consommation résidentielle à court terme, diverses méthodes de (STLF) ont été introduites récemment, y compris les réseaux de neurones combinésavec la transformée en ondelettes (WT), la machine à vecteurs de support (SVM), la logique floue (FL), et l'algorithme génétique (GA) respectivement [22–25].

Les réseaux de neurones (Artificial Neural Network, ANN en anglais) sont classés parmi les modèles paramétriques. Le plus grand avantage de l'ANN est l'efficacité de la modélisation de la dépendance temporelle. Cependant, les résultats du calcul sont susceptibles de fluctuer entre les optima parce que le modèle ANN simple est sensible à la topologie, à la sélection des valeurs de poids et aux valeurs seuil. On cite par exemple le modèle réseau neuronal de régression généralisée (GRNN) [26], le modèle autorégressif de réseau de neurones (NNANR) [27] et le modèle de perceptron multicouche MLP [28].

- **Modèle non paramétrique :** Les approches non paramétriques ne spécifient pas de forme stricte pour la relation entre les entrées et les sorties des modèles. Elles peuvent réduire l'intervention humaine, car elles fonctionnent efficacement avec un minimum d'informations. Ces techniques sont un moyen efficace de capturer la nature stochastique de la consommation d'énergie, car elles ne sont pas limitées à un nombre fixe de paramètres [29]. Ces méthodes sont capables de traiter la relation non linéaire entre la charge agrégée et les facteurs exogènes, malgré de

fortes incertitudes. Une solution courante consiste à utiliser des méthodes basées sur le Support Vector Regressor (SVR) en tant que classificateur. Par exemple, [30] et al. ont appliqué cette méthode de prévision de la consommation d'électricité en utilisant le classificateur SVR, formé à l'aide des modéles Kernel ridge regression et second exponential smoothing. En fait, les documents mentionnés font face à des défis lors du traitement de volumes de données significatifs liés à de multiples facteurs, accompagnés de leurs informations historiques. Parmi ces modèles, on retrouve ceux qui reposent sur les chaînes de Markov, l'estimation de la densité par noyau (Kernel Density Estimation, KDE), ainsi que PG. Un PG est défini comme une collection de variables aléatoires telles que tout sous-ensemble fini de celles-ci soit distribué selon une loi gaussienne [31]. Il permet de spécifier une distribution a priori sur les fonctions réelles f, représentées par $f(\mathrm{x}) \sim \mathcal{PG}(m(x), k(\mathrm{x}, \mathrm{x}'))$, où $m(x)$ est la fonction moyenne et $k(\mathrm{x}, \mathrm{x}')$ fournit la covariance entre les valeurs de fonction à deux points de données x et x$'$. La fonction noyau détermine diverses propriétés de la fonction telles que la stationnarité, la régularité, etc. Une fonction noyau populaire est la fonction noyau exponentielle quadratique (RBF) (noyau exponentiel carré), car elle peut modéliser n'importe quelle fonction régulière. Elle est donnée par $\sigma_f^2 \exp\left(-\frac{1}{2\kappa}\|\mathrm{x} - \mathrm{x}'\|^2\right)$, où l'échelle de longueur κ détermine les variations des valeurs de fonction à travers les entrées. L'algorithme de découverte de structure utilisé par le modèle PG démontre son efficacité en récupérant à la fois des structures connues et plausibles à partir de données synthétiques et réelles. Dans le contexte des ensembles de données de séries temporelles, les noyaux appris par le modèle PGs fournissent des décompositions perspicaces de la fonction sous-jacente, facilitant une extrapolation précise au-delà des plages observées. De plus, les noyaux découverts automatiquement par le modèle PG

présentent des performances supérieures par rapport à diverses classes de noyaux largement utilisées et des méthodes de combinaison de noyaux dans la modélisation de prédiction supervisée.

Un processus gaussien (PG) est un modèle statistique utilisé pour décrire une collection de variables aléatoires, dont n'importe quel sous-ensemble a une distribution conjointe gaussienne (normale). Les processus gaussiens sont largement utilisés dans les domaines tels que l'apprentissage automatique, les statistiques bayésiennes et la modélisation des séries temporelles en raison de leur flexibilité et de leur capacité à modéliser des phénomènes complexes [32]. Le PG peut être défini par la fonction de covariance qui est utilisée pour décrire les relations entre les entrées. Rasmussen et Williams ont fourni une description mathématique détaillée des modèles de PG et de leur mise en œuvre. Les modèles PG ont déjà été utilisés pour prévoir l'énergie éolienne [33] et prédire le prix de l'électricité [34]. Il a également été utilisé pour classifier les profils de consommation d'électricité [35], et pour modéliser la réponse des ménages au signal de réponse à la demande (DR) d'un agrégateur [36]. Les principaux travaux réalisés dans la littérature considérant de façon très détaillée le PG, mais ces travaux sont relativement peu nombreux. Ces auteurs ont proposé plusieurs types d'analyse dans le but d'avoir un cadre général pour fournir une prévision de la densité de probabilité de la charge électrique en appliquant le modèle de régression quantile du processus gaussien [33]. D'autres approches proposées comportent la prévision avec la régression basée sur PG [37]. Les modules composant l'algorithme PG général sont spécifiés dans [32]. Les études réalisées par [38], [39], [40] ont utilisé un processus gaussien pour prévoir la consommation d'électricité pour l'optimisation ainsi que l'apprentissage actif multitâche et de

l'apprentissage par renforcement (Reinforcement learning) dans la littérature. Cette méthode du processus gaussien inscrit dans le cadre du formalisme bayésien, elle génère des données permettant une solution basée sur les maximum posteriori (MAP) afin de prédire la consommation d'un ensemble de résidences [41]. D'autre part, [42] utilise le PG pour prédire la consommation résidentielle, mais considérant seulement la température et le jour de la semaine comme facteurs d'entrées. Tandis que [41] utilise le PG, mais ne prend pas en considération l'effet des variables climatiques. Dans la littérature il existe différents modèles PG pour prédire la puissance le PG généraliste, PG pondéré [43], le double PG, hiérarchique PG [44]. Par ailleurs, afin de déterminer la méthode la plus adaptée au problème de prévision concerné dans cette thèse, nous avons basé l'analyse sur l'approche du processus gaussien additif qui sera décrite dans la suite.

2.2.2 Apprentissage non supervisé

Le principal bénéfice de l'apprentissage non supervisé est qu'il peut être utilisé avec des données non étiquetés. Il en résulte des solutions qui peuvent être déployées en fonction de l'environnement dans lequel elles fonctionnent, ce qui simplifie considérablement l'ensemble du processus.

2.3 Algorithmes de prévisions d'un agrégat de résidences

2.3.1 Approches de développement des modèles de prévision

La conception d'un système de prévision de la charge à court terme (STLF) efficace requiert une compréhension approfondie de la relation entre la consommation d'électricité et ses éléments descriptifs [45]. Dans les zones résidentielles, la charge associée aux conditions météorologiques occupe une part significative de la demande totale, étant

donné la forte corrélation entre les conditions climatiques et la consommation d'électricité des ménages. Cette corrélation peut être particulièrement prononcée dans des zones géographiques spécifiques, où elle devient essentielle pour l'estimation et la prévision de la demande. Par exemple, dans des régions comme le Québec, objet de cette étude, une part importante de la charge totale est influencée par les conditions météorologiques, en raison des longues saisons froides [46]. Cette interaction confère une priorité à la composante sensible au climat par rapport aux autres facteurs dans le contexte d'un système STLF [47].

D'autre part, la charge liée au calendrier représente également une portion notable de la consommation énergétique résidentielle. Bien que la nature incertaine de ce facteur non météorologique puisse complexifier le processus de prévision, sa relation significative avec la consommation d'énergie justifie son inclusion dans le cadre d'une STLF efficace [48]. Les variables calendaires sont généralement influencées par les activités des occupants, telles que l'ajustement des températures de consigne (en lien avec les niveaux de confort) et les gains de chaleur internes. En conséquence, les approches de STLF s'efforcent d'estimer le comportement de la consommation d'énergie en modélisant les impacts de ses composants explicatifs, principalement en tenant compte des facteurs sensibles au climat et des facteurs calendaires [49]. La littérature propose diverses approches pour la prévision STLF résidentielle, souvent basées sur des méthodes paramétriques et non paramétriques. La catégorie paramétrique englobe les modèles statistiques qui déduisent les informations sous-jacentes d'un ensemble de données en utilisant un nombre fini de paramètres. En revanche, la catégorie non paramétrique, une autre classe statistique, ne considère pas de modèle avec une structure finie.

Dans le domaine de la STLF, divers algorithmes issus de ces modèles statistiques ont été étudiés. Les analyses de régression, telles que les moyennes mobiles intégrées

autorégressives (ARIMA) et les moyennes mobiles intégrées autorégressives saisonnières (SARIMA), sont populaires dans la classe paramétrique [50]. Des schémas communs, tels que les régressions multiples, polynomiales et linéaires simples, ont également été utilisés à cet effet [51]. Ces méthodes visent à améliorer la STLF en abordant les défis liés à la nature dynamique de la charge résidentielle, comme illustré par [52].

Par ailleurs, des algorithmes typiques de la famille non paramétrique, tels que la régression de noyau (Nadaraya-Watson), l'algorithme k-Nearest Neighbors (k-NN), et la régression vectorielle de soutien (SVR), ont également été explorés pour la STLF [53–55]. Les avantages des méthodes non paramétriques, notamment la réduction de l'intervention humaine et leur capacité à capturer la nature stochastique de la consommation d'énergie, les rendent attrayantes. Les modèles de processus gaussiens (PG) sont particulièrement prometteurs pour l'estimation, la prévision et le contrôle de la consommation d'électricité [56], [37].

Plusieurs études ont exploré l'application des modèles PG pour la STLF. Dans [57], un apprentissage par transfert basé sur un modèle PG a été utilisé pour améliorer la précision de la prédiction. Cependant, ce modèle n'a pas pris en compte la dépendance de la consommation d'électricité à l'égard de facteurs externes tels que les conditions météorologiques. D'autres travaux, [41] et [42], ont combiné PG avec différentes approches pour aborder divers aspects de la STLF. Une limitation fréquemment rencontrée dans ces études est la petite quantité de données historiques utilisées pour l'entraînement.

Il est important de souligner que les processus hiérarchiques peuvent parfois négliger l'interaction entre les éléments descriptifs de la consommation d'énergie. Dans ce contexte, la composition des noyaux pour caractériser l'interaction entre les variables d'entrée s'est révélée efficace, notamment dans les modèles PG pour la STLF [58], [59].

Des études, telles que [60], ont exploité des PG basés sur des noyaux de composition pour effectuer une régression non paramétriques pour la prédiction de la charge, bien que leurs interactions d'entrée étaient limitées. En résumé, malgré les promesses des systèmes STLF basés sur les PGs, de nouveaux développements sont nécessaires pour résoudre les défis mentionnés ci-dessus.

2.3.2 Modélisation avec le Processus Gaussien Additif

Outre le choix du modèle, le niveau d'agrégation de la consommation électrique est un autre élément essentiel qui influence l'efficacité d'un cadre de prévision à court terme (STLF). Normalement, la prévision de la charge est appliquée à la demande totale d'électricité des maisons individuelles ou combinées [61]. Deux scénarios différents ont été proposés pour traiter ce dernier cas. Dans le premier scénario, une prévision globale est obtenue à partir de la somme des prévisions au niveau des maisons. À ce sujet, les auteurs de [62] ont déclaré que le premier scénario était supérieur. En revanche, dans la seconde stratégie, une prévision unique est obtenue à partir d'une prévision globale, comme indiqué par [63], qui a affirmé l'inverse. Néanmoins, le choix entre ces deux stratégies dépend du comportement spécifique de la consommation d'énergie dans les maisons cibles. La prévision unique peut être une option judicieuse pour les habitations présentant des schémas de consommation similaires. D'autre part, les prévisions globales peuvent être plus appropriées pour les habitations dont les comportements de consommation sont différents. En effet, la sommation des charges individuelles dans ce cas peut conduire à une prévision unique imprécise [64].

Tableau 2-1 Les éléments pertinents des procédures de prévision basées sur le PG selon la littérature.

	Utilisation des big data	Prétraitement des données	Estimation des hyperparmètres	Modélisation de l'interaction des noyaux	Analyse de données périodiques (calendrier)	Point de changement
[65] [66]	✘	✘	✘	✔	✔	✔
[41] [59]	✘	✔	✘	✘	✔	✘
[60] [67]	✔	✔	✔	✘	✘	✘
[68] [69]	✘	✘	✔	✔	✔	✔
PG	✔	✔	✔	✔	✔	✘

Par conséquent, les méthodes STLF proposées pour estimer la consommation totale d'énergie d'un ensemble de résidences doivent tenir compte du comportement de chaque profil de charge. L'impact du niveau d'agrégation sur la performance des prévisions peut être étudié plus en détail dans [70]. Cependant, le tableau2-1 présente un aperçu des éléments pertinents des procédures de prévision basées sur les PG selon la littérature. Chaque référence spécifique met l'accent sur certains aspects tels que l'utilisation des big data, le prétraitement des données, l'estimation des hyperparamètres, la modélisation de l'interaction des noyaux, l'analyse de données périodiques (calendrier), et la détection de points de changement (la détection de points de changement dans les procédures de prévision basées sur les Processus Gaussiens). Les symboles ✘ ou ✔ signalent si une référence particulière englobe un aspect particulier. Si la référence prend en considération

le point mentionné, alors c'est un ✔, sinon, c'est l'autre. Selon ce tableau, nous pouvons constater que notre travail englobe tous les critères mentionnés excepte le point de changement. Cette diversité souligne l'adaptation de ces éléments en fonction du contexte spécifique de chaque approche basée sur les PG dans la littérature.

2.4 Algorithmes de classification pour un agrégat de résidences

Dans la section précédente, nous avons examiné les impacts de la température extérieure sur la consommation énergétique, soulignant sa forte influence sur la consommation électrique québécoise, notamment en raison de la prépondérance du chauffage électrique pendant les périodes de froid hivernal. Cependant, anticiper la demande énergétique est essentielle pour administrer efficacement la consommation d'énergie des habitations dans les quartiers résidentiels. Ces zones sont soumises à une dynamique de demande d'électricité influencée par divers phénomènes, dont les conditions climatiques et les préférences individuelles de confort des occupants. La nature incertaine de ces circonstances, associée à des paramètres aléatoires, se traduit par des profils de puissance présentant des caractéristiques diverses.

Dans ce contexte, la prévision de la charge globale est préconisée l'aide de la méthode de la prévision agrégée basée sur les classes (Cluster-based Aggregate Forecast - CBAF). Ainsi, durant cette thèse, nous proposons une approche novatrice qui combine des techniques de machine learning non supervisées pour élaborer un système de classification de séries temporelles. Cette structure de classification utilise un algorithme basé sur la distance dynamique, appelée Dynamic Time Warping (DTW), pour mesurer la similarité entre les profils temporels. Par la suite, nous exploitons les avantages des processus gaussiens (AGP), une technique conçue pour la prévision non paramétrique, afin de prédire la charge globale à chaque niveau de regroupement résidentiel. Notamment, un processus

gaussien à composition est employé pour fournir des prévisions efficaces pour un agrégat de résidences. Dans le contexte des prévisions pour un agrégat de résidences, cela signifie que le processus gaussien est capable de prendre en compte les relations et les dépendances entre les différentes résidences de manière plus sophistiquée.

Une étude comparative démontre que cette approche combinée peut anticiper la charge résidentielle totale avec une précision accrue. De plus, elle souligne l'importance de mécanismes efficaces à la fois pour le groupement initial et la phase de prévision, soulignant ainsi les nécessités d'un CBAF adéquat. Dans le secteur résidentiel, les prévisions globales sont couramment réalisées pour des groupes de résidences. Cependant, lors de ces processus d'agrégation, l'information concernant les caractéristiques des modèles de charge peut souvent être perdue. Les fluctuations des données font référence aux variations ou aux changements dans les données au fil du temps ou de l'espace. Ces fluctuations peuvent être dues à une variété de facteurs, tels que des événements aléatoires, des tendances saisonnières, des changements dans les comportements des consommateurs, ou des perturbations dans l'environnement. Elles sont souvent une caractéristique inhérente des données réelles et peuvent avoir un impact significatif sur l'efficacité du processus de prévision [12]. Des travaux connexes démontrent que les approches CBAF peuvent améliorer la précision des prévisions pour les charges agrégées en réduisant les caractéristiques stochastiques des profils de charge, permettant ainsi un entraînement plus performant de l'algorithme sur des données fortement corrélées [13]. L'augmentation de la précision des prévisions grâce à la CBAF contribue clairement à l'optimisation de l'énergie transactionnelle, améliorant ainsi la gestion des revenus. Étant donné que les prévisions de séries temporelles et la classification sont des méthodes standard facilitant la prise de décision, il est avantageux de regrouper les consommateurs en groupes plus restreints en fonction de leurs caractéristiques communes. Les stratégies

de classification présentent le potentiel d'apporter une amélioration statistiquement significative de la précision par rapport aux méthodes traditionnelles de prévision agrégée, dépendamment du nombre de classifications et de la taille de la base de données. Ainsi, le CBAF peut offrir un éclairage supplémentaire aux praticiens cherchant à mettre en œuvre cette stratégie dans des scénarios réels, conduisant à des améliorations substantielles des prévisions de charge pour des ensembles de données classifiées. De plus, pour intégrer des perspectives complémentaires des données en une partition plus stable, la classification par consensus émerge comme un élément robuste [15] [16]. Cette approche permet d'obtenir des partitions stables tout en gérant la diversité, générant ainsi des sous-ensembles d'attributs cohérents [17].

Contrairement aux méthodes traditionnelles qui utilisent l'agrégation pour un ensemble de maisons présentant des comportements différents [20], l'établissement de limites fines pour la classification des maisons en fonction de leur comportement s'avère complexe [71]. Ceci souligne la nécessité d'une méthode de classification fiable. Un aspect préoccupant est que la prévision de la charge doit s'effectuer sur des données de séries temporelles [72]. La robustesse face aux séries temporelles incomplètes devient donc essentielle, en particulier compte tenu de la possibilité que l'ensemble de données couvre différentes périodes. Cette complexité limite la capacité de regrouper un groupe d'utilisateurs selon des séries temporelles synchronisées [73]. De plus, les données des séries temporelles, naturellement bruyantes ou stochastiques, entravent la distinction claire des classes. Les techniques de classification bien connues montrent une prévision inefficace en raison de l'accumulation d'erreurs liées au comportement stochastique des utilisateurs [74].

Tableau 2-2 Les éléments pertinentes de la prévision des charges avec classification selon la littérature

Références	**Type des données**	**Méthodes de classification**	**Critères de distance**	**Méthodes de prévision**
[75] [76]	Séquence de motifs de classification	Classification SOM	Euclidienne	ANN/PSF
[77] [78]	Structure multizone	Basé sur une grille	Euclidienne	LSTM
[12] [79] [80]	Stratégie de température pour HVAC	Hiérarchique CBAF	Euclidienne	ARIMA/DNN
[81]	Différents paramètres du système électrique	Micro clustering	Rapport de distance du cluster	RNN/LSTM
[82] [82]	Séries chronologiques	k-means et CBAF	Euclidienne	Bootstrapping/Autres modèles
[83] [10]	Immeubles résidentiels à plusieurs étages	Spatial-Temporal et k-means	Distance Hybride	LSTM-GRU/e-HMM
[84]	Données de charge mesurées toutes les heures	k-means and kernel k-means	Dynamic time warping	LightGBM
Ce travail	Accumulation de séries temporelles	Regroupement consensuel et CBAF	DTW	Processus Gaussian Additif

La littérature offre diverses approches pour la prévision à court terme de la demande d'électricité résidentielle, en se concentrant sur la CBAF. Les travaux de recherche menés par Cini et al. [12] et Wijaya et al. [79] mettent en œuvre une prévision agrégée plus précise en regroupant les charges, prévoyant les groupes de manière distincte, puis agrégeant les estimations. Cependant, d'importantes fluctuations de charge peuvent entraîner des écarts prévisionnels réduisant la précision. Chen et al. [84] ont proposé une méthode hybride en plusieurs étapes pour améliorer la performance prévisionnelle dans un cadre stochastique.

Les techniques de classification basées sur les réseaux neuronaux artificiels [75], [76] ont également trouvé leur place dans la prévision de la charge. Tandis que Mandal et al. [75] dépendent de la norme euclidienne, Jin et al. [76] utilisent une carte auto-organisatrice pour la classification. Kong et al. [78] ont classé la charge en différents niveaux en fonction de la proportion correspondante dans la charge totale, exploitant la méthode LSTM (Long Short-Term Memory) pour sélectionner les jours similaires et agréger les prévisions. Un modèle basé sur le réseau neuronal à fonction de base radiale pour la STLF a trouvé application [85], utilisant un système d'inférence adaptatif pour ajuster les résultats en fonction des changements récents des prix en temps réel. Cao et al. [80] ont proposé un mécanisme basé sur la méthode des jours similaires pour prédire la charge en regroupant le jour cible avec des similitudes météorologiques historiques.

Étant donné la variété des profils de consommation d'énergie des bâtiments, l'incorporation d'étiquettes à diverses routines peut améliorer la précision de la prédiction [86]. Intégrer des étiquettes aux différentes routines ou habitudes de consommation peut aider à mieux caractériser ces schémas et à capturer les variations dans les données de consommation d'énergie. Cela pourrait être bénéfique par l'identification des schémas spécifiques. Les PG sont un choix populaire pour la modélisation des séries

chronologiques en raison de leur capacité à gérer des structures complexes [33]. Bien que les PG offrent des avantages attrayants, leur mise en œuvre pour des performances optimales reste un défi, nécessitant une sélection judicieuse des caractéristiques d'entrée [79], [87]. Rouwhorst et al. [88] utilisent des algorithmes de machine learning pour des approches de classification basées sur des prévisions de charge agrégées en se focalisant sur le moyen terme (MTLF) avec un horizon de prédiction entre 2 semaines et 2 ans. Ils appliquent différentes techniques de sélection de caractéristiques à la charge des transformateurs, privilégiant une approche basée sur la distance euclidienne.

Le tableau 2-2 récapitule les études récentes dans la prévision de la demande et les techniques de classification associées. Ces études révèlent que les approches de classification par consensus. Ces techniques de classification par consensus sont conçues pour combiner les résultats de plusieurs méthodes de classification afin de produire des prédictions plus robustes et plus fiables. Ces approches exploitent la diversité des méthodes de classification pour compenser les faiblesses individuelles de chaque méthode et exploiter leurs forces respectives. Cependant, plusieurs techniques nécessitent encore une validation sur des données de consommation réelles. De plus, les méthodes traditionnelles de classification basées sur la distance euclidienne présentent des limitations, ne prenant en compte que les distances ponctuelles, tandis que l'utilisation de Dynamic time warping (DTW) peut résoudre les problèmes de distorsion temporelle inhérents aux données de séries temporelles. Il est important d'évaluer régulièrement la qualité des clusters pour s'assurer qu'ils capturent correctement la structure des données. Cela peut être fait en utilisant des mesures telles que la silhouette, l'indice Davies-Bouldin, ou d'autres mesures de similarité entre clusters. En résumé, lors de l'utilisation de méthodes de clustering dans le temps, il est essentiel de prendre en compte la stabilité des clusters, la dynamique des données et d'utiliser des méthodes appropriées pour modéliser

et évaluer les clusters dans le temps. Cela garantira que les clusters identifiés capturent correctement la structure des données et sont utiles pour l'analyse et la prise de décision. non seulement ceci mais aussi la performance de regroupement dans l'amélioration des previsions.

2.5 Synthèse du chapitre

La réalisation de l'objectif visant à développer une approche probabiliste pour la modélisation des profils résidentiels est le point principale de ce chapitre. La proposition avancée repose sur la conception d'un modèle probabiliste novateur, prenant en considération la dynamique temporelle et la variabilité inhérente aux profils de présence des individus dans leur logement. Cette approche offre une perspective globale, permettant la simulation de la performance des bâtiments, la prévision de la demande énergétique, ainsi que la génération de profils de consommation des logements selon la caractérisation comportementale. L'évaluation rigoureuse de la méthode proposée a été réalisée à travers une comparaison de sa performance avec d'autres approches existantes dans la littérature. Les résultats obtenus mettent en évidence le potentiel distinctif de notre approche, démontrant de manière convaincante sa supériorité dans le contexte spécifique étudié. En outre, l'applicabilité de cette approche ne se limite pas à la modélisation résidentielle, mais s'étend également à d'autres domaines d'analyse, ouvrant ainsi la porte à des applications diverses telles que la modélisation de comportements dans des contextes variés. Une avenue prometteuse de recherche future suggérée par cette étude concerne l'exploration de l'application de la méthode pour modéliser le comportement énergétique dans des bâtiments commerciaux et institutionnels. Cette extension potentielle souligne la polyvalence de l'approche proposée, indiquant qu'elle pourrait être adaptée pour aborder des questions énergétiques dans des contextes transactionnelles.

Finalement, ce chapitre présente l'état de l'art des prévisions, démontrant la pertinence et l'applicabilité de notre approche probabiliste novatrice dans le domaine spécifique de la modélisation des profils résidentiels. Le modèle, fondé sur l'apprentissage automatique, intègre les cinq points suivants : l'utilisation d'une grande base de données, le prétraitement des données, l'optimisation des hyperparamètres, l'interaction entre les noyaux du modèle de processus gaussien additif et la prise en compte des données périodiques. En ce qui concerne la modélisation des charges résidentielles, la classification des résidences et l'analyse des incertitudes, notre méthodologie se distingue des autres travaux par son intégration de la prévision agrégée, l'agrégation de la prévision CBAF et la stabilité des classes grâce à une méthode de classification non supervisée prenant en compte la distance DTW.

Enfin, en ce qui concerne la rétrospection sur les modèles de prévision et de classification, les contraintes, les méthodes alternatives et les perspectives pour l'avenir, notre méthodologie se distingue des autres méthodes de prévision à court terme en considérant tous les facteurs exogènes, la gestion de la flexibilité et sa quantification, non seulement pour améliorer la précision, mais aussi pour gérer les attentes des utilisateurs et réduire les pics de consommation.

Chapitre 3 - Présentation des résultats par articles

Les résultats issus de la méthodologie élaborée pour atteindre les objectifs de ce projet de recherche ont été divisés en trois publications distinctes. Tout d'abord, le développement d'un modèle de prévision à court terme pour un agrégat de résidences. Ensuite, la procédure ainsi élaborée est intégrée à la caractérisation comportementale des résidences, utilisant des méthodes de classification. Enfin, la dernière publication traite des méthodes d'incertitude dans les prévisions, lesquelles incluent des informations sur la variance de la prévision appliquée sur les données, et détaille leur application spécifique dans le cadre de la flexibilité énergétique.

3.1 Modélisation des prévisions des charges résidentielles

3.1.1 Contexte

Dans cette première partie, on considère le problème de développement d'un modèle de prévision où on peut exploiter le maximum de l'historique de données sans avoir beaucoup de connaissances sur la structure des données ou les informations des occupants. Dans ce cas, le modèle proposé vise à utiliser un ensemble de composantes météorologiques et non météorologiques prédéfinies de la demande d'électricité résidentielle dans la région du Québec. En fait, différents facteurs peuvent influencer le comportement de la consommation électrique. Cependant, la prise en compte de tous ces facteurs, en particulier ceux qui sont sensibles au climat, peut augmenter considérablement la complexité de l'analyse. De plus, une telle analyse peut n'apporter que des améliorations mineures puisque certains de ces éléments n'ont qu'un très faible impact sur le profil de la demande totale. Dans cette étude, la température, l'humidité et les radiations solaires sont les variables climatiques qui ont été sélectionnées. Cependant, notre choix des composantes climatiques est le résultat d'une étude précédente, où nous avons

appliqué des analyses de sensibilité et de prévisibilité pour identifier les facteurs les plus significatifs de la demande pour une prévision efficace dans la région de Québec [89]. Le temps calendaire, qui explique les variations temporelles de l'utilisation de l'électricité, est un autre facteur pris en compte. Les variable qui sont considérés dans cette thèse sont l'heure de la journée le jour de la semaine et le temps triangulaire comme entrée de la composante calendaire. Comme mentionné, cette composante reflète le comportement des occupants à l'égard de l'utilisation de leurs appareils. D'une part, différentes mesures peuvent être envisagées pour traiter le processus de conception de STLF au moyen du modèle proposé. D'autre part, comme la taille et la diversité des données (big data)

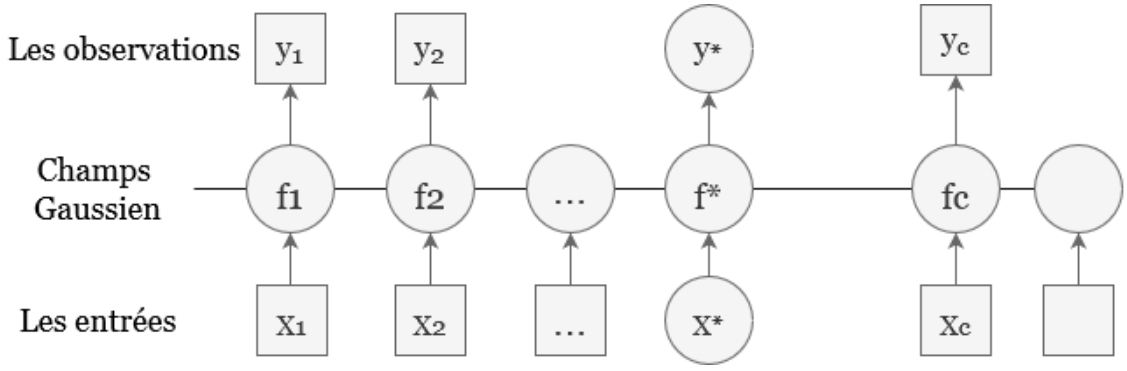

FIG. 3.1 Modèle graphique du Processus Gaussien pour régression

influencent considérablement ce processus, les informations sur les données ont également été analysées. L'originalité de notre schéma algorithmique est d'intégrer presque tous les éléments efficaces d'une prévision basée sur le processus gaussien additif (AGP) en résolvant les défis associés aux méthodes existantes pour l'inférence dans les modèles de processus gaussiens, la difficulté de l'entraînement évolutif des réseaux d'inférence et le temps de calcul élevé. Il s'agit d'un modèle graphique comme le montre la Figure 3.1, également appelée graphique en chaîne, utilisé pour la régression. La figure est reproduite en se basant sur le modèle proposé par Rasmussen et al. [32]. Les cercles symbolisent les variables à prédire. La barre horizontale épaisse représente un ensemble

de nœuds entièrement connectés.L'objectif de l'inférence bayésienne dans le modèle AGP est de calculer la distribution postérieure sur la fonction f évaluée à des entrées de test arbitraires x. Pour les vraisemblances gaussiennes telles que dans la régression, la distribution postérieure prend une forme fermée. Ainsi, la distribution prédictive en un endroit de test x_*. Il est important de noter que lorsqu'une observation y_i est prise en considération, elle est conditionnellement indépendante de tous les autres nœuds, à condition de connaître la variable latente correspondante, f_i. Cette propriété découle de la capacité de marginalisation des processus gaussiens. Par conséquent, l'ajout d'entrées supplémentaires, x, de variables latentes,la variable latente correspondante, f, et de variables non observées, y, n'affecte pas la distribution des autres variables du modèle [32].

3.1.2 Méthodologie

Le schéma fonctionnel du modèle AGP à noyau compositionnel dont on a appliqué l'addition des trois noyaux est présenté dans la figure 1 du premier article. Dans cette figure, les processus d'apprentissage et de prédiction sont représentés dans deux blocs distincts. Dans le bloc de gauche, la base de données fournit les informations qui sont utilisées pour la phase d'apprentissage. À cette étape, nous rencontrons une grande quantité de données météorologiques ou temporelles historiques qui peuvent mettre à rude épreuve la capacité de mise à l'échelle de AGP. Afin de résoudre ce problème, un processus d'estimation basé sur des sous-ensembles de données est développé. Le sous-ensemble de données (SoD) est sélectionné à l'aide d'une technique de classification qui tire parti de l'algorithme de décalage moyen (en anglais le Mean Shift Clustering). Plus précisément, le SoD présente les points d'induction comme le nombre effectif de données d'entrée, exploitées pour l'apprentissage. Par la suite, ces points sont utilisés pour approximer la matrice de covariance du modèle AGP à l'aide de la méthode Fully Independent

Training Conditional (FITC) [39]. La fonction latente et les hyperparamètres de ce modèle sont estimés par la méthode du Maximum A Posterior (MAP). L'estimation MAP est réalisée au moyen de l'algorithme MCMC en tenant compte des hyperparamètres. Comme mentionné, le modèle AGP proposé est structuré autour d'un objectif clé qui consiste à extraire les interactions d'entrée les plus efficaces en fonction de leur corrélation avec la consommation d'énergie. La formule de AGP est défini comme suit:

$$f(\mathrm{x}) \sim \mathcal{PG}\left(0, k(\mathrm{x}, \mathrm{x}')\right) \tag{3.1}$$

$$y_t | f(\mathrm{x}_t) \sim \mathcal{N}\left(f(\mathrm{x}_t), \sigma^2\right), \tag{3.2}$$

où σ^2 et $f(\mathrm{x})$ sont la variance de l'erreur d'estimation et la fonction latente, respectivement. Le terme $k(\mathrm{x}, \mathrm{x}')$ représente la fonction de covariance (ou noyau).

L'un des avantages des processus gaussiens réside dans leur capacité à conserver toutes les informations nécessaires pour décrire la distribution dans les matrices de moyenne et de covariance. De plus, ces processus bénéficient de méthodes bien établies dans la littérature pour calculer l'estimateur du maximum a posteriori (MAP) [32]. En général, ces méthodes nécessitent l'inversion d'une matrice de covariance. Dans cette optique, la décomposition de Cholesky est couramment utilisée pour cette inversion en raison de son efficacité à calculer supérieure aux méthodes conventionnelles pour les matrices définies positives. Lorsque nous avons la connaissance des choix précédemment faits par l'utilisateur dans diverses situations, la première étape du système de recommandation consiste à entraîner le processus gaussien avec ces données. Ensuite, il s'agit de déterminer dans quelles circonstances l'estimateur MAP constitue une recommandation appropriée et quand il est préférable d'explorer d'autres options.

Par conséquent, à ce niveau du processus d'apprentissage, les meilleures compositions de noyaux qui présentent ces interactions sont conservées. En conséquence, les noyaux sélectionnés sont utilisés pour construire le modèle AGP prédictif, ciblé par le mécanisme proposé. Ensuite, un processus de prévision quotidienne est effectué par l'espérance conditionnelle totale de la fonction latente de l'AGP, présentée dans le bloc de droite de la même figure. La conception de cette méthode a été focalisée sur l'aspect pratique de son intégration dans le cas d'un agrégat résidentiel avec des données réelles.

3.1.3 Résultats

L'approche proposée est basée sur un modèle de régression multivariée non paramétrique pour prédire la consommation d'énergie agrégée d'un ensemble de résidences. Un système de prévision basé sur un processus gaussien additif bayésien a été développé pour saisir les relations hautement non linéaires entre la demande de charge et les composantes sensibles au climat et au calendrier. En particulier, le modèle probabiliste proposé vise à expliquer la nature dynamique et stochastique du facteur calendaire. Afin d'évaluer la performance du modèle, les données réelles de la consommation électrique agrégée des maisons situées à Trois-Rivières ont été exploitées. Un grand avantage de l'approche proposée est qu'elle repose sur différents types de variables explicatives et catégorielles, notamment un ensemble d'entrées multivariées. Les résultats significatifs démontrent l'efficacité et la pertinence de l'approche basée sur les processus gaussiens (AGP) avec noyau compositionnel pour la prévision de la charge résidentielle. La méthode développée, centrée sur la combinaison de différents noyaux, permet une modélisation plus précise et flexible des variations complexes dans les profils de charge résidentielle. Les avantages de cette approche sont surtout mis en lumière à travers des indicateurs clés tels que la diminution des erreurs de prévision, la réduction des écarts par rapport aux

données réelles, et une meilleure prise en compte des caractéristiques comportementales. Ces résultats renforcent l'importance de l'approche basée sur les processus gaussiens avec noyau compositionnel dans le contexte de la prévision de la charge résidentielle à court terme, ouvrant ainsi la voie à des applications plus précises et adaptatives dans le domaine de la gestion de l'énergie.

Nous avons pu déterminer le comportement de la demande électrique durant les périodes froides en appliquant cette approche non paramétrique. En outre, une analyse comparative avec d'autres méthodes paramétriques et non paramétriques a été montrée pour démontrer l'efficacité du modèle proposé. Ces techniques consistent en un Long short terme memory (LSTM), une régression vectorielle de soutien (SVR) et une régression random forest (RFR). Le choix de ces méthodes spécifiques peut découler de la nécessité de couvrir différentes approches de modélisation, en mettant l'accent sur la représentation temporelle (LSTM) qui est souvent utilisé pour modéliser des séquences temporelles en raison de leur capacité à conserver la mémoire à long terme. Cela les rend adaptés à des problèmes où les données passées ont un impact significatif sur les prédictions futures. D'autre part, les méthodes paramétriques comme SVR et non paramétriques comme RFR offrent des approches différentes pour modéliser des relations complexes entre les variables d'entrée et de sortie. Cela permet de tester la performance du modèle proposé AGP par rapport à ces approches pour évaluer la performance globale de notre modèle AGP.

Energy & Buildings 254 (2022) 111459

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

A compositional kernel based gaussian process approach to day-ahead residential load forecasting

Khansa Dab [a,*], Kodjo Agbossou [a], Nilson Henao [a], Yves Dubé [b], Sousso Kelouwani [b], Sayed Saeed Hosseini [a]

[a] Department of Electrical and Computer Engineering, UQTR, Canada
[b] Department of Mechanical Engineering, UQTR, Canada

ARTICLE INFO

Article history:
Received 2 July 2021
Revised 30 August 2021
Accepted 11 September 2021
Available online 20 September 2021

Keywords:
Gaussian process
Kernels interactions
Bayesian inference method
Aggregated load forecasting
Non-parametric regression approach

ABSTRACT

Load forecasting is an expected ability of electric power networks to enable effective capacity planning. This paper proposes a probabilistic approach to short-term load forecasting (STLF) of residential power consumption. The proposed method is based on Bayesian regression modeling. It utilizes an additive Gaussian Process (GP) to estimate climate-sensitive and calendar factors of power demand. The GP model is constructed by using a set of compositional kernels that represent the most significant interactions between input variables. Such collection is built up through a sampling method, capable of selecting the n-upmost order-based interactions. Moreover, a technique is performed to deal with challenges related to multivariate input and large dataset training complexity. The forecasting model is applied to actual power consumption data of a set of houses, located in Quebec, during winter. The results demonstrate that the suggested scheme is highly efficient to model and predict residential electricity use. Furthermore, it is competitive with other forecasting algorithms, as manifested by a comparative analysis.

1. Introduction

Short Term Load Forecasting (STLF) plays an important role in power system scheduling and time-ahead electricity markets. Particularly, a daily-basis STLF is acknowledged as an essential prerequisite for day-ahead energy markets, which are used to enhance grid reliability and price stability [1].

Designing an efficient STLF system requires a sensible comprehension of the relationship between electricity consumption and its descriptive elements [2]. In the residential areas, weather-related load holds a notable share of total demand due to a strong correlation between climate conditions and household power consumption. This relationship can be intensified in specific geographical locations and become undeniably essential for demand estimation and forecasting. For instance, in geographic regions like Quebec, where this study is conducted, a significant portion of overall load is manipulated by weather circumstances [3]. In such districts, there is a consequential relationship between electricity consumption and meteorological variables due to long cold seasons [4]. This interaction gives the climate-sensitive component a priority over other factors regarding a STLF system [5]. Calendar-related load presents another notable portion of household power usage. Although the uncertain nature of this non-meteorological factor can make the forecasting process complicated, its sensible relation with power usage should also be considered by STLF [6]. The calendar variables are normally caused by occupants' activities due to for example temperature set-point adjustment (regarding comfort levels) and internal heat gain. Accordingly, STLF methods aim to estimate the power consumption behavior by modeling the impacts of its explanatory components mainly accounting for climate-sensitive and calendar factors [7].

1.1. Background

Many approaches have been proposed in the literature for residential STLF. These propositions have generally utilized data-based methods on the basis of parametric and non-parametric techniques. The parametric family is a class of statistical models that infers the underlying information of a data space by using a finite number of parameters. On the other hand, the non-parametric family, as another category of statistics, does not consider a model with finite structure [8]. STLF has been studied through different algorithms of these statistical models. Regression analyses such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) are popular statistics of parametric class, used for household STLF [9]. Multiple, polynomial, and

* Corresponding author.
E-mail address: khansa.dab@uqtr.ca (K. Dab).

https://doi.org/10.1016/j.enbuild.2021.111459
0378-7788/

straightforward linear regressions are other common schemes of this category, utilized for the same purpose [10]. These methods have been employed to improve STLF by addressing its related issues. For example, [11] has attempted to handle forecasting challenges related to dynamic nature of residential load. On the other side, Kernel regression (Nadaraya-Watson), k-Nearest Neighbors algorithm (k-NN), and Support Vector Regression (SVR) are typical algorithms of non-parametric family, exploited for STLF [12–14]. Indeed, both parametric and non-parametric schemes have proven to be successful in estimating and predicting household power consumption [15]. However, the latter is stimulated due to different advantages over the former regarding a practical household STLF. Non-parametric methods can reduce human intervention since they are able to perform with a minimum number of hyperparameters. These algorithms are good fit for capturing the stochastic nature of power consumption since they are not restricted to a fixed quantity of parameters. They are able to deal with nonlinear relationship between load and exogenous factors with high uncertainties. Gaussian process (GP) models are regarded as promising algorithms of non-parametric group for electricity consumption estimation, forecasting, and control [16,17]. Several studies have explored STLF by using GP models. In [18], the authors have intended to improve the prediction accuracy by using a transfer learning based on a GP model. In order to address time complexity, they have used converted matrices with smaller orders of prediction inferences. Nonetheless, their regression problem has been formulated in an auto-regressive manner that has not addressed the dependency of electricity consumption on external factors such as weather. [19] has combined GP with Log-Normal distribution of load profile patterns based on an auto-regressive approach. However, it has not considered temperature as a predictor. [20] has studied a forecasting system in which temperature is the only input factor. Nevertheless, it has not faced the problem of complexity since its analysis has been performed with a small amount of historical data. The study in [21] has presented hierarchical GP as another scheme for load forecasting. It has explored a simple prediction process by use of GP models based on temperature information from a weather station. However, only the last 250 data points (approximately ten days) have been used for training. It is essential to know that hierarchical processes can neglect the interaction between descriptive elements of power consumption. In this regard, the composition of the kernels to characterize the interaction between input variables is an effective technique that has been employed for STLF through GP models [22,23]. [24] has utilized compositional kernel-based GPs to carry out a non-parametric regression for load prediction. Nonetheless, it has limited the input interactions discovery by compositional kernels through exploiting only outside temperature and calendar variables. Indeed, the promise of GP-based STLF systems motivates the development of new designs to address their related issues, mentioned above.

In addition to the choice of model, the aggregation level of power consumption is another matter that influences the efficiency of a STLF framework. Normally, load forecasting is applied to total power demand of individual or combined houses [25]. Two different scenarios have been proposed to deal with the latter case. In the first scheme, an overall prediction is obtained from the summation of house-level forecasts. In the second strategy, a single prediction is obtained from an aggregate-level forecast. The authors of [26] have stated that the former scenario is superior while those of [27] have claimed the inverse. Nevertheless, this depends on power consumption behaviour across targeted houses. Single forecasting can be a decent choice for dwellings with similar usage patterns. Besides, overall forecasting can be a suitable option for homes with dissimilar consumption behaviour. In fact, the summation of individual loads in this case can alter their actual patterns and result in an inaccurate single prediction [28]. Therefore, STLF manners that are proposed to estimate total power consumption of a set of residences should take into account their individual load profile behaviour. The impact of aggregation level on forecasting performance can be further studied in [29].

1.2. Motivation and contribution

This study is encouraged by potential improvements in STLF methods that can be acknowledged from the aforementioned issues. It contributes a STLF design based on non-parametric techniques due to their advantages, albeit higher complexity. The proposed forecasting system in terms of a regression analysis is aimed at handling high dimension data associated with several input factors and their historical information [30]. It exploits main categories of meteorological and non-meteorological variables as the inputs to predict load profile as the output. This framework utilizes a probabilistic model on the basis of GP because of its ability to realize powerful tools for residential demand prediction [31,32]. The GP model is designated by compositional kernels that are intended to explain the additive nature of the input variables through an order-based examination and capture their non-linear relationship with the output. Furthermore, the suggested design employs an approximation procedure to address time and space complexity of the resultant GP while maintaining a desirable performance. Subsequently, the proposed forecasting approach whose detailed contributions are presented below is evaluated by the actual data of a set of houses, located in Quebec.

1) A structure with considerably reduced complexity that is able to deal with challenges related to processing big data.
2) An efficient method that takes advantage of a Bayesian inference with a Monte Carlo Markov Chain (MCMC) sampling technique to effectively estimate the GP model hyperparameters.
3) A scalable compositional kernel-based GP model that is capable of handling high dimensional data by selecting the most significant input interactions. The selection procedure that takes advantage of the Thompson Sampling (TS) algorithm leads to the most competitive kernel pairs discovery.

The remainder of the paper is organized as follows: Section 2 discusses the proposed methodology in details and presents an overview of the probabilistic graphical models. Section 3 describes the modeling evaluation of the proposed compositional GP model. The results and discussion of the proposed model are presented in Section 4, which is followed by conclusions in Section 5.

2. Methodology

The proposed approach performs the following methodology in order to realize an efficient load forecasting structure. The suggested GP model is aimed at explaining a set of predefined meteorological and non-meteorological components of residential power demand in the Quebec region. In fact, different factors can influence power consumption behavior. However, considering all of them, specifically climate-sensitive ones, can notably increase the analysis complexity. Besides, such consideration can result in minor improvements since some of these elements have very poor impacts on total demand pattern. In this study, temperature, humidity, and solar radiation are the weather variables that have been selected. Our choice of climate components has its roots in a previous study, where we have utilized both sensitivity and predictability analyses to identify the most significant factors of demand for an effective forecasting in the same area i.e. Quebec [37]. Calendar, which explains the time-related variations of power

Table 1
The effective elements of GP-based forecasting procedures according to the relevant literature.

	Big data utilization	Data pre-processing	Hyperparameters estimation	Kernel interaction modeling	Periodic (calendar) data analysis	Changing point
[33,34]	×	×	×	✓	✓	✓
[19,23]	×	✓	×	×	✓	×
[24,35]	✓	✓	✓	×	×	×
[31,36]	×	×	✓	✓	✓	✓
This work	✓	✓	✓	✓	✓	×

usage, is another factor that is considered. As mentioned, this component reflects occupants' behavior towards utilizing their appliances. On the other side, there are different measures that can be considered to deal with the designing process of STLF by means of GP models. Table 1 shows significant features that have been employed in the relevant literature to enhance such a procedure. Since data size and diversity (big data) notably influences this process, its related information has been also added to this table. The uniqueness of our algorithmic outline is to integrate nearly all effective elements of a fruitful GP-based forecasting with no increase in complexity. This arrangement is further promoted by the utilization of big data. The mathematical notions of Table 1 are detailed within this section. The block diagram of the compositional kernel-based GP model is presented in Fig. 1. In this figure, the training and prediction processes are shown in two separate blocks. In the left block, the database supplies the information that is utilized for the training phase. At this step, we encounter a large amount of historical weather data that can remarkably challenge GP scaling capacity. In order to address this issue, an estimation process based on data subsets is developed. The Subset of Data (SoD) is selected by using a clustering technique that takes advantage of the mean shift algorithm. Specifically, the SoD presents the inducing points as the effective number of input data, exploited for learning. Subsequently, these points are utilized to approximate the covariance matrix of the intended GP model through the Fully Independent Training Conditional (FITC) method [38]. The latent function and hyper-parameters of this model are estimated by Maximum A Posterior (MAP). The MAP estimation is realized by means of the MCMC algorithm considering hyper-priors. As mentioned, the proposed GP model is structured upon a key objective that is extracting the most efficient input interactions according to their correlation with power consumption. Therefore, at this level of learning process, the best kernel compositions that present these interactions are captured. As a result, the selected kernels are used to construct the predictive GP model, targeted by the proposed mechanism. Afterwards, a daily forecasting exercise is carried out by the full conditional expectation of the latent function of the GP, presented in the right block of Fig. 1.

2.1. GP model formulation

Let at discrete-time $t, \mathrm{x}_t = [x_t^1, x_t^2, \ldots, x_t^D] \in \mathbb{R}^D$ be a set of measurements of D exogenous variables that present weather and calendar information as the input.

Furthermore, let $y_t \in \mathbb{R}$ be aggregated power consumption that defines the output. $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\} \in \mathbb{R}^{N\times D}$ and $\mathbf{y} = \{y_1, \ldots, y_N\} \in \mathbb{R}^N$ contain the historical information of the last N measurements of both input and output variables. Accordingly, a GP can be formulated based on,

$$f(\mathrm{x}) \sim \mathcal{GP}(0, k(\mathrm{x}, \mathrm{x}\prime)) \tag{1}$$

$$y_t|f(\mathrm{x}_t) \sim \mathcal{N}\left(f(\mathrm{x}_t), \sigma^2\right), \tag{2}$$

where σ^2 and $f(\mathrm{x})$ are the variance of the estimation error and the latent function, respectively. The term $k(\mathrm{x}, \mathrm{x}\prime)$ represents the covariance function (or kernel). The Gaussian Process regression is a non-parametric Bayesian estimation that places a $\mathcal{GP}$ prior over a set of latent functions $\mathbf{f} = [f(\mathrm{x}_1), \ldots, f(\mathrm{x}_N)]$ [39]. Considering a set of test data $X^* = \{\mathrm{x}_1^*, \ldots, \mathrm{x}_L^*\}$, where L is the forecasting horizon, the predictive posterior distribution at $\mathbf{f}^* = [f(\mathrm{x}_1^*), \ldots, f(\mathrm{x}_L^*)]$ can be calculated based on,

$$P(\mathbf{f}^*|X, y) = \mathcal{N}\left(\mu(X^*), \sigma^2(X^*)\right) \tag{3}$$

where the prediction expectation, $\mu(X^*)$ is computed by,

$$\mu(X^*) = K(X^*, X)\left[K(X, X) + \sigma^2\,\mathbb{I}\right]^{-1}\mathbf{y}$$

and the prediction variance, $\sigma^2(X^*)$ is estimated through,

$$\sigma^2(X^*) = K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X^*, X)^T$$

In the above sub-equations, $\mathbb{I}$ denotes the identity matrix and K stands for the covariance matrix. Each element of K presents the kernel evaluation between two sample points of different variables from the input set. Consequently, the related set of power predictions, $\hat{\mathbf{y}} = \{\hat{y}_1, \ldots, \hat{y}_L\}$ can be estimated by calculating the expected values of the latent function through their posterior predictive distribution i.e. $\hat{\mathbf{y}} = \mu(X^*)$.

Prediction with GP models can result in high accuracy since they exploit all historical data. In addition, a GP naturally models uncertainty and allows computations to be extended into multidimensional inputs. However, scaling GP to large data brings about computational complexity mainly due to the calculation of the inversion matrix. Such complication has an order of $\mathcal{O}\left(N^3\right)$ and $\mathcal{O}\left(N^2\right)$ for time and storage, respectively. Different approximation methods have been used to reduce this high computation cost while preserving an accurate prediction. For this purpose, the proposed methodology takes advantage of the Fully Independent Training Conditional (FITC) approximation technique.

2.1.1. FITC approximation

This technique, proposed by Candela and Rasmussen [40], is a sparse approximation method. It uses a set of latent variables, $\bar{f}$, associated with a set of inducing points, $\bar{X}$ to approximate the posterior predictive, defined by the Eq. (3). Specifically, it computes the mean of the prediction through [41],

$$\mu_{\mathrm{FITC}}(X^*) = K(\bar{X}, \bar{X})\,\Sigma^{-1}K(\bar{X}, X)\left(\Lambda + \sigma^2\,\mathbb{I}\right)^{-1}\mathbf{y} \tag{4}$$

where

$$\Sigma = K(\bar{X}, \bar{X}) + K(\bar{X}, X)\left(\Lambda + \sigma^2\,\mathbb{I}\right)^{-1}K(X, \bar{X})$$

and Λ is a diagonal matrix, computed by

$$\Lambda = \mathrm{diag}\left(K(X, X) - K(X, \bar{X})\,K(\bar{X}, \bar{X})^{-1}K(\bar{X}, X)\right)$$

2.1.2. Mean-shift clustering

The inducing points, required by the FITC approximation, are produced by a clustering means. This method takes advantage of the mean-shift algorithm due to its excellent convergence proper-

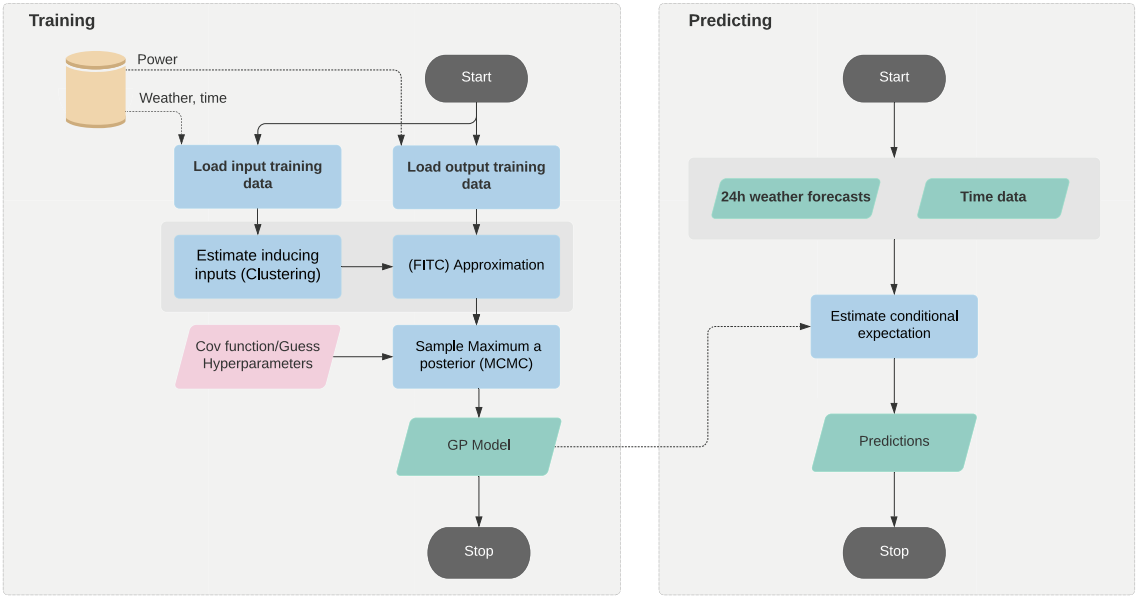

Fig. 1. The block diagram of the predictive additive GP model.

ties, implementation ease, and robust initialization. The mean-shift formula does not require a predefined number of clusters and its only prerequisite is the scale parameter [42]. Algorithm 1 explains the iterative process of this clustering technique that is executed to specify a set of inducing points as the SoD.

Algorithm 1. Gaussian mean shift clustering algorithm

for $n \in 1, 2, \ldots N$ ◁ For each data point
  $\mathrm{x} \rightarrow \mathrm{x}_n$ ◁ Start point
  repeat ◁ Iteration Loop
    $\forall n : p(n|\mathrm{x}) \rightarrow \frac{\exp\left(-\frac{1}{2}(\|\mathrm{x}-\mathrm{x}_n\|/\varepsilon)^2\right)}{\sum_{n\prime=1}^{N}\exp\left(-\frac{1}{2}(\|\mathrm{x}-\mathrm{x}_{n\prime}\|/v)^2\right)}$
    $\mathrm{x} \rightarrow \sum_{n=1}^{N} p(n|\mathrm{x})\mathrm{x}_n$ ◁ Update x
  until x update $< tol$
  $\bar{\mathrm{x}}_n \rightarrow \mathrm{x}$
end for
connected components $\{\bar{\mathrm{x}}_n\}_{n=1}^{N}, \epsilon$ ◁ Clusters

In this algorithm, a Gaussian kernel is used to estimate the density of each point in the input set (see iteration loop). It should be noted that the bandwidth $v > 0$ influences the number of clusters by intuitively scaling the distance between inducing points. To be precise, a higher/lower value of v results in a smaller/bigger number of clusters. The mean shift technique requires a stopping condition. A practical way to stop the iteration is setting a threshold (tol) for the magnitude of changes in the mean vector. A typical value for this threshold is calculated based on a small fraction of the bandwidth. This amount is $1 \times 10^{-3} v$ in Algorithm 1. The utilized clustering procedure results in a reduced complexity order of $\mathrm{O}(Nq\log(q))$, where q expresses the number of iterations.

2.2. Additive GP

A GP can be designed by an ensemble of kernel functions to enhance its potentials for STLF purposes. The functions that are used to develop a composite structure are mainly formulated based on two different forms. The first represents individual kernels that are used for exploring one-dimensional variables. The second symbolizes compositional kernels that are intended for searching interactions between multi-dimensional variables. This formation allows the GP to uncover influential properties of combined variables with regard to forecasting procedure. A framework that is constructed by the addition of a set of former/latter functions results in a first/multi-order additive kernel. For example, a second-order composite presents the summation of two-kernel functions multiplications. In addition, an additive structure can be created by combining both individual and compositional kernels. The general formulation of an additive kernel is defined by [43],

$$k_{\mathrm{so}}(\mathrm{x}) = \sum_{i=1}^{S} k_i(\mathrm{x}[\mathrm{O}_i]) \tag{5}$$

where $k_{\mathrm{so}}(\mathrm{x})$ is computed over all S possible 2-order compositions based on D number of dimensions. Accordingly, O_i stands for ith-order pair of inputs. Input pairs can be created in $S = \binom{D}{2}$ different ways, where $\binom{\cdot}{\cdot}$ denotes the binomial coefficient. For instance, with $D = 3, k_{\mathrm{so}}(\mathrm{x}) = k_1(x_1, x_2) + k_2(x_1, x_3) + k_3(x_2, x_3)$. The proposed GP model is outlined by such architecture. However, the number of additive elements in the Eq. (5) can increase dramatically with the number of inputs, which makes the posterior estimation computationally expensive. Therefore, two critical modifications are incorporated into the model to avoid a costly calculation due to a high dimensional input. The model is modified to contain only the first and second order kernel functions. Furthermore, it is enhanced by a procedure that selects the n-upmost pairs among all second-order kernels. Subsequently, the GP model, resulted from these adjustments, can be explained by,

$$k(\mathrm{x}) = k_w(\mathrm{x}_w) + k_c(\mathrm{x}_c) + k_{\mathrm{so}}(\mathrm{x}) \tag{6}$$

that k_w and k_c describe weather and calendar-related kernels, respectively, and $k_{\mathrm{so}}(\mathrm{x})$ stands for the second-order kernel compounds. According to the input dimensions, the two first elements hold all first-order composites related to temperature, humidity, solar radiation, time of day, and day of week. In this regard, the vectors x_w and x_c contain weather and calendar variables, respectively.

Incorporating the compositional kernels into the model is an important improvement in comparison with the relevant designs in the literature. The combined kernels enhance the forecasting system by revealing correlations between different variables, especially weather and calendar ones. Indeed, previous studies have mainly relied upon the utilization of basic forms i.e. first-order kernels, which formulate input dimensions individually. Various kernel are aimed at learning the characteristics of input variables via different covariance functions. Exponentiated Quadratic (EQ), Polynomial, Radial Basis, Matérn, and Spectral Mixture are common covariance bases, investigated in the literature. The additive GP structure takes advantage of EQ and Matérn covariance functions for weather and calendar variables, respectively. Specifically, the Matérn is applied to the sine and cosine functions of time of day and day of week indexes in order to explain their periodicity [44].

2.2.1. Best kernel pairs discovery

Uncovering a set of M best kernel matches from S candidates is a complex problem since the number of possible mixtures increases by input dimensions. The application of a sampling strategy can address this issue since it can eventually select the most competitive kernel pairs. In the proposed method, the number of best combinations, M, is fixed arbitrary and their initial set is chosen randomly from a uniform distribution. Besides, the TS is the sampling technique that is employed for the selection procedure. Within the training process, TS is repeated certain times per day, for the possible number of second-order kernels in order to construct a GP model based on the Eq. (6) and evaluate its performance in power consumption forecasting.

The TS technique has been described in Algorithm 2. To be precise, the role of TS is to sample a kernel combination T times each day within the learning period. According to evaluation results, this sample replaces a pair that has provided the least contribution to the prediction among M combinations in the previous iteration. Subsequently, the new sample is rewarded/punished if it improves/deteriorates the GP performance in the next repetition. The reward/penalty of a selected kernel match is executed by updating its prior (probability of success). The probability parameters for all S candidates are denoted by $(\theta_1, \ldots, \theta_S)$. Furthermore, the update rule is formulated by adding the output of the reward function to the parameters of a Bernoulli distribution. The value of this function for the j^{th} kernel is defined by,

$$r_j = \begin{cases} 1 & \text{if } e_i < e_{\min} \\ 0 & \text{if } e_i > e_{\min} \end{cases} \tag{7}$$

where e_i denotes the error metric at the iteration i and $e_{\min}$ is the smaller error observed in the past. The distribution parameters are not updated in case of $e_i = e_{\min}$.

Algorithm 2. The Thompson Sampling Algorithm

```
for i = 1,...,T
  for j = 1,...,S
    Sample θ_j ~ beta(α_j, β_j)
  end for
  j* → argmax_j θ_j
  Apply j* and observe reward (r_j)
  (α_j,β_j)→(α_j,β_j) + (r_j,1-r_j)
end for
```

Consequently, the best kernel pairs of current day create the initial set of next one. As a result, the M selected combinations at the end of the training phase are used to perform the test stage, which is next day forecasting. The block diagram in Fig. 2 shows the selection process of M best kernel pairs by using the TS. The iterative sampling process causes the subset of second-order compositional kernels to contain the most efficient pairs, and in turn, results in a GP model with significant forecasting efficiency.

2.2.2. Hyper-parameters learning

The covariance functions of the GP model include hyper-parameters that are tuned within the learning step. These parameters describe the impact of input variables on the power demand. As a Bayesian regression, a prior distribution over every hyper-parameter is required to provide its posterior inference. Accordingly, Gamma and ChiSquared distributions are used as the prior over the hyper-parameters of the EQ and Matérn functions, respectively. Particularly, they are used to explain the length-scales, ℓ_{EQ} and ℓ_M, of these convariance bases. Besides, the HalfCauchy distribution is utilized as the prior over the variance, σ of both functions. The parameters of the aforementioned distributions are described by,

$$\ell_{EQ} \sim \text{Gamma}(\alpha = 2, \beta = 1) \tag{8}$$

$$\ell_M \sim \text{ChiSquared}(\beta = 3) \tag{9}$$

$$\sigma \sim \text{HalfNormal}(\beta = 5). \tag{10}$$

In this work, the optimal values of σ is estimated by the MAP. Specifically, a No-U-Turn-Sampler (NUTS) that is an MCMC class is exploited to execute this task [45]. It should be noted that the prior over the weight of each covariance function is determined by the HalfCauchy distribution.

3. Evaluation

3.1. Real world data

Real data of 30 houses, located in the Quebec region, has been exploited for this study. This data was collected from 1st December 2017 to 31th March 2018. The data consists of power consumption information with a sampling interval of 15 min. The aggregated demand corresponds to the summation of all-houses total power usage with a maximum load of 250 kW. It is worth mentioning that the input variables do not include the information related to the building envelope and inside temperature.

Fig. 3 shows weekly behavior of aggregated power during 18 weeks starting from 1st December. The dark-blue line depicts the average power use. As it can be seen, weekdays have similar consumption patterns that are different from those of weekends. Unlike weekends, weekdays illustrate two distinct peaks in morning and evening. The difference in weekly power consumption patterns is related to occupants' behavior towards utilizing electrical

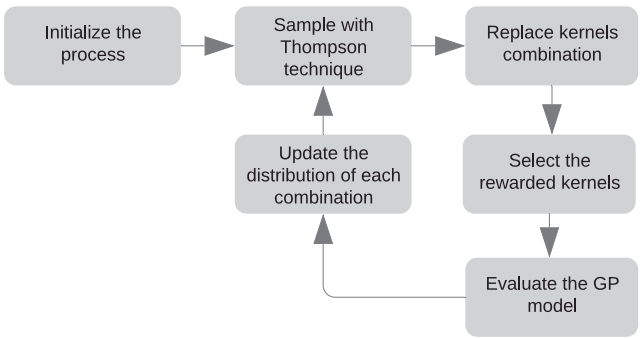

Fig. 2. The block diagram of the TS procedure, used to select the M best kernel pairs for constructing the proposed GP model.

loads. Particularly, this figure reveals that daily and weekly demands have a periodic nature. In order to characterize such periodicity, the time index is mapped onto a two-dimensional input through $\begin{pmatrix}\cos(g(t))\\ \sin(g(t))\end{pmatrix}$. The periodicity is controlled by the function $g(t)=\frac{2\pi t}{\tau}$, where τ is a period. Normally, τ is fixed to match weekly and daily patterns. In this work, we use two different periods $\tau=24$ and $\tau=24\times 7$ to generate 4 calendar variables as proposed in [44].

3.2. Uncertainty of weather data

The stochastic nature of weather data causes uncertainty for the prediction process. Historical data of the OpenWeatherMap website as a weather information provider[1] is used to estimate the standard deviation of hourly forecasting error across a time horizon of 24*h*. The prediction records correspond to the city of Trois Rivieres in Quebec, Canada during 2017. The perturbations related to meteorological variables are generated based on $\Delta_{\mathrm{T}}^{h}\sim\mathcal{N}(0,\sigma_{\mathrm{T}}^{h})$, $\Delta_{\mathrm{H}}^{h}\sim\mathcal{N}(0,\sigma_{\mathrm{H}}^{h})$, and $\Delta_{\mathrm{R}}^{h}\sim\mathcal{N}(0,\sigma_{\mathrm{R}}^{h})$, where $\Delta_{\mathrm{T}}^{h},\Delta_{\mathrm{H}}^{h}$, and Δ_{R}^{h} present Gaussian noises of temperature, humidity, and solar radiation, respectively. h defines the forecasting horizon and σ denotes the standard deviation of the forecasting error of each variable. It should be noted that the forecasting values of the solar radiation are collected from cloud [46]. The forecasting is simulated by $\widetilde{T}_h=T_h+\Delta_{\mathrm{T}}^{h}$, $\widetilde{H}_h=H_h+\Delta_{\mathrm{T}}^{h}$, and $\widetilde{R}_h=e^{\Delta_{\mathrm{R}}^{h}}R_h$. $\widetilde{T}_h,\widetilde{H}_h$, and $\widetilde{R}_h$ are the forecasting values of temperature, humidity, and solar radiation, while T_h, H_h, and R_h are the actual measurements of these elements, respectively. In order to estimate the multiplicative noise of solar radiation $\Delta_{\mathrm{R}}^{h}=\log\left(\frac{\widehat{R}_h}{R_h}\right)$ has been used.

In this transformation, $\widehat{R}_h$ and R_h stand for prediction and actual amounts, respectively. Fig. 4 depicts the standard deviation of the forecasting error of each climate variable according to the prediction horizon.

3.3. Evaluation metrics

Three different accuracy metrics are used to evaluate the results of the proposed daily forecasting exercise. The scores account for symmetric mean absolute percentage error (sMAPE), mean absolute error (MAE), and root mean squared error (RMSE), explained by,

$$\mathrm{sMAPE}=\frac{100\%}{N}\sum_{t=1}^{N}\frac{|\widehat{y}_t-y_t|}{(|\widehat{y}_t|+|y_t|)/2} \tag{11}$$

$$MAE=\frac{1}{N}\sum_{t=1}^{N}|y_t-\widehat{y}_t| \tag{12}$$

$$\mathrm{RMSE}=\left\{\frac{1}{N}\sum_{t=1}^{N}(\widehat{y}_t-y_t)^2\right\}^{1/2} \tag{13}$$

where $\widehat{y}$ and y present predicted and actual power usages for N discrete time samples. The RMSE score gives a higher weight to a larger error since the error is squared. Therefore, it is appropriate for reflecting bigger failures in forecasting results.

[1] https://openweathermap.org/.

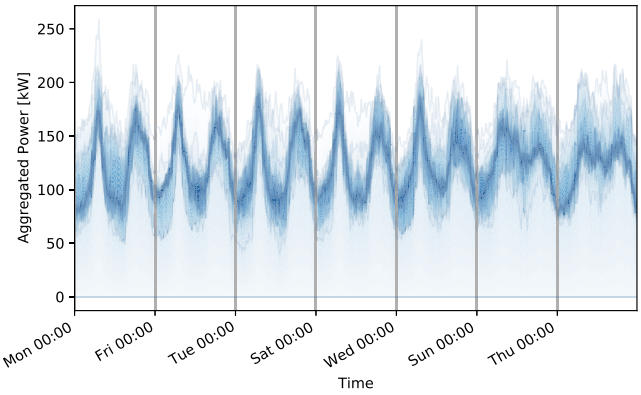

Fig. 3. Aggregated power consumption behavior across hours of the day and days of the week.

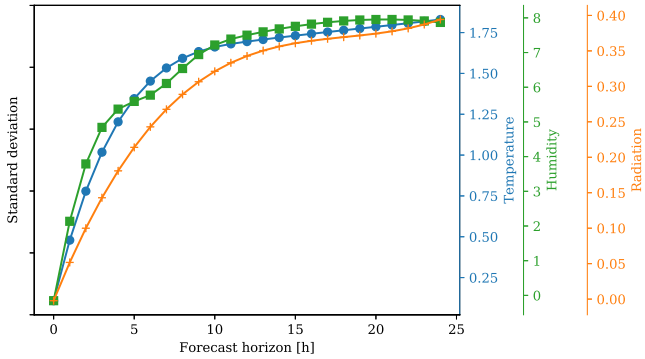

Fig. 4. Standard deviation of outdoor temperature, humidity, and solar radiation forecasting values over a 24 *h* horizon.

4. Validation and results

4.1. Data processing

The data used for the forecasting practice is managed based on the look-back window technique [47]. In this regard, a window size of 25 days has been selected. This choice that provides the best outcomes considering running time and prediction results has been made through several tests. The day-ahead forecasting practice is started with one-week data. This initial window is gradually expanded on a daily basis as new observations arrive. The dataset expansion is stopped when the look-back window length i.e., 25 days is reached. Afterwards, the daily shift of the fixed window is carried on in order to perform the forecasting for all the data. Since Gaussian process regression is assumed to have a zero-mean output, the data is standardized to respect this assumption. The standardization can provide a better estimation of covariance functions due to the fact that these bases normally include scale parameters. Such data processing can also alleviate numerical issues of this type of regression.

Subsequently, the training phase of the GP utilizes 25-days historical data of weather and calendar as the input and power demand as the output. Afterwards, the test stage exploits the forecasting data of the next day to estimate the power usage of that day. This results in an iterative modeling process where in every iteration, the GP model is re-trained and re-estimated by historical and next-day data, respectively. It can be deduced that the most competitive kernel pairs, decided by the sampling procedure, are the only properties

of the GP that are passed along across the dataset. With regard to the length of look-back window, the GP is trained with an amount of 2400 samples (4 quarters of one hour $\times$ 24 h $\times$ 25 days) in every iteration. This can bring about a complexity of 2400^3 (13824×10^6). The computational complexity of a normal GP is N^3. Computing the Cholesky decomposition of such data is very complex for a normal computer. This situation promotes the proposed GP design that has been intended to manipulate big data.

4.2. Results of clustering algorithm

The mean-shift clustering algorithm can manage the inducing points and thus, decrease the running time. Fig. 5 shows the results of the clustering procedure. It can be observed that the number of clusters does not proportionally decrease with more data accumulation and generally varies between 70 and 80 per day. In addition, the method is able to maintain the quantity of clusters less than 85 across the whole dataset. In fact, it has decreased the number of inputs by a ratio of around 96% (85/2400) in every iteration with regard to big data training complexity.

It can be understood that the clustering method can successfully capture similarities in data points and collect them in the same clusters. Therefore, it avoids the complication of the information space and accelerates the algorithmic process of the proposed GP model. It should be mentioned that the clustering algorithm has been applied only to the meteorological variables.

4.3. Best compositional kernels selection

As mentioned, the TS technique is used to capture a subset of n-best second-order kernels among all possible pairs. The number of selected composites has been set to 5. The 5-upmost compositional kernels are used to construct the additive GP model in the Eq. (6). Fig. 6 presents the result of TS during one day. It can be deduced that this method iteratively offers more competitive compositional kernels that in turn, improve the daily forecasting. This implies the effectiveness of TS to enhance the additive GP model by choosing the best additive structure in each execution.

The TS algorithm is executed successively within the whole forecasting period. This strategy can progressively increase the performance of the forecasting system as shown in Fig. 7. Besides, it can be observed that the accuracy rate is converged by the progress of the prediction task. This demonstrates the ability of the developed design to extract the best kernel pairs, which result in not only significant but also consistent forecasts. It should be note that a number of 10 iterations has been considered to create an effective subset of compositional kernels for every daily run of the TS algorithm.

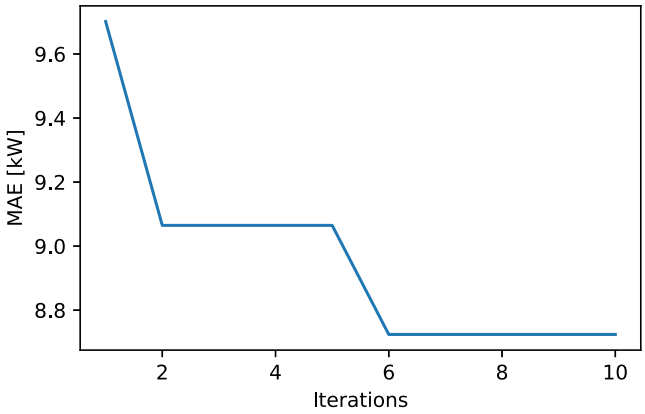

Fig. 6. Model prediction accuracy based on the efficiency of compositional kernels, selected by TS during one day.

4.4. Benchmark models

A comparison study is conducted to demonstrate the effectiveness of the proposed approach to STLF. Accordingly, basic GP, Support Vector Regression (SVR), Random Forest Regression (RFR), and multi-step LSTM (Long Short Term Memory) network, as reliable forecasting methods, are employed. As discussed, the basic GP is the composition of first-order kernels of all input variables, which in this case, account for weather and calendar. This choice is critical for the comparative study since it can assist in evaluating the significance of the compositional kernels integration as the essence of the developed structure. LSTM is chosen due to its promising performance in load forecasting tasks. In fact, one of the significant applications of LSTM is time-series prediction. As an advanced recurrent neural network, LSTM is able to effectively learn long-term dependencies in sequential data. Notwithstanding, four efficient LSTM models are considered since defining the optimal network configuration is a challenging exercise for any deep learning model. Among these four configurations, the best one is chosen for the final comparison of the methods. SVR and RFR are also exploited since they exhibit good performance in nonlinear aggregated loads forecasting [48].

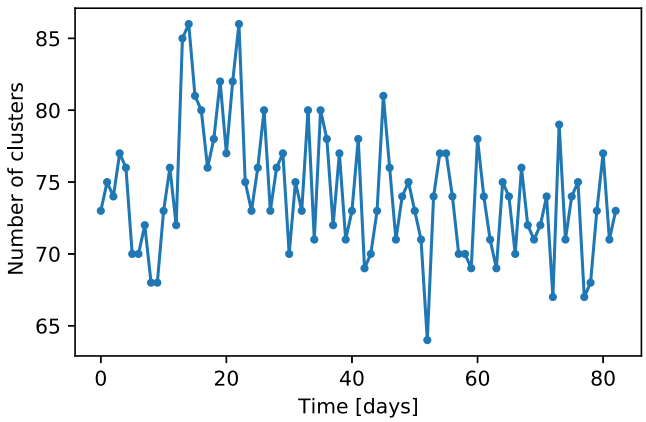

Fig. 5. Temporal evaluation of the number of clusters resulting from the mean shift clustering algorithm.

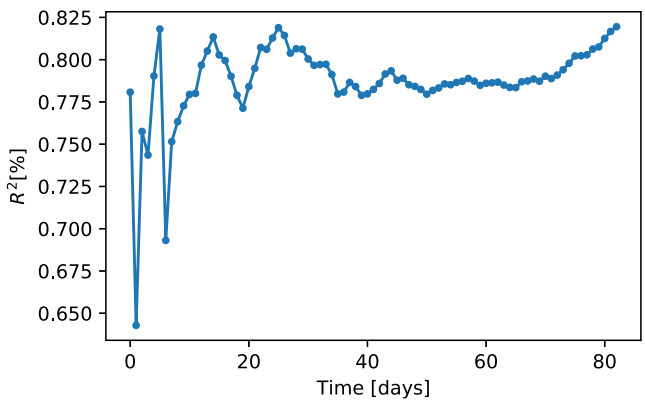

Fig. 7. Model prediction improvement due to gradual increase in the performance of compositional kernels, chosen by TS within whole dataset.

The LSTMs are designed by Keras as a powerful neural networks library, written in python [49]. Accordingly, a sequential model of Keras is chosen. In order to avoid a verbose discussion, only the configuration of the best LSTM model is detailed. This model has a fully connected structure that comprises two hidden layers of 100 units in each and a single output layer. The activation functions in the input and output layers are Hyperbolic tangent (tanh) and Rectified Linear Unit (relu), respectively. For compiling the network, Adam and mean squared error are employed in terms of the optimization algorithm and the loss function, respectively. The training phase is managed across 150 epochs by using all the 7 inputs with a sliding window of 3 samples. Nevertheless, it should be added that the other LSTM models utilize a similar structure with different number of units and epochs. On the other hand, SVR and RFR use the same training phase as the GP. For the benchmark methods, the data is also standardized to handle large boundaries.

4.5. Results and discussion

Fig. 8 presents the day-ahead forecasting results over the entire dataset, managed by the look-back window technique. Regarding the values of the accuracy metrics, it can be observed that our proposed approach is superior to other techniques within all time. Particularly, the additive GP model outperforms the basic one. This signifies the existence of meaningful interactions between influential components of power demand that has been successfully described by the subset of compositional kernels. In fact, the kernel selection procedure can not only manage the time and space complexity but also augment the GP model. Specifically, the remarkable decrease in the RMSE score of the additive GP demonstrates its capability to fit a regression line that can effectively represent the data points. On the other side, the lower performance in the starting periods, especially the 15 first days is due to the less amount of available data. This is a common issue of learning-based prediction systems with insufficient training data. Nevertheless, all the methods have achieved efficient and stable outcomes after 40 days. Moreover, Table 2 presents the average values of accuracy metrics during 4 months of daily forecasting (final comparison). It can be noticed that the suggested approach surpasses the benchmark models according to all the scores. In fact, the temporal management of the meteorological data, enabled by the mean-shift clustering algorithm from one side and the incorporation of best kernel pairs, facilitated by the TS procedure from the

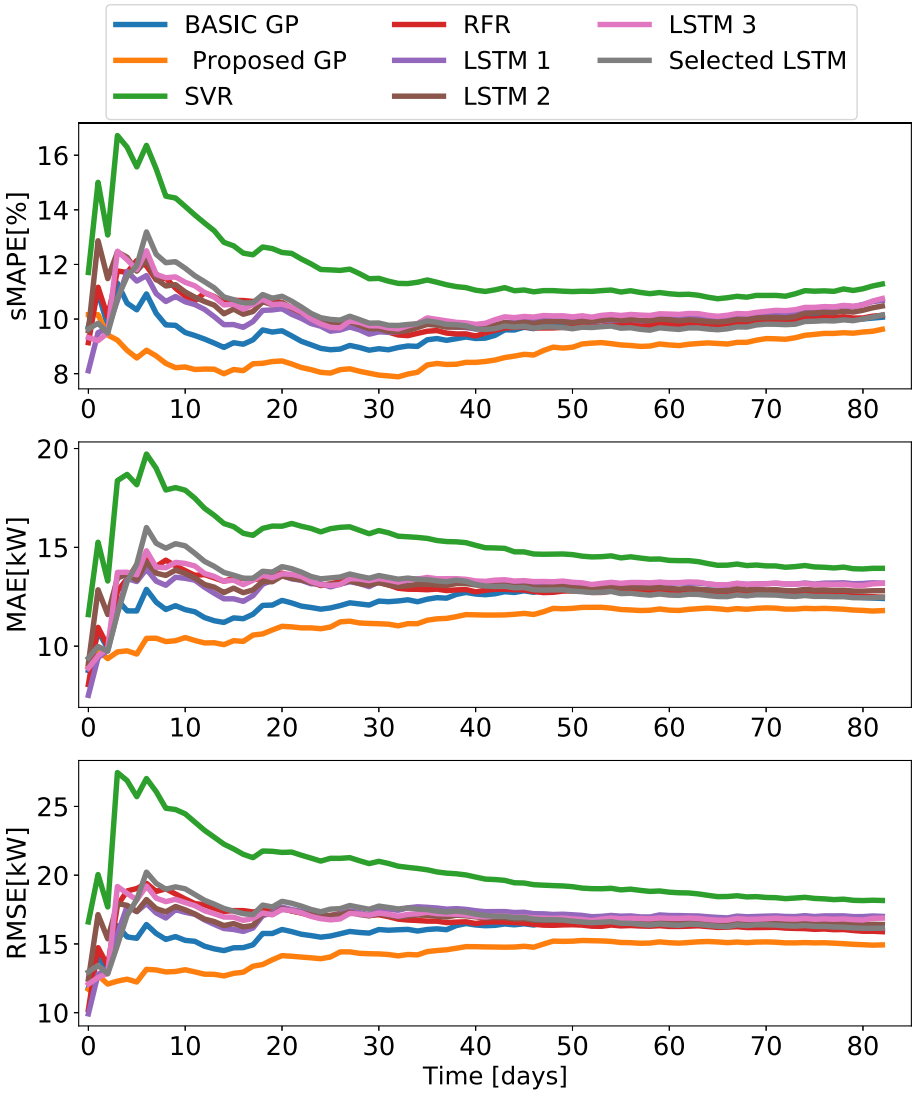

Fig. 8. The prediction error based on the sMAPE, MAE, RMSE metrics.

Table 2
The average value of MAE, RMSE, and sMAPE scores, applied to the forecasting models within 4 months

	Models	MAE [kW]	sMAPE[%]	RMSE [kW]
4 months in winter	Basic GP model	12.41	10.27	14.91
	Support Vector Regression	13.93	11.42	18.16
	Selected LSTM	13.18	11.00	17.39
	Random Forest Regression	12.47	10.31	15.88
	Proposed GP model	**11.79**	**9.88**	**12.41**

other side have resulted in a GP model that is both scalable and accurate. Indeed, throughout the results, provided in this study, the proposed method has proved to be more efficient. The additive GP design can adequately capture the dynamic stochastic nature of aggregated power consumption due to weather phenomena and occupants behaviour, represented by calendar factor.

5. Conclusions

This study has proposed a non-parametric regression model for forecasting aggregated power consumption of a set of residential buildings. It has developed a compositional kernel-based GP approach to capture the highly nonlinear relationship between load demand and its descriptive elements accounting for climate-sensitive and calendar components. Particularly, the suggested probabilistic method has been intended to incorporate the meaningful interactions between power demand factors into the forecasting practice. In this regard, an additive GP model has been constructed by using a subset of the most effective second-order kernel pairs. A sampling procedure has been developed to collect the best kernel matches. Moreover, a clustering technique has been used to manage the difficulties related to the utilization of multi-dimensional data. In order to evaluate the performance of the model, the actual data of aggregated power consumption of a set of houses, located in Quebec has been exploited. The efficiency of the proposed design has been demonstrated through a comparative analysis based on efficient benchmark models. The results have manifested that the additive GP model can outperform other methods and achieve accurate forecasting. An enhanced model is targeted in future work that considers other influential variables such as occupants' behavior and price signal.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the Laboratoire des technologies de l'énergie d'Hydro-Québec, the Natural Science and Engineering Research Council of Canada, and the Foundation of Université du Québec á Trois-Riviéres.

References

[1] K.B. Debnath, M. Mourshed, Forecasting methods in energy planning models, Renewable and Sustainable Energy Reviews 88 (March) (2018) 297–325 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032118300200. .

[2] B. Yildiz, J. Bilbao, A. Sproul, A review and analysis of regression and machine learning models on commercial building electricity load forecasting, Renewable and Sustainable Energy Reviews 73 (2017) 1104–1122. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032117302265. .

[3] F. Amara, K. Agbossou, Y. Dubé, S. Kelouwani, A. Cardenas, J. Bouchard, Household electricity demand forecasting using adaptive conditional density estimation, Energy and Buildings 156 (2017) 271–280. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778817314299. .

[4] R. Sevlian, R. Rajagopal, A scaling law for short term load forecasting on varying levels of aggregation, International Journal of Electrical Power & Energy Systems 98 (2018) 350–361 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061517306956. .

[5] Y. Wang, J.M. Bielicki, Acclimation and the response of hourly electricity loads to meteorological variables, Energy 142 (2018) 473–485 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544217317061. .

[6] P. Lusis, K.R. Khalilpour, L. Andrew, A. Liebman, Short-term residential load forecasting: Impact of calendar effects and forecast granularity, Applied Energy 205 (2017) 654–669 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261917309881. .

[7] F. Amara, K. Agbossou, Y. Dubé, S. Kelouwani, A. Cardenas, S.S. Hosseini, A residual load modeling approach for household short-term load forecasting application, Energy and Buildings 187 (2019) 132–143 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778818309228. .

[8] M. Arun, A.M. Gupta, M. Lodhe, (eds) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, Short-Term Load Forecasting Using Parametric and Non-parametric Approaches, in: Advances in Intelligent Systems and Computing, Springer Singapore, Singapore, vol. 1053, 2020, pp. 815–823 [Online]. Available: http://link.springer.com/10.1007/978-981-15-0751-9_74. .

[9] A. Tascikaraoglu, B.M. Sanandaji, Short-term residential electric load forecasting: A compressive spatio-temporal approach, Energy and Buildings 111 (2016) 380–392 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S037877881530431X. .

[10] I. Shah, H. Iftikhar, S. Ali, D. Wang, Short-Term Electricity Demand Forecasting Using ComponentsEstimation Technique, Energies 12 (13) (2019) 2532 [Online]. Available: https://www.mdpi.com/1996-1073/12/13/2532. .

[11] D. Toquica, K. Agbossou, R. Malhamé, N. Henao, S. Kelouwani, A. Cardenas, Adaptive Machine Learning for Automated Modeling of Residential Prosumer Agents, Energies 13 (9) (2020) 2250 [Online]. Available: https://www.mdpi.com/1996-1073/13/9/2250. .

[12] W. Charytoniuk, M. Chen, P. Van Olinda, Nonparametric regression based short-term load forecasting, IEEE Transactions on Power Systems 13 (3) (1998) 725–730 [Online]. Available: http://ieeexplore.ieee.org/document/708572/. .

[13] G.-F. Fan, Y.-H. Guo, J.-M. Zheng, W.-C. Hong, Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting, Energies 12 (5) (2019) 916 [Online]. Available: https://www.mdpi.com/1996-1073/12/5/916. .

[14] H. Jiang, Y. Zhang, E. Muljadi, J.J. Zhang, D.W. Gao, A Short-Term and High-Resolution Distribution System Load Forecasting Approach Using Support Vector Regression With Hybrid Parameters Optimization, IEEE Transactions on Smart Grid 9 (4) (2018) 3341–3350 [Online]. Available: https://ieeexplore.ieee.org/document/7748604/. .

[15] D. Asber, S. Lefebvre, J. Asber, M. Saad, C. Desbiens, Non-parametric short-term load forecasting, International Journal of Electrical Power & Energy Systems 29 (8) (2007) 630–635 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061507000439. .

[16] A. Jain, T. Nghiem, M. Morari, R. Mangharam, Learning and Control Using Gaussian Processes, in: 2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS), IEEE, 2018, pp. 140–149 [Online]. Available: https://ieeexplore.ieee.org/document/8443729/. .

[17] A. Zeng, H. Ho, Y. Yu, Prediction of building electricity usage using Gaussian Process Regression, Journal of Building Engineering 28 (November 2019) (2020) 101054 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S235271021930662X. .

[18] Y. Zhang, G. Luo, Short term power load prediction with knowledge transfer, Information Systems 53 (2015) 161–169 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000150. .

[19] M. Shepero, D. van der Meer, J. Munkhammar, J. Widén, Residential probabilistic load forecasting: A method using Gaussian process designed for electric load data, Applied Energy 218 (March) (2018) 159–172 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S030626191830299X. .

[20] A. Prakash, S. Xu, R. Rajagopal, H. Noh, Robust Building Energy Load Forecasting Using Physically-Based Kernel Models, Energies 11 (4) (2018) 862 [Online]. Available: http://www.mdpi.com/1996-1073/11/4/862. .

[21] J.R. Lloyd, GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes, International Journal of Forecasting 30 (2) (2014) 369–374 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169207013000757. .

[22] D. Duvenaud, J.R. Lloyd, R. Grosse, J.B. Tenenbaum, Z. Ghahramani, Structure Discovery in Nonparametric Regression through Compositional Kernel Search,

in: 30th International Conference on Machine Learning, ICML 2013, vol. 28, no. PART 3, 2013, pp. 2203–2211 [Online]. Available: http://arxiv.org/abs/1302.4922. .

[23] S. Fan, S. Member, R.J. Hyndman, Short-term load forecasting based on a semi-parametric additive model, IEEE Transactions on Power Systems (August) (2010) 1–8 [Online]. Available: https://robjhyndman.com/papers/2010STLF-FinalR1.pdf.

[24] G. Xie, X. Chen, Y. Weng, An Integrated Gaussian Process Modeling Framework for Residential Load Prediction, IEEE Transactions on Power Systems 33 (6) (2018) 7238–7248 [Online]. Available: https://ieeexplore.ieee.org/document/8400492/. .

[25] S. Humeau, T.K. Wijaya, M. Vasirani, K. Aberer, Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households, in: 2013 Sustainable Internet and ICT for Sustainability (SustainIT), IEEE, 2013, pp. 1–6 [Online]. Available: http://ieeexplore.ieee.org/document/6685208/. .

[26] P. Goncalves Da Silva, D. Ilic, S. Karnouskos, The Impact of Smart Grid Prosumer Grouping on Forecasting Accuracy and Its Benefits for Local Electricity Market Trading, IEEE Transactions on Smart Grid 5 (1) (2014) 402–410 [Online]. Available: http://ieeexplore.ieee.org/document/6684330/. .

[27] S. Bandyopadhyay, T. Ganu, H. Khadilkar, V. Arya, Individual and Aggregate Electrical Load Forecasting, in: Proceedings of the 2015 ACM Sixth International Conference on Future Energy Systems, ACM, New York, NY, USA, 2015, pp. 121–130 [Online]. Available: https://dl.acm.org/doi/10.1145/2768510.2768539. .

[28] K. Nikolopoulos, A.A. Syntetos, J.E. Boylan, F. Petropoulos, V. Assimakopoulos, An aggregate-disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis', Journal of the Operational Research Society 62 (3) (2011) 544–554 [Online]. Available: https://www.tandfonline.com/doi/full/10.1057/jors.2010.32. .

[29] G. Zotteri, M. Kalchschmidt, F. Caniato, The impact of aggregation level on forecasting performance, International Journal of Production Economics 93–94 (SPEC.ISS.) (2005) 479–491 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S092552730400266X. .

[30] E. Schulz, M. Speekenbrink, A. Krause, A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions, Journal of Mathematical Psychology 85 (2018) 1–16 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022249617302158. .

[31] J.-B. Fiot, F. Dinuzzo, Electricity Demand Forecasting by Multi-Task Learning, IEEE Transactions on Smart Grid 9 (2) (2016) 544–551 [Online]. Available: http://ieeexplore.ieee.org/document/7467578/. .

[32] L.-L. Li, J. Sun, C.-H. Wang, Y.-T. Zhou, K.-P. Lin, Enhanced Gaussian process mixture model for short-term electric load forecasting, Information Sciences 477 (2019) 386–398 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S002002551830865X. .

[33] D. Duvenaud, Expressing Structure with Kernels, phd-thesis, 2014. [Online]. Available: https://raw.githubusercontent.com/duvenaud/phd-thesis/master/kernels.pdf. .

[34] H. Keshavarz, G. Michailidis, Y. Atchade, Sequential change-point detection in high-dimensional Gaussian graphical models, arXiv 21 (2018) 1–57 [Online]. Available: http://arxiv.org/abs/1806.07870. .

[35] F. Massa Gray, M. Schmidt, Thermal building modelling using Gaussian processes, Energy and Buildings 119 (2016) 119–128 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778816300494. .

[36] R. Skagestad, Electricity Demand Forecasting with Gaussian Process Regression, Ph.D. dissertation, Norwegian University of Science and Technology, 2018 [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2566721/20123_FULLTEXT.pdf?sequence=1. .

[37] K. Dab, K. Agbossou, A. Cardenas, Y. Dube, S. Kelouwani, Sensitivity Analysis of Exogenous Variables for Load Forecasting Using Polynomial Regression, in: IECON 2019–45th Annual Conference of the IEEE Industrial Electronics Society, IEEE, vol. 1, 2019, pp. 2560–2565 [Online]. Available: https://ieeexplore.ieee.org/document/8927167/. .

[38] H. Liu, J. Cai, Y.-S. Ong, Y. Wang, Understanding and comparing scalable Gaussian process regression for big data, Knowledge-Based Systems 164 (2019) 324–335 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705118305380. .

[39] S. Park, S. Choi, Hierarchical gaussian process regression, Journal of Machine Learning Research 13 (2010) 95–110 [Online]. Available: http://proceedings.mlr.press/v13/park10a.html. .

[40] J. Qui nonero-Candela, C.E. Rasmussen, A unifying view of sparse approximate Gaussian process regression, Journal of Machine Learning Research 6 (2005) 1939–1959 [Online]. Available: http://www.jmlr.org/papers/v6/quinonero-candela05a.html.

[41] E. Snelson, Z. Ghahramani, Sparse Gaussian Processes using pseudo inputs, Advances in Neural Information Processing Systems 18 (NIPS 2005) (2005) 1–8 [Online]. Available: https://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf. .

[42] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619 [Online]. Available: http://ieeexplore.ieee.org/document/1000236/. .

[43] J.R. Gardner, C. Guo, K.Q. Weinberger, R. Garnett, R. Grosse, Discovering and exploiting additive structure for Bayesian optimization, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, vol. 54, 2017 [Online]. Available: https://www.cs.toronto.edu/rgrosse/aistats2017-additive.pdf. .

[44] M. Blum, M. Riedmiller, Electricity demand forecasting using gaussian processes, in: AAAI Workshop - Technical Report, vol. WS-13-15, 2013, pp. 10–13 [Online]. Available: https://dl.acm.org/doi/10.5555/2908259.2908261. .

[45] H. Liu, Y.-S. Ong, X. Shen, J. Cai, When Gaussian Process Meets Big Data: A Review of Scalable GPs, IEEE Transactions on Neural Networks and Learning Systems (2020) 1–19 [Online]. Available: https://ieeexplore.ieee.org/document/8951257/. .

[46] W.F. Holmgren, A.T. Lorenzo, and C. Hansen, A Comparison of PV Power Forecasts Using PVLib-Python", in 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC), IEEE, 2017, pp. 1127–1131 [Online]. Available: https://ieeexplore.ieee.org/document/8366724/. .

[47] B. Kim, Dithering Loopback-Based Prediction Technique for Mixed-Signal Embedded System Specifications, IEEE Transactions on Circuits and Systems II: Express Briefs 63 (2) (2016) 121–125 [Online]. Available: http://ieeexplore.ieee.org/document/7277065/. .

[48] N. Huang, W. Wang, S. Wang, J. Wang, G. Cai, L. Zhang, Incorporating Load Fluctuation in Feature Importance Profile Clustering for Day-Ahead Aggregated Residential Load Forecasting, IEEE Access 8 (2020) 25198–25209 [Online]. Available: https://ieeexplore.ieee.org/document/8978655/. .

[49] S. Hosseini, N. Henao, S. Kelouwani, K. Agbossou, A. Cardenas, A Study on Markovian and Deep Learning Based Architectures for Household Appliance-level Load Modeling and Recognition, in: 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), vol. 2019-June, IEEE, 2019, pp. 35–40 [Online]. Available: https://ieeexplore.ieee.org/document/8781186/. .

3.2 Classification comportementale de l'agrégat des résidences

3.2.1 Introduction

Une fois que le modèle basé sur AGP a été construit ses performance ont été optimisées. Cependant, le défi majeur réside dans la variabilité des profils de consommation d'énergie, posant ainsi des difficultés pour la prédiction précise de la charge globale. Le prochain objectif est donc d'incorporer une analyse comportementale sur les profils de charges agrégées dans le but de développer une méthode de prévision plus robuste. Avec l'apparition de nouveaux concepts tels que le *«big data»* et leurs applications étendues ces dernières années, la recherche s'est intensifiée autour des solutions non supervisées. Les algorithmes de classification, en particulier, offrent la possibilité d'extraire des connaissances approfondies des données, exploitables par les analystes pour diverses applications. Ce chapitre présente une méthodologie pour la classification des charges agrégées pour l'ensemble de 1000 résidences.

3.2.2 Contexte

Dans le contexte d'ensembles de données massifs, l'utilisation de solutions de classification supervisée devient difficile, tandis que le recours aux classifications, via des approches non supervisées, se présente comme une solution réalisable. Cette étude se focalise spécifiquement sur les données de séries temporelles. Pour relever, ce défi et contrairement aux méthodes existantes, une nouvelle approche est proposée. Nous avons combiné deux techniques majeure dans cette analyse appelée *«Cluster Based Aggregate Forecast»*(CABF) et la classification consensuelle (Consensus clustering en anglais). Dans cette dernière approche, au lieu de se fier uniquement à un algorithme de classification, Nous avons généré diverses partitions des données. Ensuite, nous avons appliqué la

classification consensuelle pour identifier les zones de consensus entre ces partitions. Cela permet d'obtenir une classification robuste et stable, moins sujette aux variations causées par le bruit ou les choix arbitraires dans les méthodes de classification individuelles. L'objectif est d'éviter la redondance d'informations, conduisant ainsi à des solutions de classification plus stables et fiables. Diverses méthodes élaborées pour la classification des bâtiments résidentiels dans la littérature seront examinées, mettant l'accent sur des approches statistiques regroupées en quatre grandes familles : les modèles de réduction de dimension, les algorithmes de classification ou les prototypes, et enfin les mesures de distance et de similarité. La classification consensuelle offre ainsi une base solide pour la compréhension continue et l'amélioration du modèle de prévision, en soulignant les différentes approches utilisées dans la littérature pour résoudre des problèmes.

3.2.3 Méthodologie

La méthode proposée, tel que présentée dans la Figure 3.2, est basée sur l'énoncé de recherche exposé. Elle débute par l'utilisation d'une base de données de profils de charge de consommation quotidienne des ménages, incluant divers appareils tels que les systèmes de chauffage, les lave-vaisselle, laveuses, sécheuses et les éclairages. L'objectif initial est de déterminer un modèle basé sur le comportement de chaque ménage, réalisé par une classification à l'aide de l'algorithme k-medoids. Cette architecture de classification CABF, incorpore deux types de données des données réelles et données simulées à partir de données réelles décrivant le comportement des occupants [82]. Le premier type concerne les données réelles, tandis que le deuxième type comprend des données simulées provenant du parc virtuel fourni par Hydro-Québec. Ces données sont échantillonnées sur une période de vingt-quatre heures avec un pas d'échantillonnage de 15 minutes. La classification consensuelle est ensuite appliquée entre les résultats de la classification

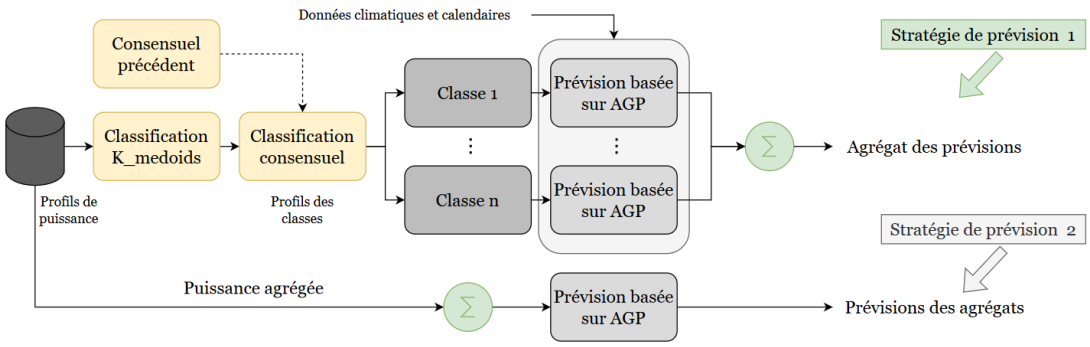

FIG. 3.2 Méthodologie proposée pour l'approche de la prévision basée sur la classification hiérarchique.

k-medoids sur des périodes de temps consécutives, conduisant à des classifications de modèles similaires.

Par la suite, le modèle de prévision proposé (AGP) est utilisé pour prévoir la puissance agrégée pour chaque classe. Le processus de classification consensuel se révèle efficace pour distinguer les sous-ensembles de données dans l'ensemble de données présentant des caractéristiques similaires [16]. L'utilisation d'une approximation de fonction, plutôt que l'application d'un modèle global à tous les ensembles de données, permet de capturer les propriétés distinctes de chaque sous-ensemble. La prévision par classes a pour objectif d'utiliser les similarités entre les données pour améliorer la précision de la prévision. Dans le contexte de la prévision de la puissance agrégée, ces informations sont extraites des agrégats au niveau des classes, supposés présenter des schémas plus réguliers avec une autocorrélation plus élevée, rendant ces signaux plus prévisibles que ceux représentant la consommation d'énergie globale. Pratiquement, la CBAF divise l'ensemble de données des profils de charge en classes individuelles, déterminant ainsi la consommation des bâtiments en différents types d'agrégats. Une technique de prévision potentielle implique la création d'un ensemble de prédicteurs pour chaque agrégat au niveau de la classe,

utilisant ces prédicteurs pour prévoir la charge totale. Plus précisément, la technique utilise des classes basées sur les k-medoids pour les données temporelles sous des mesures significatives de distance temporelle basée sur DTW. En examinant les données, Nous avons observé que même si les foyers présentaient des schémas similaires, ces schémas pouvaient être décalés dans le temps. Par exemple, deux foyers peuvent avoir des profils de comportement similaire, mais un foyer peut commencer son activité plus tôt ou plus tard que l'autre. Alors le choix du DTW sert à mesurer la similarité entre deux séries temporelles et calculer l'alignement optimal entre eux ainsi que détecter le décalage temporel. Ceci est suivie d'un algorithme de classification consensuelle intégrant la mesure de similarité de Jaccard pour produire les meilleures classes représentant des modèles similaires parmi les classes quotidiennes résultantes.

3.2.4 Résultats

La méthodologie élaborée repose sur un processus en trois étapes visant à déterminer une analyse comportementale des profils de charge des résidences. Dans une première phase, elle débute par la création de profils de puissance à travers une approche unifiée visant à améliorer la prévision de la charge à court terme pour la demande d'électricité résidentielle. En particulier, une méthode de classification de séries temporelles basée sur le consensus, utilisant la classification k-medoids, est employée. Ces profils regroupés font ensuite l'objet d'une classification supplémentaire en utilisant l'algorithme du consensus, déterminant le nombre de classes en fonction de la méthode de classification et du critère de similarité de Jaccard [90]. Cette classification aboutit à la catégorisation des résidences en différentes classes. Finalement, par le biais de AGP, un modèle de régression non paramétrique englobant le noyau des conditions météorologiques et calendaires a été utilisé pour prévoir l'agrégat de chaque classe. Une étude de simulation a été effectuée

pour révéler l'efficacité de la méthode de prévision suggérée sur des ensembles de données de séries temporelles.

Les résultats mettent en évidence l'efficacité de cette approche, qui permet d'adapter les profils de charge aux besoins spécifiques de chaque client, indépendamment de l'heure de consommation, tout en améliorant les prévisions à court terme de la puissance agrégée. La validation de la méthodologie à travers des simulations numériques, intégrant à la fois des données réelles et simulées, démontre que l'application d'une stratégie de classification horaire similaire à une classe de clients partageant des comportements de consommation semblables offre une compréhension plus approfondie de leur comportement et augmente la précision des prévisions de consommation pour ces classes de résidences.

Energy & Buildings 299 (2023) 113550

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild

Consensus-based time-series clustering approach to short-term load forecasting for residential electricity demand

Khansa Dab [a,*], Nilson Henao [a], Shaival Nagarsheth [a], Yves Dubé [b], Simon Sansregret [c], Kodjo Agbossou [a]

[a] Laboratoire d'Innovation et de Recherche en Énergie Intelligente (LIREI), Département de génie électrique et génie informatique, Université du Québec à Trois-Rivières, QC G9A 5H7, Canada
[b] Département de génie mécanique, Université du Québec à Trois-Rivières, QC G9A 5H7, Canada
[c] Laboratoire des Technologies de l'Énergie, Institut de Recherche Hydro-Québec, Shawinigan, QC G9N 7N5, Canada

ARTICLE INFO

Keywords:
Short-term load forecasting
Aggregated load forecast
Consensus clustering
Gaussian process
Residential load patterns
Time-series

ABSTRACT

Load forecasting could play a crucial role in energy management and control of buildings in residential neighborhoods. In these areas, electricity demand is influenced by different phenomena accounting for climate conditions and comfort preferences. The uncertain nature of these circumstances results in power profiles with diverse patterns. Under this condition, overall load prediction is suggested by utilizing Cluster-based Aggregate Forecasting (CBAF). Accordingly, this paper proposes a unified approach to such a practice. The proposed scheme employs an unsupervised machine-learning algorithm to develop a time-series clustering scheme that performs the classification task through the k-medoids-based clustering incorporating the Dynamic Time Warping (DTW) algorithm. Subsequently, a consensus is achieved among the resultant clusters where the Jaccard similarity index adjudges the similarity measurement. The Additive Gaussian Process (AGP), a powerful non-parametric forecasting technique, is exploited to predict aggregated load at each cluster level. With low complexity and high scalability, AGP is particularly utilized to provide effective forecasting. Numerical simulations on synthetic as well as real datasets have been carried out to illustrate the effectiveness of the proposed methodology. Additionally, two comparative studies are carried out with forecasts without clustering and with the benchmark non-parametric models employing a cluster-based technique. The proposed method demonstrates significant improvement in forecasting accuracy for both datasets by reducing the error metrics and achieving 7% improvement in the coefficient of determination (R^2) value as compared to the aggregated load forecast achieved without clustering. The comparative study demonstrates that the proposed method with AGP can forecast the total residential load more accurately than other benchmark models with an improvement of 26% and 21% in R^2, respectively, for both datasets.

1. Introduction

1.1. Background & motivation

Short-Term Load Forecasting (STLF) plays a vital role in power systems stability [1] as it analyses network operations over a typical time horizon. It assists with the safe and reliable operation of the power grid by maintaining the dynamic balance between supply and demand [2]. The STLF provides valuable information on energy consumption for robust economic residential network dispatching [3]. Accordingly, it is crucial to reform electricity markets aimed at grasping energy-saving opportunities to benefit both utility and consumers [4].

In the residential sector, aggregate forecasting is usually performed for a group of residences. However, the essential knowledge of the characteristics of load patterns is frequently lost through aggregation processes. Data fluctuations reveal changes that can deteriorate the effectiveness of the forecasting process [5]. Related research shows that CBAF techniques can increase the forecasting accuracy for aggregated loads by reducing the stochastic characteristics of load profiles, allowing the forecasting algorithm to be trained on highly correlated data [6]. Evidently, the increased forecasting accuracy via CBAF contributes to

* Corresponding author.
E-mail address: khansa.dab@uqtr.ca (K. Dab).

https://doi.org/10.1016/j.enbuild.2023.113550
Received 17 March 2023; Received in revised form 12 September 2023; Accepted 14 September 2023
Available online 29 September 2023

adaptive transactive energy for improving revenue management. Time-series forecasting and clustering are standard methods used to facilitate decision-making; thus, dividing existing consumers into smaller groups is advantageous based on their common characteristics.

Clustering strategies applied to forecasting procedure could achieve a statistically significant improvement in accuracy compared to the traditional aggregate forecasts depending on the number of clusters and the database size [7]. Thus, CBAF can offer additional insight to practitioners who wish to implement the strategy in the real world and improve load forecasting for a classified dataset. Additionally, to merge complementary perspectives of the data into a more stable partition, consensus clustering is a robust element [8] [9]. It results in stable partitions and manages diversity by generating partitions with different subsets of consistent attributes [10].

The traditional methods use aggregation for a set of houses with different behavior [11]; however, setting boundaries to divide the houses finely is challenging depending on their behavior [12]. That necessitates a reliable load clustering method with a structured aggregation such that the information is preserved. In this regard, data clustering facilitates the discovery of independent variables with similar patterns. As a result, the unorganized data are converted into similar clusters, allowing the machine learning algorithm to examine and extract critical information efficiently [13]. Energy consumption forecasting accuracy can be improved by including valuable features and other techniques. Since building energy consumption has numerous daily patterns, incorporating labels into various routines can enhance prediction accuracy [14]. A variety of clustering techniques to identify different user groups in the electricity usages and patterns of subgroups could be employed for these issues [13]. They provide differentiation for distinct clusters dependently on the preferences and comfort of occupants [15]. Cluster-based techniques have been the most used to detect similar load patterns in the time-series data to reduce the complexity of modeling forecast [16].

One of the significant aspects of concern is that load forecasting has to be performed on streams of time-series evolution data [17]. Given that the dataset may have occurred over various time periods, robustness against incomplete time series is crucial in this situation. Specifically, it restricts the possibility of clustering a group of users with synchronized time series [18]. Additionally, time-series data could be naturally noisy or stochastic, affecting clusters' distinctness. The commonly known clustering techniques suffer from inefficient forecasting as a consequence of accumulating errors due to the stochastic behavior of users [19].

1.2. Related work

The literature provides various approaches to STLF for residential electricity demand via CBAF. Research in [5], [24] implements a higher-accuracy aggregate forecast employing CBAF by clustering the households, forecasting the clusters separately, and aggregating the estimates. [28] evaluated clustering approaches that consist of efficient preprocessing of data obtained from smart meters by a model-based representation and the k-means clustering method. Their evaluation revealed that clustering results depend on the number of clusters and the size of the database. In the load forecasting task of [29], [30], the algorithm establishes the nonlinear relationship of spatial and temporal between the load and the features; however, the traditional technique of clustering was applied based on k-means. Nevertheless, large load fluctuations would cause forecasting deviations to decrease the forecasting accuracy. Chen et al. [31] proposed a hybrid multi-step forecasting method to improve the forecasting performance for the stochastic framework. Baker et al. [32] worked on the robustness of the electricity demand forecasting under uncertainty, whereas [33] worked towards the robustness of forecasting models against data integrity and its importance in the power industry.

Artificial neural network-based clustering techniques [20], [21] have also found a place for load and price forecasting in the day-ahead market. While [20] depends on the traditional Euclidean norm, [21] utilizes a self-organizing map for clustering. Kong et al. [23] classified load into various levels according to the corresponding power load proportion in the total load. They exploited the LSTM (Long Short-Term Memory) method to select similar days of all load categories by aggregating forecasts versus forecasting the aggregate. The model formulated on the radial basis function neural network approach for STLF has found application [34]. They utilized an adaptive neural fuzzy inference system to adjust the load forecasting results obtained based on the recent changes in the real-time price. Cao et al. [25] proposed a mechanism of the similar day method to group the target day with meteorological similarities of historical data and then predict the load based on the average demand of those days. An adopted auto-regressive integrated moving average (ARIMA) model was employed to evaluate the load forecasting problem for the residential sector. Li et al. [35] used the grid method for short-term day-ahead forecasting comparable to the similar day method. Still, instead of grouping days with similar weather measurements, it groups the load profiles based on the load's localization, nature, and size. For each class, they formed a neural network to forecast load demand.

Since building energy consumption has numerous daily patterns, incorporating labels to various routines can enhance prediction accuracy [14]. Standard statistical techniques need an input set of specified characteristics to estimate the stationary features of a time series. However, unlike traditional linear regressions, Machine Learning (ML) systems can map the relationships between characteristics and stochastic time series without making complex dependencies among the inputs [14]. Accordingly, for time series modeling, Gaussian processes (GPs) are one of the most popular and advanced choices in the current state of the art for non-parametric regression since they are naturally able to handle complex structures [36]. This incorporates physical insights about load data characteristics to improve accuracy while reducing training requirements [37], [38]. Despite the appealing attributes of GP approaches, implementing GP to achieve sufficient performance remains difficult because data-driven models become computationally expensive without an appropriate input feature selection method [24], [39]. Rouwhorst et al. [27] utilize ML algorithms to propose clustering approaches based on aggregated load forecasting focusing on Medium Term Load Forecasting (MTLF). A feature selection of six different clustering algorithms has been applied to transformer loading. Their clustering algorithm is based on the Euclidean distance, which computes only the point-wise distances (one-to-one).

1.3. Contributions & organization

Table 1 presents the recent studies in the demand forecasting domain along with the clustering techniques adopted. These relevant studies on load forecasting reveal that popular consensus clustering techniques dispatch the result as a congruence of various clustering techniques. Besides, several forecasting techniques are tested only on synthetic data and are yet to be validated on real-life consumption data. Additionally, most traditional clustering techniques utilize Euclidean distance, which computes only the point-wise distances (one-to-one) and lacks one-to-many or many-to-one distancing that may increase the outcome level. The ubiquitous Euclidean distance has a plethora of evidence proving its poor accuracy for classification and clustering [40]. Also, it is prone to distortion in the time axis. Exploiting DTW can resolve the distortion problem in the time axis. As discussed, the time-series data increases the complexity while accumulating over time with changes in the data characteristics. Nevertheless, there is no benchmark methodology to find consensus within accumulating time-series data for aggregate short-term load forecasting of residential demand in the transactive energy context.

Table 1
The components of load forecasting according to the relevant literature.

References	Type of Data	Clustering Methods	Distance Criteria	Forecasting Method
[20][21]	Clustering pattern sequence	SOM clustering	Euclidean	ANN/PSF
[22][23]	Multi-zone structure	Grid-based	Euclidean	LSTM
[5][24][25]	Temperature strategy for HVAC	Hierarchical CBAF	Euclidean	ARIMA/DNN
[26]	Different power system parameters	Micro clustering	Distance ratio of cluster	RNN/LSTM
[27]	Time duration of load	Spectral clustering	Vector-to-set	Gradient boosting
[28] [28]	Model-based time series	k-means and CBAF	Euclidean	Bootstrapping/Others models
[29] [30]	Multi-storied residential buildings	Spatial-Temporal and k-means	Hybrid distance	LSTM-GRU/e-HMM
[31]	Hourly metered load data	k-means and kernel k-means	Dynamic time warping	LightGBM
This Work	Accumulating time-series	Consensus clustering and CBAF	Dynamic time warping	Additive Gaussian process

DNN: Deep Neural Network, RNN: Recurrent Neural Network, PSF: Pattern Sequence Forecasting, e-HMM: ensemble Hidden Markov Model, LightGBM: Light gradient boosting machine

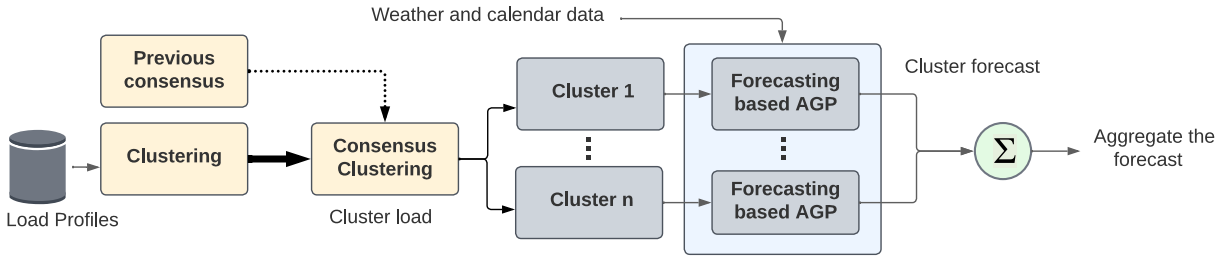

Fig. 1. The block diagram of the proposed methodology.

To address the aforementioned challenge, this paper proposes a unified methodology to aggregate residential load forecasting based on AGP models for each cluster of groups of residential houses. Particularly, a k-medoids clustering algorithm integrating dynamic time warping is utilized to cluster the power profiles of an ensemble of houses. Subsequently, to ameliorate the detrimental effects of accumulating time-series consumption data, the results from k-medoids clustering are adjudged by consensus clustering to create of residential houses based on daily energy consumption patterns. Apparently, the consensus-based approach improves the scalability since it labels the daily pattern of building energy consumption using the time-series clustering algorithm. Establishing a consensus framework helps identify key targets for the high similarity of consumption levels to improve the accuracy of the forecast. The proposed methodology for load forecasting is effectuated on synthetic and real data sets of the energy consumption of a group of houses. Also, the load forecast results of the proposed methodology are compared with other benchmark models for accuracy measures.

The remainder of this paper is organized as follows: Section 2 discusses the proposed methodology in detail. Section 3 provides the details about datasets, benchmark models and accuracy metrics utilized. Section 4 is about results and discussion covering comparative results of the proposed method. Finally, Section 5 concludes the paper.

2. Methodology

The main objective of this study is to suggest a successful demand forecasting strategy with improved accuracy. Based on the presented research statement, an overview of the proposed approach is shown in Fig. 1. First, a database of daily household consumption load profiles, such as heating systems, dishwashers, washing machines, dryers, and lightning, is used to determine a pattern based on the behaviour of each household that results in a cluster using the k-medoids clustering algorithm. Note that the data is sampled within a window of twenty-four hours. Then, employing consensus clustering, a consensus between k-medoids clustering results over consecutive periods of time is found, leading to clusters of similar patterns. Subsequently, AGP is utilized to forecast the aggregate load for each cluster.

2.1. Cluster-based aggregate load forecasting

The process of clustering is an effective method of distinguishing smaller data points from more extensive data sets with similar characteristics. As a result, exploiting a function approximation, instead of applying a global model to all data sets, captures the distinct properties of each subset.

Cluster-based forecasting aims to utilize the similarities between data points to improve the accuracy of the forecast. In the aggregate load forecasting context, this information is extracted regarding cluster-level aggregates that are supposed to have more regular patterns with higher auto-correlation and therefore are more predictable signals than those that represent the aggregate energy consumption [12]. Practically, CBAF segregates the dataset of load profiles into individual clusters. As a consequence, clustering classifies building consumption into different building types. Accordingly, a possible forecasting technique would be to create an ensemble of predictors for each cluster-level aggregate and use them to forecast the total load. Specifically, the technique first uses k-medoids-based clusters for temporal data under considerable time warp measures based on DTW and thereafter, a consensus clustering algorithm incorporating Jaccard similarity measure to yield the best clusters representing similar patterns among the resulting daily clusters.

2.1.1. k-medoids clustering

The power consumption behaviour could be influenced by different factors, which significantly increases the complexity of the study. This section outlines the clustering based on k-medoids clustering, which is a more efficient variant of k-means [41]. Instead of using the mean point as the center of a cluster as k-means, k-medoids use an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with the minimum sum of distances to other points. This algorithm is a powerful tool for identifying load usage patterns that could decrease the specified time relationship. Based on DTW distance, it defines the dissimilarity coefficients to discover the consumption similarities of the considered dataset. In this work, we propose subsequence clustering of a time series extracted via a window of twenty-four hours, i.e. clustering in segments from a single long time series.

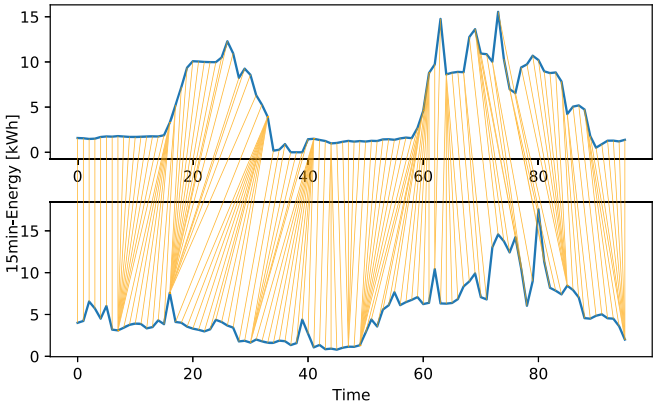

Fig. 2. Distance between clusters probabilities Dynamic Time Warping (DTW). The blue line is the data slot from clusters 1 and 2, and the orange line is the distance presented by the alignment-based common metrics. It proves useful when dealing with temporally shifted time series. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Dynamic time warping DTW, a distance measure first introduced by Berndt and Clifford [42], is a widely recognized algorithm that calculates an optimal match between two given temporal sequences. Fig. 2 illustrates nonlinear alignments between two-time series to accommodate sequences. According to Euclidean distance, the i^{th} point of one sequence is aligned with the i^{th} point of the other to provide a pessimistic measure of dissimilarity. In contrast, DTW delivers a more intuitive distance measure (one-to-many or many-to-one) to be calculated. Thus, comparing the electricity load forecast with observation is beneficial to correlate it at multiple time points for extracting meaningful information. Since Euclidean distance is unsuitable for high-dimensional time series because the data is lost [40], this method provides better accuracy in formed clusters than Euclidean.

Let $D = \{\mathrm{D}_i\}_{i=1}^{N}$ be a set of time series $\mathrm{D}_i = \{a_{i1}, \ldots, a_{iT}\}$ assumed of length T. Let W be a set containing *warping path* defined as $W = \{w_1, \ldots, w_k, \ldots, w_K\}$, where $T \leq K \leq 2T + 1$. Then, the DTW between time series $\mathrm{D}_i = \{a_{it}\}$ and time series $\mathrm{D}_j = \{a_{jt'}\}$, with the aim of minimization of the mapping cost [41], is defined by:

$$DTW(\mathrm{D}_i, \mathrm{D}_j) \overset{def}{=} \min_{w_k \in W} \frac{1}{|w_k|} \sum_{(t,t') \in w_k} \varphi(a_{it}, a_{jt'}), \tag{1}$$

where w is a warping function that realizes a mapping from the time axis of D_i onto the time axis of D_j and $\varphi \in \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a positive real-valued divergence function (generally Euclidean) [41]. As shown in Fig. 2, this technique enables non-linear alignments between 15-minute data slots from separate households to measure the similarities between clusters of various lengths and take into account those that are similar but locally out of phase.

2.1.2. Consensus clustering algorithm

Generally, the consensus clustering algorithm is dominantly used to combine the output from multiple runs of a clustering algorithm [43]. Specifically, a consensus is found between the results of the k-medoids clustering algorithm to yield clusters containing houses with similar power consumption patterns. Consider a dataset D consisting of N instances, i.e. $D = \{\mathrm{D}_i\}_{i=1}^{N} = \{\mathrm{D}_i, \ldots, \mathrm{D}_d \ldots, \mathrm{D}_N\}$, and let $\tilde{\mathrm{C}}_d$ be a set of clusters for day d resulting from k-medoids clustering method, such that

$$\tilde{\mathrm{C}}_d = \{\tilde{c}_d^i\}_{i=1}^{M} = Cluster(\mathrm{D}_d). \tag{2}$$

Now, let C_d be the set containing clusters resulting from the consensus algorithm. Here, the consensus is formed between the consecutive outputs of the previous day's consensus results and the present-day k-medoids clustering results (2), i.e.

$$\mathrm{C}_d = Consensus(\mathrm{C}_{d-1}, \tilde{\mathrm{C}}_d) \tag{3}$$

where C_{d-1} denotes the consensus clustering results from the previous day. Note that for the first day, there is no consensus, i.e. $\mathrm{C}_1 = \tilde{\mathrm{C}}_1$. In (3), Jaccard similarity [44] is utilized to evaluate the output of the clustering algorithm for the predefined cluster numbers; it is the measure of ratio or tradeoff between similar consensus. Consequently, that information is utilized to decide the merger of the consensus. Besides, it is capable of handling categorical data, independence of specific distance metrics, and usefulness in assessing clustering algorithm performance and stability. It is particularly relevant in the context of measuring cluster similarity because it focuses on the overlap between data points assigned to different clusters. This similarity index quantifies how effectively the clusters are bifurcated. The Jaccard similarity index is found between the clusters of previous day consensus and present day's k-medoids clusters [44], i.e.

$$\mathrm{c}_d^* = \underset{\tilde{c}_d^i \in \tilde{\mathrm{C}}_d}{\operatorname{argmax}} \sum_{i=1}^{M} Jaccard(c_{d-1}^i, \tilde{c}_d^i)$$

where $\mathrm{c}_d^* \in \mathrm{C}_d$. In summary, the consensus clustering results from the previous day C_{d-1} are fed to the algorithm to contrast with the subsequent clusters to reach a consensus, see (3). Hence, the clusters resulting from the consensus algorithm are updated every 24 hours to achieve accurate forecasting.

2.2. Forecasting engine

The amount of electricity consumed by each user among all clusters is determined. Additionally, an AGP is utilized to perform the model forecast for each cluster. Specifically, the basic model formulation is applied to the data set for each cluster as the first step in constructing the regression model. The forecasting continuously increases the value of cluster-based forecasts. The predictions for the clusters are then combined. Thus, the regression model examines the relationship between energy use and climatic variables. Also, to enhance the model, different day types are included, after which the set comprises the hours and day of the week.

This work uses a probabilistic method to perform the STLF strategy with minimal information available from the houses based on GP regression. Complex responses can be modeled using a Gaussian process with additive functions while retaining interpretability [45]. The data is modeled as the output of the Bayesian non-parametric model. A function that follows a GP is defined as a probability distribution over functions, i.e.,

$$f \sim \mathcal{GP}(m, k), \tag{4}$$

where mean function $m : \mathcal{X} \to \mathbb{R}$ and covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are defined $\forall x \in \mathcal{X}$ as follows:

$$m(x) = \mathbb{E}[f(x)], \tag{5}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x)) \cdot (f(x') - m(x'))]. \tag{6}$$

$k(x, x')$ presents the covariance function or kernel. Additionally, combining multiple kernel functions enhances the forecasting system by revealing correlations between different variables, especially weather and calendar ones. Combining kernel functions to obtain a complex structure necessitates a positive semi-definite covariance matrix. Since kernel functions are significant for capturing complex relationships, we take advantage of three commonly used kernel functions. Then the model forecasting results under different quantiles are input into the kernel density estimation function, and thus the probability density function can be obtained. The main advantage of the GP model is that it can maintain the uncertainty about the variance associated with each point, which is rather crucial for probabilistic load forecasting [36]. Accordingly, this paper chooses an AGP structure incorporating Squared exponential (Se) and Matérn covariance functions for weather and calendar variables, respectively. Notably, the Matérn is applied to the sine

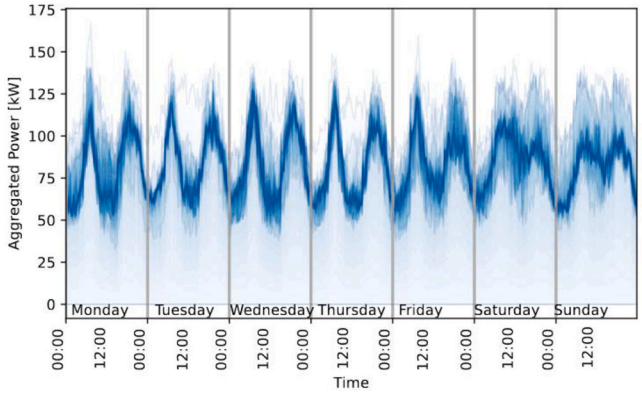

Fig. 3. Aggregated power consumption behavior across hours and days of the week for the case of 17 real data sampled each 15 min.

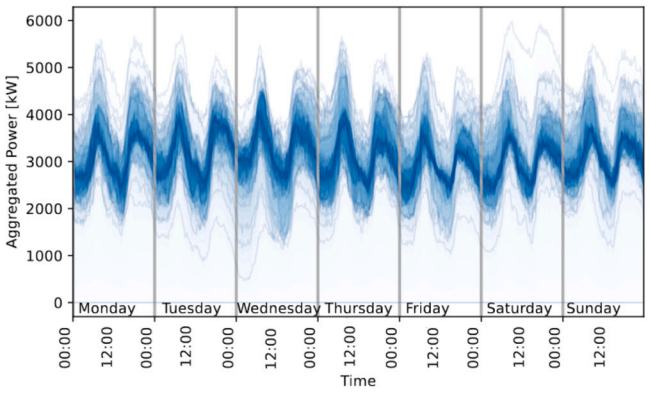

Fig. 4. Aggregated power consumption behavior across hours and days of the week for the case of 1000 houses sampled each 15 min.

and cosine functions of time of day and day of week indexes to explain their periodicity [13]. The commonly used covariance function associated with the weather has the form:

$$k_{Se}(\mathrm{x},\mathrm{x}') = \eta^2 \exp\left(-\frac{(||\mathrm{x}-\mathrm{x}'||)^2}{\ell^2}\right). \tag{7}$$

The covariance function hyperparameters η and ℓ control the length scale. Gaussian process models use kernels to describe the prior covariance between two function values. The GP is smoothened using this covariance function. The kernel function exploited to describe the calendar events is called Matérn 3/2 or Matérn 5/2, i.e.

$$k_{Ma}(\mathrm{x}-\mathrm{x}') = \sigma^2 \frac{2^{1-\nu}}{\tau(\nu)} \left(\sqrt{2\nu}\frac{\mathrm{x}-\mathrm{x}'}{\ell}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{\mathrm{x}-\mathrm{x}'}{\ell}\right), \tag{8}$$

where ℓ and ν are both positive hyperparameters, K_{ν} is the modified Bessel function. These kernels can tackle the STLF problem in this case of study [46]. Moreover, utilizing the traditional GP model involves calculating the determinant and inverse of the covariance matrix that brings formidable cubic computational complexity $O(N^3)$. In this work, the GP is trained with an amount of 2400 samples (4 quarters of one hour $\times$ 24 h $\times$ 25 days) in every iteration. This can bring the complexity of about 2400^3 (13824×10^6). Computing the Cholesky decomposition of such data is very complex for a normal computer. This situation promotes the proposed GP design intended to manipulate big data. As a result, the new data reinforcement necessitates using the sparse GP by employing the Fully Independent Training Conditional (FITC) approximation and the mean shift clustering technique. More details with a complete description can be found in Khansa et al. [4].

3. Validation framework

3.1. Data

We effectuate the proposed method on two datasets to test the performance of our proposed clustering and prediction strategies. In this study, simulations are performed using load data obtained from 17 real-life houses in the Québec, Canada region to accomplish the real possibilities of the suggested strategies. Notably, house consumption consists of electric heaters, and the data set covers winter days. Weather data from a nearby weather station were also collected. Information on power used during a 15-minute sampling interval makes up the aggregated demand data corresponding to the total power consumption of all households, as illustrated in Fig. 3. The second dataset consists of information on 1000 houses obtained from Hydro-Québec [47]. Fig. 4 depicts the aggregated power's weekly behaviour over 4 months grouped by days of the week. It is worth mentioning that the database includes information of high noise degree which is introduced in the data, to quantify the robustness of the methodology as mentioned

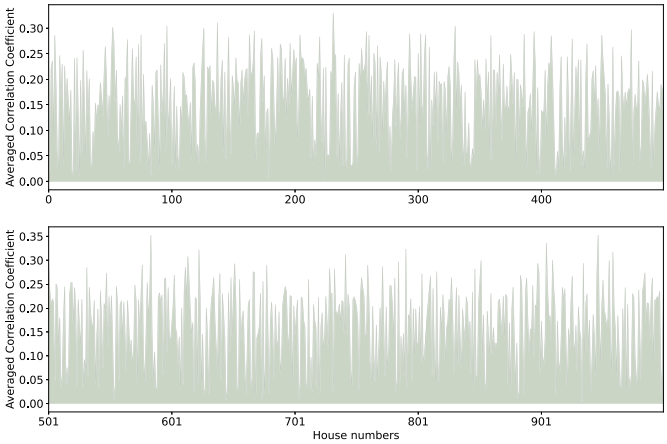

Fig. 5. Variation of the load data via averaged correlation coefficient with regards to datasize for the case of 1000 houses.

by the dataset's source. Additionally, in order to see the variation in load curves to validate if there is significant variation among different users, Fig. 5 demonstrates the correlation between each load profile in the dataset of 1000 houses. For brevity of presentation, the correlation coefficient for each house concerning other houses is averaged. The values in the figure are <0.5, indicating that the load profiles are diverse [30].

The dark-blue line in both Figs. 3 and 4 depict the average power consumption. It can be observed that weekdays have similar consumption patterns that are different from those of weekends. Unlike weekends, weekdays illustrate two distinct peaks in the morning and evening. The difference in weekly power consumption patterns is related to occupants' behaviour towards utilizing electrical loads. Moreover, temperature, humidity, and solar radiation are the most explanatory components, mainly accounting for climate-sensitive and calendar factors.

3.2. Benchmark models

Defining the optimal network structure, significantly influenced by the characteristics of the data sets, is a challenge for any deep learning model. A comparative study is conducted in this context to demonstrate the proposed approach's effectiveness across both datasets, offering a more accurate perspective on building energy demand forecasting. The proposed method is evaluated against baseline benchmark models, notably Support Vector Regression (SVR), Random Forest Regression (RFR) and Long Short Term Memory (LSTM) [29]. These non-parametric regression models are selected since they can capture

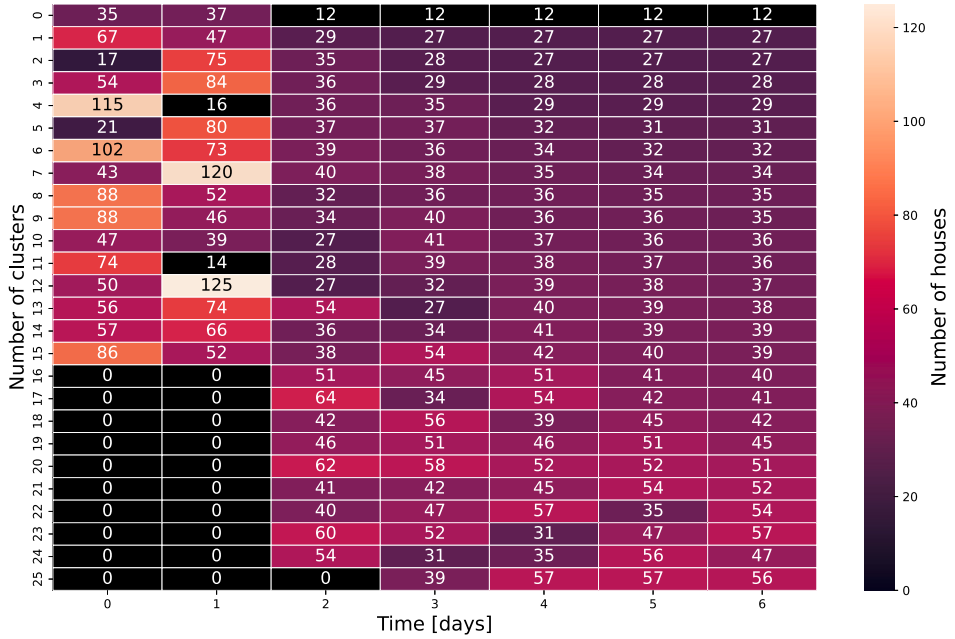

Fig. 6. Clusters labels depicting the number of houses in each cluster for the case of 1000 houses during 7 days.

long-term dependencies in the multivariate time series and extract their existing patterns for aggregated loads forecasting tasks.

As discussed, the high interpretability and non-parametric flexibility make the GP popular in resolving challenges in forecasting load and the case of big data. Accordingly, we have attempted to forecast the demand based on AGP for aggregated load from the load profile database with and without applying the proposed methodology of consensus clustering. The proposed method aggregates the energy consumption of households in each cluster into one-time series, forecasts the aggregate consumption of each cluster, and then aggregates the forecast of each cluster. This strategy addresses the efforts involved in improving the system-level intraday load forecasting by applying clustering to identify groups of customers with similar load consumption patterns to perform load forecasting. Conversely, the other strategy, without clustering, aggregates the energy consumption of all households and then forecasts the aggregate consumption. This effectively assesses the performance of the consensus clustering strategy of this paper.

3.3. Accuracy metrics

The main objective of the forecasting algorithm is to minimize the gap between the measured aggregated load and the forecasted aggregated load for the next day. The day-ahead aggregated load forecast accuracy is measured using five metrics. The most widely used statistical metric for determining forecast accuracy in the literature is Mean Absolute Percentage Error (MAPE) [48]. However, MAPE is unreliable and takes undefined values when there are zero values for the actuals, which can happen in demand forecasting. Additionally, it takes extreme values when the actuals are very close to zero. As a result, MAPE is either asymmetric or severely inaccurate for particular cases. Therefore, this paper utilizes R-squared (9), Mean Absolute Error (MAE) (10), Normalized Mean Absolute Error (NMAE) (11), Symmetric Mean Absolute Percentage Error (sMAPE) (12), and Root Mean Squared Error (RMSE) (13) that do not have the abovementioned limitations.

$$R^2 = 1 - \frac{\sum_{t=1}^{N}(y_t - \hat{y}_t)}{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2} \tag{9}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - \hat{y}_t\right|, \tag{10}$$

$$\mathrm{NMAE} = \frac{\sum_{t=1}^{N}\left|y_t - \hat{y}_t\right|}{\sum_{t=1}^{N}\left|y_t\right|}, \tag{11}$$

$$\mathrm{sMAPE} = \frac{100\%}{N}\sum_{t=1}^{N}\frac{\left|\hat{y}_t - y_t\right|}{(|\hat{y}_t| + |y_t|)/2}, \tag{12}$$

$$\mathrm{RMSE} = \left\{\frac{1}{N}\sum_{t=1}^{N}\left(\hat{y}_t - y_t\right)^2\right\}^{1/2}, \tag{13}$$

where $\hat{y}_t$ and y_t present predicted and actual power usages for N discrete-time samples, respectively.

4. Results and discussion

4.1. Results of clustering

The variation in the load pattern of households hinges on the behaviour of the individual home associated with power usage. Generally, aggregate diversity smoothens daily load patterns, making substation loads relatively more predictive, but a single client's electricity consumption depends more on underlying human behaviour. In an individual household, the daily routines and lifestyles of the residents, as well as the types of major appliances owned, may have a more direct impact on the short-term load patterns that follow. For instance, some households may have a fixed routine of turning on the dryer after using their washing machine, which involves a high probability of significant electricity consumption within an hour or two. This phenomenon explains the root causes of residential energy consumption by classifying the houses into clusters with similar behaviors.

The proposed methodology of consensus clustering is effectuated on the power consumption profile data from 1000 houses. The number of households included in each cluster is very diverse (Fig. 6), considering the same type of partition and among different time series similarity measures. Precisely, 1000 houses are classified into 26 clusters, as depicted in Fig. 6. The number of houses in each cluster changes daily for the case of 1000 houses. This implies the effectiveness of consensus to enhance the CBAF model since the index based on a Jaccard similarity reaches one on the fourth day of execution [49]. Fig. 7 presents the result of the convergence of consensus clustering during eight days. Closer to one value indicates a better clustering result and the convergence

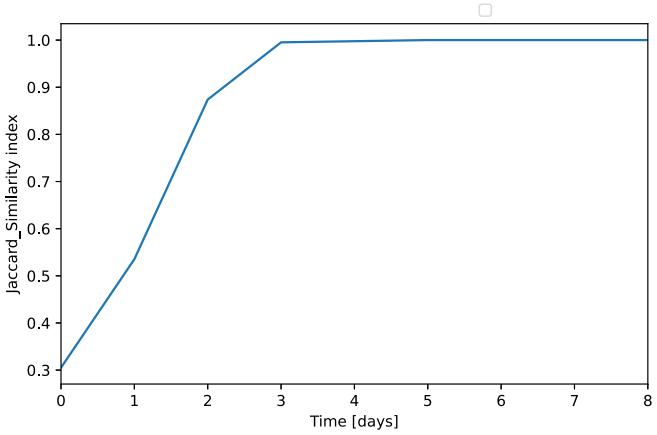

Fig. 7. Consensus clustering accuracy selected by the Jaccard similarity index.

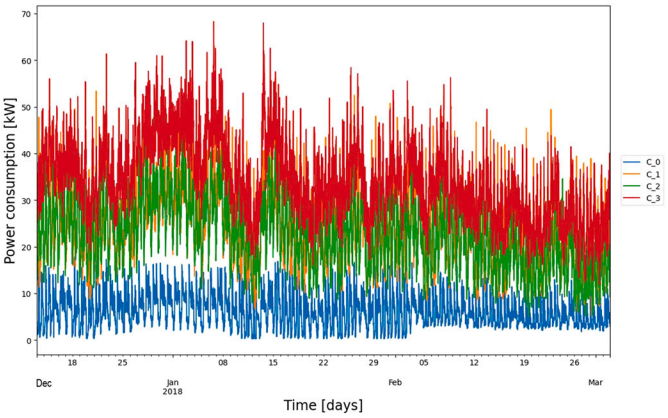

Fig. 8. Power consumption patterns resulted from the clustering for the case of 17 houses in the winter period, Blue: Cluster 1 - 2 houses, Orange: Cluster 2 - 5 houses, Green: Cluster 3 - 5 houses and Red: Cluster 4 - 5 houses.

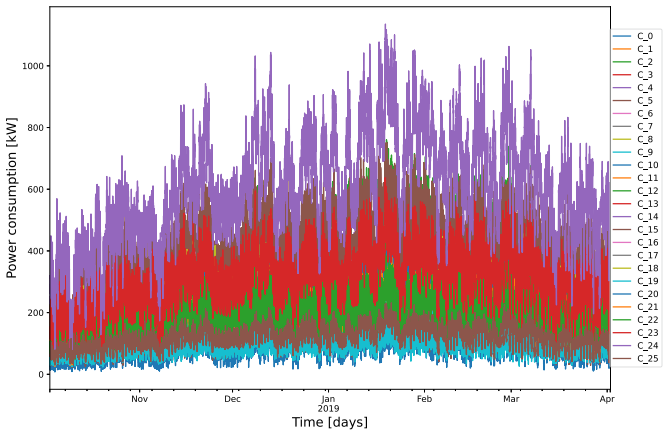

Fig. 9. Power consumption patterns resulted from the clustering for the case of 1000 houses in the winter period.

of the consensus algorithm. It can be deduced that this method iteratively offers more performance accuracy that, in turn, improves daily clustering (Fig. 7). Furthermore, Fig. 9 depicts the power consumption patterns of different clusters. The mean computed is highlighted in different colors depicting different clusters, with the dominant cluster shown in violet consuming more power in comparison to other clusters.

Similarly, the proposed methodology of consensus clustering is effectuated on the power consumption profile data from 17 houses. Fig. 8 shows the clustering results from the proposed methodology of the

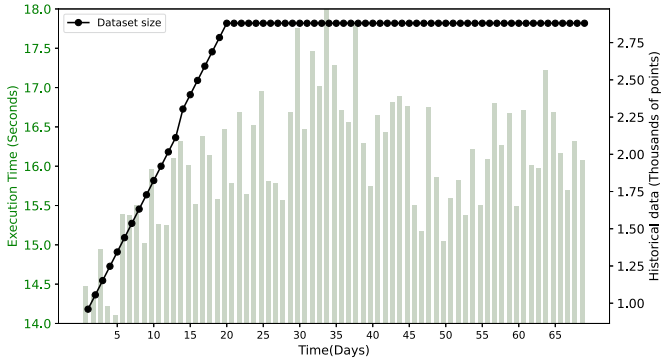

Fig. 10. The execution time towards the increase of historical data size in each iteration of forecasting based on CABF with AGP proposed model.

aggregated power profiles of all 17 households distributed in four clusters. It is clear that the aggregated load is relatively consistent, and its pattern is evident that corresponds to the different lifestyles that each household may possess. In particular, the apparent performance is achieved through the proposed methodology, which confirms that clustering can effectively improve prediction accuracy since one cluster incorporates houses with similar behavior. Besides, matching the accuracy using multiple predictors helps increase the precision and performance of the forecasting results.

Fig. 10 shows the execution time concerning the data size of the proposed AGP for each day performing the forecasting. Accordingly, the data with 15-minute granularity was exploited in this case study, which increased the training data and brought the complexity, as mentioned in section 2.2. The new data reinforced the utilization of the sparse AGP design proposed in this study, as explained in [4]. The look-back window method was adopted [50], to deal with the algorithmic complexity. Accordingly, the AGP model is retrained at each iteration by increasing historical data throughout the dataset. The window size was chosen to understand the impact of the window size on the prediction accuracy. Starting from day 1 to day 20 is the training of the dataset, which is fixed by the look-back window. During this time, the size of the dataset (black dots in Fig. 10) increases, and the execution time varies between 14 seconds and 18 seconds. The dataset expansion stops when the look-back window reaches 20 days; then, it will be constant for the rest of the period. This technique maintained the number of clusters for the primary data set at less than 85 across the whole dataset. Therefore, they decreased the number of inputs by a ratio of around 96% (85/2400) in every iteration. Consequently, the evaluation process was repeated for all the days. Note that this proposed structure does not consider auto-regressive inputs and also does not use the test data to retrain. The look-back window uses historical observations and predicts for the next day, which is effective and accurate in time-series forecasting [30].

4.2. Results of forecasting

4.2.1. Cluster based vs standard results

To determine the performance of the proposed CBAF methodology, the results of forecasting the electric demand of residential houses are presented. Here, we compare the prediction accuracy of the baseline predictor of the aggregated energy demand against the proposed CBAF methodology. As aggregated load forecasting can be performed using a direct forecasting technique, the additive GP model for aggregate forecasting is applied with and without clustering to adjudge the influence of cluster-based aggregate forecasts. The design of AGP can effectively reflect the dynamic nature of aggregated power consumption due to weather occurrences and occupants' behaviour, as represented by calendar elements. In fact, based on the time series similarity of power

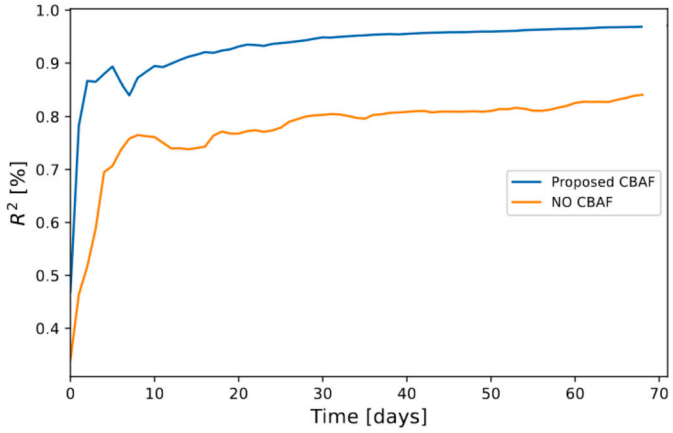

Fig. 11. Comparing the R^2 of the proposed CBAF strategy with the methodology without CBAF for the case of 17 houses.

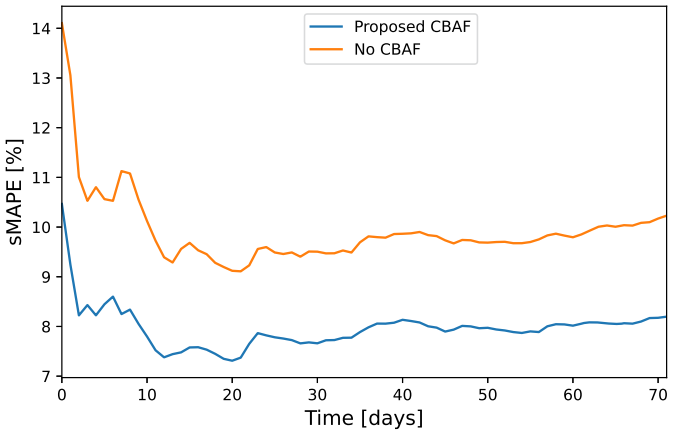

Fig. 12. Comparing the sMAPE of the proposed CBAF strategy with the methodology without CBAF for the case of 17 houses.

Table 2
Accuracy metrics.

Strategy	Dataset	sMAPE [%]	R^2 [%]
Proposed CBAF	17 Houses	8.19	0.95
	1000 Houses	5.82	0.96
No CBAF	17 Houses	10.41	0.88
	1000 Houses	5.92	0.89

demand in each cluster and the existence of meaningful interactions between influential components, the forecasting results are evaluated.

Fig. 11 compares the proposed CBAF employing the AGP model to forecast with a direct method of forecasting without clustering. The coefficient of determination R^2 is utilized to measure the accuracy of the electricity demand prediction. For prediction of the next day, the results from previous days are utilized for training a model in order to estimate the next day. Besides, to evaluate with a normalized metric, we have utilized sMAPE (see Fig. 12). Concerning the result of the cluster-based forecast, 8.19% sMAPE is noted in contrast to 10.41% without clustering. Also, R^2 of the proposed CBAF is closer to 1, indicating superior predictions in contrast to the direct forecasting without clusters (Table. 2).

The day-ahead load forecast is performed for several consecutive weeks to analyze the significance of results with much larger datasets. We observe the variation of the error of each cluster for this period. Consequently, the accuracy of the day-ahead forecast without clustering is compared to the accuracy of the day-ahead forecast with clustering for data of 1000 houses. The forecasts for each cluster are summed and compared with the average error calculated for the aggregate load.

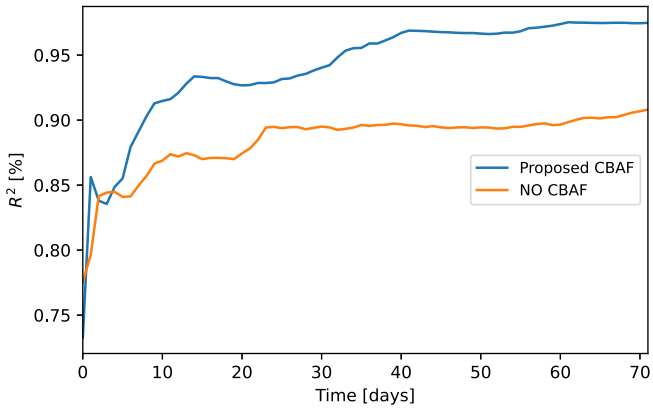

Fig. 13. Model prediction improvement due to gradual increase in the performance of R^2, for the proposed CBAF strategy compared to the case without CBAF for 1000 houses.

Intuitively, aggregating forecasts with clustering outperform forecasting without clustering since aggregate consumption patterns are more regular than individual consumption patterns. This level of accuracy is expected because forecasting accuracy without clustering has been reported to reduce with an increase in the level of aggregation, as shown in Figs. 13 and 14. Concerning the accuracy metrics result of 1000 houses, 5.82% sMAPE is noted in contrast to 8.29% without clustering. Besides, R^2 for the proposed CBAF remains closer to 1, proving better than the direct forecasting without clustering (Table. 2).

Table 3
The average values of accuracy metrics scores for different forecasting models.

Strategy	Dataset	Models	R^2 [%]	MAE [kW]	NMAE [kW]	sMAPE [%]	RMSE [kW]
Proposed CBAF	17 Houses	Support Vector Regression	0.69	9.31	0.10	10.23	11.97
		Random Forest Regression	0.71	9.17	0.15	11.36	12.39
		Long Short Term Memory	0.77	8.45	0.09	10.22	10.60
		Proposed Additive GP	**0.95**	**6.80**	**0.07**	**8.19**	**8.66**
	1000 Houses	Support Vector Regression	0.75	170.17	0.05	4.80	194.99
		Random Forest Regression	0.80	145.07	0.04	4.94	173.04
		Long Short Term Memory	0.89	110.63	0.04	6.63	127.91
		Proposed Additive GP	**0.96**	**97.29**	**0.03**	**5.82**	**120.46**
NO CBAF	17 Houses	Support Vector Regression	0.70	13.93	0.10	11.42	18.16
		Random Forest Regression	0.79	12.47	0.15	10.31	15.88
		Long Short Term Memory	0.76	13.18	0.11	11.00	17.39
		Proposed Additive GP	**0.88**	**12.41**	**0.09**	**10.41**	**12.41**
	1000 Houses	Support Vector Regression	0.75	168.93	0.15	4.98	197.58
		Random Forest Regression	0.80	155.09	0.13	5.58	199.25
		Long Short Term Memory	0.85	160.93	0.11	6.70	140.74
		Proposed Additive GP	**0.89**	**100.50**	**0.06**	**5.92**	**130.65**

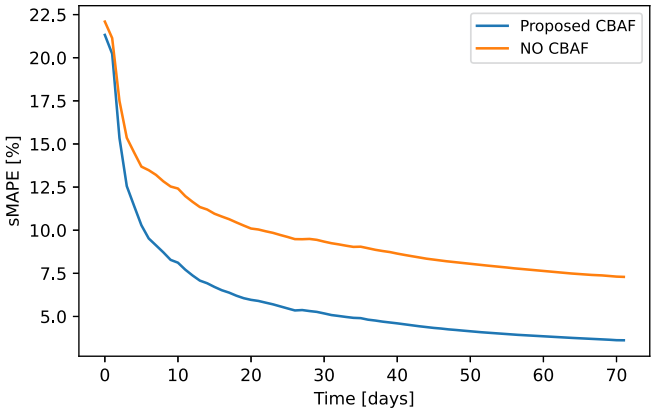

Fig. 14. Comparing the sMAPE of the proposed CBAF strategy with the methodology without CBAF for the case of 1000 houses.

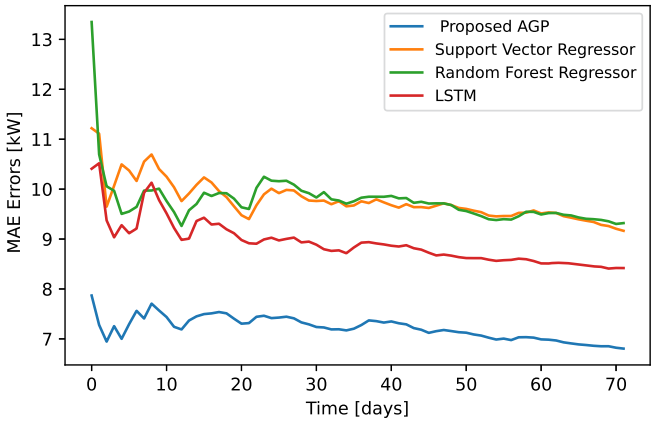

Fig. 15. Comparing the MAE of different forecasting models with the proposed CBAF strategy for the case of 17 houses.

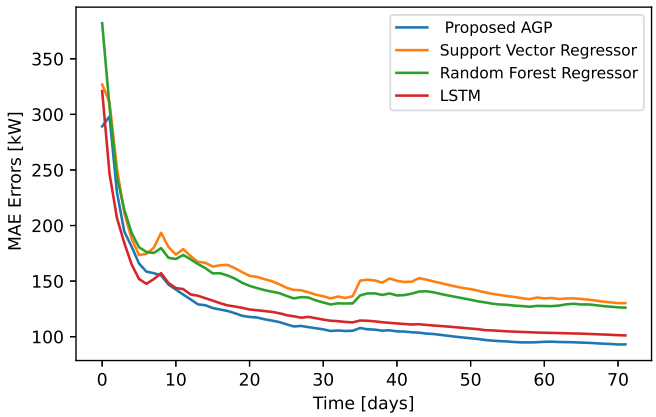

Fig. 16. Comparing the MAE of different forecasting models with the proposed CBAF strategy for the case of 1000 houses.

The concept of aggregating forecasts with a cluster-based method delivers better accurate forecasting performance than forecasting the aggregate directly. The consensus clustering performance significantly improved clusterings' robustness for both databases. Importantly, the forecasting accuracy is increased since the incorporation of the proposed consensus clustering exploits smart meter data to enhance load forecasting by clustering consumers based on their similar load consumption patterns.

4.2.2. Benchmark results

We compare the forecasting results of the proposed CBAF technique incorporating non-parametric AGP with the CBAF technique incorporating other benchmark models to validate the reliability and efficacy of the method proposed in this work. The analysis is carried out for a dataset of 17 real-life houses and a much larger dataset of 1000 houses. Proven benchmark non-parametric models as SVR, RFR, and LSTM are employed for the comparison. These models are known for their ability to capture complex patterns and dependencies in time series data, especially with historical data and a horizon of 24 h as well. This effectiveness was demonstrated for forecasting tasks/multivariate time series analysis in [4], which highlights the use of non-parametric regression models in the existing case of study. Furthermore, the proposed model uses the look-back window technique-managed day-ahead forecasting results for the entire dataset, providing accurate predictions.

Table 3 gives a breakdown of R^2, MAE, NMAE, sMAPE, and RMSE metrics for various models to compare the actual and predicted load forecast. Figs. 15 and 16 show the comparison of MAE for forecasting of 70 days with CBAF strategy accompanied with different forecasting models for 17 and 1000 houses, respectively. On the other hand, Fig. 17 shows the same comparison without the CBAF strategy. From Figs. 15, 16 and 17, it is evident that the suggested methodology consistently outperforms comparative methods for both datasets. The additive GP model performs better than the parametric SVR, RFR and non-parametric LSTM. The additive GP's significant improvement in the RMSE score mainly demonstrates its ability to fit a regression line that correctly captures the data points. It can be noticed that the suggested approach surpasses the benchmark models as illustrated in Table 3. The additive GP design can adequately capture the dynamic stochastic na-

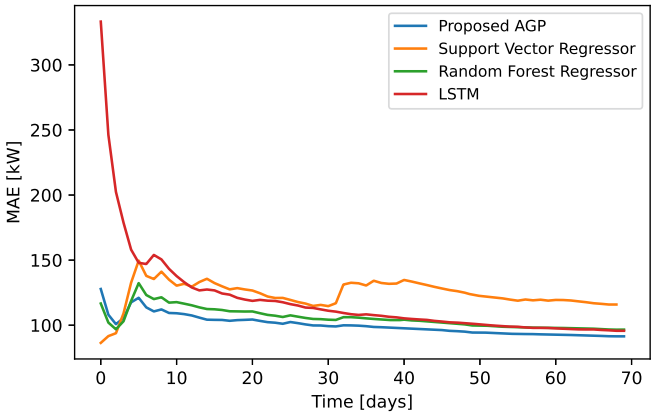

Fig. 17. Comparing the MAE of different forecasting models without the proposed CBAF strategy for the case of 1000 houses.

ture of aggregated power consumption due to weather phenomena and occupants' behaviour, represented by calendar factors. Indeed, the proposed approach has worked well for classifying time-series data, and a high level of forecast accuracy could be obtained. Additionally, with a larger dataset, more profiles were achieved per cluster, which resulted in lower volatility and more accurate load forecasting.

5. Conclusion

A unified approach to improve short-term load forecasting for residential electricity demand is proposed in this work. Particularly, a consensus-based time-series clustering approach employing k-medoids clustering is utilized. Furthermore, AGP, a non-parametric regression model encompassing the weather and calendar events kernel was utilized to forecast the aggregate of each cluster. A simulation study was carried out to reveal the efficacy of the suggested forecasting method over time-series datasets of 17 and 1000 houses. The proposed consensus clustering algorithm excelled in identifying clusters based on the 24 h power profile of residences in congruence with the accumulating time-series daily power profiles. Moreover, a comparative study employing other benchmark forecasting models with the proposed technique was also presented to demonstrate the AGP model's superiority. The accuracy metrics reveal that the proposed cluster-based aggregated forecasting utilizing the proposed AGP is a superior alternative to the existing short-term electricity demand forecasting techniques. In future work, the goal is to strike an optimal balance between the quality and performance of time series clustering for prediction analysis. Additionally, understanding the residual component related to the occupancy behaviour is an important aspect of all real-world forecasting tasks. To enhance the forecasting accuracy, concurrent prediction intervals will be utilized, and to study the effect of uncertainties in the forecasting process, occupancy behaviour will be incorporated.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgement

The authors would like to thank the Laboratoire des technologies de l'énergie d'Hydro-Québec, the Natural Science and Engineering Research Council of Canada, and the Foundation of Université du Québec à Trois-Rivières.

References

[1] X. Guo, Y. Gao, Y. Li, D. Zheng, D. Shan, Short-term household load forecasting based on Long- and Short-term Time-series network, Energy Rep. 7 (4 2021) 58–64, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S2352484721001219.

[2] K. Qian, X. Wang, Y. Yuan, Research on regional short-term power load forecasting model and case analysis, Processes 9 (9) (9 2021) 1617, [Online], Available: https://www.mdpi.com/2227-9717/9/9/1617.

[3] A. Kalakova, H.S.V.S.K. Nunna, P.K. Jamwal, S. Doolla, A novel genetic algorithm based dynamic economic dispatch with short-term load forecasting, IEEE Trans. Ind. Appl. 57 (3) (5 2021) 2972–2982, [Online], Available: https://ieeexplore.ieee.org/document/9376973/.

[4] K. Dab, K. Agbossou, N. Henao, Y. Dubé, S. Kelouwani, S.S. Hosseini, A compositional kernel based Gaussian process approach to day-ahead residential load forecasting, Energy Build. 254 (1 2022) 111459, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S037877882100743X.

[5] A. Cini, S. Lukovic, C. Alippi, Cluster-based aggregate load forecasting with deep neural networks, in: 2020 International Joint Conference on Neural Networks (IJCNN), vol. 7, IEEE, 2020, pp. 1–8, [Online], Available: https://ieeexplore.ieee.org/document/9207503/.

[6] K. Park, S. Yoon, E. Hwang, Hybrid load forecasting for mixed-use complex based on the characteristic load decomposition by pilot signals, IEEE Access 7 (2019) 12 297–12 306, [Online], Available: https://ieeexplore.ieee.org/document/8610068/.

[7] S. Aghabozorgi, A. Seyed Shirkhorshidi, T. Ying Wah, Time-series clustering – a decade review, Inf. Sci. 53 (10 2015) 16–38, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000733.

[8] G. Le Ray, P. Pinson, Online adaptive clustering algorithm for load profiling, Sustain. Energy Grids Netw. 17 (2019) 3.

[9] T. Yang, N. Pasquier, F. Precioso, Semi-supervised consensus clustering based on closed patterns, Knowl.-Based Syst. 235 (1 2022) 107599, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705121008613.

[10] X. Cheng, L. Wang, P. Zhang, X. Wang, Q. Yan, Short-term fast forecasting based on family behavior pattern recognition for small-scale users load, Clust. Comput. 25 (3) (6 2022) 2107–2123, [Online], Available: https://link.springer.com/10.1007/s10586-021-03362-9.

[11] K.B. Debnath, M. Mourshed, Forecasting methods in energy planning models, Renew. Sustain. Energy Rev. 88 (8) (2016) 297–325, 5 2018 [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032118300200.

[12] A. Sfetsos, C. Siriopoulos, Time series forecasting with a hybrid clustering scheme and pattern recognition, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 34 (3) (5 2004) 399–405, [Online], Available: http://ieeexplore.ieee.org/document/1288351/.

[13] M. Blum, M. Riedmiller, Electricity demand forecasting using Gaussian processes, AAAI Workshop - Technical Report, vol. WS-13-15, 2013, pp. 10–13, [Online], Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.649.6285&rep=rep1&type=pdf.

[14] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, H. Zareipour, Energy forecasting: a review and outlook, IEEE Open Access J. Power Energy 7 (11) (2020) 376–388, [Online], Available: https://ieeexplore.ieee.org/document/9218967/.

[15] Q. Hu, F. Li, C.-f. Chen, A smart home test bed for undergraduate education to bridge the curriculum gap from traditional power systems to modernized smart grids, IEEE Trans. Ed. 58 (1) (2 2015) 32–38, [Online], Available: http://ieeexplore.ieee.org/document/6815765/.

[16] L. Kotzur, P. Markewitz, M. Robinius, D. Stolten, Impact of different time series aggregation methods on optimal energy system design, Renew. Energy 117 (3 2018) 474–487, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148117309783.

[17] G.A. Susto, A. Cenedese, M. Terzi, Time-series classification methods: review and applications to power systems data, in: Big Data Application in Power Systems, Elsevier, 2018, pp. 179–220, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128119686000097.

[18] O. Motlagh, A. Berry, L. O'Neil, Clustering of residential electricity customers using load time series, Appl. Energy 237 (3 2019) 11–24, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261918318816.

[19] X. Ruhang, Efficient clustering for aggregate loads: an unsupervised pretraining based method, Energy 210 (11 2020) 118617, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544220317254.

[20] P. Mandal, T. Senjyu, N. Urasaki, T. Funabashi, A neural network based several-hour-ahead electric load forecasting using similar days approach, Int. J. Electr. Power Energy Syst. 28 (6) (7 2006) 367–373, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061506000275.

[21] C.H. Jin, G. Pok, Y. Lee, H.-W. Park, K.D. Kim, U. Yun, K.H. Ryu, A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting, Energy Convers. Manag. 90 (1 2015) 84–92, https://doi.org/10.1016/j.enconman.2014.11.010, [Online], Available:.

[22] E. Atam, E.C. Kerrigan, Optimal partitioning of multithermal zone buildings for decentralized control, IEEE Trans. Control Netw. Syst. 8 (3) (9 2021) 1540–1551, [Online], Available: https://ieeexplore.ieee.org/document/9409744/.

[23] W. Kong, Z.Y. Dong, Y. Jia, D.J. Hill, Y. Xu, Y. Zhang, Short-term residential load forecasting based on LSTM recurrent neural network, IEEE Trans. Smart Grid 10 (1) (1 2019) 841–851, [Online], Available: https://ieeexplore.ieee.org/document/8039509/.

[24] T.K. Wijaya, M. Vasirani, S. Humeau, K. Aberer, Cluster-based aggregate forecasting for residential electricity demand using smart meter data, in: 2015 IEEE International Conference on Big Data (Big Data), IEEE, 10 2015, pp. 879–887, [Online], Available: http://ieeexplore.ieee.org/document/7363836/.

[25] X. Cao, S. Dong, Z. Wu, Y. Jing, A data-driven hybrid optimization model for short-term residential load forecasting, in: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, IEEE, 10 2015, pp. 283–287, [Online], Available: http://ieeexplore.ieee.org/document/7363083/.

[26] H. Jahangir, H. Tayarani, S.S. Gougheri, M.A. Golkar, A. Ahmadian, A. Elkamel, Deep learning-based forecasting approach in smart grids with microclustering and bidirectional LSTM network, IEEE Trans. Ind. Electron. 68 (9) (9 2021) 8298–8309, [Online], Available: https://ieeexplore.ieee.org/document/9145791/.

[27] G. Rouwhorst, E.M.S. Duque, P.H. Nguyen, H. Slootweg, Improving clustering-based forecasting of aggregated distribution transformer loadings with gradient boosting and feature selection, IEEE Access 10 (2022) 443–455, [Online], Available: https://ieeexplore.ieee.org/document/9661304/.

[28] P. Laurinec, M. Loderer, M. Lucka, V. Rozinajova, Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption, J. Intell. Inf. Syst. 53 (2) (10 2019) 219–239, [Online], Available: http://link.springer.com/10.1007/s10844-019-00550-3.

[29] A.-N. Khan, N. Iqbal, A. Rizwan, R. Ahmad, D.-H. Kim, An ensemble energy consumption forecasting model based on spatial-temporal clustering analysis in residential buildings, Energies 14 (11) (5 2021) 3020, [Online], Available: https://www.mdpi.com/1996-1073/14/11/3020.

[30] Y. Wang, Y. Kong, X. Tang, X. Chen, Y. Xu, J. Chen, S. Sun, Y. Guo, Y. Chen, Short-term industrial load forecasting based on ensemble hidden Markov model, IEEE Access 8 (2020) 160 858–160 870, [Online], Available: https://ieeexplore.ieee.org/document/9183956/.

[31] Z. Chen, Y. Chen, T. Xiao, H. Wang, P. Hou, A novel short-term load forecasting framework based on time-series clustering and early classification algorithm, Energy Build. 251 (11 2021) 111375, https://doi.org/10.1016/j.enbuild.2021.111375, [Online], Available.

[32] M.R. Baker, K.H. Jihad, H. Al-Bayaty, A. Ghareeb, H. Ali, J.-K. Choi, Q. Sun, Uncertainty management in electricity demand forecasting with machine learning and ensemble learning: case studies of COVID-19 in the US metropolitans, Eng. Appl. Artif. Intell. 123 (8 2023) 106350, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0952197623005341.

[33] J. Luo, T. Hong, S.-C. Fang, Benchmarking robustness of load forecasting models under data integrity attacks, Int. J. Forecast. 34 (1) (1 2018) 89–104, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0169207017300900.

[34] Zhang Yun, Zhou Quan, Sun Caixin Lei Shaolan, Liu Yuming, Song Yang, RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment, IEEE Trans. Power Syst. 23 (3) (8 2008) 853–858, [Online], Available: http://ieeexplore.ieee.org/document/4509471/.

[35] Zizi Zhang, Hong Li, Yang Zhao, Xiaobo Hu, Short-Term Load Forecasting Based on the Grid Method and the Time Series Fuzzy Load Forecasting Method, International Conference on Renewable Power Generation (RPG 2015), vol. 2015, Institution of Engineering and Technology, 2015, CP679, [Online], Available: https://digital-library.theiet.org/content/conferences/10.1049/cp.2015.0382.

[36] Y. Yang, S. Li, W. Li, M. Qu, Power load probability density forecasting using Gaussian process quantile regression, Appl. Energy 213 (8) (3 2018) 499–509, https://doi.org/10.1016/j.apenergy.2017.11.035, https://linkinghub.elsevier.com/retrieve/pii/S0306261917316100, [Online], Available:.

[37] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, vol. 7 (5), MIT Press, 2006, pp. 25–51, www.GaussianProcess.org/gpml.

[38] O. Corradi, H. Ochsenfeld, H. Madsen, P. Pinson, Controlling electricity consumption by forecasting its response to varying prices, IEEE Trans. Power Syst. 28 (1) (2 2013) 421–429, [Online], Available: http://ieeexplore.ieee.org/document/6205637/.

[39] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of smart meter data analytics: applications, methodologies, and challenges, IEEE Trans. Smart Grid 10 (3) (5 2019) 3125–3148, [Online], Available: https://ieeexplore.ieee.org/document/8322199/.

[40] C.A. Ratanamahatana, E. Keogh, Making time-series classification more accurate using learned constraints, in: Proceedings of the 2004 SIAM International Conference on Data Mining, University of California - Riverside, Philadelphia, Society for Industrial and Applied Mathematics, 4 2004, pp. 11–22, [Online], Available: https://epubs.siam.org/doi/10.1137/1.9781611972740.2.

[41] S. Soheily-khah, Generalized k-means based clustering for temporal data under time warp, thesis, 2017, [Online], Available: https://tel.archives-ouvertes.fr/tel-01680370v2.

[42] D.J. Bemdt, J. Clifford, Using Dynamic Time Warping to FindPatterns in Time Series, Stern School of Business New York University, New York, 1994, Tech. Rep. [Online], Available: www.aaai.org.

[43] S. Coleman, P.D.W. Kirk, C. Wallace, Consensus clustering for Bayesian mixture models, BMC Bioinform. 23 (1) (7 2022) 290, [Online], Available: http://www.ncbi.nlm.nih.gov/pubmed/35864476, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC9306175.

[44] A. Al-Najdi, A closed patterns-based approach to the consensus clustering problem, Ph.D. dissertation, Universite Cote d'Azur, Cote D'Azur, 11 2016, [Online], Available: https://theses.hal.science/tel-01478626.

[45] D. Duvenaud, H. Nickisch, C.E. Rasmussen, Additive Gaussian processes, Mach. Learn. (stat.ML) 12 (2011), [Online], Available: http://arxiv.org/abs/1112.4394.

[46] D. Duvenaud, J.R. Lloyd, R. Grosse, J.B. Tenenbaum, Z. Ghahramani, Structure discovery in nonparametric regression through compositional kernel search, in: 30th International Conference on Machine Learning, ICML 2013, in: PART 3, vol. 28, 2 2013, pp. 2203–2211, [Online], Available: http://arxiv.org/abs/1302.4922.

[47] S. Sansregret, K. Lavigne, B. Le Lostec, L. Francois, F. Guay, High resolution bottom-up residential electrical model for distribution networks planning, in: Building Simulation Conference Proceedings, vol. 5, 2019, pp. 3540–3547, [Online], Available: http://www.ibpsa.org/proceedings/BS2019/BS2019_210716.pdf.

[48] C. Tofallis, A better measure of relative prediction accuracy for model selection and model estimation, J. Oper. Res. Soc. 66 (8) (8 2015) 1352–1362, [Online], Available: https://www.tandfonline.com/doi/full/10.1057/jors.2014.103.

[49] S.S. Hamidi, E. Akbari, H. Motameni, Consensus clustering algorithm based on the automatic partitioning similarity graph, Data Knowl. Eng. 124 (11 2019) 101754, [Online], Available: https://linkinghub.elsevier.com/retrieve/pii/S0169023X18304919.

[50] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, Tech. Rep., OTexts, 2018, [Online], Available: https://otexts.com/fpp3/.

3.3 Analyse des incertitudes dans les prévisions à court terme

3.3.1 Contexte

Les agrégateurs sur les marchés de flexibilité agissent en tant d'intermédiaires, rassemblant la flexibilité énergétique des consommateurs auprès des opérateurs de réseaux ou des gestionnaires de systèmes de distribution (DSO). Leur rôle fondamental dans la régulation du réseau est d'offrir des réductions de charge basée sur des limites de puissance en estimant la charge attendue des consommateurs dans des scénarios de réponse à la demande. Cependant, l'incertitude inhérente au comportement des consommateurs pose un défi significatif, entraînant des écarts entre la consommation d'énergie estimée et la réalité. Lors des appels d'offres de paquets d'énergie à échanger, les agrégateurs doivent tenir compte de ces écarts potentiels pour garantir des offres compétitives et réalisables. L'évaluation de la réduction potentielle repose principalement sur les agrégateurs, qui doivent trouver un équilibre entre stimuler la participation des consommateurs et estimer avec précision les réductions de charge. La sous-estimation risque de compromettre les objectifs, tandis que la surestimation peut faire perdre de revenus. Ainsi, la capacité des agrégateurs à anticiper ces déviations devient centrale pour assurer leur viabilité économique et l'efficacité des marchés de flexibilité.

Transition vers la modélisation et prévisions probabilistes: Dans ce contexte, et en lien avec les chapitres précédents, la prévision de la charge et la classification jouent un rôle essentiel dans la gestion efficace de la consommation d'énergie résidentielle. Ces éléments contribuent à la planification stratégique des futures installations de production d'énergie. Cependant, la dynamique de l'utilisation de l'énergie est soumise à diverses incertitudes, telles que les facteurs environnementaux, les exigences de confort des occupants. L'évolution prévu dans les prévisions s'oriente vers des approches

probabilistes. La prévision probabiliste, de plus en plus populaire, offre une vision complète des résultats potentiels, contrairement aux prévisions déterministes. Trois types principaux de prévisions probabilistes sont explorés: estimation des quantiles, intervalles de confiance et fonctions de densité complètes. Ces approches visent à maximiser la précision des distributions prédictives. Ainsi, les prévisions probabilistes pour une variable aléatoire Y_t à l'instant t sont définies par sa fonction de densité de probabilité f_t et sa fonction de distribution cumulative F_t. Les quantiles, les intervalles de confiance, et les fonctions de densité complètes jouent un rôle clé dans cette approche. Cette formalisation vise à équilibrer l'incitation à la participation des consommateurs et l'estimation précise des réductions de charge, assurant ainsi que les offres des agrégateurs sont compétitives et réalisables. En résumé, bien que les agrégateurs ne puissent pas éliminer l'incertitude postérieure, leur capacité à anticiper et à intégrer ces déviations dans leurs stratégies d'offre est importante. Cette anticipation garantit non seulement leur viabilité économique mais contribue également à l'efficacité et à la stabilité globales des marchés de flexibilité.

3.3.2 Méthodologie

Le but de ce travail est de quantifier l'incertitude dans la prévision de charge pour les réseaux électriques intelligents en utilisant des statistiques non paramétriques. La méthodologie proposée introduit une approche basée sur un modèle statistique pour fournir une représentation plus complète de l'incertitude et de l'étude des variations de la charge. Elle fournit des valeurs de prévision de la charge sous forme de distributions globales, qui sont ensuite échantillonnées pour générer de nouvelles données à partir desquelles la fonction de densité de probabilité est extraite pour quantifier les incertitudes dans les prévisions, exprimée par des intervalles de confiance autour de la variable de sortie. Ce processus facilite l'identification des besoins de flexibilité en matière de

consommation d’énergie, ce qui est essentiel pour une prise de décision éclairée.

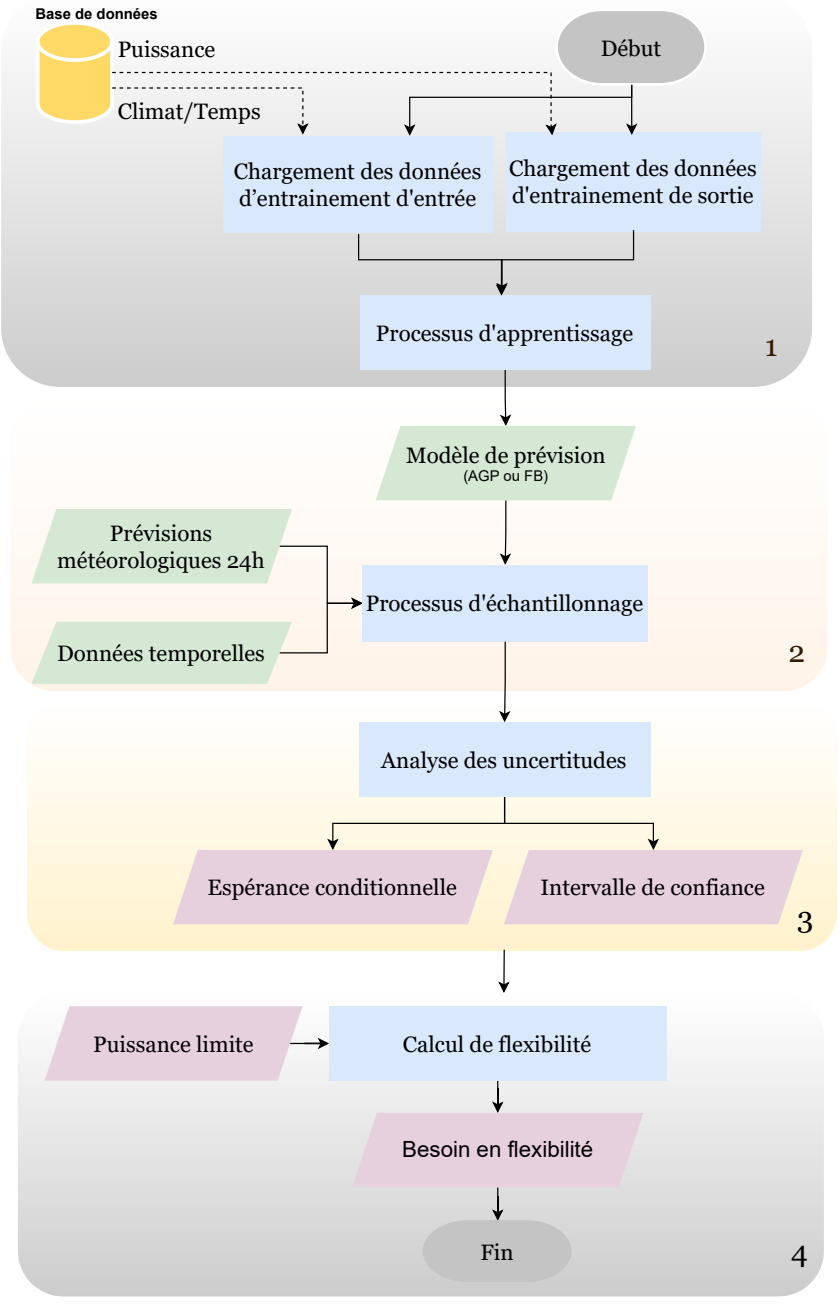

FIG. 3.3 Méthodologie proposée pour l'analyse des incertitudes des prévisions

La stratégie proposée est appliquée à un ensemble de 1000 maisons afin d'évaluer son efficacité à quantifier les incertitudes pour les prévisions. En outre, une étude de cas illustrative est présentée, centrée sur un ensemble composé de 14 maisons, toutes alimentées par un transformateur, afin de valider l'efficacité de la méthode suggérée. La méthodologie générale de la modélisation de l'incertitude est schématisée par la Figure 3.3. Sur cette figure, le premier bloc 1 et la couleur verte représente les travaux antérieurs, tandis que les autres couleurs indiquent les nouveaux travaux envisagés. La première étape consiste à identifier les données, puis à appliquer le modèle de prévision basé sur le AGP [91] ou bien le modèle *Prophet* [91]. Ces étapes sont décrites en détail dans le troisième article. Ainsi, *Prophet* est un outil de prévision développé par Facebook pour la série temporelle, souvent utilisé pour la prédiction de la consommation résidentielle d'énergie. Ce modèle sert aussi a capturer des tendances linéaires ou non linéaires dans les données. Il fournit des intervalles de confiance pour ses prédictions, ce qui est intéressant pour la planification du risque. Notons que pour le *Prophet*, il y a trois sources d'incertitude dans la prévision : l'incertitude dans la tendance, l'incertitude dans les estimations de la saisonnalité, et l'ajout d'un facteur d'incertitude dans la prévision.

La prochaine étape implique l'utilisation des prédictions du modèle de prévision dans le processus d'échantillonnage. En exploitant les principes des statistiques bayésiennes, l'AGP offre un cadre adaptable pour saisir les incertitudes et les relations présentes dans les données. Lorsqu'il est utilisé pour la prédiction, l'AGP génère des estimations ponctuelles et offre également une distribution postérieure complète pour les valeurs prédites, en tenant compte des vraisemblances et en incorporant les incertitudes. Cette distribution postérieure englobe une gamme de résultats plausibles, reflétant ainsi l'incertitude du modèle quant aux valeurs réelles des prédictions. En échantillonnant cette distribution, une multitude de scénarios potentiels sont générés, chacun tenant compte

de l'incertitude présente dans le modèle. Ces points de données échantillonnés, souvent désignés sous le terme d'échantillons prédictifs postérieurs, offrent une compréhension plus approfondie des résultats potentiels.

Dans la troisième phase de l'analyse de l'incertitude, les données échantillonnées (échantillons prédictifs postérieurs), avec leurs composantes d'espérance conditionnelle et d'intervalles de confiance, jouent un rôle central dans l'analyse de l'incertitude dans un cadre de modélisation bayésien.L'espérance conditionnelle sert de mesure de la tendance centrale, représentant la prédiction moyenne pour de nouveaux points de données. Toutefois, compte tenu des incertitudes inhérentes aux données du monde réel, les intervalles de confiance dérivés de la distribution prédictive postérieure deviennent inestimables. Ces intervalles fournissent une gamme quantifiable de valeurs plausibles pour les prédictions, exprimant efficacement l'incertitude associée au modèle. Ils englobent le spectre des résultats potentiels et offrent aux décideurs un aperçu de la variabilité inhérente aux prédictions.

La quatrième phase utilise un seuil critique de consommation maximale (limite de puissance). Ce seuil, influencé par la capacité du réseau, les modèles de demande des consommateurs et les facteurs environnementaux, est essentiel pour évaluer la flexibilité sur les marchés de l'énergie. En quantifiant la différence entre la consommation électrique prévue et ce seuil maximal, nous évaluons la flexibilité requise. Cette étude améliore le calcul de la charge de flexibilité, en prenant en compte de manière proactive les réductions potentielles de la consommation d'électricité. Ce calcul, qui s'appuie sur l'espérance conditionnelle et les intervalles de confiance, révèle la flexibilité nécessaire pour l'ensemble des maisons en tenant compte de l'incertitude inhérente. Ces informations guident les actions futures de l'agrégateur pour participer efficacement au marché de la

flexibilité, en veillant à ce que la demande d'électricité reste dans des limites gérables.

3.3.3 Résultats

Dans cette étude, une méthode d'évaluation de la flexibilité en considérant l'incertitude liée à la prévision des charges flexibles et non flexibles est développée. L'étude utilise deux approches distinctes: le AGP et le modèle *Prophet*, pour la prévision à court terme des charges agrégées. L'approche AGP, en particulier, introduit un modèle probabiliste pour prédire la distribution de l'incertitude, en incorporant des paramètres d'incertitude de puissance. Les résultats de cette analyse approfondie, portant sur les technologies de quantification des incertitudes dans les prévisions, offrent des perspectives importantes pour les gestionnaires de réseau. En outre, l'étude inclut un cas d'étude spécifique portant sur 14 ménages reliés au même transformateur de la même base des données. Les résultats soulignent la supériorité de la prévision basée sur l'AGP en termes de précision. L'évaluation comparative de l'incertitude montre que le modèle AGP surpasse le modèle *Prophet* en termes d'incertitude horaire et de calcul des besoins en flexibilité. Cette approche est particulièrement pertinente pour les gestionnaires du réseau, où une amélioration de la précision prédictive est fondamentale pour des décisions avisées. L'efficacité de cette méthodologie est prometteuse pour améliorer les pratiques de prévision de charge des opérateurs de réseau, en fournissant non seulement des prédictions plus précises, mais aussi une quantification des incertitudes associées. Cette avancée pourrait significativement optimiser la prise de décision en temps réel et la gestion globale du réseau et améliorer la flexibilité pour un agrégat de résidences.

Uncertainty Quantification in Load Forecasting for Smart Grids Using Non-parametric Statistics

Khansa Dab[1], Shaival Nagarsheth[1], Fatima Amara[3], Nilson Henao[1], Kodjo Agbossou[1], Yves Dubé[2], and Simon Sansregret[3]

Université du Québec à Trois-Rivières (UQTR), Trois-Rivières, Québec G9A5H7, Canada
[1]Laboratoire d'Innovation et de Recherche en Énergie Intelligente (LIREI), Département de génie électrique et génie informatique, UQTR
[2]Département de génie mécanique, UQTR

[3]Laboratoire des Technologies de l'Énergie, Institut de Recherche Hydro-Québec, Shawinigan, QC G9N 7N5, Canada

***Abstract*—Aggregators in flexibility markets act as intermediaries, pooling and selling consumer flexibility to grid operators or distribution system operators (DSOs). They are essential for grid management, offering load reductions based on power limits, and estimating expected consumer load in demand response scenarios. However, the inherent uncertainty in consumer behaviour poses a significant challenge, leading to deviations between projected and actual power consumption. In this context, this paper proposes a methodology for quantifying forecast uncertainties in power profiles at the aggregator level. The proposed methodology introduces a model-based approach to provide a more comprehensive representation of uncertainty and investigation of load variations. It provides load forecast values as comprehensive distributions, which are then sampled to generate newly sampled data from which the probability density function is extracted to quantify uncertainty, expressed by confidence intervals around the expected output. This process facilitates the identification of flexibility requirements regarding power consumption, which is essential for informed decision-making. The proposed strategy is effectuated on a synthetic dataset to evaluate its effectiveness in quantifying the uncertainties for probabilistic forecasts. Additionally, a potential case study with a neighbourhood of 14 houses connected to the same distribution transformer is presented to validate the proposed method. A comparative investigation of quantified uncertainties is presented by employing the Additive Gaussian Process (AGP) and the *Prophet* forecasting model, highlighting the usefulness of the proposed approach in flexibility markets. The results demonstrated the superiority of AGP-based load forecasts and flexibility needs with precise prediction accuracy.**

***Index Terms*—Additive Gaussian Process, Facebook *Prophet* model, Flexibility markets, Forecasting analysis, Uncertainty analysis.**

NOMENCLATURE

Abbreviations

AGP	Additive Gaussian Process
ANN	Artificial Neural Network
CDF	Cumulative Distribution Function
CI	Confidence Interval
DSM	Demand Side Management
EMS	Energy Management System
EQ	Exponential Quadratic kernel
KDE	Kernel Density Estimation
M	Matérn kernel
MAE	Mean Absolute Error
PDF	Probability Density Function
RMSE	Root Mean Squared Error
sMAPE	squared Mean Absolute Percentage Error
STLF	Short-Term Load Forecasting

Parameters and variables

α	Percentile
ℓ_{EQ},η_{EQ}	Hyperparameters of EQ
ℓ_M,η_M	Hyperparameters of M
$\mathcal{O}(N^3)$	Covariance matrix
x_c	Vector calendar variables
x_w	Vector weather variables
σ	Variance
D_x	Data domain of x
F^{-1}	Inverse Cumulative Distribution Function
$g(t)$	Trends of non-periodic changes
h	Horizon
N	Dimension of the covariance matrix
$s(t)$	Nonlinear function on a daily, weekly,or yearly
t	Discrete time
X	Covariates
x	Inputs
y	Aggregated power (kW)
y_{lim}	Power limit (kW)

I. INTRODUCTION

A. Background & Motivation

UNCERTAINTY in forecasts in recent years has been an important aspect of many fields, including statistics, economics, weather prediction, and machine learning. It reflects the inherent unpredictability or variability in future outcomes, and it's crucial to understand and quantify this uncertainty to make informed decisions [1]. However, uncertainty arises from various factors, including seasonal variations, weather conditions, economic fluctuations, customer behaviour, the model's parameters, and unforeseen events. Therefore, it is essential to develop forecasting models and methodologies that can handle these uncertainties and provide reliable forecasts [2]. On the other hand, electricity grids in cold regions face unique challenges compared to those in milder climates. Electric space heating significantly complicates the energy demand profile, especially in cold regions.

In these areas, the energy demand is further complicated by the heavy reliance on electric space heating and water heating. Occupant behaviours can also significantly impact electricity consumption patterns, and the unpredictability of these behaviours introduces uncertainty in load forecasting. For instance, daily routines throughout the day, changes in work schedules, sleep patterns, and other activities can vary, leading to fluctuations in electricity usage. Additionally, the increasing use of electric vehicles (EVs) adds to the fluctuating energy demands [3]. This creates a demanding landscape that experiences both seasonal and daily peaks. Cold winters contribute to significant seasonal spikes in power consumption, while daily patterns add further variability, putting a strain on local distribution networks [4], [5], [6]. In the context of local flexibility markets, it is crucial to tackle these challenges, particularly the uncertainties surrounding load [7]. These markets serve as vital channels for making real-time adjustments to electricity consumption, offering potential solutions to congestion management and peak shaving [8]. Making these real-time adjustments becomes imperative to maintain grid stability and ensure uninterrupted power supply during periods of high demand [9]. Within this ecosystem, the accuracy of the aggregated load estimation in a neighbourhood is paramount [10] for cold weather regions, as inaccuracies can lead to economic inefficiencies and potential grid challenges. If overestimated, it might fail to reduce consumption as promised, which can lead to penalties and grid instability. On the other hand, underestimating flexibility means missing out on market opportunities and potential revenue [11]. While traditional methods of load estimation have their merits, they increasingly fall short when confronted with the daily consumption intricacies in cold regions. The combined impact of electric space and water heating systems, the rising tide of EVs, and the nuanced changes in consumption caused by the vicissitudes of cold climates necessitate an approach to the prediction of aggregated load with uncertainties [12].

Short-Term Load Forecasting (STLF) serves as a pivotal tool in addressing load demand for optimal electricity market planning, as demonstrated by its extensive application in delivering usage plans [13]. Furthermore, its efficacy extends to the Energy Management System (EMS), playing a crucial role in real-time load consumption prediction for the implementation of effective Demand Side Management (DSM) programs aimed at enhancing energy efficiency [14]. The primary goal of such initiatives is the reduction of end-user electricity consumption by strategically modifying load patterns, particularly during peak times [15]. However, upon a critical examination of the current state of the art, certain gaps and inadequacies come to the forefront, particularly when viewed through the lens of modern smart grids [16]. The existing body of literature predominantly explores STLF using deterministic or probabilistic models [17]. Additionally, the surge in popularity of machine learning techniques, particularly neural networks and data-driven methods like Artificial Neural Networks (ANN) [18], Support Vector Machines (SVM) [19], and Gaussian Process (GP) [20], has been noteworthy in recent years for forecasting aggregated load. These techniques promise the ability to capture non-linear patterns but reveal a common weakness in uncertainty quantification when subjected to a critical evaluation.

Despite their proficiency in forecasting, these machine learning models often struggle to provide reliable uncertainty estimates. Their inherent "black-box" nature, combined with the risk of overfitting, introduces unpredictability into forecasts, particularly in the face of anomalous events or rapid grid changes. Furthermore, the reliance of supervised learning algorithms on training datasets with precise forecasts introduces uncertainties that limit their practical application. This leads us to a fundamental question: How can uncertainties be effectively incorporated into forecast models to enhance their reliability and applicability in dynamic energy environments?

B. Related Work

Point or deterministic forecast methods have been widely used historically because of their simplicity and understandable employment [21]. However, these deterministic methods are gradually replaced by probabilistic methods that respond to the stochastic factors corresponding to the system's flexibility [22]. The methods proposed by those works suffer from two issues: the first is the accumulation of errors due to the stochastic behaviour of end-users, and the second is the insufficiency of the model to provide reliable forecasts of users with different power patterns for an ensemble of houses since uncertainties can significantly impact the actual demand [23]. Various advanced probabilistic load forecasting methods have prominently emerged in recent years. While prior research has not explicitly addressed uncertainty propagation from systems, notable progress has been made in this forecasting domain. For instance, [17] investigates the propagation of input uncertainty, recognizing the challenges involved in predicting outputs. In this context, [16] examines the outputs of machine learning algorithms to quantify uncertainty in determining future power demand changes.

On the other hand, the application of Gaussian processes network-based models, as highlighted in [24], stands out for its ability to generate empirical distributions by sampling multiple predictions. This method is effective, and analytical distributions prove valuable for gradient-based design, by minimizing the need for extensive predictions. Focusing on a forecast horizon of 24 hours, this approach estimates load confidence intervals based on quantiles derived from past forecast errors. This method's adaptability extends to security analyses of power systems, demonstrating its capacity to generate demand scenarios at specified risk levels. The primary objective of this analysis is to understand system reactions to electricity use ramps and periods of low load [25], [26]. In practice, three strategies are often used to communicate uncertainty in load forecasts that allow a more comprehensive exploration of uncertainties in load predictions: scenario forecasting [27], interval forecasting [28], and quantile forecasting [2], [29].

Table I provides a comprehensive overview of how each reference navigates the complexities of uncertainty within the context of load forecasting. By examining the entries in the table, one can discern the diverse methodologies and approaches employed by different authors to address uncertainties. [32]

Table I
THE EFFECTIVE ELEMENTS OF UNCERTAINTY IN FORECASTING PROCEDURES ACCORDING TO THE RELEVANT LITERATURE

References	Modeling approach	Non parametric models	Uncertainty analysis	Temporal Dynamics	Exogenous factors	Application
[7]	Local flexibility market mechanism	✔	✔	✘	✔	DSOs flexibility services and quantify the financial benefits
[17]	Trajectory forecasting	✔	✔	✔	✘	Improving mean prediction accuracy
[16]	ML and ensemble learning	✔	✔	✔	✘	Future power demand changes
[21]	Deterministic	✘	✔	✘	✘	Manage user expectations
[22]	Statistical models	✔	✔	✔	✘	Wind power DM
[24]	Gaussian Process Regression	✔	✘	✘	✔	Reduce the peak energy demands and energy supply risks
[25]	Markov-chain mixture distribution (MCM)	✔	✔	✘	✘	react to ramps of electricity use
[26]	NN- based CIs	✔	✔	✘	✘	DM and risk management in energy systems
[28]	Neural Network	✔	✔	✔	✘	Required power reserve (PV)
[29]	Quantile regression neural network	✔	✔	✘	✔	✘
[30]	Five algorithms of ensemble learning	✔	✔	✔	✔	Produced good estimates of the confidence in a forecast
[31]	Bootstrap aggregating	✔	✘	✔	✘	Improving forecasting load
[32]	Combined unsupervised ensemble learning	✔	✔	✔	✔	Detect trend shift better and handle the noise in data more precisely
Our work	Additive Gaussian Process	✔	✔	✔	✔	Calculating flexibility needs

DM: Decision Making, ML: Machine Learning, CIs: confidence intervals,PV: Photovoltaic Power

incorporates neural network models and applies confidence interval-based uncertainty quantification for electricity price forecasting. The authors combine different analyses for time series analyses (statistical) by applying uncertainty to clustered data for power to better detect trend shift (concept drift) and handle the noise in data more precisely. Another study [29] builds upon methodologies from competition winners, integrating quantile regression and neural networks for load and price forecasting. Authors in [7] emphasize the robustness of a model in handling missing data and outliers and adapting to trend changes. The authors thoroughly examine the model's mechanisms for estimating uncertainty, providing confidence intervals, and evaluating reliability in scenarios where uncertainty plays a pivotal role [33].

While quantifying forecast uncertainty may support better decision-making in the energy industry, there have been few journal articles published on quantifying forecast uncertainty [34]. Uncertainty in load stems from various exogenous factors such as temperature, humidity, and solar radiation. It is also attributed to the temporal dynamics, encompassing seasonality, trends, and cyclic patterns [17]. Several works have been carried out [30], [32], [16] encompassing temporal dynamics or exogenous factors; however, a notable distinction is made regarding the application of uncertainty quantification that did not explicitly consider the practical implementation or utilization of their uncertainty quantification methods. The interplay of these factors introduces variations in the data quality and quantity incorporated into the forecasting models. As emphasized in the literature, the impact of exogenous factors on uncertainty analysis is significant, and understanding this relationship becomes paramount in refining forecasting methodologies [35]. Striking a balance between the richness of data and the potential influence of exogenous variables is essential, given that an increase in observed data may mitigate model noise, yet the inherent process noise remains linked to the underlying data-generating process, maintaining its level of uncertainty, especially when data points are scarce [36].

C. Contributions & Organization

The main objective of our study is to predict load patterns encompassing the representation of the uncertainties inherent in the power consumption forecasting process to quantify the uncertainties and provide flexibility requirement calculations for the aggregators in the energy markets. Accordingly, the contribution of this work is twofold: The first centers on proposing a methodology for load forecasting with induced uncertainties estimation. That involves generating forecasts and offering measures of uncertainty associated with each confidence interval. Secondly, the load forecasts and the associated uncertainties are analyzed for each forecast interval of 15 minutes for a 24-hour forecast window. Utilizing the probabilistic models generating posterior predictive sampled data, a more precise understanding is achieved of the reliability of the load forecasts, supporting informed decision-making processes. This transparency in reporting conditional expectations and confidence intervals is crucial, especially in effectively conveying flexibility requirements for managing loads and reducing the DSO's network operational costs. The proposed methodology utilizes the AGP for performing load forecasting [37],[38], and quantifying uncertainties as well. To evaluate its efficacy, a comparative study is presented with a modular regression model, also known as the $Prophet$ model [39]. The modeling accuracy is evaluated through several metrics for scoring the forecasting methods. Subsequently, the uncertainty quantification calculating the flexibility need is

carried out utilizing the confidence interval width and inverse CDF. Comparative analysis is effectuated on two case studies: (i) on the synthetic dataset of 1000 houses located in Quebec, and (ii) on a low voltage network consisting of 14 houses fed by the same transformer.

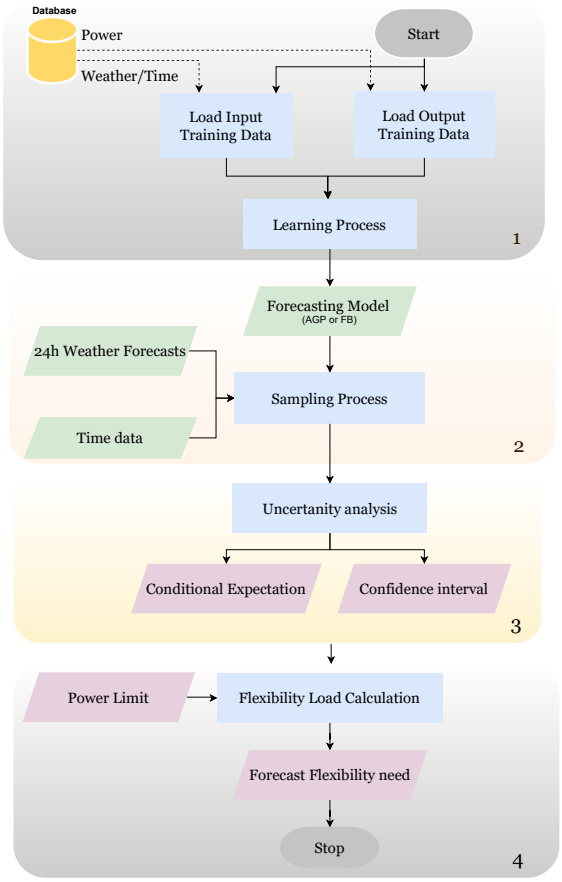

Figure 1. Flowchart of the proposed methodology

The rest of the paper is organized as follows: Section II presents the proposed methodology in detail. Section III formulates the forecasting models utilized in the methodology. Section IV presents results and discussions of the two case studies. An investigation of uncertainty quantification has been presented for two case studies, followed by the conclusion in Section V.

II. Methodology

Fig. 1 shows the methodology to tackle the task of forecasting household power consumption, focusing on quantifying uncertainties and calculating the flexibility needs. It is divided into four distinct phases. The first phase starts by gathering historical power consumption data and capturing diverse load patterns and trends during winter. The consumption data includes both flexible and non-flexible loads as well. A flexible load refers to the electricity consumption that can be adjusted or shifted in time without significant inconvenience or cost. This contributes mainly to the flexibility process and can be exploited by the aggregator to balance the grid, especially during peak demand periods. Examples of flexible loads include space and water heating, washing machines, and dryers. Non-flexible loads, on the other hand, are those that cannot be easily adjusted or shifted without causing significant disruption, for instance, essential lighting systems. In addition to consumption data, this process considers external factors such as weather conditions and calendar time. Initial data analysis is facilitated through non-parametric statistical techniques based on a forecasting model to uncover underlying patterns and variances in the dataset. Here, the learning process involves training the model using historical data, where hyperparameters are estimated to optimize forecasting accuracy. The yield of the first phase is a tuned forecast model. In this work, two models are considered for the comparison; namely, AGP [37], [38] and *Prophet* [39] for load forecasting with uncertainties.

The second phase consists of utilizing the tuned forecasting model predictions in the sampling process. As AGP is a probabilistic model, this phase is crucial in understanding the mechanics of the forecasting model that captures the inherent uncertainty in the function estimation by including probabilistic components. Leveraging the power of Bayesian statistics, AGP provides a flexible framework for capturing uncertainties and relationships in the data. When utilized for prediction, AGP generates point estimates and also provides a complete posterior distribution for the predicted values from the priors and the likelihoods by incorporating the uncertainties. The posterior distribution encapsulates a range of plausible outcomes, reflecting the model's uncertainty about the true values of the predictions. By sampling from this distribution, we generate a multitude of potential scenarios, each respecting the uncertainty present in the model. These sampled data points, often referred to as posterior predictive samples, enable a more comprehensive understanding of the potential outcomes and aid in making informed decisions. This Bayesian approach not only provides a point forecast but also equips us with the tools to assess the range of possibilities and make robust decisions based on the inherent uncertainties in the data. Note that for the *Prophet* there are three sources of uncertainty in the forecast: uncertainty in the trend, uncertainty in the seasonality estimates, and additional observation noise. The uncertainty in the seasonality estimates can be extracted by Bayesian sampling to get the posterior predictive sampled data.

In the third phase of analyzing the uncertainty, the sampled data (posterior predictive samples), with its components of conditional expectation and confidence intervals, plays a pivotal role in conducting uncertainty analysis within a Bayesian modeling framework. The conditional expectation serves as a central tendency measure, representing the average prediction for new, unseen data points. However, acknowledging the inherent uncertainties in real-world data, confidence intervals derived from the posterior predictive distribution become invaluable. These intervals provide a quantifiable range of plausible values for predictions, effectively expressing the uncertainty associated with the model. They encapsulate the

spectrum of potential outcomes and offer decision-makers insights into the variability inherent in the predictions. Consequently, in this work, this comprehensive uncertainty analysis, incorporating both conditional expectation and confidence intervals, empowers users to make informed decisions on the possible power profiles, considering the full spectrum of possibilities and acknowledging the uncertainties inherent in the underlying data and modeling assumptions.

The fourth phase utilizes a critical maximum consumption threshold (power limit). This threshold, influenced by grid capacity, consumer demand patterns, and environmental factors, is essential for evaluating flexibility in energy markets. By quantifying the difference between forecasted power consumption and this maximum threshold, we assess the required flexibility. This investigation enhances flexibility load calculation, addressing potential reductions in power consumption proactively. This calculation, relying on conditional expectation and confidence intervals, reveals the flexibility needed for the ensemble of houses by taking into account the inherent uncertainty. These insights guide further actions by the aggregator in participating effectively in the flexibility market, ensuring power demand stays within manageable bounds.

III. Forecasting Models

A. Additive Gaussian Process Forecasting Model

AGPs are a class of models that have gained popularity in machine learning and statistics. Realizations from an AGP correspond to random functions, and consequently, AGPs naturally provide a prior for an unknown regression function that is to be estimated from data. By definition, the prior probability density of AGP function values $f(X) = (f(x_1), f(x_2), \ldots, f(x_N))^T$ for any finite number of fixed input covariates $X = (x_1, x_2, \ldots, x_N)$ where $x_i \in \mathcal{X}$ is defined to have a joint multivariate Gaussian distribution [40]:

$$f(x) \sim \mathcal{N}\left(0,\, K_{X,X}(\theta)\right) \tag{1}$$

The elements of the N-by-N covariance matrix are determined by the AGP kernel function, denoted as $[K_{X,X}(\theta)]_{i,j} = k(x_i, x_j|\theta)$, where θ represents the parameters. In general, the mean in eq. 1 can depend on X, but in practice, a zero mean is often assumed. The covariance, also known as the kernel function, of the normal distribution governs the smoothness of the function f, indicating how rapidly the regression function can change. While AGP is formulated such that any finite-dimensional marginal follows a Gaussian distribution, AGP regression is considered a non-parametric method since the regression function f lacks an explicit parametric form [41]. More precisely, AGP encompasses a countably infinite number of parameters that define the regression function, corresponding to the function values f at all possible inputs.

In this case study, the forecast is based on the AGP approach, where the kernel (covariance) is expressed as a sum of kernels. In this additive structure, each kernel models the effect of individual covariates or their interactions. Intuitively, each AGP component f now represents a nonlinear function that characterizes the corresponding effect, and the cumulative impact of multiple covariates is the sum of these nonlinear functions. This is achieved by employing specific kernels tailored to different types of covariates. Subsequently, the AGP model, resulting from this configuration, can be explained by,

$$k(\mathrm{x}) = k_w(\mathrm{x}_w) + k_c(\mathrm{x}_c) + k_{\mathrm{so}}(\mathrm{x}) \tag{2}$$

Where the vectors x_w and x_c contain weather and calendar variables, respectively. The hyperparameters of the Exponential Quadratic kernel (EQ) continuous covariates are dedicated to the weather-related component. The weather kernel is depicted in (3),

$$k_{EQ}(\mathrm{x}, \mathrm{x}') = \eta_{EQ}^2 \exp\left(-\frac{(||\mathrm{x} - \mathrm{x}'||)^2}{\ell_{EQ}^2}\right). \tag{3}$$

However, Matérn 5/2 (M) functions would be the kernel for the calendar component as shown in (4),

$$k_M(\mathrm{x} - \mathrm{x}') = \sigma^2 \frac{2^{1-\nu}}{\tau(\nu)} \left(\sqrt{2\nu}\frac{\mathrm{x} - \mathrm{x}'}{\ell_M}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{\mathrm{x} - \mathrm{x}'}{\ell_M}\right), \tag{4}$$

In this kernel, the Half Cauchy is utilized as the prior over the variance σ. Gamma ℓ_{EQ} and ChiSquared ℓ_M distributions are used as the prior over the hyper-parameters of the EQ and Matérn functions, respectively. Particularly, they are used to explain the lengthscales, η_{EQ} and η_M, of these covariance bases. Additionally, The third symbolizes a compositional first-order kernel that is intended for searching interactions between multi-dimensional variables [37].

Additionally, the goal of Bayesian inference is to compute the posterior distribution over the function $f(x)$ evaluated at arbitrary test inputs x. For Gaussian likelihoods, the posterior distribution takes a convenient closed-form solution, thus the predictive distribution at a test variable. However, it is difficult to compute in practice when N is large. The computational cost of matrix inversion is in $\mathcal{O}(N^3)$. Naively, these operations each incur $\mathcal{O}(N^3)$ computations, as well as $\mathcal{O}(N^2)$ storage for each entry of the kernel (covariance) matrix, often starting with a Cholesky decomposition. To resolve those issues in the configuration-based AGP a scalability analysis has been applied. It should be noted that the choice of the values of hyperparameters is the same as in our previous work [37]. While in straightforward conditions, the Bayesian approach might be unproductive, most applications of AGPs rely on engineering sophisticated hand-crafted kernels involving many hyperparameters where the risk of overfitting is pronounced. A more robust solution is to incorporate confidence intervals that reflect these uncertainties in the model choice. Initially, we have a prior distribution that predicts the aggregated power. As the data is gathered, we refine this to include only functions that align with the observations, creating a posterior distribution. This posterior is essentially an updated prior, incorporating new data. Each new piece of data further improves this process. The AGP, in this context, describes a probability distribution across a range of potential functions that match a given set of points. This model allows us to determine mean values for these functions and assess the confidence of these predictions through variance. The function (posterior) is continuously updated with new data. The AGP represents then

a probability distribution across possible functions, where any subset of these functions follows a joint Gaussian distribution. Meanwhile, for the regression predictions, the mean function derived from the posterior distribution is used. More in-depth details of AGPs are available in [37], [42].

B. Prophet Forecasting Model

A modular regression model popularly known as the *Prophet* model was developed by Facebook [39]. It is built to handle time-series data with varied seasons [43], [44], and offers a versatile framework for deriving confidence intervals to determine the uncertainty inherent in the prediction system. Specifically, it is based on an additive model composed of three components: the trends $g(t)$ simulating non-periodic changes in the data, the seasonality $s(t)$ describing nonlinear behaviour on a daily, weekly, or yearly basis, and the third is the error term ε_t representing the distinctive features of the data improving the accuracy. Mathematically, in this study, the decomposed time-series model comprising two fundamental components is utilized to scrutinize power consumption patterns across a group of households [45]:

$$y(t) = g(t) + s(t) + \varepsilon_t, \tag{5}$$

Eq. (5) doesn't use traditional logistic regression for its growth modeling, but it employs an adaptive approach to effectively capture the growth patterns in the data. The trend function $g(t)$ is a nonlinear saturating function modeled using the logistic growth function, given by:

$$g(t) = \frac{c(t)}{1 + e^{-k(t-m)}} \tag{6}$$

where $c(t)$ is a time-varying consumption per day, k denotes a varying growth rate and m is the offset parameter. The periodic effect of yearly seasonal variations is modeled using the Fourier series; hence, an approximate smooth seasonal effect is tied with a standard Fourier series represented as:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos \frac{2\pi nt}{p} + b_n \sin \frac{2\pi nt}{p} \right), \tag{7}$$

where p is the period of the seasonality, it can be 365.25 or 7 for yearly and weekly seasonality, respectively. The modeling flow of the Prophet model is shown in Fig. 2. It is designed to auto-tune the hyperparameters, and the training splits the data in two: (i) timestamps "ds" containing the time and date details, (ii) the logged values "y".

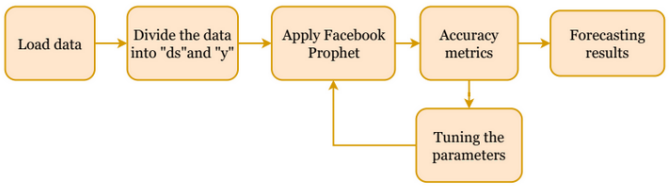

Figure 2. Prophet's modeling flow

IV. Results and Discussion

A. Case study - 1000 houses

1) Data and Analysis Setup: In this work, simulations are conducted using load data sourced from aggregate simulated end-user profiles of 1000 residential houses. The database with a sampling interval of 15 minutes is administered by Hydro-Québec, a research institution situated in Québec. The specified time covers the period from December 1, 2018, to April 31, 2019. Additionally, this study incorporates temperature, humidity, and solar radiation data from the same geographic location within demand areas. As illustrated in Figure 3, it is

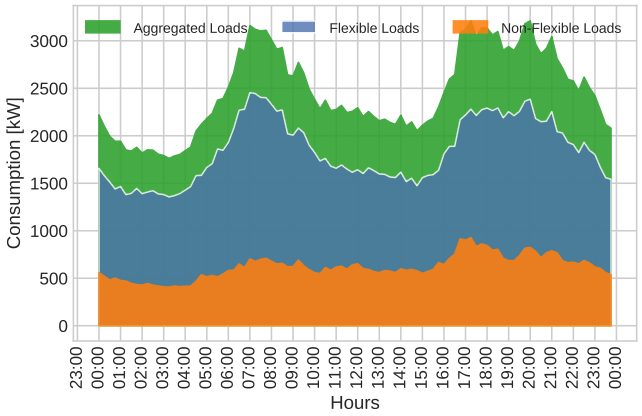

Figure 3. Power consumption by flexible and non-flexible loads within the total load of the 1000 houses for a specific day

evident that flexible loads, namely air conditioners and heating systems, constitute the major part of the rated building load. The remaining loads are assumed to be non-flexible to facilitate Demand Response (DR). Figure 3 presents the cumulative stacked graph with a peak load of 3000 kW observed at 5 AM, which encompasses the total electrical demand, including both flexible and non-flexible loads. To analyze variations in these load profiles, statistical methods were applied to the dataset. Initially, the focus of the forecasting and uncertainty estimation was on non-flexible loads, a process that introduced a certain level of additional uncertainty into the results. Later, the analysis was expanded to include the entire aggregated load, thereby covering both flexible and non-flexible load types.

2) Forecasting performance: Forecasting load demand depends on factors like the number of households and various infrastructure components. However, consumption related to non-flexible loads namely lighting, major household appliances, and electronics for and horizon of 24 hours follows a highly stochastic pattern. The overall demand in a neighbourhood can be quite uncertain, primarily due to the presence of significant non-flexible loads, posing significant challenges for grid management and the behaviors of occupants that change depending on the calendar variables. Consequently, error and uncertainty are interconnected yet separate facets in measurement characterization. An error signifies the variance between a measurement outcome and the actual value of

the power. In contrast, uncertainty gauges the confidence in the assertion that the power forecasting result accurately reflects the power value, encompassing various factors influencing reliability. These terms jointly define the precision of measurements. Hence, this work performs load forecasting with uncertainty through AGP and the Prophet model for aggregated power consumption profiles encompassing flexible and non-flexible loads and aggregated non-flexible loads for an ensemble of 1000 houses. *Accuracy's metrics:* The efficiency of the two employed models is evaluated using a variety of statistical parameters. The table presents the mean of the metrics of all the generated predicted profiles, including mean absolute error (MAE) (8), root mean square error (RMSE) (9), coefficient of determination (R^2) (10), and squared mean absolute percentage error (sMAPE) (11).

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|y_t - \widehat{y}_t|\,, \tag{8}$$

$$\mathrm{RMSE} = \left\{\frac{1}{N}\sum_{t=1}^{N}(\widehat{y}_t - y_t)^2\right\}^{1/2}, \tag{9}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{t=1}^{N}(y_t - \widehat{y}_t)}{\sum_{t=1}^{N}(y_t - \widehat{y}_t)^2}, \tag{10}$$

$$\mathrm{sMAPE} = \frac{100\%}{N}\sum_{t=1}^{N}\frac{|\widehat{y}_t - y_t|}{(|\widehat{y}_t| + |y_t|)/2}, \tag{11}$$

where $\widehat{y}$ and y present predicted and actual power usages for N discrete-time samples.

AGP model forecast: Figure 4 shows the hourly forecasting result for the aggregated load of 1000 houses in a 24-hour day-ahead scenario. It elaborates the posterior analysis through AGP to display the degree to which data generated from the model could deviate from data generated from the true distribution. Multiple sample trajectories in Figure 4 visually portray the potential range of load scenarios. The different confidence intervals ranged between 55%, 65%, 75%, 85%, and 95%, represented by the shaded area around the forecast curve, emphasizing the variability and potential outcomes. This visualization illuminates the inherent uncertainty in the forecast, offering a detailed perspective on potential load fluctuations within the specified confidence bounds, helping to achieve accurate demand response in the face of uncertainties and variations between predicted and actual electricity consumption.

Since our predictive distribution is Gaussian, this quantity enables us to form, for example, a 95% credible set representing the beliefs about the interval, which is 95% likely to contain the truth function compared to the other confidence intervals. As shown in Figure 4, the power uncertainty is higher during the midday period. Conversely, the uncertainty is less during the morning (from 6:00 to 10:00) and afternoon (from 17:00 to 21:00). The depicted Figure 5 represents a day-long hourly probabilistic forecast, focusing specifically on the Aggregated non-flexible loads. The forecasted values are presented as probabilistic distributions, with distribution shapes indicating the forecasted range of heating load values

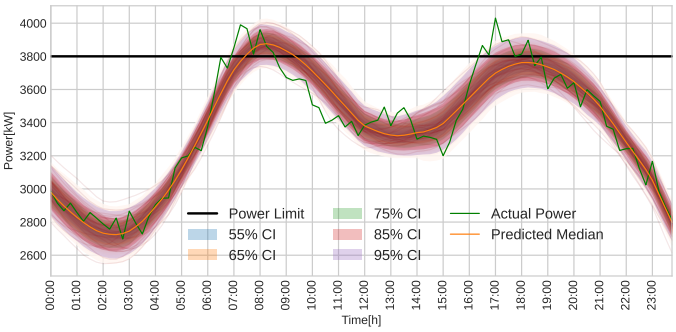

Figure 4. A sample day-long hourly AGP-based probabilistic load forecast, indicating uncertainty with different % of confidence interval (Date: DEC 1, 2019) case **aggregated loads**

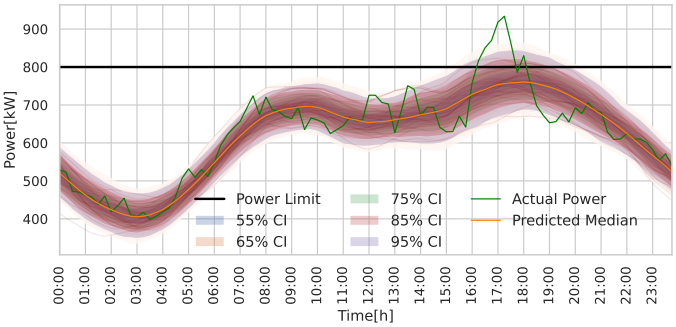

Figure 5. A sample day-long hourly AGP-based probabilistic load forecast, indicating uncertainty with different % of confidence interval (Date: DEC 1, 2019) case of **aggregated non-flexible loads**

for each hour of the day. Variations in distribution shapes reflect the level of uncertainty in the forecasts, with less sharp distributions indicating higher uncertainty. Various percentiles of the confidence interval represent distinct levels of uncertainty.

The power limit is indicated by the black line depending on the requirement for flexibility it is fixed for the aggregated and the Aggregated non-flexible loads to 3800kW and 800kW, respectively. Additionally, it is determined using granular time interval data to mimic the dynamic shape of a customer's demand. It closely follows the actual demand (black line) leading up to and following the event.

Prophet model forecast: By applying the *Prophet* model with the proposed method, the uncertainties of power forecasting, and load forecasting with various probability indices (from 55% to 95%) for a day-ahead forecast are represented in Figures 6 and 7 for total aggregated load and aggregated non-flexible load, respectively.

As shown in Figure 6, the uncertainty of power forecasting is higher in the middle of the day, when the occupants are at home utilizing more power. While in the morning and afternoon, the uncertainty is less. Power forecasting uncertainty increases and decreases with power increase and decrease, respectively. Also, the uncertainty is increased when the time horizon is larger. For example, at 10:00 and 17:00, power outputs are almost at the same level (about 5000 kW) with

Table II
ACCURACY METRICS FOR A DAY AHEAD FORECAST WITH CONFIDENCE INTERVAL 95%

Strategy	**Type Load**	MAE[kW]	RMSE[kW]	R^2[%]	sMAPE[%]
Additive Gaussian Process	Total Load	203.00	240.04	0.71	8.27
	Heating Load	141.03	161.05	0.73	15.36
	Other Load	32.61	43.89	0.83	5.19
Prophet model	Total Load	266.47	355.78	0.69	9.59
	Heating Load	230.65	238.19	0.60	24.64
	Other Load	49.81	68.50	0.72	7.69

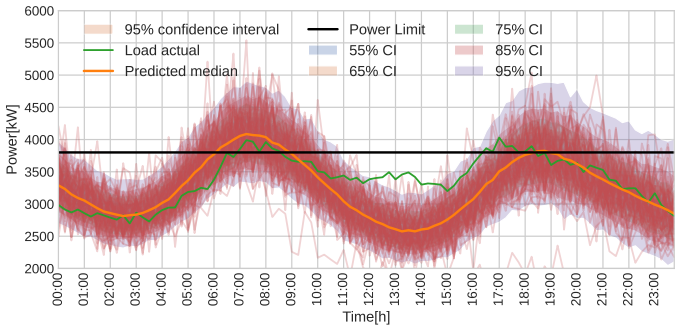

Figure 6. A sample day-long hourly Prophet-based probabilistic load forecast, indicating uncertainty with different % of confidence interval (Date: DEC 1, 2019) case of **aggregated loads**

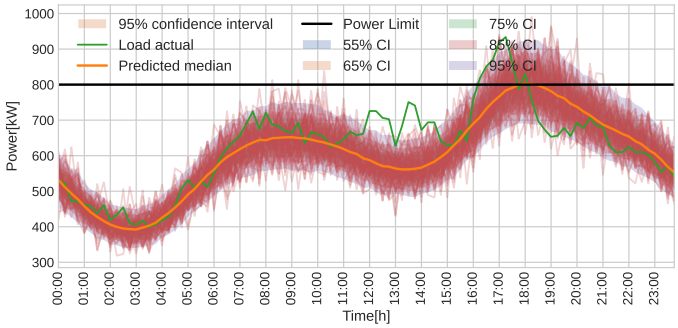

Figure 7. A sample day-long hourly Prophet-based probabilistic load forecast, indicating uncertainty with different % of confidence interval (Date: DEC 1, 2019) case of **aggregated non-flexible loads**

higher uncertainty. Figures 6 and 7 illustrate a posterior distribution over a power load variable, which represents the power consumption over time. A solid black line indicates a power limit as a reference, suggesting a constant power level of 2500 kW throughout the depicted period. The shaded regions, delineated by varying levels of transparency, represent different confidence intervals (CIs), such as 95%, 85%, 75%, 65%, and 55%, showing the uncertainty in the power load predictions. A solid black line indicates a power limit as a reference, suggesting a constant power level of 3800kW and 800kW for total aggregated and aggregate non-flexible load, respectively. Additionally, Figures 6 and 7 include two important lines: one in red representing the actual power load observed over time, and another in orange representing the predicted median load. The intersection of these lines with the shaded regions provides insight into the model's accuracy in estimating power consumption at different levels of confidence. These figures provide a comprehensive visual representation of the uncertainty associated with power load predictions and a comparison to the actual observed load.

The accuracy of these models is assessed for various load types, with the results summarized in Table II. The average sMAPE for 1-day forecasting horizons for different load types is found to be 8.27%, 15.36%, and 5.19%, respectively. Importantly, the AGP model outperforms the *Prophet* model when applied to the aggregated total load and the aggregated non-flexible loads, respectively. This work underscores the variation in electricity consumption forecast based on different load types and the importance of considering weather and calendar variables in peak load demand forecasting. The AGP model demonstrates superior performance in short-term load forecasting with uncertainties, offering valuable insights for grid management and flexibility analysis as compared to the *Prophet* model.

3) uncertainty quantification and flexibility demand calculations: The statistical analysis to assess the forecasted uncertainty and flexibility calculations can be leveraged by DSOs to evaluate the system security level or assess the demand flexibility for the considered day. Specifically, in the flexibility markets, it can often be used to manage consumption by setting capacity limit thresholds or limiting power consumption [46]. Hence, extracting uncertainty distributions for specific times, guided by selected confidence intervals by establishing a power limit, such as 3800 kW for 1000 households, becomes pivotal in making decisions to adjust consumption. The hourly probability distribution obtained from the forecast errors represents the likelihood of power consumption being less than or equal to a specific value in kilowatts (kW). Statistically, the Probability Density Function (PDF) can help determine the appropriate threshold based on the desired level of risk. The PDFs provide a visual tool for assessing the likelihood of extreme events to perform risk assessment [28]. This section provides the discussions related to the uncertainty quantification and flexibility calculations by effectuating the proposed methodology with two employed models, namely AGP and *Prophet* forecasts. Additionally, a comparative analysis is achieved by plotting the hourly distribution of the load forecasts resulting from two models

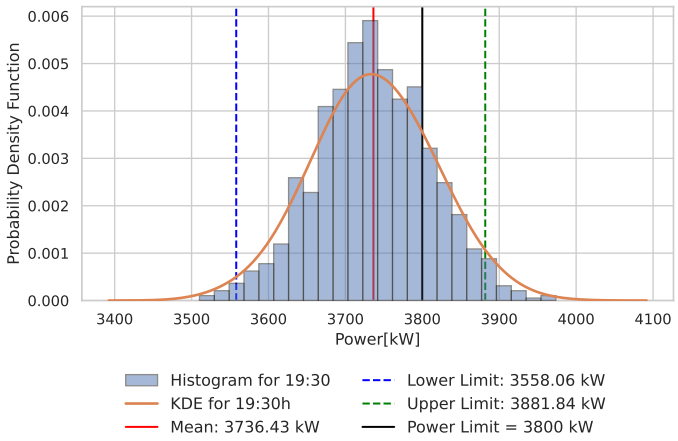

Figure 8. Probability density function of the **selected hour(19:30)** using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines with Gaussian Process model case of **aggregated loads**

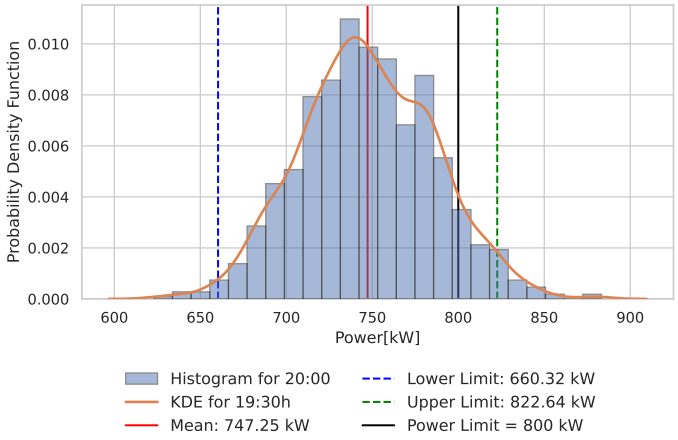

Figure 9. Probability density function of the **selected hour(20:00)** using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines with Gaussian Process model case of **aggregated non-flexible loads**

for identifying trends and anticipating the flexibility needs. Since the samples are taken from the posterior distribution, their probability density also needs estimation. Hence, we use Kernel Density Estimation (KDE) to achieve the non-parametric probability density, where we center a smooth scaled kernel function at each datapoint and then take their average [47]. Note that it is an empirical distribution that cannot be expressed analytically. The forecasting uncertainty can be represented as upper and lower-bound margins around the power forecast. The probability density can be drawn according to the selected samples from the predictive analysis since the forecast errors can be expressed as a percentage of the rated power. The bound margins are extracted from an inverse cumulative distribution function [28]. By CDF, we denote the function that returns probabilities of aggregated power y bounded lower to a value y_α, i.e.,

$$prob(y \leq y_\alpha) = F(y_\alpha), \tag{12}$$

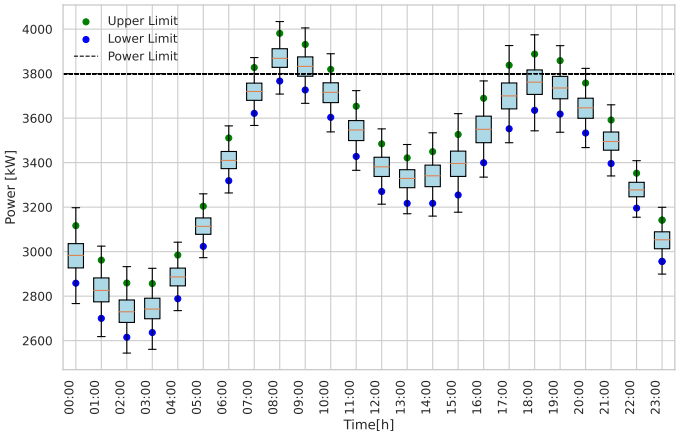

Figure 10. Hourly uncertainty based on a day-ahead forecast of **aggregated loads** from AGP-based method.

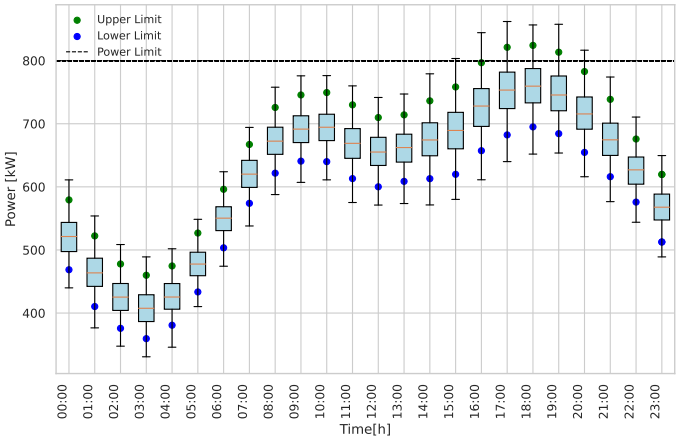

Figure 11. Hourly uncertainty based on a day-ahead forecast of **aggregated non-flexible loads** from AGP-based method.

where α is the desired percentile. Now, the inverse of the CDF gives a value y_α for which the $F(y_\alpha)$ will return α, i.e.,

$$F^{-1}(\alpha) = y_\alpha. \tag{13}$$

From (12) and (13), we can get surpassed aggregated power (flexibility requirement) by which it exceeds the power limit (y_{lim}), i.e. $\Delta y = y_\alpha - y_{lim}$.

AGP Forecasting: Figures 8 and 9 depict the probability distribution at a particular hour for aggregate total load and aggregated non-flexible load for the forecast resulting from the AGP forecasting model. The kernel density estimated value is represented by the orange line, reflecting a specific power at that point and representing how the PDF values change across the band. It comprehends data distribution, sets thresholds, evaluates risks, and calculates how much the aggregated power consumption can cross the upper capacity threshold. The upper and lower bounds are defined at 95% confidence levels of the aggregated power consumption. This confidence level quantifies the associated uncertainty of power values. This visualization aids in understanding the probability distribution of power values and the influence of confidence

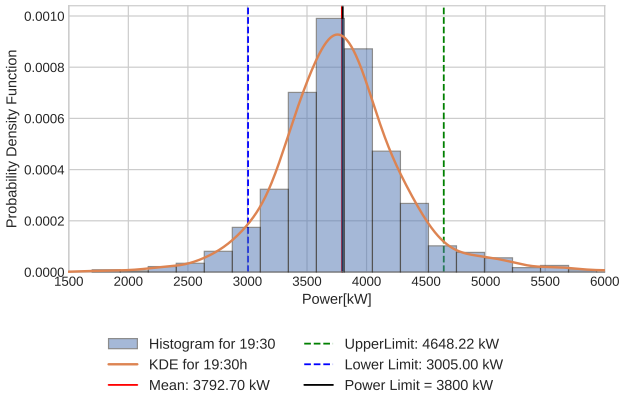

Figure 12. Probability density function of the **selected hour(19:30)** using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines case of **aggregated loads**

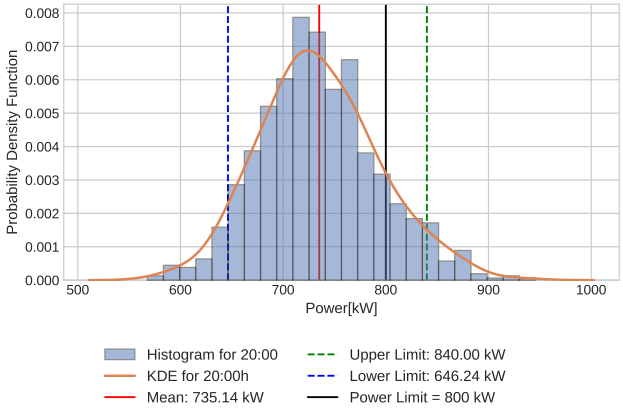

Figure 13. Probability density function of the **selected hour(20:00)** using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines in case of **aggregated non-flexible loads**

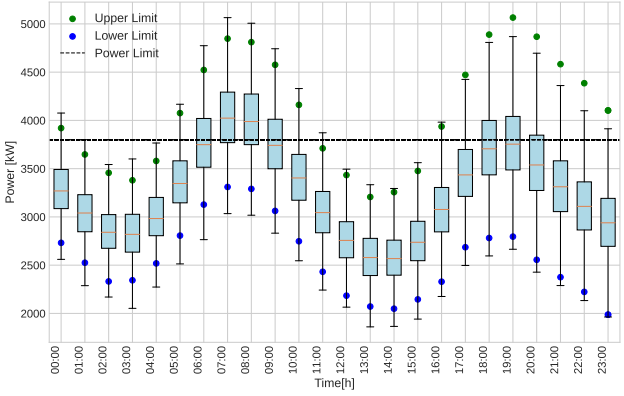

Figure 14. Hourly uncertainty based on a day-ahead forecast of **aggregated loads** from the *Prophet*-based method.

levels In Figure 8, the visualization for a one-day ahead at 19:30h with the upper bound of power 3881.55kW and the lower bound of power 3558.06kW is shown to exceed the power limit (y_{lim}) fixed to 3800kW by 81.55kW. Note that in this case study, the power limit is established at specific values: 3800 for the scenario involving an aggregated load of

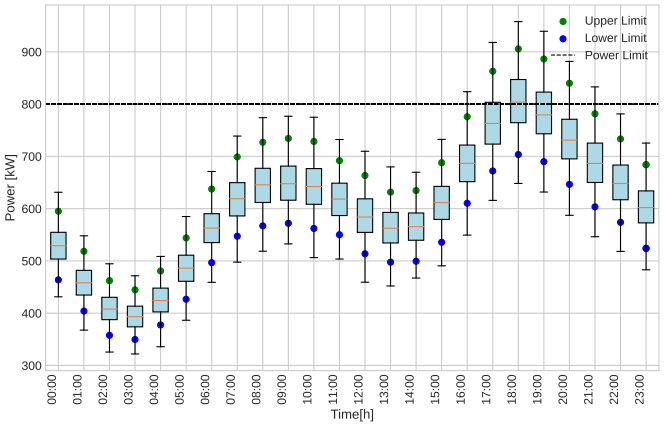

Figure 15. Hourly uncertainty based on a day-ahead forecast of **aggregated non-flexible loads** from the *Prophet*-based method.

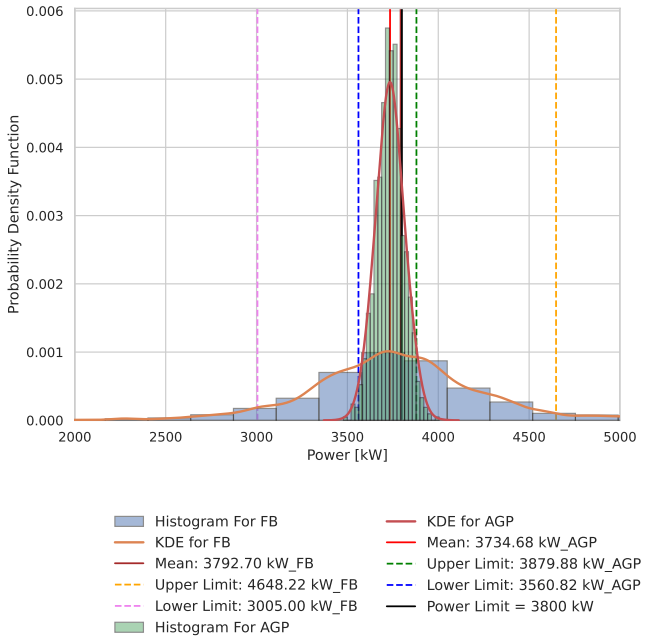

Figure 16. Probability density function of the **selected hour(20:00)** using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines for **aggregated loads**

1000 houses and 800 for the scenario concerning non-flexible load. A similar analysis is carried out in Figure 9 for the non-flexible loads.

Figures 10 and 11 illustrate uncertainty quantifications resulting from the AGP forecasted results for every hour in a 24-hour day-ahead scenario. It can be observed from Figure 10 that the capacity threshold is crossed during the peak hours of a typical day, i.e., from morning 7:00 to 10:00 a.m. and from 5:00 to 8:00 p.m. For the aggregated non-flexible load (Figure 11) the duration of crossing the capacity threshold is limited to the evening hours, indicating the heavy use of non-flexible loads during this period.

Prophet Forecasting: Similar to the uncertainty analysis

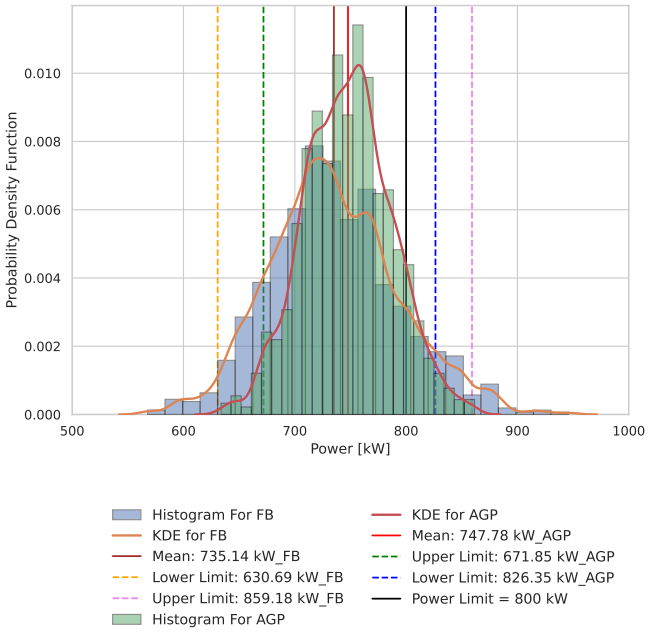

Figure 17. Probability density function of the **selected hour(20:00)** using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines for **aggregated non-flexible loads**

performed for the results for the AGP model forecast, Figures 12 to 15 correspond to the forecasting results from *Prophet* model. From Figure 12 it can be observed that the upper bound (4648.22 kW) of the aggregated total load decided by the 95 % confidence interval exceeds the capacity threshold of 800 kW by a huge margin of 848.22 kW. On the contrary, the aggregated non-flexible load upper bound (840 kW) exceeds 40 kW from its capacity threshold (power limit y_{lim}) of 800 kW. Figures 14 and 15 reveal the same results of capacity limit threshold crossing in the morning and the evening hours; however, it is visible that the *Prophet* suffers from the larger prediction errors by giving higher values of the quantified uncertainty and the resulting flexibility need for the peak hours. Comparative results for both models are depicted in Figure 16 and 17 for aggregated total and aggregated non-flexible loads, respectively. Note that for comparative analysis, both the models are trained on the same set of synthetic data of 1000 houses. The PDFs are plotted for the hour 19:30 for the case of aggregated load and 20:00 for the non-flexible load of peak usage and the quantified uncertainty for the same day-ahead predictions is displayed. It is clear that the flexibility needs prediction resulting from the AGP-based method is much lower compared to the *Prophet*-based model results. This indicates the superiority of the AGP-based methodology in short-term load forecasting with uncertainties, which results in precise uncertainty quantification.

B. Case study analysis of 14 aggregated houses

To demonstrate the performance of the proposed approach for effective demand response decision-making, we assume a case study to forecast the aggregated power consumption of

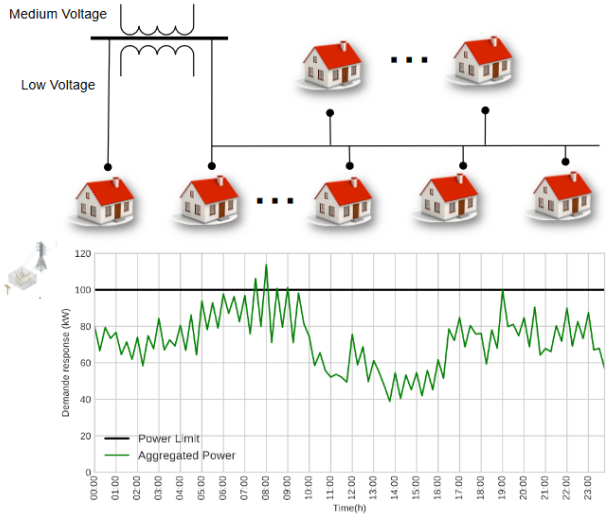

Figure 18. Load profile of 14 households in the peak morning and evening hours on Dec. 21, 2018, connected to the 100 kVA distribution transformer

14 households supplied by the same low-voltage transformer, assuming the threshold capacity of 100 kW. In this case study, the 14 consumers were randomly selected from a database of 1000 houses. Three of the houses in this group consume 10 kW each, while the rest consume between 5 and 10 kW. This study considers the peak period in winter December 2018. During peak usage periods, the network's transformer capacity is critical. To avert overloading, particularly in cold weather conditions, the aggregator could anticipate the fluctuations in demand using the forecasted power consumption to instruct and encourage consumers to adjust their usage patterns in response to changes in electricity prices, grid conditions, or environmental concerns, and regulate their electricity consumption according to the existing flexibility scope. Figure 18 illustrates the aggregated power profile of 14 households on December 21, 2018, where the black line denotes the maximum power limit (y_{lim}), set at 100 kW. To streamline our analysis and avoid repeating figures for all scenarios, we focus specifically on the aggregated load forecast analysis using AGP and *Prophet* models. Beginning with the forecasting of the first model and following the previously mentioned methodology, the uncertainty analysis results for the case of 14 houses are determined.

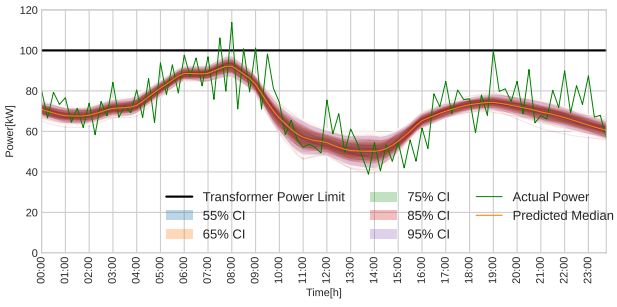

Figure 19. Predicted aggregated end-user load (solid orange line) and associated uncertainty obtained using the **AGP** on a typical winter day.

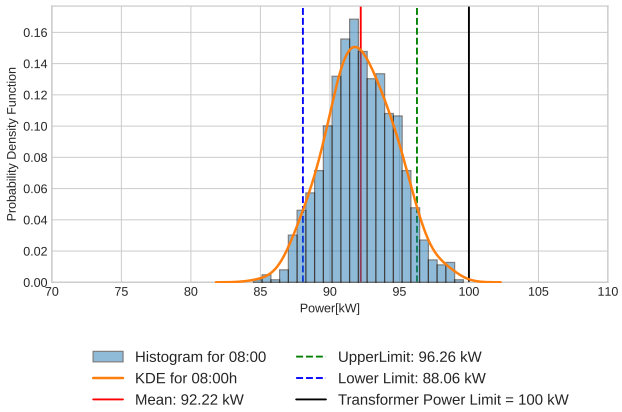

Figure 20. Probability density function of the selected hour (08:00) using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines for aggregated load

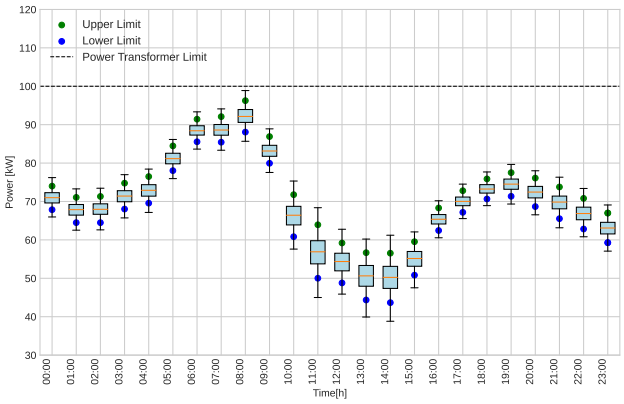

Figure 21. Hourly uncertainty based on a day-ahead forecast of aggregated loads from the **AGP-based** method.

Forecast and uncertainty analysis using AGP: The assessment of uncertainty in demand forecasting was conducted using the hourly probability density function derived from the forecast errors. Subsequently, the hourly risk curve was developed by incorporating all errors from these PDFs. Given that forecast errors can be quantified as a percentage of the rated power, the power consumption can be established based on the inverse cumulative function. Figures 19 and 20 illustrate the necessary variations, following the methodology outlined in Section II. The result indicates that the uncertainty associated with its data predictions remains within the aggregated power limit set for the specified transformer and does not exceed the upper power limit.

Forecast and uncertainty analysis using Prophet: Conversely, with the *Prophet* model, we observe (Figures 22, 23 and 24) an approximate additional load (surpassed power (Δy) of 5 kW. This value represents the flexibility requirement, as determined within the 95% confidence interval. In essence, this additional load reflects the extra capacity that the system might require to handle unforeseen fluctuations, ensuring reliability and stability in power supply to the group of 14 houses under study. As demonstrated in Figures 21 and 24, there

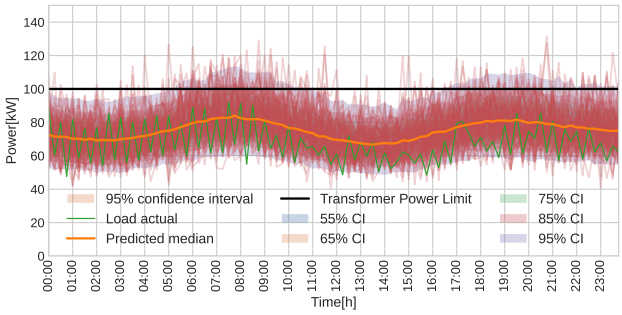

Figure 22. Hourly uncertainty based on a day-ahead forecast of aggregated loads from the **Prophet-based** method.

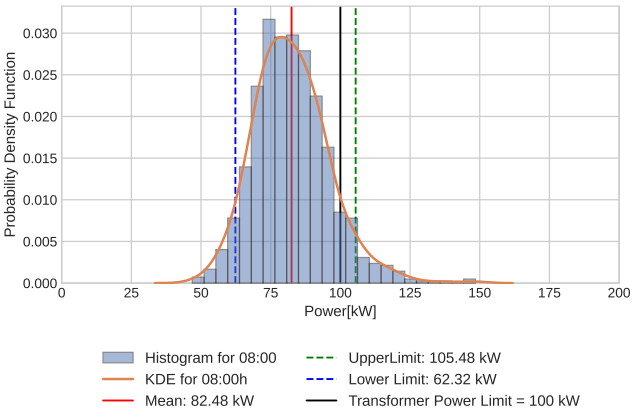

Figure 23. Probability density function of the selected hour (08:00) using both a histogram and a kernel density estimate. The upper and lower limits of power at 95% confidence levels, as well as a power limit value, with vertical lines for aggregated load

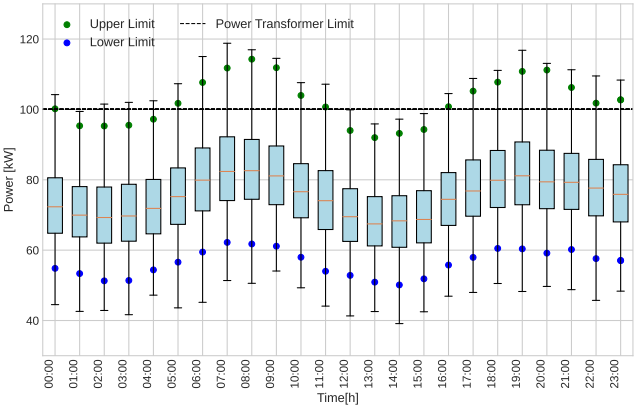

Figure 24. Hourly uncertainty based on a day-ahead forecast of aggregated loads from the **Prophet based** method.

are noticeable differences in peak load predictions during morning and afternoon periods between the two models. This analysis highlights a dual perspective: Firstly, compared to the Prophet method, the advanced forecasting approach with AGP yields superior forecasting results for the short-term horizon. Secondly, better forecasting by AGP lays a robust foundation for strategic planning. That enables organizations to craft flexible, well-equipped strategies to handle various

future scenarios. Applying this approach in a real-world case study with real-grid constraints can help create an efficient indicator for decision-making related to demand response and energy consumption, allowing energy utilities to have proactive communication with consumers ahead of time about the expected demand response events based on forecasted demand.

V. Conclusion

This study introduces a method for quantifying the uncertainties and calculating the flexibility needed for aggregated household power consumption forecasts. The methodology utilized the AGP approach to perform short-term load forecasting by considering uncertainties. Then, a statistical investigation was adopted to quantify the uncertainties present in the forecast on an hourly basis. That resulted in the flexibility need calculations for the peak hours where the power load is most vulnerable to exceeding the capacity limit threshold. To provide a comparative analysis of the proposed methodology, a well-known $Prophet$ forecasting model was employed to perform the forecasting and also accounted for the statistical investigation of quantifying uncertainties. These investigations were performed on a synthetic dataset of an ensemble of 1000 residential buildings located in Québec, Canada and the prediction window was fixed for a 24-hour day ahead. The investigation was also extended to a case study of 14 households connected to the same transformer. The results demonstrated the AGP-based forecast's superiority with precise prediction accuracy. The comparative uncertainty assessment revealed better hourly uncertainty and flexibility requirement calculations for the AGP model in contrast to the $Prophet$ model. This work can be useful to the grid's capacity limitation services where enhancement in forecasting accuracy bolsters the capacity for making more informed decisions.

Acknowledgment

The authors would like to thank Michael Fournier, Juan Carlos Oviedo, and Luis Fernando Rueda Researchers in the Laboratory of Technologies of Énergie (LTE Hydro-Quebec) for their valuable discussions and cooperation in providing the data that improved the quality of the results. This work was supported in part by the Laboratoire des Technologies de l'Energies (LTE) d'Hydro-Quebec, the program MITACS of Canada, the Natural Science and Engineering Research Council of Canada, and the Fondation de l'UQTR

References

[1] Z. Wang, Q. Wen, C. Zhang, L. Sun, and Y. Wang, "DiffLoad: Uncertainty Quantification in Load Forecasting with Diffusion Model," *International journal of forecasting*, 5 2023. [Online]. Available: http://arxiv.org/abs/2306.01001

[2] Y. Wang, G. Hug, Z. Liu, and N. Zhang, "Modeling load forecast uncertainty using generative adversarial networks," *Electric Power Systems Research*, vol. 189, p. 106732, 12 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378779620305356

[3] A. W. Danté, K. Agbossou, S. Kelouwani, A. Cardenas, and J. Bouchard, "Online modeling and identification of plug-in electric vehicles sharing a residential station," *International Journal of Electrical Power & Energy Systems*, vol. 108, pp. 162–176, 6 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061518324426

[4] M. Neukomm, V. Nubbe, and R. Fares, "Grid-interactive Efficient Buildings Technical Report Series: Overview of Research Challenges and Gaps," US-DEPARTEMENT OF ENERGY, us-, Tech. Rep., 12 2019. [Online]. Available: https://www.energy.gov/eere/buildings/articles/grid-interactive-efficient-buildings-technical-report-series-overview

[5] S. E. Ahmadi, D. Sadeghi, M. Marzband, A. Abusorrah, and K. Sedraoui, "Decentralized bi-level stochastic optimization approach for multi-agent multi-energy networked micro-grids with multi-energy storage technologies," *Energy*, vol. 245, p. 123223, 4 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544222001268

[6] M. Blum and M. Riedmiller, "Electricity demand forecasting using gaussian processes," *AAAI Workshop - Technical Report*, vol. WS-13-15, pp. 10–13, 2013. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.649.6285&rep=rep1&type=pdf

[7] C. Heinrich, C. Ziras, T. V. Jensen, H. W. Bindner, and J. Kazempour, "A local flexibility market mechanism with capacity limitation services," *Energy Policy*, vol. 156, p. 112335, 9 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0301421521002056

[8] R. E. Geneidy, B. Howard, and D. Allinson, "Implications of uncertainties in energy demand baseline estimations on building energy flexibility," International Buildibngs, LOUGHBOROUGH UK, Tech. Rep., 2020. [Online]. Available: http://www.ibpsa.org/proceedings/BSO2020/BSOV2020_ElGeneidy.pdf

[9] S. Z. Tajalli, A. Kavousi-Fard, M. Mardaneh, A. Khosravi, and R. Razavi-Far, "Uncertainty-Aware Management of Smart Grids Using Cloud-Based LSTM-Prediction Interval," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 9964–9977, 10 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9505617/

[10] C. Heinrich, C. Ziras, A. L. Syrri, and H. W. Bindner, "EcoGrid 2.0: A large-scale field trial of a local flexibility market," *Applied Energy*, vol. 261, p. 114399, 3 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261919320860

[11] C. Silva, P. Faria, Z. Vale, and J. M. Corchado, "Demand response performance and uncertainty: A systematic literature review," 5 2022.

[12] EnerNOC, "The Demand Response Baseline," *ENERNOC*, pp. 1–5, 1 2009. [Online]. Available: https://www.naesb.org/pdf4/dsmee_group3_100809w3.pdf

[13] J. Xie and T. Hong, "GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1012–1016, 7 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169207015001405

[14] W. El-Baz and P. Tzscheutschler, "Short-term smart learning electrical load prediction algorithm for home energy management systems," *Applied Energy*, vol. 147, pp. 10–19, 6 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261915001592

[15] M. Cao, J.-W. Xiao, H. Fang, Z.-W. Liu, and Y.-W. Wang, "A novel similar-day based probability density forecasting framework for residential loads," *International Journal of Electrical Power & Energy Systems*, vol. 152, p. 109253, 10 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061523003101

[16] M. R. Baker, K. H. Jihad, H. Al-Bayaty, A. Ghareeb, H. Ali, J.-K. Choi, and Q. Sun, "Uncertainty management in electricity demand forecasting with machine learning and ensemble learning: Case studies of COVID-19 in the US metropolitans," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106350, 8 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0952197623005341

[17] B. Ivanovic, Y. Lin, S. Shrivastava, P. Chakravarty, and M. Pavone, "Propagating State Uncertainty Through Trajectory Forecasting," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2351–2358, 10 2021. [Online]. Available: http://arxiv.org/abs/2110.03267

[18] D. Chaturvedi, A. Sinha, and O. Malik, "Short term load forecast using fuzzy logic and wavelet transform integrated generalized neural network," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 230–237, 5 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061514007091

[19] S. Li, X. Kong, L. Yue, C. Liu, M. A. Khan, Z. Yang, and H. Zhang, "Short-term electrical load forecasting using hybrid model of manta ray foraging optimization and support vector regression," *Journal of Cleaner Production*, vol. 388, p. 135856, 2 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652623000148

[20] J. Duan, Q. Tang, J. Ma, and W. Yao, "Operational Status Evaluation of Smart Electricity Meters Using Gaussian Process Regression With Optimized-ARD Kernel," *IEEE Transactions on Industrial Informatics*, pp. 1–11, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10121656/

[21] R. Bessa, C. Möhrlen, V. Fundel, M. Siefert, J. Browell, S. Haglund El Gaidi, B.-M. Hodge, U. Cali, and G. Kariniotakis, "Towards Improved Understanding of the Applicability of Uncertainty Forecasts in the Electric Power Industry," *Energies*, vol. 10, no. 9, p. 1402, 9 2017. [Online]. Available: http://www.mdpi.com/1996-1073/10/9/1402

[22] J. Li, L. Ren, B. Wang, and G. Li, "Probabilistic Load Forecasting of Adaptive Multiple Polynomial Regression considering Temperature Scenario and Dummy variables," *Journal of Physics: Conference Series*, vol. 1550, no. 3, p. 032117, 5 2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1550/3/032117

[23] J. Domínguez-Jiménez, N. Henao, K. Agbossou, A. Parrado, J. Campillo, and S. H. Nagarsheth, "A Stochastic Approach to Integrating Electrical Thermal Storage in Distributed Demand Response for Nordic Communities With Wind Power Generation," *IEEE Open Journal of Industry Applications*, vol. 4, pp. 121–138, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10093061/

[24] A. Zeng, H. Ho, and Y. Yu, "Prediction of building electricity usage using Gaussian Process Regression," *Journal of Building Engineering*, vol. 28, no. April 2019, p. 101054, 3 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S235271021930662X

[25] J. Munkhammar, D. van der Meer, and J. Widén, "Very short term load forecasting of residential electricity consumption using the Markov-chain mixture distribution (MCM) model," *Applied Energy*, vol. 282, p. 116180, 1 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261920315816

[26] H. Quan, D. Srinivasan, and A. Khosravi, "Uncertainty handling using neural network-based prediction intervals for electrical load forecasting," *Energy*, vol. 73, pp. 916–925, 8 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544214008032

[27] D. T. Frazier, W. Maneesoonthorn, G. M. Martin, and B. P. McCabe, "Approximate Bayesian forecasting," *International Journal of Forecasting*, vol. 35, no. 2, pp. 521–539, 4 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016920701830147X

[28] X. Yan, D. Abbes, and B. Francois, "Uncertainty analysis for day ahead power reserve quantification in an urban microgrid including PV generators," *Renewable Energy*, vol. 106, pp. 288–297, 6 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148117300228

[29] W. Zhang, H. Quan, and D. Srinivasan, "An Improved Quantile Regression Neural Network for Probabilistic Load Forecasting," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4425–4434, 7 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8419220/

[30] M. Al-Gabalawy, N. S. Hosny, and A. R. Adly, "Probabilistic forecasting for energy time series considering uncertainties based on deep learning algorithms," *Electric Power Systems Research*, vol. 196, p. 107216, 7 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378779621001978

[31] P. Laurinec, M. Lóderer, M. Lucká, and V. Rozinajová, "Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption," *Journal of Intelligent Information Systems*, vol. 53, no. 2, pp. 219–239, 10 2019. [Online]. Available: http://link.springer.com/10.1007/s10844-019-00550-3

[32] H. M. Dipu Kabir, A. Khosravi, S. Nahavandi, S. Member, and A. Kavousi-Fard, "Partial Adversarial Training for Neural Network-Based Uncertainty Quantification," *IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE*, vol. 5, no. 4, p. 595, 2021. [Online]. Available: https://www.ieee.org/publications/rights/index.html

[33] P. Jiang, R. Li, N. Liu, and Y. Gao, "A novel composite electricity demand forecasting framework by data processing and optimized support vector machine," *Applied Energy*, vol. 260, p. 114243, 2 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261919319300

[34] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 7 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169207015001508

[35] F. Amara, K. Agbossou, Y. Dubé, S. Kelouwani, A. Cardenas, and S. S. Hosseini, "A residual load modeling approach for household short-term load forecasting application," *Energy and Buildings*, vol. 187, pp. 132–143, 2019. [Online]. Available: https://doi.org/10.1016/j.enbuild.2019.01.009

[36] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 12 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253521001081

[37] K. Dab, K. Agbossou, N. Henao, Y. Dubé, S. Kelouwani, and S. S. Hosseini, "A compositional kernel based gaussian process approach to day-ahead residential load forecasting," *Energy and Buildings*, vol. 254, p. 111459, 1 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S037877882100743X

[38] K. Dab, N. Henao, S. Nagarsheth, Y. Dubé, S. Sansregret, and K. Agbossou, "Consensus-based time-series clustering approach to short-term load forecasting for residential electricity demand," *Energy and Buildings*, vol. 299, p. 113550, 11 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778823007806

[39] S. J. Taylor and B. Letham, "Forecasting at Scale," *PeerJ Preprints*, 2017. [Online]. Available: https://peerj.com/preprints/3190v2/

[40] J. Wang, "An Intuitive Tutorial to Gaussian Processes Regression," *Preprint submitted to Elsevier*, 9 2020. [Online]. Available: http://arxiv.org/abs/2009.10862

[41] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005. [Online]. Available: http://www.jmlr.org/papers/v6/quinonero-candela05a.html

[42] C. E. R. C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006, 2006, vol. 7, no. 5. [Online]. Available: www.GaussianProcess.org/gpml

[43] A. I. Almazrouee, A. M. Almeshal, A. S. Almutairi, M. R. Alenezi, and S. N. Alhajeri, "Long-Term Forecasting of Electrical Loads in Kuwait Using Prophet and Holt–Winters Models," *Applied Sciences*, vol. 10, no. 16, p. 5627, 8 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/16/5627

[44] A. Shakeel, D. Chong, and J. Wang, "Load forecasting of district heating system based on improved FB-Prophet model," *Energy*, vol. 278, p. 127637, 9 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544223010319

[45] S. Dash, C. Chakraborty, S. K. Giri, and S. K. Pani, "Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics," *Pattern Recognition Letters*, vol. 151, pp. 69–75, 11 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167865521002762

[46] L. Lind, J. P. Chaves-Ávila, O. Valarezo, A. Sanjab, and L. Olmos, "Baseline methods for distributed flexibility in power systems considering resource, market, and product characteristics," *Utilities Policy*, vol. 86, p. 101688, 2 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095717872300200X

[47] S. McDonald and D. Campbell, "A review of uncertainty quantification for density estimation," *Statistics Surveys*, vol. 15, no. none, pp. 1–71, 1 2021. [Online]. Available: https://projecteuclid.org/journals/statistics-surveys/volume-15/issue-none/A-review-of-uncertainty-quantification-for-density-estimation/10.1214/21-SS130.full

Chapitre 4 - Discussions et Conclusions

Ce chapitre examine les approches utilisées tout au long de la thèse, en analysant les perspectives des méthodes et des algorithmes employés, tout en tenant compte des défis rencontrés et en suggérant des améliorations pour les méthodes appliquées. En liaison avec les deux chapitres précédents, nous évaluons l'efficacité de nos modèles. La méthodologie proposée fournit des outils analytiques et méthodologiques pour améliorer la gestion de la demande d'énergie, facilitant ainsi une prise de décision optimisée de la part des agrégateurs et des opérateurs de réseau. Enfin, nous présentons les différentes conclusions et recommandations pour les travaux futurs de cette thèse.

4.1 Discussions et Perspectives

4.1.1 Perspectives de la modélisation des prévisions à court terme

La modélisation des prévisions à court terme, en constante évolution, offre des perspectives prometteuses tirées de plusieurs axes clés. La gestion efficace des données constitue un point central, avec une exploration approfondie des outils Python, notamment l'intégration d'Influx et d'autres logiciels spécialisés, pour faciliter la partie algorithmique du processus. Une perspective essentielle réside aussi dans le traitement des bases de données, où l'utilisation des compteurs intelligents se présente comme une opportunité significative, permettant l'extraction de données en temps réel pour des prévisions plus précises. La prise en compte des données météorologiques, essentielle pour anticiper les variations de la demande énergétique, soulève des défis algorithmiques liés à la manipulation de données massives. L'avenir de la modélisation repose sur l'amélioration de l'exploitation des données des compteurs intelligents, l'optimisation des bases météorologiques, l'intégration d'outils algorithmiques efficaces, et la recherche de solutions pour gérer la complexité informatique. Il est pertinent aussi d'explorer

des solutions qui exploitent des algorithmes efficaces, tout en optimisant l'utilisation des ressources informatiques. La simulation de longues périodes de prédiction avec des modèles tels que les PGs peut être accélérée en utilisant des outils spécialisés pour gérer les charges computationnelles lourdes, permettant ainsi une modélisation plus rapide et efficace. Ces perspectives promettent des prévisions plus réactives et précises, adaptées aux besoins dynamiques du secteur de l'énergie.

Plus précisément, par rapport au modèle AGP proposée notre principal intérêt ne réside pas dans la simple génération aléatoire des fonctions de covariance à partir d'un a priori, mais plutôt dans l'incorporation des informations que les données d'apprentissage nous fournissent concernant la fonction de covariance. Il existe plusieurs approches pour comprendre les modèles de régression à l'aide du PG. Dans la proposition de notre modèle AGP multidimensionnel, nous avons exploré différentes approches pour créer un noyau combiné ayant les propriétés souhaitées. Dans notre structure, nous avons opté pour l'addition de noyaux du deuxième ordre. Cette approche nous a permis d'intégrer autant de structures de haut niveau que nécessaire dans notre modèle. Lors de la modélisation d'applications à plusieurs dimensions, la somme de noyaux peut créer une structure additive de différentes dimensions. Plus précisément, lorsque les noyaux qui sont additionnés représentent chacun des fonctions dépendant uniquement d'un sous-ensemble de dimensions d'entrée, la configuration sera de la même manière suivante:

$$f(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3) \sim \mathcal{PG}\left(0, k_w(\mathrm{x}_1, \mathrm{x}_1{}') + k_c(\mathrm{x}_2, \mathrm{x}_2{}') + k_{so}(\mathrm{x}_3, \mathrm{x}_3{}')\right)$$

ce qui est équivalent au modèle:

$$f_w(\mathrm{x}_1) \sim \mathcal{PG}\left(0, k_w(\mathrm{x}_1, \mathrm{x}_1{}')\right) \tag{4.1}$$

$$f_c(\mathrm{x}_2) \sim \mathcal{PG}\left(0, k_c(\mathrm{x}_2, \mathrm{x}_2{}'t)\right) \tag{4.2}$$

$$f_{so}(\mathrm{x}_3) \sim \mathcal{PG}\left(0, k_{so}(\mathrm{x}_3, \mathrm{x}_3{}')\right) \tag{4.3}$$

$$f(\mathrm{x}) = f_w(\mathrm{x}_1) + f_c(\mathrm{x}_2) + f_{so}(\mathrm{x}_3) \tag{4.4}$$

Il est important de noter que dans cette décomposition, l'addition de deux noyaux ne possède pas une interprétation similaire à celle de la multiplication de deux fonctions pour effectuer une prédiction. Une autre approche envisageable dans ce cadre est la multiplication des noyaux. Pour sélectionner la combinaison appropriée de noyaux pour l'apprentissage non supervisé, une méthode consiste à utiliser un algorithme Greedy appliqué à une classe de modèles compositionnels pour l'apprentissage non supervisé, une démarche qui est parallèle à notre approche de recherche [58]. Bien que le temps de calcul soit indéniablement prolongé, il permet d'obtenir des résultats performants et potentiellement plus précis. Pour évaluer l'efficacité de notre méthode dans la découverte de structures améliorées, nous avons entrepris une recherche de noyaux sur plusieurs séries temporelles. En réalité, un PG dont le noyau est la somme de noyaux peut être interprété comme la combinaison de fonctions issues de PG constitutifs. Cette approche offre une autre méthode pour visualiser les structures apprises. Notamment, tous les noyaux dans notre espace de recherche peuvent être considérés comme équivalents à des sommes de produits de noyaux de base grâce à l'application de la distributivité. Dans ce contexte, notre modèle peut adopter deux configurations distinctes, tout en conservant son aspect

additif:

$$f(\mathrm{x}) = f_w(\mathrm{x}_1) \times [f_c(\mathrm{x}_2) + f_{so}(\mathrm{x}_3)] = f_w(\mathrm{x}_1) \times f_c(\mathrm{x}_2) + f_w(\mathrm{x}_1) \times f_c(\mathrm{x}_3) \tag{4.5}$$

Dans une deuxième configuration on peut rendre le modèle multiplicatif au au lieu d'additif et la configuration serait comme suit:

$$f(\mathrm{x}) = f_w(\mathrm{x}_1) \times f_c(\mathrm{x}_2) \times f_{so}(\mathrm{x}_3) \tag{4.6}$$

Le modèle développé présente néanmoins certaines limites dans son applicabilité. L'un des problèmes majeurs est la complexité algorithmique incluant des méthodes qui ajustent dynamiquement les priorités au fil du temps pour s'adapter aux changements. Cependant, le modèle AGP présente deux contraintes majeures :

1. La complexité globale de calcul est en $O(N^3)$, où N représente la dimension de la matrice de covariance K.
2. La consommation de mémoire est quadratique: en raison de la complexité de calcul, le modèle standard des PGs atteint rapidement ses limites. Pour les tâches de régression impliquant de grands ensembles de données, un modèle de PG (*"sparse PG"*) est utilisé pour réduire la complexité computationnelle.

Dans notre étude, nous avons appliqué des techniques de classification et d'approximations pour remédier à ces problèmes. La technique du Thompson Sampler, appliquée sur le troisième noyau $f_{so}(\mathrm{x}_3)$, constitue l'une des contributions qui nous a permis d'améliorer les résultats des prévisions. À mesure que l'information est recueillie, la composition des noyaux, combinant des données météorologiques et calendaires concernant les récompenses des actions, est définie. En échantillonnant les actions selon

la probabilité a posteriori qu'elles soient optimales, l'algorithme continue d'échantillonner toutes les actions qui pourraient être optimales, tout en éloignant l'échantillonnage de celles qui sont peu susceptibles d'être optimales. On est donc face au problème de l'adaptabilité à la non-stationnarité. La sélection appropriée des distributions a priori peut être déterminant. Des méthodes plus sophistiquées d'incorporation de connaissances a priori, telles que l'apprentissage bayésien itératif, peuvent aider à ajuster les priors au fur et à mesure de l'apprentissage . En substance, l'algorithme explore toutes les actions prometteuses tout en éliminant progressivement celles considérées comme sous-performantes. Cette intuition est formalisée dans des analyses théoriques récentes du Thompson sampling. Néanmoins, certains problèmes persistent dans ce contexte. Pour pallier ces difficultés, deux algorithmes peuvent être suggérés, notamment l'Upper Confidence Bound (UCB) ou le Bootstrapped Thompson Sampling.

L'algorithme proposé a démontré son efficacité sur différents profils de résidences, illustrant son applicabilité dans la prévision de profils avec des performances significatives. Bien que la méthode AGP soit plus rapide et efficiente, elle nécessite davantage de données et de temps pour converger pendant l'entraînement. En termes de fiabilité face aux perturbations, les AGP surpassent les deux autres techniques examinées. De plus, leur compatibilité avec l'apprentissage semi-supervisé les rend plus adaptés au déploiement dans le secteur résidentiel. Par conséquent, nous avons exploré les fonctions de covariance et les hypersparamètres dans le cadre de notre modèle, identifiant ainsi des aspects clés de l'algorithme qui influent sur la qualité des prévisions. L'ajustement approprié de ces paramètres a été importante pour optimiser les performances du modèle. D'autre part, l'inférence bayésienne des hyperparamètres dans les PGs est cependant difficile et il est nécessaire de trouver un équilibre entre les objectifs de coût computationnel, de précision de prédiction et de robustesse des

intervalles d'incertitude. Alors que dans des conditions simples, l'approche bayésienne complète pourrait être contre-productive, la plupart des applications réelles des PGs reposent sur l'ingénierie de noyaux sophistiqués faits à la main impliquant de nombreux hyperparamètres, où le risque de surajustement est prononcé et plus difficile à détecter. Une solution plus robuste consiste à incorporer des intervalles de prédiction qui reflètent ces incertitudes dans le choix du modèle. Dans cette situation, plutôt que de simplement utiliser l'approche bayésienne complète pour optimiser les hyperparamètres du modèle, une approche plus prudente serait d'incorporer des intervalles de prédiction pour tenir compte des incertitudes associées au choix du modèle. Ces intervalles de prédiction fournissent une indication de la fiabilité des prédictions du modèle, permettant ainsi aux praticiens de mieux évaluer la qualité de leur modèle et de prendre des décisions plus éclairées. En étudiant l'inférence bayésienne complète dans des modèles PG plus sophistiqués, tels que les PG profonds, déformés et convolutionnels, nous pourrions mieux comprendre comment gérer ces incertitudes de manière plus efficace et développer des modèles plus fiables pour la prédiction de la consommation d'énergie ou d'autres applications similaires [92], déformés [93] et convolutionnels [94] offrira une meilleure compréhension de cette question et constitue une direction imminente des travaux futurs.

Comparé à d'autres algorithmes de classification des sous-ensembles de données, nous avons examiné différentes approches d'approximation, mettant en avant l'approche du sous-ensemble de données (SoD). Malgré sa complexité réduite, SoD peut rencontrer des défis liés à la génération de variances de prédiction précises, en particulier dans des cas impliquant des données redondantes. SoD représente une stratégie simple pour approximer le AGP. Concernant la sélection du SoD, différentes méthodes peuvent être employées, notamment la sélection aléatoire, l'utilisation de techniques de clustering telles que le K-means [15] et K-dimensional tree [95], ou l'application de critères d'apprentissage actif,

chacune ayant ses avantages et inconvénients.

En conclusion, bien que notre modèle ait montré des résultats prometteurs, il existe encore des opportunités d'amélioration. Les travaux futurs pourraient se concentrer sur l'exploration de nouvelles fonctions de covariance, la recherche de stratégies de sélection de sous-ensemble plus robustes, et l'extension du modèle à des contextes spécifiques pour évaluer sa généralisabilité.

4.1.2 Perspectives de la classification des profils de charges

En ce qui concerne les défis et les perspectives liés à la classification de profils des charges agrégées pour la prévision à court terme, plusieurs obstacles sont identifiés, notamment la dimensionnalité élevée, la forte corrélation des données, et la présence significative de bruit dans les séries temporelles. Le choix des méthodes de classification, notamment les mesures de similarité et les critères de sélection, constitue un défi majeur. Un exemple notable de ces derniers est le critère d'inertie utilisé dans l'algorithme de K-moyennes. Il mesure la dispersion des points au sein d'un même groupe, cherchant à minimiser cette dispersion pour obtenir des groupes plus cohérents. Ce critère influence directement la qualité de la classification en évaluant la compacité des classes formés. Ainsi, le choix judicieux du critère d'inertie peut grandement influencer les performances et la pertinence des méthodes de classification dans divers contextes d'analyse de données.

Par ailleurs, les algorithmes de classification non supervisées présentent des limitations, notamment la nécessité de pré-allouer le nombre de classes, tandis que le clustering hiérarchique, bien que flexible, est restreint aux petits ensembles de données en raison de sa complexité quadratique. Les algorithmes basés sur des modèles et sur la densité sont moins utilisés en raison de leur temps de calcul élevé et de leur sensibilité

aux hypothèses de l'utilisateur sur les paramètres. Les perspectives de notre méthode proposée reposent sur l'exploitation des caractéristiques spécifiques des sous-groupes de données où plusieurs types de données contribuent aux prévisions. On pourrait également intégrer des mécanismes d'adaptation dynamique pour ajuster automatiquement les classes en fonction des changements de comportement dans les données, par exemple, en utilisant des algorithmes de détection de rupture tels que le changepoint detection ou l'algorithme Pelt (Pruned Exact Linear Time). Ces algorithmes sont capables d'assurer une meilleure adaptation aux évolutions temporelles. D'autre part, plusieurs perspectives spécifiques émergent pour les algorithmes de prévision basés sur des classes et la CC. Les axes comprennent l'amélioration de la précision grâce à des méthodes d'agrégation avancées, l'intégration de données multiples, l'interprétabilité des clusters et l'exploration d'applications sectorielles spécifiques. Les aspects clés pour la CC comprennent plusieurs éléments cruciaux. Premièrement, il est essentiel de renforcer la stabilité du système. Ensuite, assurer la robustesse face aux données bruitées est un autre enjeu majeur. La capacité à étendre le processus au traitement de données temporelles est également importante. Par ailleurs, l'affinement des mesures de similarité joue un rôle clé dans l'amélioration de la précision du clustering. De plus, il est nécessaire d'améliorer l'évolutivité du système pour gérer efficacement des volumes de données plus importants. Enfin, l'intégration de l'incertitude dans le processus de CC représente un aspect fondamental pour améliorer la fiabilité et l'efficacité du système.

En résumé, bien que des progrès aient été réalisés dans la représentation des séries temporelles, ainsi que dans les mesures de distance et les critères de sélection, il est impératif d'intensifier les efforts pour améliorer les approches de classification. De plus, travailler sur des techniques visant à améliorer la évolutivité de l'algorithme de CC est essentiel, permettant ainsi son application à des ensembles de données plus vastes

tout en maintenant des performances de qualité. Une attention particulière portée à des comparaisons approfondies dans divers contextes peut orienter les futures recherches vers une meilleure compréhension des besoins spécifiques de la classification comportementale des charges.

4.1.3 Perspectives de l'analyse des incertitudes dans les prévisions à court terme

Il est essentiel d'inclure et de communiquer clairement les informations sur l'incertitude dans les prévisions à court terme des charges pour optimiser les opérations des systèmes énergétiques. La conscience de cette incertitude est importante pour permettre aux utilisateurs d'anticiper les variations imprévisibles auxquelles les systèmes énergétiques sont confrontés et de prendre des décisions éclairées en tenant compte des scénarios possibles. Notre étude a examiné les incertitudes liées aux charges flexibles et non flexibles, ajoutant une dimension de complexité à nos prévisions. Ces aspects, combinés à l'utilisation de deux techniques distinctes, le modèle FB et le AGP, ont enrichi notre compréhension des sources d'incertitude. Cependant, deux perspectives ont été confrontées au cours de ces analyses, présentées comme suit : Premièrement, les charges non contrôlables, souvent imprévisibles, peuvent introduire des fluctuations inattendues dans la demande d'énergie. En tenant compte des incertitudes dans les prévisions de ces charges, nous pouvons développer des stratégies plus robustes pour ajuster la production d'énergie en temps réel, améliorant ainsi la résilience du système face aux variations imprévues. En outre, la résolution des défis liés à l'exploitation réelle des marchés locaux de ressources énergétiques distribuées (DER) implique de relever plusieurs défis clés. Tout d'abord, il est nécessaire de résoudre la question des puissance limite P_l ou le baseline. Par conséquent, le risque horaire prend en compte toutes les erreurs des fonctions de distribution cumulatives (CDF). La fonction de densité de probabilité (PDF) peut être

dérivée à partir des échantillons de l'analyse prédictive, exprimant les erreurs de prévision en pourcentage de la puissance nominale. Un paramètre clé pour évaluer la fiabilité est le loss of load probability (LOLP), qui mesure la probabilité que la demande de charge (L_h) dépasse la limite de puissance (P_h) à un moment donné. Des études approfondies sont nécessaires pour déterminer l'utilisation efficace de cette puissance dans des conditions changeantes, telles que les fluctuations du prix de l'électricité et la planification du réseau par les opérateurs. Il est également essentiel de déterminer si l'accent doit être mis sur les services de limitation de capacité uniquement, en tenant compte notamment des défis liés à la définition de la capacité pour de très petits agrégats de ressources énergétiques distribuées (DER), ce qui peut introduire des incertitudes et potentiellement réduire la valeur globale des services fournis.

Deuxièmement, il est essentiel de prêter une attention spécifique à la conception du mécanisme de marché. Compte tenu de la possibilité qu'un nombre limité d'agrégateurs participent aux enchères, l'examen des questions d'équité, des stratégies potentielles et de la manipulation des prix de l'électricité devient primordial. Dans cette situation, l'utilité pourra mieux comprendre la quantification de la flexibilité. Cela lui permettra de négocier ses besoins en termes de flexibilité de manière plus éclairée.

En conclusion, la gestion de l'incertitude dans les prévisions à court terme des charges sont des aspects cruciaux pour le secteur énergétique. En adoptant une approche proactive et en intégrant ces informations dans les opérations des systèmes, nous pouvons améliorer la fiabilité, la résilience et l'efficacité globale des réseaux énergétiques. L'analyse des incertitudes explique la diversité des comportements des consommateurs et les variations potentielles dans les schémas de consommation. L'utilisation de deux modèles distincts dans les analyses statistiques a permis de comparer les performances et les limitations de

chaque approche de prévision dans la gestion de l'incertitude. Cette analyse comparative ouvre la voie au développement de modèles combinées qui pourraient capitaliser sur les avantages de chaque méthode.

4.2 Conclusions et recommandations

4.2.1 Conclusions

Il est intéressant d'anticiper le comportement énergétique du réseau résidentiel global ainsi que de chacun de ses composants afin d'optimiser ses opérations. Les recherches durant cette thèse ont principalement porté sur le développement de modèles de prévision à court terme de la demande électrique pour un ensemble de résidences. La méthodologie statistique adoptée dans cette étude vise à modéliser la consommation électrique des habitations, permettant ainsi de prédire la demande sur une période de vingt-quatre heures. Les modèles proposés étaient principalement non paramétriques. Plus spécifiquement, dans cette thèse, nous avons abordé la question de la prévision à court terme sous trois angles distincts. Le premier axe s'est concentré sur l'élaboration d'un modèle de prévision répondant à une demande fréquemment exprimée dans la littérature qui est la modélisation des prévisions pour un ensemble de résidences en tenant compte de cinq variables exogènes, à savoir la température extérieure, l'humidité extérieure, le rayonnement solaire global, le jour de la semaine et l'heure du jour. Dans cette partie de la thèse, nous avons examiné spécifiquement les limites des méthodes existantes, en particulier en ce qui concerne la modélisation non paramétrique multivariée de la consommation de résidences. Nous avons introduit une approche basée sur l'analyse statistique bayésienne, reposant sur des processus gaussiens additifs. Nous avons également ajusté les hyperparamètres et modifié la composition des noyaux, en mettant en avant la contribution significative de Processus Gaussien Additif (AGP). Par la suite, le modèle AGP non paramétrique

intégrant le noyau des événements météorologiques et calendaires présenté dans cette thèse, a été utilisé pour prédire la consommation d'un agrégat de maisons. Une comparaison a été réalisée en incluant d'autres modèles de prévision a également été présentée pour démontrer la supériorité du modèle AGP par rapport au modèle paramétriques et non paramétriques.

Dans le deuxième axe, nous avons développé une stratégie de classification des profils de charges résidentiels. L'analyse des données souligne que la diversité de ces profils exerce une influence significative sur l'hétérogénéité de la consommation énergétique. À cet égard, nous avons conçu une approche visant à classer, à un pas de 15 minutes, la probabilité qu'une maison appartienne à une classe spécifique. Cette démarche permet d'obtenir une stratégie de classification comportementale plus cohérente. Une approche de classification de séries temporelles basée sur la classification consensuelle et utilisant la classification K-médoïdes est adoptée. Enfin, la classification consensuelle est appliquée aux classes résultantes de K-médoïdes. L'algorithme de classification consensuelle proposé s'est distingué dans la classification des profils de puissance des résidences sur 24 heures. Les prévisions basées sur l'AGP sont améliorées en intégrant la nouvelle stratégie de CBAF. Une étude de simulation a été réalisée pour démontrer l'efficacité de la méthode de prévision suggérée sur des ensembles de données de séries temporelles de 17 puis 1000 maisons. Les résultats de validation démontrent que la prévision agrégée basée sur CBAF, constitue une alternative supérieure aux techniques existantes de prévision de la demande d'électricité à court terme.

Le troisième axe s'est concentré sur la conception d'un mécanisme de quantification des incertitudes de prévision. Dans le contexte des marchés de flexibilité, les agrégateurs jouent un rôle essentiel en agissant comme des intermédiaires entre les consommateurs

et les opérateurs de réseaux. Leur fonction essentielle est de gérer le réseau en proposant des réductions de charge basées sur des limites de puissance. Cependant, l'incertitude liée au comportement des consommateurs présente des contraintes significatives. Pour surmonter ces défis, la méthodologie adoptée dans cette thèse vise à produire des prévisions qui résonnent profondément avec les comportements des réseaux électriques modernes, offrant aux opérateurs de réseau non seulement des prévisions ponctuelles, mais également une vision plus claire des incertitudes associées. En comprenant les intervalles de confiance entourant ces prévisions, les opérateurs peuvent prendre des décisions plus éclairées, garantissant des opérations optimales du réseau. L'efficacité de ce mécanisme proposé est évaluée sur des données synthétiques des charges flexibles et non flexibles, comparant les résultats avec des modèles de prévision couramment utilisés tels que le modèle Facebook Prophet (FB). Cette comparaison démontre l'utilité de l'approche proposée basée sur AGP dans le contexte des marchés de flexibilité par rapport à FB. Il est important de souligner que cette stratégie trouve une application concrète dans notre démarche de modélisation tout au long de la thèse illustrant ainsi son utilité et sa pertinence dans les marchés de flexibilité.

4.2.2 Recommandations

Les résultats obtenus ainsi que les limites identifiées dans nos travaux actuels nous conduisent à suggérer plusieurs recommandations pour les travaux futurs. Étant donné la nature aléatoire du comportement humain et le fait que nos méthodes actuelles ont été évaluées dans des cas d'étude spécifiques, nous suggérons d'explorer davantage les limites de ces approches.

- Dans des applications futures, il serait pertinent d'intégrer d'autres charges au sein de l'architecture de prévision, comme les véhicules électriques [96]. En outre,

les véhicules électriques ont un impact significatif sur la demande d'énergie, notamment lors des périodes de recharge. Premièrement, intégrer ces charges dans les prévisions permettrait de mieux anticiper et gérer les variations de la demande énergétique liées à la mobilité électrique. Deuxièmement, en tenant compte des véhicules électriques dans les prévisions, les gestionnaires de réseau peuvent développer des stratégies pour gérer efficacement la charge, évitant ainsi des pics soudains de demande qui pourraient surcharger le réseau.

Par ailleurs, étant donné la nature aléatoire du comportement humain et la spécificité des cas d'étude évalués, il serait judicieux d'approfondir la compréhension des limites des approches de prévisions. Alors, il serait intéressant de prendre en considération les températures intérieures par zone de chaque résidence en tenant compte des comportements d'occupation. Ces éléments peuvent apporter une flexibilité supplémentaire aux réseaux électriques [97]. À cet effet, il serait envisageable de classer les résidences en fonction du type de bâtiment et du nombre de zones à l'intérieur de chacune d'entre elles. Dans cette optique, il serait enrichissant d'explorer des méthodes permettant de sélectionner l'ordre des hyperparamètres du modèle et de déterminer leur nombre optimal à inclure dans l'approche du AGP. Cette démarche viserait à améliorer la performance des prévisions. Un élément fondamental dans l'approche multivariée consiste à trouver des solutions pour analyser les données dans des espaces de dimensions différentes.

- Une autre recommandation serait d'incorporer le prix de l'électricité en tant que nouvelle variable exogène. Il convient de noter que l'augmentation de la dimension de l'estimateur peut rendre les calculs plus complexes et exiger une plus grande capacité de mémoire [32]. Par conséquent, il serait approprié d'étudier la manière d'adapter nos méthodes de noyaux additives pour gérer des données

multidimensionnelles de grandes dimensions. Ainsi, on peut progresser vers un système entièrement dirigé vers des approches bayesiens robustes ou même par l'IA pour avoir des meilleurs prévisions à court-terme.

- Les prévisions probabilistes des heures atypiques reposent sur un algorithme spécialisé que nous avons développé à cet effet. Il est important de noter que ces prédictions ne sont pas disponibles pour les jours inhabituels par exemple le jour de Noël, ou la variation de la consommation des résidences à cause du Covid 19, etc. car ce sont des évènements rares. Il est nécessaire de mener des recherches afin d'identifier ces heures inhabituelles et d'explorer leur pertinence. Par exemple, l'utilisation de techniques de détection d'anomalies peut permettre d'automatiser ce processus, facilitant ainsi la prise de décision informée et la gestion efficace des variations exceptionnelles de la demande énergétique. Cependant, l'un des principaux défis est que la méthodologie utilisée dans la thèse dépend de vastes ensembles de données qui ne contiennent pas nécessairement les informations requises pour la création de ces modèles. Néanmoins, une approche hybride pourrait être une voie prometteuse.
- Durant cette thèse, nous avons traité la problématique du regroupement des profils de charges au niveau agrégé des résidences. Cependant, pour les travaux futurs, d'autres analyses pourraient se concentrer sur la caractérisation des incertitudes au sein des différentes classes de résidences. Ces analyses pourraient bénéficier des résultats obtenus dans notre deuxième article, où une classification efficace des profils de charges résidentielles a été réalisée. Ainsi, il serait intéressant d'appliquer des techniques de quantification de la flexibilité spécifiquement aux profils de charges de chaque classe résultante de la technologie de classification proposée dans les travaux précédents. Finalement, il serait aussi pertinent d'explorer

l'exploitation de la méthode AGP présentée dans la thèse dans une stratégie de contrôle, surtout en tirant parti de l'amélioration de l'incertitude.

- Un autre aspect à considérer dans les futurs développements est la manière de résoudre le problème couramment rencontré en statistique, connu sous le nom de *fléau de la dimension* ou *malédiction de la dimension*. Nous devons explorer des approches pour atténuer les défis posés par la haute dimensionnalité des données. Les perspectives d'amélioration que nous avons évoquées ci-dessus sont particulièrement envisageables lors de l'utilisation de données supplémentaires. Ces nouvelles sources de données pourraient enrichir nos modèles et améliorer la précision de nos prévisions. Nous avons déjà utilisé l'analyse en composantes principales (PCA) dans notre analyse de sensibilité. Cependant, d'autres algorithmes telles que des réseaux de neurones auto encodeurs ou bien Linear Discriminant Analysis (LDA) [98] peuvent être utilisés pour éliminer les caractéristiques redondantes et réduire ainsi la dimension partir de la variance.

Références

[1] P. Gass, D. Echeverría, and A. Asadollahi, "Cities and Smart }Grids in Canada," International Institute for Sustainable Development, canada, Tech. Rep., 9 2017. [Online]. Available: www.iisd.org

[2] Johanne Whitmore, Pierre-Olivier Pineau, and Sylvain Audette, *RÉGLEMENTATION DE L'ÉNERGIE AU QUÉBEC QUELLES OPTIONS POUR ACCÉLÉRER LA TRANSITION ÉNERGÉTIQUE ET LA DÉCARBONATION ?*, 2021st ed. canada: Chaire de gestion du secteur de l'énergie, HEC Montréal, 9 2021, vol. 2021. [Online]. Available: https://transitionenergetique.gouv.qc.ca/plan-directeur-en-transition-energetique

[3] M. Grabner, Y. Wang, Q. Wen, B. Blažič, and V. Štruc, "A Global Modeling Framework for Load Forecasting in Distribution Networks," *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4927–4941, 11 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10092804/

[4] Natural Resources Canada (NRCan), "Energy Factbook 2023-2024," Natural Resources Canada, canada, Tech. Rep., 2023. [Online]. Available: copyright-droitdauteur@nrcan-rncan.gc.ca

[5] Johanne Whitmore and Pierre-Olivier Pineau, *DONNÉES SUR L'ÉNERGIE AU CANADA*, 2021st ed. Canada: Statistique Canada, 9 2022. [Online]. Available: https://energie.hec.ca/wp-content/uploads/2022/09/RAPPORT_DonneesEnergie_WEB.pdf

[6] A. Wadhera, J. Ayoub, and M. Roy, "Smart Grid in Canada," *Natural Resources Canada*, 4 2019. [Online]. Available: https://natural-resources.canada.ca/sites/www.nrcan.gc.ca/files/canmetenergy/pdf/Smart%20Grid%20in%20Canada%20Report%20Web%20FINAL%20EN.pdf

[7] Société d'habitation du Québec, "Les baby-boomers et le logement (Habitation Québec)," *Le bulletin d'information de la société d'habitation du Québec*, vol. 5, no. 1, pp. 1–16, 2010. [Online]. Available: http://habitation.gouv.qc.ca/fileadmin/internet/publications/H01051.pdf

[8] J. Whitmore and P.-O. Pineau, "Energie a Quebec," État de l'énergie au Québec 2019, Chaire de gestion du secteur de l'énergie, Tech. Rep., 2018.

[9] Martel Eric, "Annual Report HQ 2019," HYDRO-QUÉBEC, Canada, Tech. Rep., 2 2020. [Online]. Available: https://www.hydroquebec.com/data/documents-donnees/pdf/annual-report.pdf

[10] P. Wang, B. Liu, and T. Hong, "Electric load forecasting with recency effect: A big data approach," *International Journal of Forecasting*, vol. 32, no. 3, pp. 585–597, 7 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169207015001557

[11] T. Hong and M. Shahidehpour, "Load Forecasting Case Study," *U.S. Department of Energy*, 2015. [Online]. Available: https://pubs.naruc.org/pub.cfm?id=536E10A7-2354-D714-5191-A8AAFE45D626

[12] A. Cini, S. Lukovic, and C. Alippi, "Cluster-based Aggregate Load Forecasting with Deep Neural Networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 7 2020, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9207503/

[13] K. Park, S. Yoon, and E. Hwang, "Hybrid Load Forecasting for Mixed-Use Complex Based on the Characteristic Load Decomposition by Pilot Signals," *IEEE Access*, vol. 7, pp. 12 297–12 306, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8610068/

[14] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," *Information Systems*, vol. 53, pp. 16–38, 10 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000733

[15] G. Le Ray and P. Pinson, "Online adaptive clustering algorithm for load profiling," *Sustainable Energy, Grids and Networks*, vol. 17, 3 2019.

[16] T. Yang, N. Pasquier, and F. Precioso, "Semi-supervised consensus clustering based on closed patterns," *Knowledge-Based Systems*, vol. 235, p. 107599, 1 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705121008613

[17] X. Cheng, L. Wang, P. Zhang, X. Wang, and Q. Yan, "Short-term fast forecasting based on family behavior pattern recognition for small-scale users load," *Cluster Computing*, vol. 25, no. 3, pp. 2107–2123, 6 2022. [Online]. Available: https://link.springer.com/10.1007/s10586-021-03362-9

[18] C. Tarmanini, N. Sarma, C. Gezegin, and O. Ozgonenel, "Short term load forecasting based on ARIMA and ANN approaches," *Energy Reports*, vol. 9, pp. 550–557, 5 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352484723000653

[19] K. Jeong, C. Koo, and T. Hong, "An estimation model for determining the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network)," *Energy*, vol. 71, pp. 71–79, 7 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S036054421400440X

[20] K. B. Debnath and M. Mourshed, "Forecasting methods in energy planning models," *Renewable and Sustainable Energy Reviews*, vol. 88, no. March, pp. 297–325, 5 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032118300200

[21] C. Sandels, J. Widén, and L. Nordström, "Forecasting household consumer electricity load profiles with a combined physical and behavioral approach," *Applied Energy*, vol. 131, pp. 267–278, 10 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261914006308

[22] Y. Liang, D. Niu, M. Ye, and W.-C. Hong, "Short-Term Load Forecasting Based on Wavelet Transform and Least Squares Support Vector Machine Optimized by Improved Cuckoo Search," *Energies*, vol. 9, no. 10, p. 827, 10 2016. [Online]. Available: http://www.mdpi.com/1996-1073/9/10/827

[23] J. Luo, T. Hong, and S.-C. Fang, "Benchmarking robustness of load forecasting models under data integrity attacks," *International Journal of Forecasting*, vol. 34, no. 1, pp. 89–104, 1 2018. [Online]. Available: http://dx.doi.org/10.1016/j.ijforecast.2017.08.004https://linkinghub.elsevier.com/retrieve/pii/S0169207017300900

[24] Z. Tavassoli-Hojati, S. Ghaderi, H. Iranmanesh, P. Hilber, and E. Shayesteh, "A self-partitioning local neuro fuzzy model for short-term load forecasting in smart grids," *Energy*, vol. 199, p. 117514, 5 2020. [Online]. Available: https://doi.org/10.1016/j.energy.2020.117514https://linkinghub.elsevier.com/retrieve/pii/S0360544220306216

[25] A. Baliyan, K. Gaurav, and S. K. Mishra, "A Review of Short Term Load Forecasting using Artificial Neural Network Models," *Procedia Computer Science*, vol. 48, no. C, pp. 121–125, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2015.04.160https://linkinghub.elsevier.com/retrieve/pii/S1877050915006699

[26] T. Alquthami, M. Zulfiqar, M. Kamran, A. H. Milyani, and M. B. Rasheed, "A Performance Comparison of Machine Learning Algorithms for Load Forecasting in Smart Grid," *IEEE Access*, vol. 10, pp. 48 419–48 433, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9765492/

[27] L. P. Clark Velasco, D. L. Lou Polestico, G. O. Paolo Macasieb, M. V. Bryan Reyes, and F. B. Vasquez Jr, "Load Forecasting using Autoregressive Integrated Moving Average and Artificial Neural Network," International Journal of Advanced Computer Science and Applications, Tech. Rep. 7, 2018. [Online]. Available: www.ijacsa.thesai.org

[28] F. Wahid, R. Ghazali, A. S. Shah, and M. Fayaz, "Prediction of Energy Consumption in the Buildings Using Multi-Layer Perceptron and Random Forest," *International Journal of Advanced Science and Technology*, vol. 101, pp. 13–22, 4 2017. [Online]. Available: http://article.nadiapub.com/IJAST/vol101/2.pdf

[29] E. Erişen, C. Iyigun, and F. Tanrısever, "Short-term electricity load forecasting with special days: an analysis on parametric and non-parametric methods," *Annals of Operations Research*, pp. 1–34, 12 2017.

[30] C. Rao, Y. Zhang, J. Wen, X. Xiao, and M. Goh, "Energy demand forecasting in China: A support vector regression-compositional data second exponential smoothing model," *Energy*, vol. 263, 1 2023.

[31] J. Wang, "An Intuitive Tutorial to Gaussian Processes Regression," *arxiv.org*, 9 2024. [Online]. Available: https://arxiv.org/pdf/2009.10862.pdf

[32] C. E. R. C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006, 2006, vol. 7, no. 5. [Online]. Available: www.GaussianProcess.org/gpml

[33] Y. Yang, S. Li, W. Li, and M. Qu, "Power load probability density forecasting using Gaussian process quantile regression," *Applied Energy*, vol. 213, no. November 2017, pp. 499–509, 3 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261917316100

[34] M. Alamaniotis, S. Chatzidakis, and L. Tsoukalas, "Monthly load forecasting using kernel based gaussian process regression," in *MedPower 2014*, vol. 2014, no. CP665. Institution of Engineering and Technology, 2014, pp. 60 (8 pp.)–60 (8 pp.). [Online]. Available: https://digital-library.theiet.org/content/conferences/10.1049/cp.2014.1693

[35] C. Leysen, M. Verbeke, P. Dagnely, and W. Meert, "Energy consumption profiling using Gaussian processes," in *2016 IEEE 8th International Conference on Intelligent Systems (IS)*. IEEE, 9 2016, pp. 470–477. [Online]. Available: http://ieeexplore.ieee.org/document/7737463/

[36] T. X. Nghiem and C. N. Jones, "Data-driven demand response modeling and control of buildings with Gaussian Processes," in *2017 American Control Conference (ACC)*. IEEE, 5 2017, pp. 2919–2924. [Online]. Available: http://ieeexplore.ieee.org/document/7963394/

[37] A. Zeng, H. Ho, and Y. Yu, "Prediction of building electricity usage using Gaussian Process Regression," *Journal of Building Engineering*, vol. 28, no. November 2019, p. 101054, 3 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S235271021930662X

[38] D. Foreman-Mackey, E. Agol, S. Ambikasaran, and R. Angus, "Fast and Scalable Gaussian Process Modeling with Applications to Astronomical Time Series," *The Astronomical Journal*, vol. 154, no. 6, p. 220, 11 2017. [Online]. Available: https://iopscience.iop.org/article/10.3847/1538-3881/aa9332

[39] H. Liu, J. Cai, Y.-S. Ong, and Y. Wang, "Understanding and comparing scalable Gaussian process regression for big data," *Knowledge-Based Systems*, vol. 164, pp. 324–335, 1 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705118305380

[40] I. R. Goumiri, B. W. Priest, and M. D. Schneider, "Reinforcement Learning via Gaussian Processes with Neural Network Dual Kernels," *arXiv.*, pp. 1–22, 2020. [Online]. Available: http://arxiv.org/abs/2004.05198

[41] M. Shepero, D. van der Meer, J. Munkhammar, and J. Widén, "Residential probabilistic load forecasting: A method using Gaussian process designed for electric load data," *Applied Energy*, vol. 218, no. March, pp. 159–172, 5 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S030626191830299X

[42] A. Prakash, S. Xu, R. Rajagopal, and H. Noh, "Robust Building Energy Load Forecasting Using Physically-Based Kernel Models," *Energies*, vol. 11, no. 4, p. 862, 4 2018. [Online]. Available: http://www.mdpi.com/1996-1073/11/4/862

[43] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression," *IEEE Transactions*

on Industrial Electronics, vol. 65, no. 1, pp. 300–308, 1 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7945510/

[44] S. Park and S. Choi, "Hierarchical gaussian process regression," *Journal of Machine Learning Research*, vol. 13, pp. 95–110, 2010. [Online]. Available: http://proceedings.mlr.press/v13/park10a.html

[45] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," pp. 1104–1122, 6 2017.

[46] R. Sevlian and R. Rajagopal, "A scaling law for short term load forecasting on varying levels of aggregation," *International Journal of Electrical Power & Energy Systems*, vol. 98, pp. 350–361, 6 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061517306956

[47] Y. Wang and J. M. Bielicki, "Acclimation and the response of hourly electricity loads to meteorological variables," *Energy*, vol. 142, pp. 473–485, 1 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544217317061

[48] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Applied Energy*, vol. 205, pp. 654–669, 11 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261917309881

[49] F. Amara, K. Agbossou, Y. Dubé, S. Kelouwani, A. Cardenas, and S. S. Hosseini, "A residual load modeling approach for household short-term load forecasting application," *Energy and Buildings*, vol. 187, pp. 132–143, 3 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778818309228

[50] A. Tascikaraoglu and B. M. Sanandaji, "Short-term residential electric load forecasting: A compressive spatio-temporal approach," *Energy and Buildings*, vol. 111, pp. 380–392, 1 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S037877881530431X

[51] I. Shah, H. Iftikhar, S. Ali, and D. Wang, "Short-Term Electricity Demand Forecasting Using ComponentsEstimation Technique," *Energies*, vol. 12, no. 13, p. 2532, 7 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/13/2532

[52] D. Toquica, K. Agbossou, R. Malhamé, N. Henao, S. Kelouwani, and A. Cardenas, "Adaptive Machine Learning for Automated Modeling of Residential Prosumer Agents," *Energies*, vol. 13, no. 9, p. 2250, 5 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/9/2250

[53] W. Charytoniuk, M. Chen, and P. Van Olinda, "Nonparametric regression based short-term load forecasting," *IEEE Transactions on Power Systems*, vol. 13, no. 3, pp. 725–730, 1998. [Online]. Available: http://ieeexplore.ieee.org/document/708572/

[54] G.-F. Fan, Y.-H. Guo, J.-M. Zheng, and W.-C. Hong, "Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting," *Energies*, vol. 12, no. 5, p. 916, 3 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/5/916

[55] H. Jiang, Y. Zhang, E. Muljadi, J. J. Zhang, and D. W. Gao, "A Short-Term and High-Resolution Distribution System Load Forecasting Approach Using Support Vector Regression With Hybrid Parameters Optimization," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3341–3350, 7 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7748604/https://ieeexplore.ieee.org/document/7748604/

[56] A. Jain, T. Nghiem, M. Morari, and R. Mangharam, "Learning and Control Using Gaussian Processes," in *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 4 2018, pp. 140–149. [Online]. Available: https://ieeexplore.ieee.org/document/8443729/

[57] Y. Zhang and G. Luo, "Short term power load prediction with knowledge transfer," *Information Systems*, vol. 53, pp. 161–169, 10 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000150

[58] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani, "Structure Discovery in Nonparametric Regression through Compositional Kernel Search," *30th International Conference on Machine Learning, ICML 2013*, vol. 28, no. PART 3, pp. 2203–2211, 2 2013. [Online]. Available: http://arxiv.org/abs/1302.4922

[59] S. Fan, S. Member, and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Transactions on Power Systems*, no. August, pp. 1–8, 2010. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5985500

[60] G. Xie, X. Chen, and Y. Weng, "An integrated Gaussian process modeling framework for residential load prediction," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7238–7248, 2018.

[61] S. Humeau, T. K. Wijaya, M. Vasirani, and K. Aberer, "Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households," in *2013 Sustainable Internet and ICT for Sustainability (SustainIT)*. IEEE, 10 2013, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/6685208/

[62] P. Goncalves Da Silva, D. Ilic, and S. Karnouskos, "The Impact of Smart Grid Prosumer Grouping on Forecasting Accuracy and Its Benefits for Local Electricity Market Trading," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 402–410, 1 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6684330/

[63] S. Bandyopadhyay, T. Ganu, H. Khadilkar, and V. Arya, "Individual and Aggregate Electrical Load Forecasting," in *Proceedings of the 2015 ACM Sixth International Conference on Future Energy Systems*. New York, NY, USA: ACM, 7 2015, pp. 121–130. [Online]. Available: https://dl.acm.org/doi/10.1145/2768510.2768539

[64] K. Nikolopoulos, A. A. Syntetos, J. E. Boylan, F. Petropoulos, and V. Assimakopoulos, "An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis," *Journal of the*

Operational Research Society, vol. 62, no. 3, pp. 544–554, 3 2011. [Online]. Available: https://www.tandfonline.com/doi/full/10.1057/jors.2010.32

[65] D. Duvenaud, "Expressing Structure with Kernels," *phd-thesis*, 2014. [Online]. Available: https://raw.githubusercontent.com/duvenaud/phd-thesis/master/kernels.pdf

[66] H. Keshavarz, G. Michailidis, and Y. Atchade, "Sequential change-point detection in high-dimensional Gaussian graphical models," *Journal of Machine Learning Research*, vol. 21, 6 2018. [Online]. Available: http://arxiv.org/abs/1806.07870

[67] F. Massa Gray and M. Schmidt, "Thermal building modelling using Gaussian processes," *Energy and Buildings*, vol. 119, pp. 119–128, 5 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778816300494

[68] J.-B. Fiot and F. Dinuzzo, "Electricity Demand Forecasting by Multi-Task Learning," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 544–551, 3 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7467578/

[69] R. Skagestad, "Electricity Demand Forecasting with Gaussian Process Regression," Ph.D. dissertation, Norwegian University of Science and Technology, 2018. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2566721/20123_FULLTEXT.pdf?sequence=1

[70] G. Zotteri, M. Kalchschmidt, and F. Caniato, "The impact of aggregation level on forecasting performance," *International Journal of Production Economics*, vol. 93-94, no. SPEC.ISS., pp. 479–491, 1 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S092552730400266X

[71] A. Sfetsos and C. Siriopoulos, "Time Series Forecasting with a Hybrid Clustering Scheme and Pattern Recognition," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 34, no. 3, pp. 399–405, 5 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1288351/

[72] G. A. Susto, A. Cenedese, and M. Terzi, "Time-Series Classification Methods: Review and Applications to Power Systems Data," in *Big Data Application in Power Systems*. Elsevier, 2018, no. January, pp. 179–220. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128119686000097

[73] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series," *Applied Energy*, vol. 237, pp. 11–24, 3 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261918318816

[74] X. Ruhang, "Efficient clustering for aggregate loads: An unsupervised pretraining based method," *Energy*, vol. 210, p. 118617, 11 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544220317254

[75] P. Mandal, T. Senjyu, N. Urasaki, and T. Funabashi, "A neural network based several-hour-ahead electric load forecasting using similar days approach," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 6, pp. 367–373, 7 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061506000275

[76] C. H. Jin, G. Pok, Y. Lee, H.-W. Park, K. D. Kim, U. Yun, and K. H. Ryu, "A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting," *Energy Conversion and Management*, vol. 90, pp. 84–92, 1 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0196890414009662

[77] E. Atam and E. C. Kerrigan, "Optimal Partitioning of Multithermal Zone Buildings for Decentralized Control," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 3, pp. 1540–1551, 9 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9409744/

[78] 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*, vol. 10, no. 1, pp. 841–851, 1 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8039509/

[79] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 10 2015, pp. 879–887. [Online]. Available: http://ieeexplore.ieee.org/document/7363836/

[80] X. Cao, S. Dong, Z. Wu, and Y. Jing, "A Data-Driven Hybrid Optimization Model for Short-Term Residential Load Forecasting," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. IEEE, 10 2015, pp. 283–287. [Online]. Available: http://ieeexplore.ieee.org/document/7363083/

[81] H. Jahangir, H. Tayarani, S. S. Gougheri, M. A. Golkar, A. Ahmadian, and A. Elkamel, "Deep Learning-Based Forecasting Approach in Smart Grids With Microclustering and Bidirectional LSTM Network," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 9, pp. 8298–8309, 9 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9145791/

[82] P. Laurinec, M. Lóderer, M. Lucká, and V. Rozinajová, "Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption," *Journal of Intelligent Information Systems*, vol. 53, no. 2, pp. 219–239, 10 2019. [Online]. Available: http://link.springer.com/10.1007/s10844-019-00550-3

[83] K. Dab, K. Agbossou, N. Henao, Y. Dubé, S. Kelouwani, and S. S. Hosseini, "A compositional kernel based gaussian process approach to day-ahead residential load forecasting," *Energy and Buildings*, vol. 254, p. 111459, 1 2022. [Online]. Available: https://doi.org/10.1016/j.enbuild.2021.111459https://linkinghub.elsevier.com/retrieve/pii/S037877882100743X

[84] Z. Chen, Y. Chen, T. Xiao, H. Wang, and P. Hou, "A novel short-term load forecasting framework based on time-series clustering and early classification

algorithm," *Energy and Buildings*, vol. 251, p. 111375, 11 2021. [Online]. Available: https://doi.org/10.1016/j.enbuild.2021.111375

[85] Zhang Yun, Zhou Quan, Sun Caixin, Lei Shaolan, Liu Yuming, and Song Yang, "RBF Neural Network and ANFIS-Based Short-Term Load Forecasting Approach in Real-Time Price Environment," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 853–858, 8 2008. [Online]. Available: http://ieeexplore.ieee.org/document/4509471/

[86] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy Forecasting: A Review and Outlook," *IEEE Open Access Journal of Power and Energy*, vol. 7, no. November, pp. 376–388, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9218967/

[87] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 5 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8322199/

[88] G. Rouwhorst, E. M. S. Duque, P. H. Nguyen, and H. Slootweg, "Improving Clustering-Based Forecasting of Aggregated Distribution Transformer Loadings With Gradient Boosting and Feature Selection," *IEEE Access*, vol. 10, pp. 443–455, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9661304/

[89] K. Dab, K. Agbossou, A. Cardenas, Y. Dube, and S. Kelouwani, "Sensitivity Analysis of Exogenous Variables for Load Forecasting Using Polynomial Regression," in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1. IEEE, 10 2019, pp. 2560–2565. [Online]. Available: https://ieeexplore.ieee.org/document/8927167/

[90] S. S. Hamidi, E. Akbari, and H. Motameni, "Consensus clustering algorithm based on the automatic partitioning similarity graph," *Data & Knowledge Engineering*, vol. 124, p. 101754, 11 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169023X18304919

[91] Q. Huang, J. Li, and M. Zhu, "An improved convolutional neural network with load range discretization for probabilistic load forecasting," *Energy*, vol. 203, p. 117902, 7 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544220310094

[92] A. C. Damianou and N. D. Lawrence, "Deep Gaussian Processes," *arXiv:1211.0358v2*, 11 2012. [Online]. Available: http://arxiv.org/abs/1211.0358

[93] E. Snelson and Z. Ghahramani, "Sparse Gaussian Processes using pseudo inputs," *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pp. 1–8, 2005. [Online]. Available: https://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf

[94] A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman, "GPflow: A Gaussian process

library using TensorFlow," *arXiv:1610.08733v1*, 10 2016. [Online]. Available: http://arxiv.org/abs/1610.08733

[95] E. S. Kashani, S. Bagheri Shouraki, Y. Norouzi, and B. De Baets, "A density-grid-based method for clustering k-dimensional data," *Applied Intelligence*, vol. 53, no. 9, pp. 10 559–10 573, 5 2023. [Online]. Available: https://link.springer.com/10.1007/s10489-022-03711-0

[96] A. W. Danté, K. Agbossou, S. Kelouwani, A. Cardenas, and J. Bouchard, "Online modeling and identification of plug-in electric vehicles sharing a residential station," *International Journal of Electrical Power & Energy Systems*, vol. 108, pp. 162–176, 6 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0142061518324426

[97] L. Rueda, S. Sansregret, B. Le Lostec, K. Agbossou, N. Henao, and S. Kelouwani, "A Probabilistic Model to Predict Household Occupancy Profiles for Home Energy Management Applications," *IEEE Access*, vol. 9, pp. 38 187–38 201, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9367182/

[98] P. Singh, S. D. Joshi, R. K. Patney, and K. Saha, "The Fourier decomposition method for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2199, p. 20160871, 3 2017. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspa.2016.0871

Annexe A -

Dans le premier chapitre, nous avons souligné l'importance de la prévision de la charge pour une gestion efficace de la consommation d'énergie dans les résidences. Les défis associés à l'analyse des prévisions de charge, notamment la difficulté à modéliser ou à trouver des similarités entre les profils de charge, ont été identifiés comme des obstacles majeurs dans le développement et l'application de ces systèmes. Pour surmonter ces défis, Le chapitre 2 introduit une méthode basée sur la modélisation des profils stochastiques. Cette approche vise à améliorer les processus de prévision et de classification des profils de charge pour une meilleure compréhension de la consommation agrégée. En outre, la méthode développée et expliquée ici repose sur un modèle de régression non paramétrique utilisant un processus gaussien additif (AGP) pour estimer les influences climatiques et calendaires sur la demande électrique. En complément, l'Annexe présente une analyse de sensibilité axée sur les facteurs exogènes, avec pour objectif d'identifier les habitudes de consommation, ainsi que leur corrélation avec les périodes de consommation résidentielle agrégée. Cette analyse a permis de déterminer les variables les plus significatives, telles que la température, l'humidité, le rayonnement solaire et le temps calendaire, qui sont cruciales pour les prévisions à court terme. L'analyse de sensibilité s'intéresse particulièrement à ces facteurs pour déterminer leur influence sur les prévisions de puissance. Elle repose sur une méthode d'échantillonnage sélectionnant les interactions les plus pertinentes. En parallèle, une technique spécifique a été développée pour répondre aux défis posés par les entrées multivariables et la complexité de grands ensembles de données. Ce modèle de prévision a été appliqué à des données réelles de consommation électrique d'un groupe de maisons au Québec durant l'hiver.

Sensitivity Analysis of Exogenous Variables for Load Forecasting Using Polynomial Regression

Conference Paper · October 2019

DOI: 10.1109/IECON.2019.8927167

5 authors, including:

Khansa Dab
Université du Québec à Trois-Rivières
4 PUBLICATIONS **13** CITATIONS
SEE PROFILE

K. Agbossou
Université du Québec à Trois-Rivières
241 PUBLICATIONS **6,474** CITATIONS
SEE PROFILE

Alben Cardenas
Université du Québec à Trois-Rivières
78 PUBLICATIONS **1,344** CITATIONS
SEE PROFILE

Yves Dubé
Université du Québec à Trois-Rivières
72 PUBLICATIONS **2,435** CITATIONS
SEE PROFILE

Sensitivity Analysis of Exogenous Variables for Load Forecasting using Polynomial Regression

Khansa Dab*, Kodjo Agbossou*, Alben Cardenas*
Yves Dube†, Sousso Kelouwani†
* *Department of Electrical and Computer Engineering*
† *Department of Mechanical Engineering*
Intelligent Energy Research and Innovation Laboratory
Research Hydrogen Institute
University of Quebec at Trois-Rivieres
Trois-Rivieres, Canada
{khansa.dab, kodjo.agbossou, alben.cardenasgonzalez, yves.dube, sousso.kelouwani}@uqtr.ca

***Abstract*—The choice of explicative variables could influence the efficiency of power consumption prediction. Different techniques based on a sensitivity analysis permit to choose the best set of variables among all candidates. Therefore, studies introduce a different set of factors like the temperature to construct a useful prediction system. The goal of this paper is to define the best variable candidate that can efficiently describe power consumption. Accordingly, a model is developed to integrate both meteorological and complementary variables for load forecasting. We use, therefore, powerful tools based on, namely ANOVA, ANCOVA, and Backward Elimination to identify the significant factors that will be inserted in the Principal Component Analysis (PCA) for household and aggregated levels. Mainly, the objective of PCA is to minimize the dimension of the data-set and maximize the information entropy over distinct un-correlated principal extractions. Consequently, the application of polynomial regression to principal components for load forecasting conducts to better results providing a useful forecast by capturing the best explanatory variables.**

***Keywords*—Sensitivity analysis; polynomial regression; forecast aggregated power;**

I. Introduction

In Quebec, Canada, residential electricity consumption is highly influenced by environmental factor [1]. Analyzing these factors, is an essential prerequisite for developing an efficient power consumption prediction framework, required by energy management systems. In fact, climatic variables have different impacts on electricity consumption due to varying geographical conditions. Then, the critical issue of a logical analysis should be determining the significant meteorological factors that mostly influence power consumption [2].

Besides, the performance of the prediction model can be improved by a combination of several factors. Although this combination can enhance the prediction accuracy, it can notably increase the complexity of forecasting due to the non-stationary nature of meteorological variables. The fact is that the evaluation of computational time [3] of these variables requires a large amount of data. Therefore, identifying an effective combination of significant factors is another important issue.

There are several methods that have been proposed in the literature to address the above issues. Those methods have mainly used times series analysis, in the context of a nonlinear and non-stationary problem [4]. Dahl et al. have exploited selected features corresponding to the different days of the week [5]. Lusis et al. have considered holidays and calendar effects as crucial variables for load forecasting [2]. Beccali et al. have intended to improve forecasting by using humidity for cooling and heating demands [6]. Soliman et al. have employed the direction and speed of the wind to predict the load by also including the air temperature [7]. On the other hand, other authors analyze the forecasting of residential power consumption using only the temperature. It is considered in many studies the most variable that influences the electricity demand [8].

The other point is that electricity consumption of single residence is less predictable than an aggregated load [2]. This difference is due to the geographical place of houses and their exposition to the weather not only this but also the calendar schedule of the occupants that makes a difference in power consumption. The inconvenience of the methods, including some of the factors, is the lack of concurrently analysis of the role of both meteorological information and calendar variables. Indeed, calendar variables such as working days, weekends, and holidays, can noticeably change power consumption pattern [9], [10].

In this paper, we propose a combined method to select the most efficient set of variables for residential power consumption prediction. Different statistical techniques are applied to a set of meteorological and calendar variables to identify the significant ones. These techniques take advantage of ANOVA, ANCOVA, and Backward Elimination methods. The significant correlated factors resulted from sensitivity analysis techniques are used as input on the Principal Component Analysis (PCA). Subsequently, PCA is employed to manage the complexity of the chosen variables by reducing the dimension of the dataset [11]. Through a correlation analysis of the significant variables PCA results in a set of uncorrelated components [12], [13]. Afterward, in the context of a short term forecasting, a polynomial regression method is applied to these uncorrelated components to predict power consumption. We provide an

analysis that elaborates important remarks to improve the choice of variables and thus, the performance of prediction. We organize this paper in different sections. We introduce our methodology in section II. Section III, presents sensitivity analysis methods starting with ANOVA, ANCOVA, Backward Elimination, PCA, polynomial regression. Finally, section IV shows our main results and a brief conclusion.

II. Methodology

We introduce a technique based on sensitivity analysis for the data by taking its advantage to forecast the consumption of the electricity. The sensitivity analysis of the factors (exogenous variables) used for load forecast is done; the reduction of the dimension of variables data set in order to guarantee this accuracy is provided. Then the new dataset is inserted into the model described in Fig.1 to forecast the residential electricity consumption using a polynomial regression model. The methodology consists of three stages:

1) Acquisition and statistical analysis of data source among electricity demand, climatic conditions, and calendar ones.
2) Feature selection with ANOVA/ANCOVA test and Backward Elimination algorithm. Then dimension reduction with PCA of best candidates uncorrelated variables.
3) Develop a forecast model, so we reinforce the use of forecast considering the exogenous climatic variables and describe the dependent variables for one house individually and then for a group of houses.

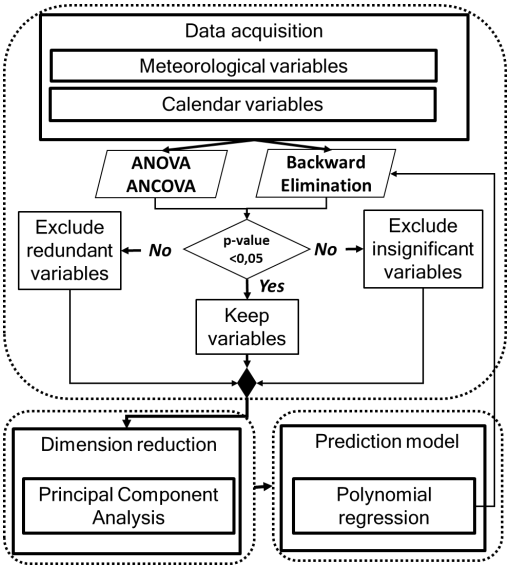

Fig. 1. A conceptual diagram for the methodology

III. Sensitivity analysis methods

A. Description of Data

The collected dataset contains the residential power consumption, meteorological variables such as temperature, humidity, wind speed, atmospheric pressure, irradiance which is the summation of vertical and horizontal solar radiation. In the overall mathematical notation of this paper, the inputs are denoted by a multidimensional variable $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Where $\mathbf{x}_t$ is a realization of X at the discrete time t, N denotes the length of the sequence. Furthermore, a point $\mathbf{x}_t = [x_{1,t}, x_{2,t}, \ldots, x_{n,t}]$ is a vector of dimension n for which each element $x_{i,t}$ corresponds to the reading of the i^{th} explanatory variable at time t. Moreover, the power profile is an unidimensional variable represented as $Y = \{y_1, y_2, \ldots, y_N\}$. The data set contain also the calendar variables. They are either categorical in increasing order case of day of the week and hour of day or binary case of holidays. To include holidays as numeric variable, we construct a binary time series x_t, for holidays $x_t = 1$ and otherwise $x_t = 0$. Besides, to consider the time of the day effects, we introduce periodic regressors defined by trigonometric time series alternating in a daily period. The same approach is done for the day of the week in those references [14]–[16]. By including periodic signals, the model can fit repetitive daily patterns in household power consumption. Those factors explain the sensible part of space heating in total power. Trigonometric explanatory variables are defined as follows:

$$\gamma_{i,t} = \sin(\omega_i t) \quad (1)$$

$$\phi_{i,t} = \cos(\omega_i t) \quad (2)$$

Where $\omega_i = \frac{2\pi}{T}$, $T = 24$ is the period for the time defined in hours. $T = 7$ is the period for the time in days unit. Because the time of the day affects power consumption, it is considered in the input set. For an individual and total houses, we consider a box that is composed of all the exogenous variables as input and the power consumption as output starting with the temperature which is between -31°C and 34°C here in Canada. The mean of the temperature is 10°C. The humidity is in the range of 42%, 92%. The mean for both solar radiation(vertical and horizontal) is 9.41 and 20.14 (Wh/mA). The mean and the max for the atmospheric pressure is 911 and 1043 (hPa) respectively. Some preprocessing statistical test was done to clean the data from outliers and missing values. The ANOVA and ANCOVA test analysis later are used to exclude the redundant explanatory variables. Backward Elimination is used to remove insignificant variables. The best candidates variables are employed to forecast electricity consumption as a new input on the model based on PCA.

B. ANOVA and ANCOVA

ANOVA is a statistical analysis that is used to determine the differences in the averages of a set of data when grouped under several certain categorical factors. It compares the influence of all these factors over power consumption by detecting the interaction between them. It is applied to define the mean difference based on a continuous independent response, which is the power by the analysis of variance of exogenous variables. The originals variables taking into consideration are 13 in our case of study, as shown in Table I. ANCOVA (Multichannel variance analysis) [17] is used for the continuous and categorical variables, which are the holidays, the working days, and the weekend. ANCOVA test is a combination of the ANOVA test and regression analysis. ANCOVA helps to test if there is a factor that affects the dependent variable in our case, the power consumption after eliminating the variance due to the covariates.

C. Backward Elimination

The sensitivity analysis is a measure of the significance of variables in explaining the output, which is the power in our case. The goal here is to build a polynomial regression model that includes few variables without compromising the ability to predict power consumption. It is essential to select the best variable that gives less prediction error. It is done by the Backward Elimination that calculates the significance of each exogenous input variable associated with time [18]. To detect the contribution of each variable in the model, we start by all the variables, and we test the effect of eliminating each one of them using the criteria of adjustment which is, in this case, the polynomial regression [19]. At degree 3, the polynomial regression achieves the minimum of error. There is no need to increase more the degree of the polynomial because it captures the optimal results which provide the variable that will be taken if its miss gives deterioration of statistic significance. The process is repeated until all variables get eliminated one after another without losing the relevancy of the system. In the first step, all the input variables are introduced. In this step, the p-value, which its threshold is set as 0.05, is used to determinate the significance of variables. In the last step of the algorithm, only the most significant variable in the model is kept. Results will be presented in the discussion part.

D. Principal Component Analysis (PCA)

PCA is a dimension reduction procedure used for orthogonal transformation when there are some correlated variables into linearly uncorrelated variables called the principal components [10]. The PCA could be viewed in different ways. It performs a nonlinear dataset. In this case, the calendar factors are presented as categorical and binary variables. To get good results after applying PCA to the combination of all those variables, we need to verify the relationship between continuous and categoricals variables [20]. The main technique used in this multivariate statics is the Pearson correlation. It measures the direction of a linear relationship between two variables. It is done by the determination of the covariance and the square root of the sample variance. We apply the polynomial regression, and we calculate the coefficient of determination R^2 that presents the linear fit between input and output. After that, the correlation between each of them is determined. PCA can minimize the correlation between variables and maximize the information entropy over distinct principal extractions. The new components are deduced from the explanatory variables according to the next linear model:

$$z_{j,t} = \mathbf{q}_j \mathbf{x}_t \tag{3}$$

Where $\mathbf{q}_j = [q_1, q_2, \ldots, q_n]$ is a row vector of PCA coefficients and $\mathbf{x}_t = [x_{1,t}, x_{2,t}, \ldots, x_{n,t}]^\intercal$ is a column vector for which each element correspond to the value of each exogenous variable at time t.

E. Polynomial Regression

The polynomial regression is the method used in all this paper for all the algorithms used in our analysis. It is a nonlinear relationship between the input variables and the mean of the output, which is the dependent variable (power) in the houses. Polynomial regression allows finding an adequate prediction function for data that does not seem to have a linear relationship. Our goal is to find a prediction function which will model this correlation between the output and input. We will apply all the next methods with polynomial regression to find out how to extend the effect of each variable in the model and how the effect may vary for a house or aggregated profile of a house. We consider a time step of 15 minutes for all the data. The polynomial model is constructed by the next equation:

$$y_t = \alpha + \mathbf{w}_1 \tilde{\mathbf{x}}_{1,t} + \mathbf{w}_2 \tilde{\mathbf{x}}_{2,t} + \cdots + \mathbf{w}_n \tilde{\mathbf{x}}_{n,t} \tag{4}$$

Where α is the power offset, and where $\tilde{\mathbf{x}}_{i,t} = [(x_{i,t})^1, (x_{i,t})^2, \ldots, (x_{i,t})^p]^\intercal$ are column vectors where each element corresponds to a power evaluation of the explanatory variable $x_{i,t}$ from the order one until an arbitrary integer defined by p. Furthermore, $\mathbf{w}_i = [w_{1,i}, w_{2,i} \ldots, w_{p,i}]$ is a row vector where each element is a coefficient multiplying each component of the polynomial expansion of each explanatory variable.

F. Performance Metrics

The first metric we discuss is called the coefficient of determination R^2 [21]. R^2 explains how well the polynomial regression explains the variance of the response variable, which is the power consumption in the houses. The R^2 is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{5}$$

The SSR is the sum of the squares due to regression, and the SST is the sum of squares total, and SSE is the sum of squares due to error are respectively:

$$SSR = \sum_{t=1}^{N} (\hat{y}_t - \bar{y})^2 \tag{6}$$

$$SST = \sum_{t=1}^{N} (y_t - \bar{y})^2 \tag{7}$$

$$SSE = \sum_{t=1}^{N} (\hat{y}_t - y_t)^2 \tag{8}$$

Where $\hat{y}_t$ the predicted value at time t, y_t is the actual value of the predicted variable at time t and where $\bar{y}$ is the mean value of the predicted variable:

$$\bar{y} = \frac{1}{N} \sum_{t=1}^{N} y_t \tag{9}$$

The other metric used is the cross-validation CV; it is a simple realization method that provides a direct estimation of the predictor error. We split our data into training data that estimate the parameter of the model, which is 70% of the whole data, and the predictor error is evaluated with 30% of the validation data. The other metric used is the mean absolute error (MAE):

$$MAE = \frac{1}{N} \sum_{t=1}^{N} |\hat{y}_t - y_t| \tag{10}$$

IV. Results

This section presents the results of the analysis. Initially, Fig.2 summarizes Pearson correlation between predicted and exogenous variables. As illustrated, some of the inputs are strongly correlated with each other and with output. For instance, temperature and relative humidity (RH) are correlated at 86%, and 70% with the power, respectively. However, the fact of including both exogenous variables does not improve the accuracy of the prediction significantly. To illustrate this behavior, we performed simulations by adding variables sequentially according to the absolute value of the Pearson correlation between the studied input and the power (lower row in the heatmap of Fig.2). Accordingly, the first time the polynomial model is performed with just one variable having the highest negative correlation value with the power for this case is the temperature with a value equal to -0.8. However, the power with the day of the week (t4) has a positive correlation equal to 0.4. Subsequently, we increased the dimension of the input by adding more significant variables. The result of this procedure is shown in Fig.3. In this image, bars represent the coefficient of determination R^2, and the red line shows the R^2 improvement after each new variable is considered. For this analysis, a polynomial model of order $p = 3$ was used.

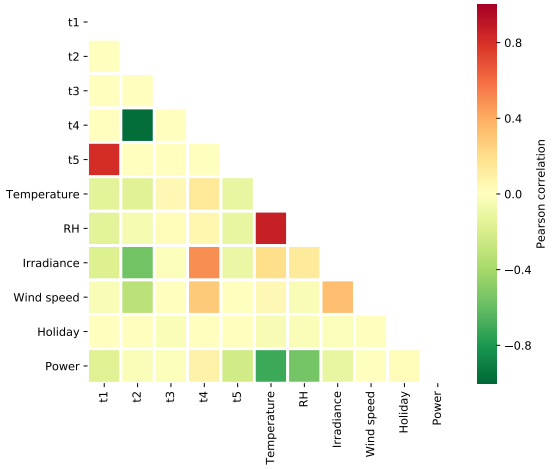

Fig. 2. Pearson correlation between different studied variables

A. ANOVA and ANCOVA Results

By applying the p-value which explains the level of significance within a statistical test representing the probability of the occurrence of a given factor. The p-value is used as an alternative to providing the smallest level of significance at which the null hypothesis would be rejected. The results of ANOVA and ANCOVA tests demonstrate that temperature, relative humidity, irradiance, wind speed, holidays, the day of the week, the hour of the day are sufficient to be predictor variables. However, the null hypothesis is rejected for pressure and direction of the wind. The other variables can be accepted with a smaller p-value $\leq$0.05. They produce a significant and positive impact on power, as shown in Table I. Thus the variable of atmospheric pressure in the station has been eliminated from the model to obtain a model that fits better the regression model also it does not have effects due to

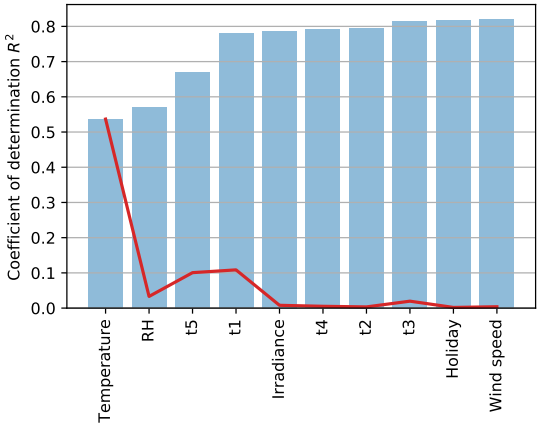

Fig. 3. Sequential model improvement and coefficient of determinationR^2

TABLE I. Analyse of ANOVA/ANCOVA

Source	MS	Contribution%	f_value	p_value
Temperature (°C)	5.96	35.19	9.615	0.002
Humidity (%)	6.08	21.52	6.210	0.003
HOR.Solar radiation (Wh/mA°^{2})	6.12	8.86	7.596	0.016
VER.Solar radiation (Wh/mA°^{2})	5.09	3.23	13.476	0.013
Pressure (hPa)	5.96	1.00	96.10	0.070
Speed.of wind (m/s)	5.96	12.20	7.359	0.007
Direction.of wind (A°)	5.00	1.22	4.225	0.060
Holidays	24.08	9.82	8.432	0.014
Hour of day (t1)	9.53	2.14	4.921	0.027
Hour of day (t2)	3.30	0.11	0.993	0.004
Hour of day (t3)	4.20	2.12	17.03	0.010
Day of week (t4)	3.29	1.14	0.462	0.013
Day of week (t5)	5.30	1.45	4.329	0.041

(MS)Mean Square

the p-value $>$ 0.05, the same for the direction of the wind. The temperature (p-value $\leq$ 0.05) and the irradiance (p-value $\leq$ 0.05) produce a significant and negative effect on the dependent variable. Between both of these variables, the temperature has a more substantial impact compared with other variables 35.19%. The important two other variables which have signification are the speed of the wind and the humidity with a value equal to 21.52% and 12.20% respectively. The results of the ANOVA test are that we eliminate the pressure and the direction of the wind as non-significant variables. In the model with ANCOVA test and while grouping them by a calendar factor, we eliminate the variables hour of the day (t2). The categorical variables holidays (p-value =0.014) and hour of the day (t1)(p-value =0.027) have a significant positive effect on the power consumption. Between the two, holidays have a bigger impact on power, as shown by their F-value (8.432 vs. 4.921). The efficient variables candidates will be the temperature, the humidity, the wind, the holidays, for the hour of the day (t1), (t2) and, the day of the week (t5). The introduction of the time of the day and the day of the week as input increase the performance of the prediction. This is true because the behavior of the occupants depends on the time being in their houses and depending on if it is a working day or it is a weekend.

B. Backward Elimination Results

The significance of the variable, targeted by Backward Elimination, is evaluated through a polynomial regression technique. As shown in Fig.4, the results are provided for

a single house. As for entire homes, the meteorological and calendar variables based on the third degree of polynomial regression are utilized. The most significant variables are temperature, humidity, and time. The result of Backward Elim-

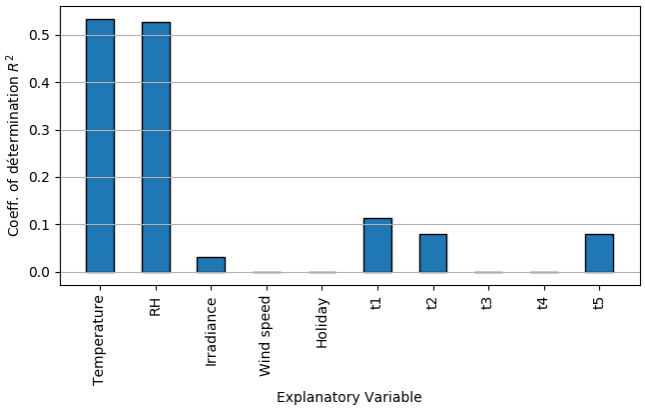

Fig. 4. Coefficient of Determination, Backward Elimination case of single house

ination is presented by using the coefficient of determination R^2. Accordingly, R^2 for the individual house is 62%, the total houses is 82%, as shown in Fig.5 and Fig.6. Also, the MAE values for the individual house and total houses are 1.02 kW and 3.02 kW, corresponding to a peak value equal to 15 and 40 kW, respectively, as shown in Fig.7 and Fig.8. The optimal model parameter in the polynomial order has been selected by using the cross-validation technique. It has the best averaged predictive performance. The cross-validation divides the data into training and validation sets to evaluate the model. The metrics, used for both are the MAE and R^2. In Fig.6, the optimal degree of polynomial regression occurs in 3. Afterward, the decrease in R^2 is due to system sensibility to the increases of parameters that result in over-fitting at the polynomial degree of 4 [22]. As shown in Fig.8, over-fitting also increases the MAE value. The estimation of error after removing a variable by Backward Elimination demonstrates the influence of that factor. In our analysis, the temperature and humidity are identified as the most significant factors, since their removal notably increases the error. We reach a similar percentage of R^2 for the individual houses and the total houses. Irradiance, time of the day (t1), (t2), and day of the week (t5) are other variables that influence the consumption but with less percentage. Considering the Backward Elimination results, holidays, and speed of wind are excluded from the analysis since they have almost no impact on power consumption. Nevertheless, to improve the prediction accuracy, the rest of the variables (t3) and (t4), even with lower impacts are considered. However, these results increase the dimension of 10 input variables. Therefore, a PCA method is utilized to reduce the dimension of the original dataset. In fact, in short-term forecasting by using meteorological variables, it is advantageous to select the key variables that have a stronger influence on power consumption. In our case, three uncorrelated components resulted from PCA analysis are utilized for prediction. In fact, these components can explain $80\,\%$ of the cumulative variance of all input variables. Subsequently, the polynomial regression is applied to three principal components to predict power consumption. As shown in Fig.9, 10 the predicted power efficiently fit the actual one.

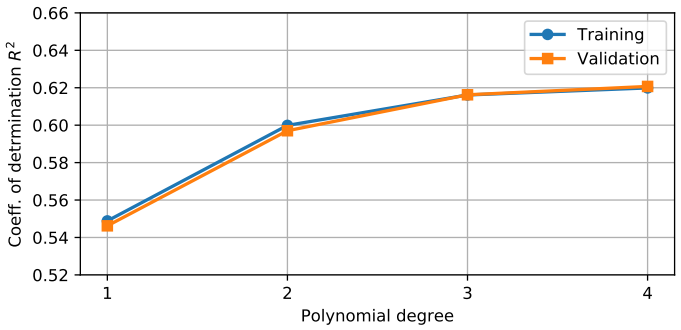

Fig. 5. R^2 case of single house

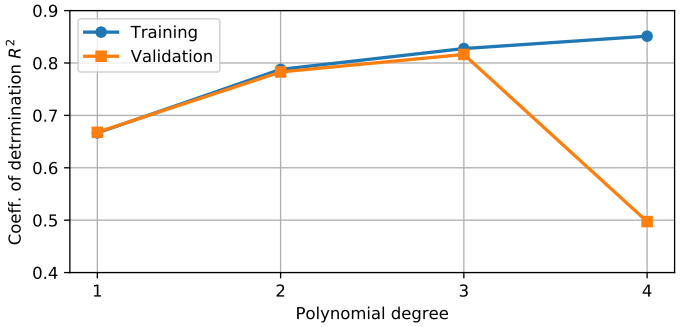

Fig. 6. R^2 case of total houses

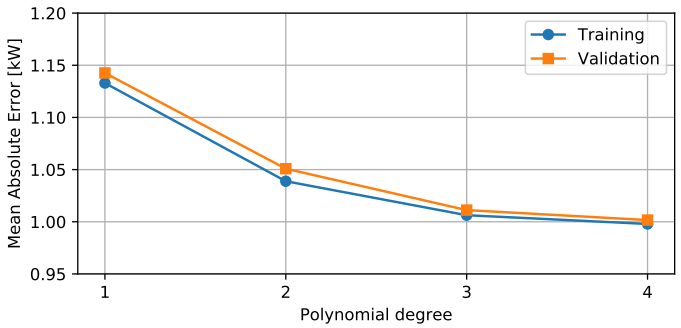

Fig. 7. MAE case of single house

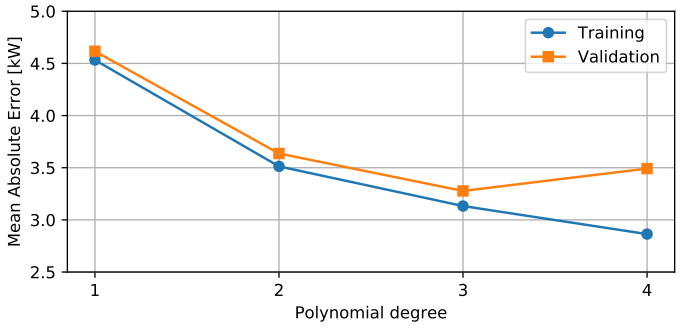

Fig. 8. MAE case of total houses

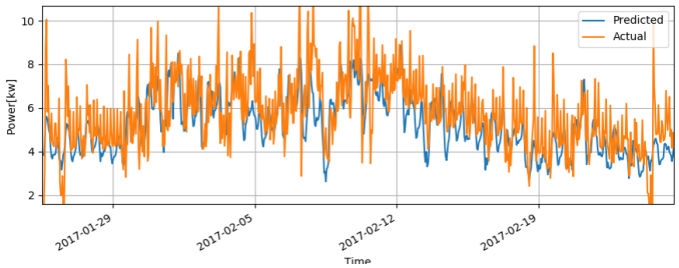

Fig. 9. Actual power consumption vs predicted using polynomial regression: case of single house

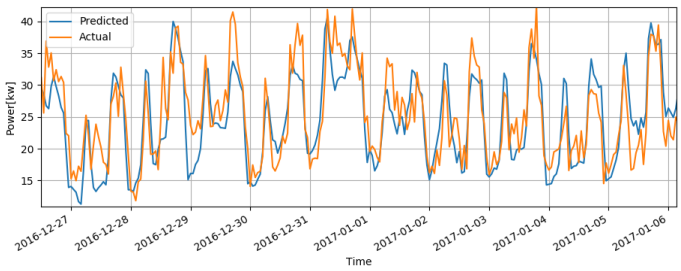

Fig. 10. Actual power consumption vs predicted using polynomial regression: case of total houses

V. Conclusion

In this paper, a polynomial regression model for predicting the electricity demand is used based on sensitivity analysis. The techniques applied for feature selection of best candidate are ANOVA, ANCOVA, cross-validation, analysis of variance and Backward Elimination processes. ANOVA eliminates some of the variables with low p-value, e.g., the pressure and direction of the wind. The Backward Elimination of polynomial regression eliminates the day of the week and hour of the day. Applying PCA to the polynomial regression model permits to reduce the dimension of the data and achieve uncorrelated component. The considerable variables are used into the regression model to predict the power consumption in individual and expanded residential network. Based on the obtained results, the significant variables have been employed for the modeling of power consumption by a polynomial regression model with lower complexity. As a perspective, it is essential to apply this analysis to an adaptive process using the non-parametric model by providing the dynamic component considering the PCA to reduce the complexity in the load forecasting.

Acknowledgment

This work was supported in part by the Laboratoire des technologies de l'énergie (LTE) d'Hydro-Quebec, the Natural Science and Engineering Research Council of Canada and the Fondation of UQTR.

References

[1] "Clean energy to power us all. Hydro-Quebec. Available at: http://www.hydroquebec.com/data/documents-donnees/pdf/annual-report.pdf," 2018.

[2] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Applied Energy*, vol. 205, no. November, pp. 654–669, 2017.

[3] S. Girard, T. Romary, J.-M. Favennec, P. Stabat, and H. Wackernagel, "Sensitivity analysis and dimension reduction of a steam generator model for clogging diagnosis," *Reliability Engineering & System Safety*, vol. 113, pp. 143–153, 2012.

[4] M. Cai, M. Pipattanasomporn, and S. Rahman, "Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques," *Applied Energy*, vol. 236, no. November 2018, pp. 1078–1088, 2019.

[5] M. Dahl, A. Brun, O. S. Kirsebom, and G. B. Andresen, "Improving short-term heat load forecasts with calendar and holiday data," *Energies*, vol. 11, no. 7, pp. 1–16, 2018.

[6] M. Beccali, M. Cellura, V. Lo Brano, and A. Marvuglia, "Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area," *Renewable and Sustainable Energy Reviews*, vol. 12, no. 8, pp. 2040–2065, 2008.

[7] S. A. Soliman and A. M. Al-kandari, *Electrical Load Forecasting. Modeling and Model Construction.* Springer Berlin Heidelberg, 2010.

[8] F. Amara, K. Agbossou, Y. Dube, S. Kelouwani, and A. Cardenas, "Estimation of temperature correlation with household electricity demand for forecasting application," *IECON Proceedings (Industrial Electronics Conference)*, pp. 3960–3965, 2016.

[9] J. Moral-Carcedo and J. Pérez-García, "Time of day effects of temperature and daylight on short term electricity load," *Energy*, vol. 174, pp. 169–183, 2019.

[10] F. Ziel, "Modeling public holidays in load forecasting: a German case study," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 191–207, 2018.

[11] J. Zhang, X.-y. Yang, F. Shen, Y.-w. Li, H. Xiao, H. Qi, H. Peng, and S.-h. Deng, "Principal Component Analysis of Electricity Consumption Factors in China," *Energy Procedia*, vol. 16, pp. 1913–1918, jan 2012.

[12] T. H. Dang-Ha, F. M. Bianchi, and R. Olsson, "Local short term electricity load forecasting: Automatic approaches," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2017-May, pp. 4267–4274, 2017.

[13] J. Shlens, "A tutorial on principal component analysis: Derivation, Discussion and Singular Value Decomposition," Tech. Rep., 2003.

[14] R. Weron, "Modeling and ForecastingElectricity Loads and Prices," pp. 1–195, 2013.

[15] W. H. B, "Weather and Harvest Cycles," *The Economic Journal/Wiley on behalf of the Royal Economic Society*, vol. 31, no. 124, p. 429, 2006.

[16] J. Moral-Carcedo and J. Pérez-García, "Integrating long-term economic scenarios into peak load forecasting: An application to Spain," *Energy*, vol. 140, pp. 682–695, 2017.

[17] A. E. Permanasari, D. R. A. Rambli, and P. D. D. Dominic, "Forecasting method selection using ANOVA and Duncan multiple range tests on time series dataset," *Proceedings 2010 International Symposium on Information Technology - Engineering Technology, ITSim'10*, vol. 2, no. May 2015, pp. 941–945, 2010.

[18] R. Chauhan and H. Kaur, *Predictive Analytics and Data Mining*, 2nd ed. Amsterdam • Boston • Heidelberg • London New York • Oxford • Paris • San Diego San Francisco • Singapore • Sydney • Tokyo Morgan Kaufmann is an imprint of Elsevier: Elsevier, 2015.

[19] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables," *Applied Energy*, vol. 140, pp. 385–394, 2015.

[20] J. A. L. M. Verleysen and Nonlinear, *Nonlinear Dimensionality Reduction(Information Science and Statistics)*, 2nd ed. USA: Springer Science + Business Media, LLC, New York, NY , USA, 2012.

[21] F. Emmert-Streib and M. Dehmer, "Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 521–551, 2019.

[22] G. B. D. Costa, A. Z. Bertoletti, A. P. D. Morais, and G. C. Junior, "Curve Fitting Analysis of Expulsion Fuse Links through the Cross-Validation Technique," *Proceedings of the 2018 IEEE PES Transmission and Distribution Conference and Exhibition - Latin America, T and D-LA 2018*, pp. 1–5, 2018.