

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE DE L'IMPACT DES VARIABLES MÉTÉOROLOGIQUES SUR LA
PRÉVISION DE LA DEMANDE ÉNERGÉTIQUE AU QUÉBEC

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR

CHARLIE HÉBERT-PINARD

NOVEMBRE 2023

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.04-2020). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je tiens à exprimer ma sincère gratitude envers toutes les personnes qui ont contribué à la réalisation de ce mémoire. Tout d'abord, je souhaite remercier mon directeur, Alexandre Blondin Massé, ainsi que mon codirecteur Alejandro Di Luca, pour leurs conseils éclairés, leur expertise et leur soutien constant qui ont été d'une valeur inestimable tout au long de cette étude.

Mes remerciements vont également à Stéphane Dellacherie, collaborateur et ingénieur dans l'unité de Prévisions des apports et de la demande à Hydro-Québec, pour ses précieux conseils et sa disponibilité tout au long de ce projet. Ses apports et ses discussions enrichissantes ont grandement contribué à l'aboutissement de ce travail.

J'en profite pour souligner la contribution de Arnaud Zinflou, chercheur à l'Institut de recherche d'Hydro-Québec (IREQ), Rachel Bazile et Fabian Tito Arandia Martinez, membres de l'équipe de Prévisions des apports et de la demande à Hydro-Québec lors du transfert de connaissance sur l'acquisition de données et le fonctionnement des modèles de prévisions utilisés à l'interne.

Finalement, je souhaite remercier le programme Accélération de Mitacs ainsi que le programme de bourses d'études supérieures du Canada au niveau de la maîtrise pour les bourses qui m'ont été attribuées. Cette aide financière a été essentielle pour la poursuite de mes études de maîtrise.

TABLE DES MATIÈRES

LISTE DES TABLEAUX	v
LISTE DES FIGURES	vi
LISTE DES ACRONYMES ET VARIABLES	ix
RÉSUMÉ	xi
INTRODUCTION	1
ÉTAT DE L'ART	4
CHAPITRE I ANALYSE DESCRIPTIVE DE LA CONSOMMATION D'ÉLEC- TRICITÉ DANS LA PROVINCE DU QUÉBEC	9
1.1 Consommation énergétique	9
1.1.1 Tendance	10
1.1.2 Périodicité	12
1.1.3 Résiduel	16
1.2 Variables météorologiques	17
1.2.1 Température	20
1.2.2 Vitesse du vent	22
1.2.3 Humidité relative	24
1.2.4 Couverture nuageuse	26
1.2.5 Rayonnement solaire	28
CHAPITRE II ÉVALUATION DE L'EFFET DES INTRANTS MÉTÉO- ROLOGIQUES SUR LES MODÈLES DE PRÉVISION	30
2.1 Modèle statistique linéaire	30
2.1.1 Description des expériences	34
2.1.2 Analyse des expériences	37
2.2 Modèle neuronal	43

2.2.1	Description des expériences	46
2.2.2	Analyse des expériences	48
CHAPITRE III ÉVALUATION DE LA RÉPARTITION SPATIALE DES INTRANTS		62
3.1	Exploration des stations	64
3.2	Performances du modèle neuronal	68
3.3	Pondération des stations	71
CHAPITRE IV COMBINAISON DES STATIONS ET DES VARIABLES MÉTÉOROLOGIQUES		74
4.1	Évaluation du modèle de prévision résultant	74
4.2	Comparaison de l'architecture	79
CONCLUSION		82
RÉFÉRENCES		85

LISTE DES TABLEAUX

Tableau		Page
1.1	Valeurs critiques pour les niveaux de confiance du test de Dickey-Fuller augmenté.	12
2.1	Résultats du score MAPE journalier des expériences sur des données vues 2013-01-01 à 2017-01-01.	35
2.2	Résultats du score MAPE journalier des expériences sur des nouvelles données 2017-01-01 à 2019-12-31.	35
2.3	Scores MAPE journaliers moyens sur les données de test, de validation et d'entraînement pour les expériences réalisées avec le réseau de neurones.	47
3.1	Scores MAPE journaliers des simulations sur les ensembles de test, validation et entraînement en fonction du sous-ensemble de température passé en intrant.	69
3.2	Score MAPE des simulations sur les ensembles de test, validation et entraînement en fonction du sous-ensemble de températures passé en intrant.	71
3.3	Coefficients de pondérations calculés en fonction de la population et de la valeur SHAP pour les 5 régions initiales.	72
3.4	Scores MAPE journaliers des simulations sur les ensembles de test, validation et entraînement pour les températures globales calculées.	72
4.1	Statistiques de précision du modèle de prévision sur les données de 2012 à 2019 avec séparation 60% entraînement, 20% validation et 20% test.	76
4.2	Architecture et utilisation des ressources des modèles de prévisions.	80

LISTE DES FIGURES

Figure	Page
1.1 Évaluation de la tendance du BQ	11
1.2 Spectre de l'amplitude absolue des fréquences de la série temporelle du BQ	13
1.3 Autocorrélation de la série du BQ pour 730 et 30 jours de décalage.	15
1.4 Autocorrélation partielle de la série de BQ pour les décalages d'au plus 200 heures.	16
1.5 Signal résiduel de la série temporelle du BQ.	17
1.6 Matrice de corrélation entre les variables météorologiques et le BQ à Montréal pour les données horaires de 2012 à 2019	19
1.7 BQ en fonction de la température à Montréal sur la période com- plète de 2012 à 2019 avec estimation polynomiale des moindres carré en rouge.	21
1.8 Corrélation croisée de la température avec le BQ calculée avec l'en- semble de la période 2012-2019	22
1.9 Graphique de densité du BQ en fonction de la vitesse du vent . . .	23
1.10 BQ en fonction de la nébulosité à Montréal	27
1.11 BQ sans la variation due à la température en fonction de la nébu- losité pour la nuit et pour le jour	27
1.12 BQ sans la variation due à la température en fonction de la radia- tion solaire incidente pour les journées avec $T > 0^{\circ}\text{C}$ et les journées avec $T < 0^{\circ}\text{C}$	28
2.1 Score MAPE mensuel moyen pour chacune des expériences sur la période non-vue de 2017 à 2019.	36
2.2 Score MAPE horaire moyen pour chacune des expériences sur la période non-vue de 2017 à 2019.	36

2.3	Coefficients calculés pour la vitesse du vent.	37
2.4	BQ en fonction de la température réelle à Montréal (gauche) et de la température du point de rosée (droite)	39
2.5	Valeurs de charge observées et prévues par le modèle de contrôle, avec la température de rosée pour les journées du 1er au 3 juillet 2018	39
2.6	Coefficients calculés pour les trois paliers de nébulosité.	41
2.7	Coefficients calculés pour la radiation normalisés en fonction de la radiation solaire horaire moyenne.	42
2.8	Architecture du modèle neuronal, avec n intrants (vert), m neurones par couche cachée (bleu) et k neurone en sortie (rouge). . .	44
2.10	Score MAPE horaire moyen pour chacune des expériences sur les données de test de 2012 à 2019.	48
2.9	Score MAPE mensuel moyen pour chacune des expériences sur les données de test de 2012 à 2019.	48
2.11	Consommation énergétique réelle et prévue et conditions météorologiques pour les journées du 18 février 2019 et 8 janvier 2014 . .	51
2.12	Consommation énergétique réelle et prévue avec et sans inertie du vent et conditions météorologiques pour les journées du 7 février 2018 et 10 octobre 2016	53
2.13	Moyenne de la valeur absolue des valeurs de Shapley pour les entrées de température et de vitesse du vent pour les journées du 7 février 2018 et 10 octobre 2016.	54
2.14	Moyenne de la valeur absolue des valeurs de Shapley pour les entrées de température et de température de rosée pour la journée du 1 juillet 2018	55
2.15	Contribution des intrants de température et d'humidité sur la prévision de 400 échantillons distribués aléatoirement sur l'ensemble de données	57
2.16	Moyenne de la valeur absolue des valeurs de Shapley pour les variables de température (Tyul) et de nébulosité (Nyul) au temps de prévision et au pas de temps d'inertie.	58

2.17	Moyenne de la valeur absolue des valeurs de Shapley pour les variables de température (Tyul) et de radiation (Rad) au temps de prévision et au pas de temps d'inertie.	59
2.18	Score MAPE mensuel moyen des différentes combinaisons sur les données de test de 2012 à 2019	61
3.1	Matrice de corrélation du BQ et de la température aux 5 stations initiales	63
3.2	Positionnement géographique des 15 stations météorologiques sélectionnées pour l'étude combinatoire.	64
3.3	Score MAPE moyen pour la période de janvier 2015 et janvier 2017 obtenu avec chaque station météorologique individuelle.	66
3.4	Score MAPE de chaque station météorologique moyenné pour toutes les combinaisons testées pour la période de janvier 2015 et janvier 2017.	66
3.5	Ratio d'occurrence de chaque station parmi les combinaisons ayant obtenues un score MAPE inférieur à 1.95.	67
3.6	Valeurs de Shapley moyennes obtenues pour la simulation avec les 15 températures de l'ensemble.	70
4.1	Scores MAPE journalier et mensuel moyens sur les données de test de 2012 à 2019. Occurrence des écarts de prévision supérieurs à 1000 MW pour chaque mois.	77
4.2	Moyenne de la valeur absolue des valeurs de Shapley sur 400 échantillons pour les intrants météorologiques au temps de prévision et pour la moyenne des 24 pas de temps d'inertie.	78

LISTE DES ACRONYMES ET VARIABLES

Variables

BQ	Besoin Québécois
HR	Humidité relative
Neb, N	Nébulosité
Rad	Radiation solaire incidente
T	Température
Td, Tdew	Température du point de rosée
Tress	Température ressentie
Vent, V	Vitesse du vent
_yul	de Montréal
_ow	de Ottawa
_qc	de Québec
_sh	de Sherbrooke
_bc	de Baie-Comeau
_shaw	de Shawinigan
_rim	de Rimouski
_beau	de Beauce
_sag	de Saguenay
_gas	de Gaspé
_cha	de Charlevoix
_tr	de Trois-Rivières
_drum	de Drummondville
_mtrem	de Mont-Tremblant
_chap	de Chapais

Acronymes

ARMA	Autorégressif avec moyenne mobile
CDD	Degrés de climatisation
HDD	Degrés de chauffage
MAPE	Erreur moyenne absolue en pourcentage
MSE	Erreur quadratique moyenne
RNA	Réseau de neurone artificiel
SHAP	SHapley Additive exPlanation

RÉSUMÉ

À Hydro-Québec, la prévision précise de la demande énergétique est essentielle pour la gestion des réseaux électriques, car une surestimation peut entraîner des pertes économiques et écologiques, tandis qu'une sous-estimation peut conduire à des pannes majeures. Au Québec, la situation est particulièrement influencée par la météo en raison principalement de l'utilisation de systèmes de chauffage et de climatisation électriques qui sont soumis à des conditions variables et sur un large étendue géographique.

Le but de ce mémoire est d'améliorer la précision des prévisions à court terme de la demande énergétique au Québec en considérant l'impact de plusieurs variables météorologiques. Le mémoire présente deux contributions. Dans un premier temps, le comportement d'un modèle statistique linéaire et d'un réseau de neurones artificiel est analysé en présence des différentes entrées. Les variables météorologiques considérées sont la température, l'humidité, la vitesse du vent, la nébulosité et la radiation solaire incidente.

Dans un deuxième temps, on généralise les travaux à plusieurs stations. Une analyse combinatoire sur 15 stations météorologiques de la province a permis d'isoler les meilleures combinaisons de ces variables, alliant efficacité et précision. Les résultats montrent une amélioration de la précision des modèles de prévision, avec une erreur moyenne absolue en pourcentage (MAPE) finale de 1.32% sur la période de 2012 à 2019 par rapport à 1.69% pour le contrôle, en ajoutant les variables de température du point de rosée, de radiation solaire incidente et de vitesse du vent provenant des stations de Montréal, Québec et Shawinigan.

Ce mémoire fournit des informations pertinentes pour améliorer la fiabilité et la précision des prévisions de la demande énergétique au Québec en tenant compte des variables météorologiques clés, permettant ainsi de mieux gérer et optimiser le réseau électrique face aux conditions météorologiques changeantes.

INTRODUCTION

La prévision de la demande énergétique est une préoccupation répandue dans le domaine de l'énergie. Les quantités grandissantes d'électricité à produire à l'échelle d'un pays jumelées aux coûts environnementaux qui y sont associés laissent très peu de place à l'incertitude (IEA, 2019). La production doit être précise, et ce malgré la contrainte de stockage que présente l'électricité. Une sous-production peut mener à des bris de service qui devront être évitées en s'approvisionnant à des sources moins écoresponsables et plus coûteuses. À l'inverse, une surproduction mènera à des pertes importantes en termes de ressources et d'argent. Il est donc essentiel d'être préparé, ce qui nécessite une bonne connaissance du besoin des consommateurs.

La consommation énergétique d'une population peut être divisée en fonction de l'activité des secteurs industriel, commercial et résidentiel (Hobby et Tucci, 2011; Apadula *et al.*, 2012; Amato *et al.*, 2005). Ces activités sont influencées par le contexte socio-économique de la région, mais également par la météo qui amène une forte variabilité de la charge avec l'utilisation de systèmes de chauffage et de climatisation électriques (Grenier, 2006; Gross et Galiana, 1987). La planification est de ce fait primordiale à long terme pour le développement d'installations adaptées (Nateghi et Mukherjee, 2017), à moyen terme pour la préparation des ressources nécessaires ou la négociation de contrats et finalement, à court terme pour la production et la distribution de l'énergie (Hahn *et al.*, 2009). La prévision à très court terme, typiquement sur un horizon de 24h, est aussi un élément clé du problème de plan de production qui tente d'optimiser l'état d'engagement

des centrales en respectant les contraintes de capacité, de coût et de temps de démarrage et d'arrêt des unités de production (Gross et Galiana, 1987).

Dans ce mémoire, nous nous concentrons sur les activités de la société d'état Hydro-Québec, plus spécifiquement à la prévision à très court terme de la demande énergétique québécoise. La prévision de la charge au Québec présente certains défis en raison des conditions météorologiques variables, parfois extrêmes, et de son étendue géographique. L'objectif est donc d'évaluer l'impact de différentes variables météorologiques sur la consommation énergétique de la province et d'améliorer la performance des modèles utilisés en choisissant adéquatement les intrants. Ce mémoire décrit cette étude en quatre chapitres.

Le premier chapitre fait une analyse détaillée des données de consommation énergétique et des données météorologiques sur une période de 8 ans, de 2012 à 2019. La décomposition de la série temporelle de consommation énergétique permet d'identifier les saisonnalités et d'estimer la tendance sur cette période. Une analyse de corrélation croisée permet de repérer la relation entre les variables météorologiques et la consommation selon différents décalages temporels.

Le deuxième chapitre évalue l'effet des intrants météorologiques dans les modèles de prévision. Les variables de température, d'humidité, de vitesse du vent, de nébulosité et de radiation solaire incidente sont évaluées. Pour cette étape, un modèle statistique autorégressif linéaire (Clements *et al.*, 2016) a été utilisé afin d'obtenir des résultats interprétables permettant une compréhension des impacts générés par les variables météorologiques. L'étude est ensuite étendue à un modèle basé sur un réseau de neurones artificiel (RNA) afin de comparer avec les simulations générées par l'équipe de Prévision des apports et de la demande à Hydro-Québec et pour permettre la considération des relations entre les variables, sans linéarisation prédéterminée. Le résultat des expériences montre un impact positif pour

chacune des variables considérées lorsqu'elles sont intégrées adéquatement dans le modèle choisi, mais leur combinaison permet une réduction maximale de l'erreur de prévision.

Le troisième chapitre s'intéresse plutôt à la position géographique des variables utilisées. Le territoire desservi par Hydro-Québec est très large et les conditions météorologiques sont variables d'une région à une autre. L'information de plusieurs villes peut donc être pertinente, mais un mauvais choix ou une trop grande quantité d'intrants pourrait alourdir et dégrader les performances d'un modèle de prévision (Blum et Langley, 1997). Le défi de ce chapitre est d'identifier les meilleures stations météorologiques à utiliser et la quantité optimale pour améliorer la prévision du besoin énergétique du Québec.

Le quatrième chapitre combine les résultats récoltés dans les chapitres 2 et 3. La combinaison des stations et des variables météorologiques est évaluée pour ses capacités de précision et de performance de calcul.

ÉTAT DE L'ART

La hausse de la consommation et des prix de l'électricité depuis les années 1980 (Grenier, 2006; Mukherjee et Nateghi, 2017) a motivé un nombre de recherches grandissant dans les milieux académique et opérationnel (Hong *et al.*, 2020). Ces recherches se différencient principalement par la situation socio-économique et climatique de la région, le secteur ciblé, l'horizon de prévision et le modèle de prévision.

La majorité des études reconnaissent l'importance de la température dans le calcul de la consommation énergétique, mais d'autres variables météorologiques ont également été relevées comme ayant un impact significatif. Parmi les études qui évaluent l'impact des intrants météorologiques sur la consommation résidentielle, on retrouve les suivantes.

Amato *et al.* se sont intéressés à la réponse régionale de la demande énergétique aux changements climatiques en évaluant la composante sensible à la température de la consommation au Commonwealth du Massachusetts aux États-Unis (Amato *et al.*, 2005). Pour ce faire, les degrés de climatisation, "Cooling Degree Days" (CDD), et les degrés de chauffage, "Heating Degree Days" (HDD), sont dérivés des valeurs de température en calculant la différence entre la température et le point d'équilibre correspondant à une zone de confort. Les auteurs considèrent également les heures d'ensoleillement dans leur étude. Les résultats obtenus confirment une sensibilité de la demande énergétique à la température avec une augmentation de 4.7% de consommation pour une accumulation mensuelle de 100 unités de HDD et 8.8% pour 100 unités de CDD dans le secteur résidentiel.

Mukherjee et Nateghi ont évalué la sensibilité climatique de la consommation énergétique en Floride, aux États-Unis (Mukherjee et Nateghi, 2017). Les auteurs ont établi que les variables de degrés de chauffage et de climatisation n'étaient pas des mesures adéquates puisqu'elles ne prennent pas en compte l'humidité dans l'atmosphère. De meilleurs résultats sont obtenus en utilisant la température du point de rosée. La vitesse du vent et le taux de précipitation sont aussi évalués. Les résultats démontrent une augmentation de la consommation mensuelle de 200 GW pour une augmentation de la température du point de rosée moyenne de 1°F, une diminution jusqu'à 1500 GW/mois avec l'augmentation de la vitesse du vent et une diminution générale avec les précipitations.

Dans une seconde étude, Nateghi et Mukherjee ont développé un modèle de prévision à long terme pour évaluer les projections de consommation énergétique en Indiana, aux États-Unis, en fonction de deux scénarios d'émission de CO₂ (Nateghi et Mukherjee, 2017). Les variables climatologiques utilisées dans leur étude sont la température minimale et maximale, la vitesse du vent et la précipitation. L'analyse de dépendance partielle des intrants du modèle entraîné a révélé une plus grande importance accordée à la température minimale et maximale en hiver, suivi de la vitesse du vent à l'automne, alors que les précipitations ont démontré un impact moindre.

Apadula et al. ont analysé l'effet des variations météorologiques sur la demande énergétique en Italie (Apadula *et al.*, 2012). Les variables considérées sont la température, la vitesse du vent, l'humidité relative et la couverture nuageuse. La vitesse du vent et l'humidité relative ont été prises en compte dans le calcul de la température ressentie avec l'effet de refroidissement éolien et l'indice de chaleur. Les résultats ont montré une légère amélioration de la précision des modèles en ajoutant ces deux variables, mais l'impact majeur provenait de la température. La couverture nuageuse a montré une influence inférieure à 1% tout au long de

l'année.

Aisyah et al. ont fait une analyse de l'impact de différentes variables météorologiques sur la prévision de la demande énergétique à Bali en Indonésie (Aisyah *et al.*, 2022). Dans cette étude, la température, la radiation solaire, la vitesse du vent, le taux de précipitation, la pression et l'humidité relative ont été examinés. Les résultats ont démontré de plus fortes corrélations avec la demande énergétique pour la température, la radiation solaire et la vitesse du vent, alors que le taux de précipitation, la pression et l'humidité relative sont moins corrélés. Lorsqu'inclus dans les modèles de prévisions, seules la température et la radiation solaire ont démontré un impact positif sur la précision des résultats obtenus.

Grenier et al. ont présenté en 2006 le développement de la prévision de la demande énergétique à court terme au Québec (Grenier, 2006). Dans cet article, les auteurs mentionnent l'importance de la météorologie pour estimer les normales horaires, mensuelles et les pointes de charge. Ils indiquent que l'effet moyen d'une variation de 1 °C sur la demande de chauffage est d'environ 400 MW/h et que l'impact de la couverture nuageuse peut aller jusqu'à 3000 MW/h. Le modèle de prévision décrit considère les variables de température, vitesse du vent, couverture nuageuse et type de précipitation.

Les modèles de prévisions utilisés dans la littérature proviennent majoritairement de méthodes statistiques, de méthodes d'apprentissage automatique ou d'une combinaison hybride de ces dernières (Lago *et al.*, 2018).

On retrouve parmi les méthodes statistiques, une approche à équations multiples basée sur la structure ARMA proposée par Clements et al. qui combine la composante autorégressive avec le terme de perturbation et les variables exogènes de température et jours spéciaux en isolant les périodicités journalière et hebdomadaire propres à la consommation énergétique (Clements *et al.*, 2016). Le terme

de perturbation correspond au facteur de correction basé sur l'erreur d'estimation des pas de temps passés et les variables dites exogènes réfèrent aux variables extérieures à la série temporelle qui est prédite. Les techniques de régression multivariée par spline adaptative (MARS) et modèle d'autorégression intégré et moyenne mobile (ARIMA) ont aussi été appliquées au problème de prévision de la demande à court terme et comparées par Al-Musaylh et al. (Al-Musaylh *et al.*, 2018). Un modèle de régression linéaire multiple paramétrique a été appliqué pour l'étude de Apadula et al. (Apadula *et al.*, 2012). Les auteurs s'accordent pour dire que ces méthodes sont généralement plus simples à comprendre, à appliquer et à calculer.

Les approches utilisant l'apprentissage automatique sont aussi très populaires dans ce milieu de recherche. On retrouve entre autres le modèle d'arbre de régression additif bayésien (BART) appliqué aux études de Nateghi et Mukherjee (Nateghi et Mukherjee, 2017; Mukherjee et Nateghi, 2017) qui fait la somme de plusieurs arbres de décision simple et qui permet une mesure de l'importance des intrants considérés. Les méthodes de réseau neuronal de régression généralisée (GRNN) et de séparateurs à vaste marge (SVM) ont été utilisées dans la recherche de Aisyah afin de comparer l'impact des variables météorologiques (Aisyah *et al.*, 2022). Le modèle NBEATSx, une version modifiée de l'architecture NBEATS avec l'ajout de variables exogènes, permettant une prévision précise des composantes de tendance et saisonnalité d'une série temporelle a également été appliqué à la prévision électrique (Olivares *et al.*, 2023).

Une étude comparative sur 27 méthodes statistiques et d'apprentissage automatique pour la prévision d'électricité a démontré que les modèles d'apprentissage automatique obtiennent généralement de meilleurs résultats que les modèles statistiques avec des scores MAPE significativement meilleurs selon le test de Diebold-Mariano (Lago *et al.*, 2018). L'étude a également conclu que les versions hybrides ne performant pas forcément mieux que leur version simplifiée, c'est-à-dire

la composante statistique ou la composante d'apprentissage automatique individuelle. Les résultats générés par un modèle d'apprentissage automatique ne sont toutefois pas toujours interprétables.

D'une manière plus générale, le développement d'un modèle de prévision temporelle est un défi qui génère beaucoup d'intérêt dans la communauté scientifique comme peuvent en témoigner les compétitions Makridakis qui visent à identifier les moyens d'améliorer la précision des prévisions temporelles depuis 40 ans (Makridakis *et al.*, 2022). Parmi les propositions de modèles élaborés et explicables qui pourraient être applicables au problème à l'étude, on retrouve l'architecture de *Temporal Fusion Transformer* (TFT) qui utilise un mécanisme d'attention pour obtenir des résultats précis et explicables (Lim *et al.*, 2021) et le modèle *Temporal Convolutional Networks* (TCN) avec mécanisme d'attention intégré (Pantiskas *et al.*, 2020). Ces modèles ont été testés au sein de l'équipe de Prévision des apports et de la demande à Hydro-Québec et présentent des résultats prometteurs.

CHAPITRE I

ANALYSE DESCRIPTIVE DE LA CONSOMMATION D'ÉLECTRICITÉ DANS LA PROVINCE DU QUÉBEC

Ce mémoire s'intéresse à la prévision de l'utilisation d'électricité au Québec. Comme première étape, une analyse de la consommation électrique des années passées est incontournable. De plus, les conditions météorologiques ont montré avoir un impact important sur les habitudes de consommation des Québécois (Grenier, 2006). C'est pourquoi la section suivante s'intéresse aux relations existantes entre les variables météorologiques et la consommation énergétique. Ces connaissances permettront ensuite de bien guider le choix et l'apprentissage des modèles de prévision.

1.1 Consommation énergétique

La variable principale qu'on souhaite estimer est appelée besoin québécois (BQ) et signifie la puissance nécessaire pour approvisionner la population québécoise. Bien que le terme consommation soit utilisé dans ce mémoire, cette quantité englobe l'utilisation des abonnés, mais également les pertes encourues lors du transport et de la distribution de la charge. Cette quantité est mesurée en mégawatt (MW). Nous nous basons sur un échantillon assemblé à l'interne par l'unité Prévision des apports et de la demande à Hydro-Québec. Il s'agit de mesures horaires de 2012 à

2019 qui représentent la consommation électrique en MW des secteurs résidentiel et commercial combinés pour le Québec en entier.

Ces données, distribuées dans le temps, forment une série temporelle. Typiquement, une série temporelle peut être décomposée en trois composantes appelées *tendance*, *périodicité* et *résidu*.

1.1.1 Tendances

La tendance d'une série temporelle correspond à la variation de sa moyenne sur une longue période de temps. Dans le cas de la consommation par exemple, certaines études constatent des augmentations d'année en année en raison de la croissance de la population et du développement économique de leur région (Apadula *et al.*, 2012; Nateghi et Mukherjee, 2017; Mukherjee et Nateghi, 2017). Comme notre échantillon s'étend sur une période de 8 ans, il est possible qu'un tel effet soit observable et qu'il doive être pris en compte lors de l'entraînement de modèles de prévision. Parmi les méthodes d'analyse de tendance des séries temporelles, le calcul d'une moyenne mobile ou d'une régression linéaire simple devrait permettre une bonne estimation (Mudelsee, 2019). Le calcul d'une moyenne mobile étendue sur une année permet de voir l'évolution de la moyenne annuelle de la série. La régression linéaire des moindres carrés identifie la pente qui s'ajuste au mieux à la distribution des données dans les temps.

La figure 1.1 montre l'évolution de la moyenne annuelle et le meilleur ajustement de régression linéaire. Nous obtenons pour la droite des moindres carrés, une pente de -0.0054 MW/h, ce qui représente une baisse de 375 MW entre le début et la fin de l'intervalle de 2012 à 2019. Considérant la valeur moyenne de 23000 MW, cette baisse de moins de 2% est petite, de sorte que la tendance semble négligeable.

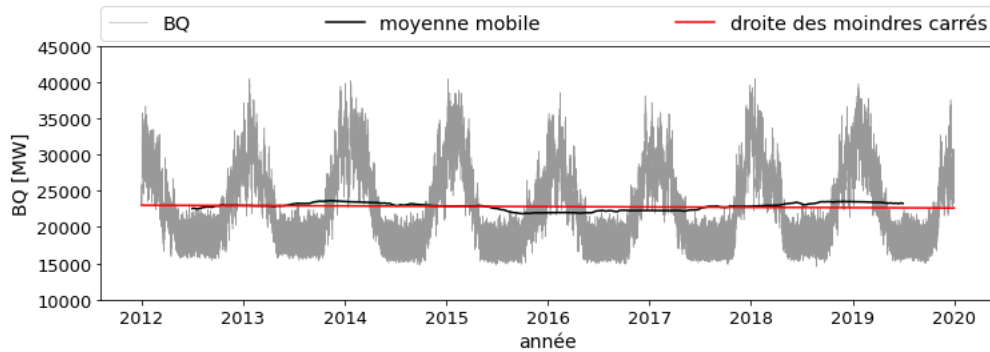


FIGURE 1.1 Évaluation de la tendance du BQ en gris avec une moyenne mobile de 1 an en noir et une estimation des moindres carrés de degré 1 en rouge.

Une façon plus robuste de vérifier la présence d'une tendance dans une série consiste à effectuer un test statistique de stationnarité. Pour qu'une série soit considérée stationnaire, elle doit avoir une moyenne et une variance constantes dans le temps. Il n'y a donc pas de tendance lorsque les conditions de stationnarité sont rencontrées.

Une méthode couramment utilisée pour tester la stationnarité est le test de Dickey-Fuller augmenté, qui s'applique aux séries temporelles afin de déterminer si elles ont une racine unitaire (Dickey et Fuller, 1979). En présence d'une racine unitaire, la variance de la série temporelle augmente avec le temps ce qui permet à la série de s'éloigner de sa valeur moyenne. Si la racine est inférieure à 1, il est possible de démontrer que les conditions de stationnarité sont rencontrées. Ainsi le test se compose de la manière suivante :

- Hypothèse nulle : Une racine unitaire est présente et la série temporelle est non stationnaire.
- Hypothèse alternative : La racine est inférieure à 1, la série temporelle est stationnaire.

Les intervalles de confiance en dessous desquels il est recommandé de rejeter l'hypothèse nulle sont donnés au tableau 1.1 (MacKinnon, 2010).

TABLEAU 1.1 Valeurs critiques pour les niveaux de confiance du test de Dickey-Fuller augmenté.

Niveau de confiance	Valeur critique
1%	-3.430
5%	-2.862
10%	-2.567

Dans le cas de notre échantillon, nous obtenons une valeur de -6.27 qui est significativement inférieure à la limite de l'intervalle de confiance de 1%. Ainsi, on peut affirmer avec moins de 1% d'incertitude que l'hypothèse nulle est rejetée et que la série temporelle du BQ est stationnaire. Une autre validation pour consolider ce résultat est la valeur-p de 4×10^{-8} qui représente la probabilité d'accepter l'hypothèse alternative si l'hypothèse nulle vraie. On en conclut donc qu'aucune tendance n'est présente sur cette période de 8 ans.

1.1.2 Périodicité

La périodicité ou saisonnalité d'une série temporelle correspond à un signal récurrent à un intervalle de temps régulier. On peut penser par exemple que les fluctuations de consommation au fil des saisons seront similaires d'année en année. Une connaissance des périodes d'une série peut grandement améliorer l'entraînement d'un modèle de prévision.

Transformée de Fourier

Afin d'identifier la fréquence et l'intensité des saisonnalités qui composent la série temporelle du BQ, on peut effectuer une analyse en série de Fourier (Musbah *et al.*, 2019). Comme nos données sont discrètes, on utilise la transformée de Fourier discrète (FFT) (Cooley et Tukey, 1965). Celle-ci nous fournit une représentation spectrale, en somme de sinus et cosinus, contenant les termes de fréquence de la série initiale. Les fréquences dominantes sont alors déterminées par une amplitude absolue très marquée.

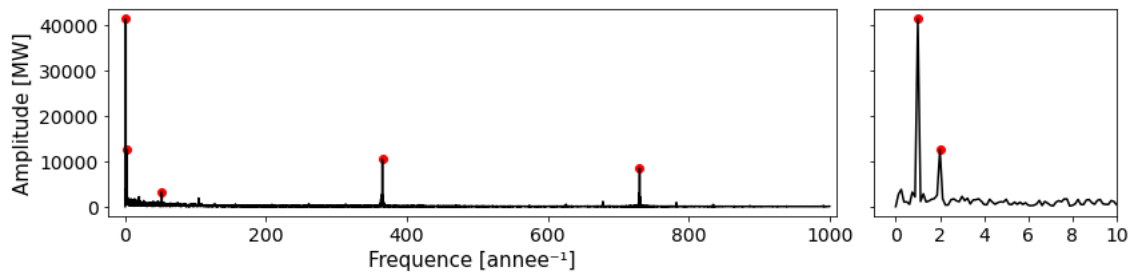


FIGURE 1.2 Spectre de l'amplitude absolue des fréquences de la série temporelle du BQ en noir avec identification des pointes d'amplitude importante avec un point rouge. Spectre allant jusqu'à 1000 fréquences à gauche et vue détaillée sur les 10 premières fréquences à droite.

La figure 1.2 présente le spectre résultant de cette transformation. La fréquence présentée correspond au nombre de fois qu'un cycle apparaît en un an. On retrouve sur la figure quatre démarcations d'amplitude importante aux fréquences de 1, 2, 365 et 730 $annee^{-1}$. Le décalage symétrique et la perte d'intensité nous permettent toutefois d'éliminer la pointe à 2 $annee^{-1}$ puisqu'il s'agit d'une harmonique de la première pointe. Cet harmonique apparaît durant le calcul du spectre en raison de la taille de l'échantillon ou du filtre appliqué qui ne permet pas une reconstruction de l'onde avec une seule fréquence. On ne peut toutefois pas éliminer assurément

celle à 730 annee^{-1} , malgré son décalage symétrique, car son amplitude est presque aussi élevée que celle à 365 annee^{-1} . Une cinquième pointe est également présente à 52 annee^{-1} , mais son amplitude est modérée.

On peut donc en déduire que les fréquences fondamentales de la série de consommation sont de 1, 52, 365 et 730 annee^{-1} , ce qui correspond à des périodes de 1 an, 1 semaine, 1 jour et 12h. On peut aussi remarquer que la variation annuelle est environ 4 fois plus forte que la variation journalière et que la variation hebdomadaire est la moins prononcée. Il est également intéressant de noter que le cycle annuel de cette série est celui qui est le plus influencé par la météorologie, avec les variations saisonnières de température, ce qui confirme l'importance de cette composante. Les cycles de 12h et 1 semaine semblent plutôt associés au comportement humain selon les heures de pointe et les jours ouvrables. Le cycle diurne est, quant à lui, affecté par ces deux composantes.

Autocorrélation

En plus des saisonnalités, il est possible de repérer les effets d'inertie dans la série, c'est-à-dire les relations contenues entre points successifs. On regarde d'abord l'autocorrélation, soit la corrélation de Pearson entre la série et le décalage de celle-ci. Un coefficient de corrélation d'une valeur de 1 ou -1 représente une concordance parfaite, directe ou inverse, alors qu'une valeur de 0 signifie plutôt une absence de relation.

On remarque sur la figure 1.3 l'effet de saisonnalité qui revient avec des périodes de 12h, 24h et 1 an. La similitude est plus grande aux quelques heures précédant un point puis aux multiples de 24h qui suivent avec une perte d'amplitude à partir du 7e jour.

Une autre évaluation qui peut être faite sur notre série temporelle est l'auto-

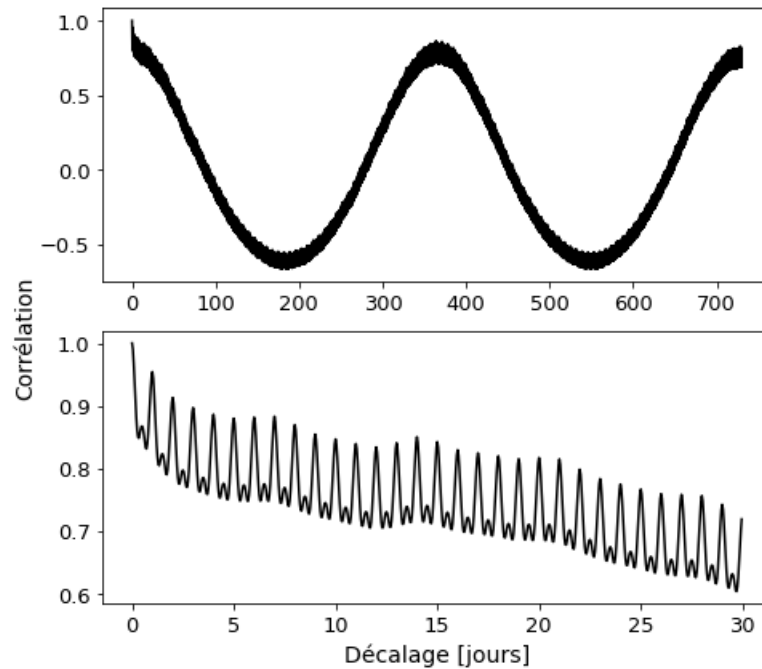


FIGURE 1.3 Autocorrélation de la série du BQ pour 730 et 30 jours de décalage.

corrélation partielle de Yule-Walker. Il s'agit d'une méthode autorégressive qui évalue la contribution des points dans le calcul des points suivants. L'implémentation utilisée calcule les coefficients de l'équation avec un ordre équivalent au décalage. Ainsi, on définit les composantes du vecteur de paramètres \mathbf{b} tel que $t_0 = b_0 + \sum_{i=1}^n b_i t_i$ et n le décalage. Le coefficient b_n nous donne alors la contribution du décalage t_n en tout point.

En fixant une limite pour les autocorrélations partielles, dont la valeur absolue est supérieure à 0.2, les points de décalage pertinents sont : 0, 1, 2, 3, 7, 15, 16, 24, 25, 26, 144, 145 et 169 heures. La fonction considère le point zéro, mais il ne s'agit que d'une comparaison de la série avec elle-même, sans décalage, d'où la corrélation parfaite. On constate des similitudes plus fortes au cours de la première journée, particulièrement pour les 3 premières heures, mais il y a aussi des pointes pour

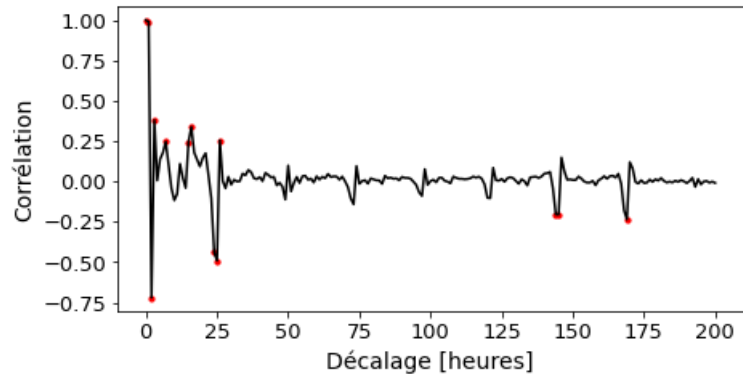


FIGURE 1.4 Autocorrélation partielle de la série de BQ pour les décalages d'au plus 200 heures.

7, 15 et 16 heures. On remarque ensuite une valeur marquée à tous les décalages multiples de 24h respectant la saisonnalité de la série, mais ces valeurs sont plus importantes pour 1, 6 et 7 jours.

1.1.3 Résiduel

Le résiduel de la série est la variation qui ne peut être associée ni à la tendance, ni à la périodicité. Celle-ci est généralement aléatoire et plus difficile à prévoir.

En utilisant une décomposition qui soustrait la tendance et les saisonnalités calculées à l'aide d'une moyenne mobile, on peut obtenir le signal résiduel présenté à la figure 1.5.

Ce signal, dont l'amplitude s'étend jusqu'à 4000 MW, soit près de 20% su signal original, est significatif. Ces variations marquées ne pouvant être expliquées par la série elle-même nécessitent la contribution de variables exogènes pour être comprises et apprises par un modèle de prévision.

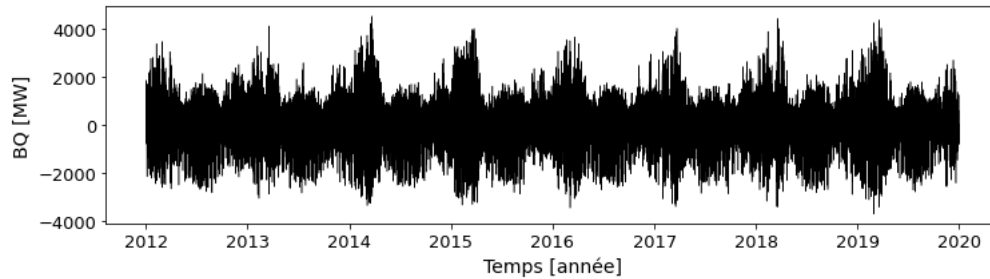


FIGURE 1.5 Signal résiduel de la série temporelle du BQ.

1.2 Variables météorologiques

Le choix des variables météorologiques à considérer a été basé sur des études antérieures (Apadula *et al.*, 2012; Gross et Galiana, 1987) ainsi que sur l'intuition, en se basant sur des principes physiques. On souhaite fournir le maximum d'information au modèle lors de son entraînement, mais une quantité trop grande d'attributs peu pertinents ou corrélés entre eux peut limiter les performances de ce dernier (Blum et Langley, 1997).

Si on s'intéresse à la consommation résidentielle, l'énergie est utilisée principalement pour le chauffage de l'air, mais également pour le chauffage de l'eau, la climatisation, l'éclairage et l'utilisation des électroménagers ou des appareils électroniques (Hydro-Québec, 2023). Parmi ces facteurs, plusieurs dépendent des habitudes de vies résidentielles et donc de l'heure de la journée. On sait par exemple que l'utilisation des électroménagers sera maximale avant et après les heures de travail ou la fin de semaine, alors que le chauffage et la climatisation dépendront davantage des conditions extérieures, soit majoritairement la température (Savić *et al.*, 2014). Une étude a démontré en 2006 que l'impact moyen d'un changement de température de 1°C pouvait générer une variation de 400 MW de la demande pour le chauffage et que la couverture nuageuse pouvait avoir un impact allant jusqu'à 3000 MW (Grenier, 2006). Les études sur les conditions de confort hu-

main montrent également l'influence de l'humidité relative, la vitesse du vent et le rayonnement solaire (Song *et al.*, 2022). L'éclairage est une combinaison des habitudes des gens et de la luminosité extérieure altérée par la couverture nuageuse et la position du soleil dans son cycle journalier.

Évidemment, la nature des bâtiments influence grandement la consommation avec entre autres la disposition des fenêtres, la qualité de l'isolement et la taille. Un bâtiment dit standard pourra consommer plus de $65 \text{ kWh}/\text{m}^2/\text{an}$, soit le double d'un bâtiment écoénergétique (Ionescu *et al.*, 2015). Cependant, ces aspects varient très peu dans un court intervalle de temps, ce qui nous permet de les supposer constants pour les prévisions à court terme.

En plus de prendre en compte la variation temporelle suivant les habitudes des consommateurs, on peut donc s'intéresser aux conditions météorologiques et leur relation avec le besoin énergétique. Les variables considérées sont la température à 2m (T), l'humidité relative à 2m (HR), la vitesse du vent à 10m, le rayonnement solaire incident à la surface (Rad) et la couverture nuageuse (Neb) qui pourraient avoir un impact direct ou indirect en influençant les habitudes de vie.

Les données météorologiques de température, humidité relative, vitesse du vent et couverture nuageuse proviennent de l'historique d'observations de la station YUL d'Environnement et Changements Climatiques Canada (ECCC) qui est située à l'aéroport Pierre-Elliott Trudeau de Montréal (ECCC, 2023a). La résolution temporelle de ces données est horaire et la période utilisée couvre de 2012 à 2019. Un post-traitement est appliqué aux variables pour combler les valeurs manquantes en utilisant la valeur moyenne sur les 8 ans à l'heure, la journée et le mois correspondants. La nébulosité subit également une transformation développée par les météorologues d'Hydro-Québec pour changer les unités et considérer l'épaisseur des nuages. Les valeurs initiales sont des observations de la fraction du ciel, en octa,

qui est couvert par un nuage et le résultat du post-traitement est un ratio sur 10 du niveau de couverture normalisé sur la période sélectionnée. Puisque les mesures du rayonnement solaire incident n'étaient pas accessibles, cette variable est tirée des réanalyses ERA5 (Hersbach *et al.*, 2020), soit les simulations du modèle global de l'European Centre for Medium-Range Weather Forecasts (ECMWF) corrigées avec l'assimilation d'observations. Bien qu'il s'agisse de simulations, les réanalyses offrent des valeurs très proches des valeurs réelles puisqu'elles sont ajustées à l'aide d'une base de données volumineuses de mesures terrestres et satellitaires. Le jeu de données s'étend sur le globe entier avec une résolution de 0.25 degré, mais seulement le point de grille le plus proche de l'aéroport de Montréal a été sélectionné afin de rester cohérent avec la source des autres variables. L'échantillon utilisé couvre également la période de 2012 à 2019 avec un pas de temps horaire.

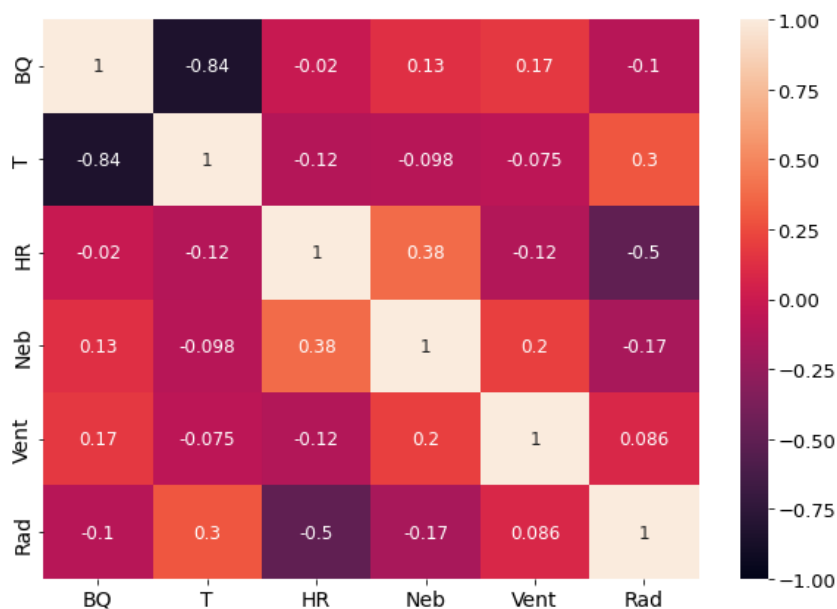


FIGURE 1.6 Matrice de corrélation entre les variables météorologiques et le BQ à Montréal pour les données horaires de 2012 à 2019. Variables : besoin québécois (BQ), température (T), humidité relative (HR), nébulosité (Neb), vitesse du vent (Vent) et radiation (Rad).

La figure 1.6 représente la matrice de corrélation pour l'ensemble de ces variables. On note une forte corrélation inverse entre le BQ et la température. Les autres variables, avec un coefficient sous 0.2, ne présentent pas de corrélations évidentes avec le BQ. On peut toutefois constater des relations entre les différentes variables. Il y a d'abord la radiation et la température, ce qui s'explique par le fait que le rayonnement solaire incident influence fortement le bilan énergétique à la surface qui domine dans les variations de température journalière et annuelle. Il y a également un rapport inverse entre ce rayonnement et l'humidité relative qui pourrait simplement s'expliquer par le fait que l'humidité relative augmente pendant la nuit puisque l'air se refroidit et sa pression de vapeur saturante diminue, contrairement au rayonnement solaire qui est absent pendant cette période et qui augmente durant la journée. Finalement, la couverture nuageuse, ou nébulosité, présente une relation significative avec l'humidité relative, ce qui peut s'expliquer par le fait qu'une humidité relative élevée participe à la formation de nuage.

Suivant la logique de ces observations, certaines de ces variables sont dépendantes et contiennent des informations redondantes dont on doit tenir compte.

1.2.1 Température

La première variable météorologique d'intérêt est la température, dont la corrélation globale avec le BQ est négative. La figure 1.7 présente plus en détail la relation directe entre ces deux variables.

On note d'abord que la corrélation négative est due à la forte relation inverse pour les basses températures en hiver et que la relation positive pour les températures élevées à l'été est plus faible. Cette relation polynomiale met en évidence la séparation entre l'énergie nécessaire au chauffage et celle à la climatisation. De plus, on remarque un plancher entre les températures de 13 à 20°C approximativement,

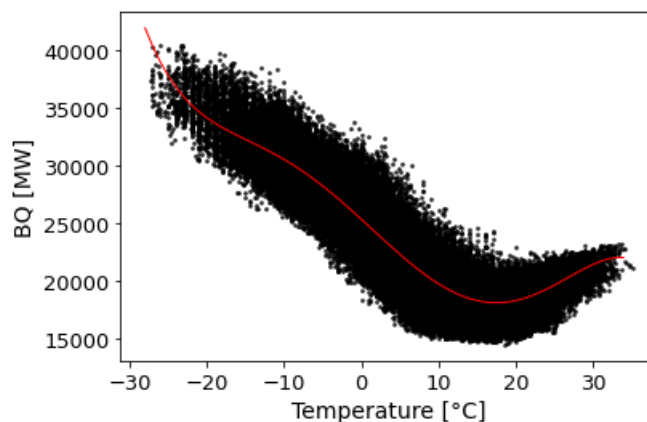


FIGURE 1.7 BQ en fonction de la température à Montréal sur la période complète de 2012 à 2019 avec estimation polynomiale des moindres carrés en rouge.

indiquant l'absence d'impact lorsque les conditions sont confortables.

Une seconde évaluation intéressante est une analyse de corrélation croisée du signal, qui identifie les décalages entre la température et le BQ. L'idée est donc d'observer si la température en un point t peut avoir un impact sur la demande en un point $t + \Delta t$.

La figure 1.8 met en évidence les périodicités communes aux deux séries temporelles, soit de 24h et de 365 jours. Ainsi, la corrélation entre un décalage de 1 an, le même jour un 1 an plus tôt, est meilleure qu'avec un décalage de 2 mois seulement par exemple. On voit également que l'amplitude diminue en s'éloignant dans le temps, mais dans des signes opposés à l'hiver et à l'été, causant une diminution globale très lente, mais peu représentative. La variation journalière est largement plus importante en été, passant d'une corrélation forte aux décalages multiples de 24 heures à une absence de corrélation entre ces décalages. Cette périodicité est plus subtile en hiver et elle amplifie la corrélation inverse déjà présente. Ainsi, le maximum d'amplitude se trouve près du milieu de la période avec 11 heures de décalage. L'amplification peut s'expliquer par le fait qu'habituellement le re-

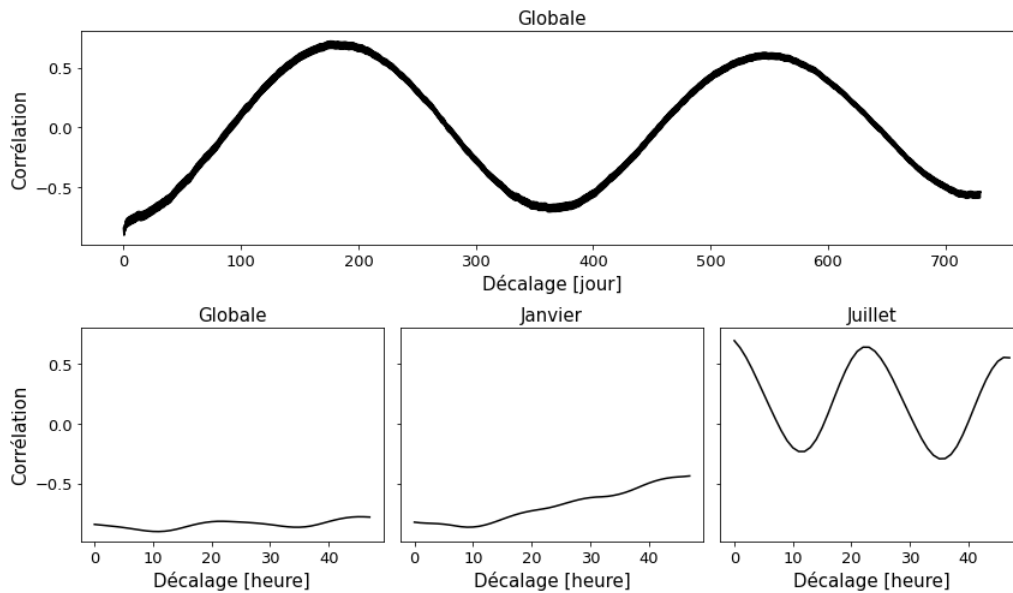


FIGURE 1.8 Corrélation croisée de la température avec le BQ calculée avec l'ensemble de la période 2012-2019 et observée sur un décalage allant jusqu'à 2 ans avec vue détaillée sur les décalages des 48 premières heures pour les séries temporelles complètes et pour les mois de janvier et juillet séparément.

froidissement extérieur est maximal pendant la nuit alors que la consommation l'est pendant le jour. Ce principe est toujours vrai en été, mais l'effet est inverse puisque la source d'énergie consommée est due au refroidissement de l'air et non le chauffage. Globalement, il y a tout de même une amélioration de la corrélation initiale lorsque le décalage est égal à 11 heures puisque la relation entre les deux variables d'intérêt est plus forte à cette saison.

1.2.2 Vitesse du vent

Une deuxième variable météorologique d'intérêt est la vitesse du vent. On émet l'hypothèse que l'interaction du vent avec le BQ repose sur l'effet de refroidissement de celui-ci. Toutefois, on ne remarque aucune relation directe avec la tem-

pérature et une très faible corrélation de 0.17 avec le BQ.

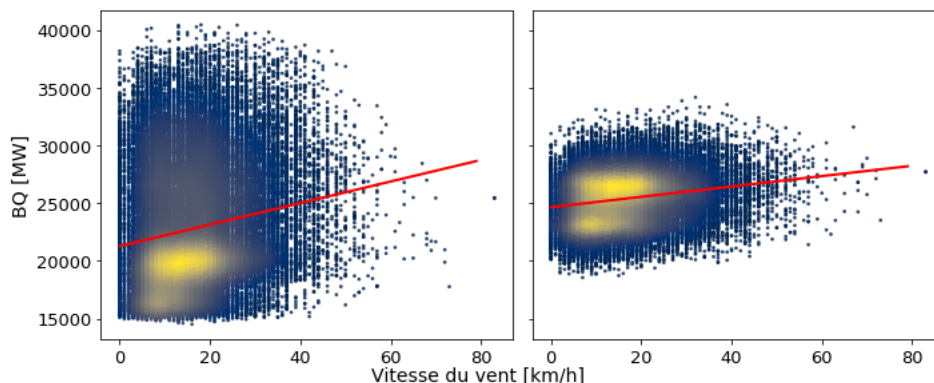


FIGURE 1.9 Graphique de densité du BQ en fonction de la vitesse du vent (gauche) et du BQ moins la variation due à la température en fonction de la vitesse du vent (droite) avec estimation linéaire des moindres carrés (rouge). Retrait de la température avec les degrés plus grands que 0 du polynôme de la figure 1.7.

Si on évalue la relation directe de ces composantes sur la figure 1.9, la droite des moindres carrés montre tout de même une tendance positive, c'est-à-dire une augmentation de la consommation avec l'augmentation de la vitesse du vent. Par contre, la variance est très élevée, accordant un faible niveau de confiance à cette observation. Si on tente de limiter la variance de l'échantillon en retirant l'effet de la température, le signal résiduel montre une meilleure corrélation avec la vitesse du vent, mais la pente de la droite se rapproche de zéro. Cette observation suggère que la vitesse du vent aurait un impact positif sur le BQ, et ce indépendamment de la température.

Une autre option pour prendre en compte l'effet du vent, serait de le combiner avec la température en calculant l'effet de refroidissement éolien (Osczevski et Bluestein, 2005). Cet effet de refroidissement est causé par les échanges de chaleur entre la peau et l'air qui l'entoure. Si la chaleur se déplace de la peau vers l'air, cet

air sera réchauffé et le transfert ralentira jusqu'à atteindre un état d'équilibre dans lequel la température est uniforme. Cependant si cet air est déplacé et remplacé par de l'air froid, alors la peau devra à nouveau dégager une grande quantité de chaleur. On peut supposer que l'effet sera comparable pour les interactions entre l'air extérieur et les bâtiments, dont l'isolation n'est pas parfaite.

On peut calculer l'indice de refroidissement ressenti par l'humain avec les formules suivantes pour des vents forts, supérieurs à 4.8 km/h, et des vents faibles, inférieurs à 4.8 km/h (Osczevski et Bluestein, 2005) :

$$R = \begin{cases} 13.12 + 0.6215T + (0.3965T - 11.37) * V^{0.16} & \text{si } V \geq 4.8 \\ T + 0.2 * (0.1345T - 1.59) * V & \text{si } V < 4.8, \end{cases} \quad (1.1)$$

où T est la température en °C et V la vitesse du vent en km/h.

1.2.3 Humidité relative

Plusieurs études, évaluant les relations entre les variables météorologiques et la demande énergétique, mentionne l'importance de l'humidité relative (Aisyah *et al.*, 2022; Apadula *et al.*, 2012; Mukherjee et Nateghi, 2017). Cette relation n'est toutefois pas évidente dans le cas des données d'Hydro-Québec, avec une corrélation quasi nulle de -0.019. L'hypothèse sur l'influence de l'humidité relative repose dans l'indice de chaleur qu'elle provoque à des températures supérieures à 27°C. Dans son mécanisme de refroidissement, le corps humain produit de la sueur qui prend l'énergie de la peau lors de son évaporation. Toutefois, plus l'air est humide, moins l'évaporation de la sueur sera possible. Ainsi, une sensation de chaleur plus grande que la température réelle est perçue. On peut donc supposer que la consommation d'énergie pour la climatisation sera augmentée pour satisfaire les conditions

de confort. Pour vérifier cette hypothèse, on peut calculer l'indice de chaleur pour les températures supérieures à 27°C à l'aide de la formule suivante (Rothfus et Headquarters, 1990).

$$HI = c_1 + c_2T + c_3R + c_4TR + c_5T^2 + c_6R^2 + c_7T^2R + c_8TR^2 + c_9T^2R^2, \quad (1.2)$$

où

$$\begin{aligned} c_1 &= -8.78469475556 & c_6 &= -0.0164248277778 \\ c_2 &= 1.61139411 & c_7 &= 2.211732 \times 10^{-3} \\ c_3 &= 2.33854883889 & c_8 &= 7.2546 \times 10^{-4} \\ c_4 &= -0.14611605 & c_9 &= -3.582 \times 10^{-6} \\ c_5 &= -0.012308094 \end{aligned}$$

et T est la température en °C et R l'humidité relative en %.

Un autre impact que l'humidité relative pourrait avoir sur la consommation repose dans la déshumidification de l'air lors de la climatisation (Zhao *et al.*, 2011). Comme la quantité de vapeur d'eau que l'air peut contenir augmente avec la température, en diminuant la température et en maintenant la même quantité de vapeur d'eau dans une parcelle d'air, on augmente son humidité relative. Lorsqu'on arrive à saturation, la parcelle atteint le maximum de vapeur d'eau qu'il lui est possible de contenir. Une baisse de température à ce niveau provoque alors la condensation de l'eau, ce qui résulte en un dégagement de chaleur latente et donc un réchauffement de la parcelle. Cela signifie que l'air humide nécessite davantage d'énergie à refroidir que l'air sec. Ainsi, une autre façon de prendre en compte l'humidité relative consisterait à utiliser la température du point de rosée. Cette mesure correspond à la température à laquelle l'air doit être refroidi pour atteindre

la saturation (100% d'humidité relative).

1.2.4 Couverture nuageuse

Une autre variable météorologique pouvant avoir une incidence sur la consommation énergétique est la couverture nuageuse ou nébulosité (Apadula *et al.*, 2012). Celle-ci nous informe sur les besoins en éclairage durant le jour et peut avoir un effet sur la température réelle et ressentie. Il est toutefois important de noter que le besoin énergétique pour l'éclairage ne représente que 5% de la consommation d'un ménage selon les informations récoltées par Hydro-Québec (Hydro-Québec, 2023). De plus, l'assombrissement causé par une forte couverture nuageuse n'affecte que quelques heures d'ensoleillement puisqu'en absence de rayonnement solaire, la luminosité sera déjà minimale.

On peut donc s'intéresser à la relation avec la température. La corrélation entre la nébulosité et la température est nulle si on se fie à la figure 1.6. On sait toutefois que les nuages interceptent une portion du rayonnement solaire, ce qui limite le réchauffement par radiation à la surface. Cela dit, les nuages interceptent également le rayonnement terrestre, produisant ainsi un effet de serre qui réchauffe la surface (Ahrens, 2015). Une grande nébulosité sera alors associée à une diminution du réchauffement durant le jour et une diminution du refroidissement durant la nuit.

La figure 1.10 présente une relation directe positive entre la nébulosité et le BQ. Cette relation est plus évidente lorsque l'effet de la température est retiré, mais la pente est moins prononcée. On consommerait donc davantage en moyenne lorsque la nébulosité est élevée, ce qui concorde avec l'effet sur la température hivernale et sur la luminosité.

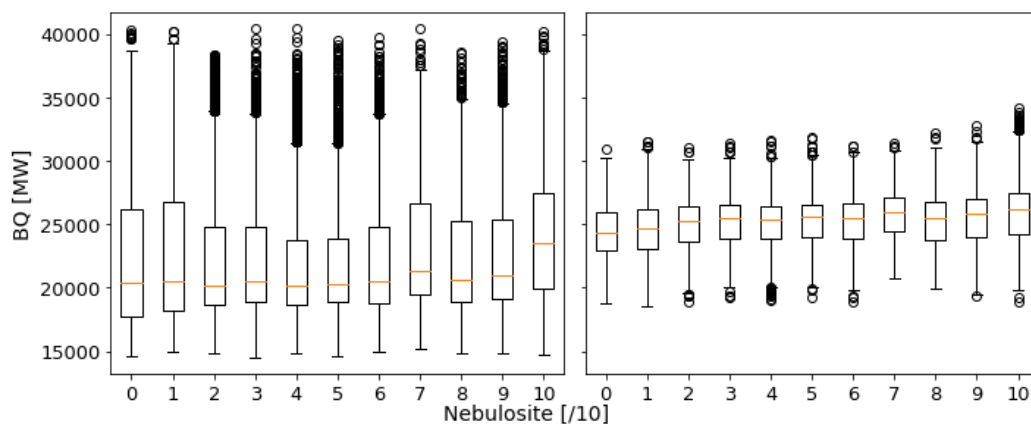


FIGURE 1.10 BQ en fonction de la nébulosité (gauche) et BQ moins la variation due à la température en fonction de la nébulosité (droite).

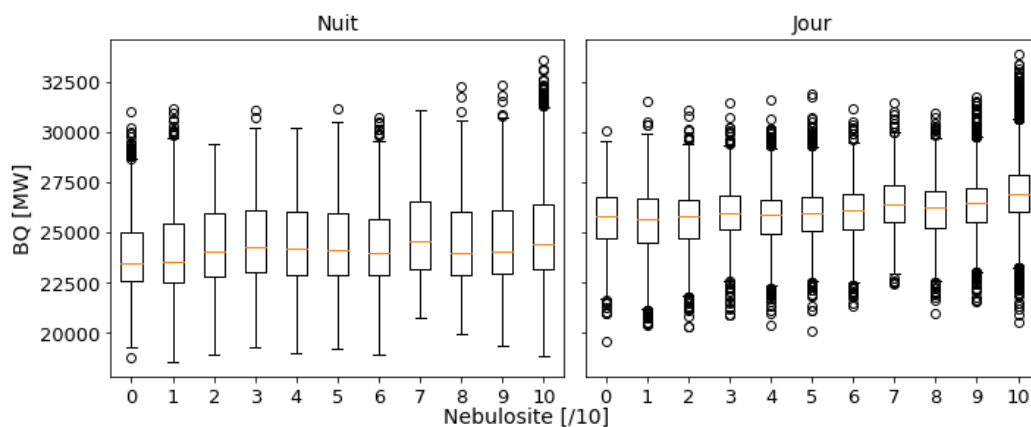


FIGURE 1.11 BQ sans la variation due à la température en fonction de la nébulosité pour la nuit de 18h à 6h (gauche) et pour le jour de 6h à 18h (droite).

Si on observe ensuite la relation séparément pour la nuit et le jour sur la figure 1.11, les diminutions de refroidissement et de réchauffement ne sont pas observées. On constate également qu'il y a une légère augmentation de la consommation pour un couvert nuageux élevé même durant la nuit contrairement au résultat attendu. Cette variation pourrait être due au comportement humain ou à un effet d'inertie.

1.2.5 Rayonnement solaire

Le rayonnement solaire incident à la surface de la Terre, en W/m^2 , est une mesure qui pourrait également contenir les informations de luminosité. Contrairement à la couverture nuageuse, cette mesure permet de faire la distinction entre le jour et la nuit puisque le rayonnement solaire reçu à la surface est nul durant la nuit. Ainsi, le cycle de luminosité saisonnier serait aussi pris en compte. De plus, il contient de l'information sur la couverture nuageuse, car une partie du rayonnement sera interceptée par celle-ci.

Cette variable nous donne également de l'information sur la température qui est partiellement redondante lorsque nous connaissons la température de l'air extérieur. Toutefois, elle nous informe aussi sur la température intérieure des bâtiments. En effet, selon le type de fenêtre utilisée, une partie du rayonnement est transmise à l'intérieur de la pièce permettant un réchauffement. Ce réchauffement peut se représenter par des variations allant jusqu'à 5°C supplémentaires (Besoudo *et al.*, 2010), ce qui réduit l'utilisation d'appareils de chauffage.

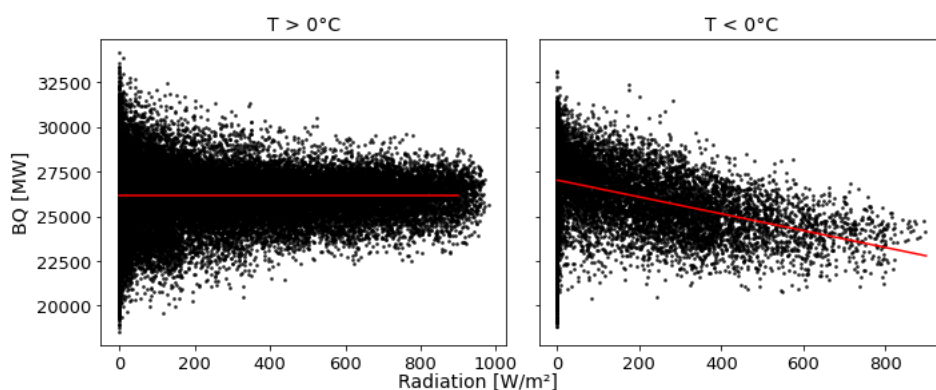


FIGURE 1.12 BQ sans la variation due à la température en fonction de la radiation solaire incidente pour les journées avec $T > 0^\circ\text{C}$ et les journées avec $T < 0^\circ\text{C}$.

La figure 1.12 présente la relation entre le BQ et la radiation solaire après le retrait de la variation due à la température. On peut y voir une relation négative pour les journées plus froides ($T < 0^{\circ}\text{C}$), ce qui implique une diminution de la consommation avec l'augmentation du rayonnement. On ne remarque toutefois pas d'impact pour les températures plus élevées, ce qui pourrait indiquer un effet négligeable sur la climatisation des bâtiments ou une contrebalance avec la diminution de consommation pour l'éclairage.

CHAPITRE II

ÉVALUATION DE L'EFFET DES INTRANTS MÉTÉOROLOGIQUES SUR LES MODÈLES DE PRÉVISION

La prochaine étape de ce mémoire est d'évaluer l'influence des variables météorologiques sur les modèles de prévision. L'objectif de cette étape est de déterminer les variables qui permettent d'améliorer la précision des prévisions, mais également de comprendre les relations apprises par les modèles. Pour ce faire, un modèle statistique et un modèle d'apprentissage automatique sont utilisés et comparés.

2.1 Modèle statistique linéaire

Afin de tester les hypothèses posées quant à l'influence des différentes variables météorologiques sur la consommation électrique, une première série d'expériences a été effectuée avec un modèle statistique de type équations multiples. Cette approche, d'abord présentée par Clements et al. (Clements *et al.*, 2016) puis retravaillée pour la prévision au Québec (David *et al.*, 2023), est un modèle autorégressif qui permet de prendre en compte des variables exogènes et de l'inertie. L'inertie est un concept qui fait référence au contexte d'une valeur qu'on détermine avec les pas de temps précédents de la variable. Celle-ci est considérée pour les périodes journalières et hebdomadaires pour s'ajuster aux périodicités de la charge énergétique. Les variables exogènes, quant à elles peuvent être caté-

gorielles, comme le jour de la semaine ou les jours fériés, ou continues comme la température. Le modèle comprend également une composante de moyenne mobile, c'est-à-dire un paramètre de correction de l'erreur sur la dernière prévision. Cela dit, bien que ce modèle soit spécialisé au problème de prévision de la demande, une limite s'impose en raison de sa structure linéaire. En effet, à moins d'appliquer une transformation de linéarisation des intrants, le modèle ne peut apprendre de relation non explicite. Il est également important de noter que malgré la liberté du choix des intrants permis par l'architecture, le temps d'entraînement tout comme la difficulté d'atteindre la convergence du calcul de coefficients augmentent avec le nombre d'intrants. L'avantage de cette approche repose surtout sur sa simplicité qui permet une bonne compréhension des résultats, ce qui est essentiel dans le cadre de ce mémoire pour comprendre l'impact des différentes variables.

Le calcul de la charge prévue à l'heure h du jour d suit l'équation suivante :

$$L_{h,d} = \theta_{h0} + \theta_{h,d1}L_{h,d-1} + \theta_{h,d2}L_{h,d-7} + \phi_{h1}\varepsilon_{h,d-1} + \phi_{h2}\varepsilon_{h,d-7} + \varepsilon_{h,d} + f_{h,d}^S + f_{h,d}^T + f_{h,d}^X \quad (2.1)$$

où $f_{h,d}^S$, $f_{h,d}^T$ et $f_{h,d}^X$ contiennent respectivement l'effet des jours spéciaux, de la température et des variables exogènes que nous détaillons plus bas. La variable $L_{h,d}$ correspond au logarithme de la charge en MW et $\varepsilon_{h,d}$, le terme de perturbation du modèle déterminé par l'erreur de prévision du pas de temps précédent. Les variables θ_{hi} et ϕ_{hi} , avec $i = 0, 1, 2$, sont les paramètres entraînaibles du modèle.

La gestion des jours spéciaux diffère de la méthode référée. Les jours fériés ne sont pas considérés dans cette version du modèle et la variable $f_{h,d}^S$ contient plutôt une indication du jour de la semaine et du fuseau horaire de la période donnée. Plus précisément,

$$f_{h,d}^S = \alpha_{h1}dst_{h,d} + \alpha_{h2}dst_{h,d-1} + \alpha_{h3}dst_{h,d-7} + \sum_{j=0}^6 (\alpha_{j,h4}S_{j,h,d}L_{h,d-1}) \quad (2.2)$$

où $S = 1$ pour $j =$ le jour de la semaine de la prévision et $S = 0$ pour les autres j . La variable $dst_{h,d}$ est une variable comprise entre 0 et 1 qui indique s'il s'agit de l'heure d'été (*Daylight Saving Time*) et α_{hi} correspond aux paramètres entraînaibles. Ensuite,

$$f_{h,d}^T = \sum_{k=1}^2 (\beta_{k,h1}H_{k,h,d} + \beta_{k,h2}C_{k,h,d} + \beta_{k,h3}H_{k,h,d-1} + \beta_{k,h4}C_{k,h,d-1} + \beta_{k,h5}H_{k,h,d-7} + \beta_{k,h6}C_{k,h,d-7}) \quad (2.3)$$

avec $H_{k,h,d}$ et $C_{k,h,d}$, les effets de chauffage et de refroidissement obtenus par linéarisation en 4 paliers de la température et $\beta_{k,hi}$, les paramètres entraînaibles. Les 4 paliers utilisés ont été déterminés dans les recherches de David et al. (David et al., 2023) pour s'ajuster aux pentes de la relation entre la température et la consommation au Québec.

$$C_{1h,d} = \begin{cases} 0 & T_{h,d} \leq 21 \\ T_{h,d} - 21 & 21 < T_{h,d} \leq 33 \\ 33 - 21 & T_{h,d} > 33 \end{cases} \quad (2.4)$$

$$C_{2h,d} = \begin{cases} 0 & T_{h,d} \leq 28 \\ T_{h,d} - 28 & 28 < T_{h,d} \leq 33 \\ 33 - 28 & T_{h,d} > 33 \end{cases} \quad (2.5)$$

$$H_{1h,d} = \begin{cases} 0 & T_{h,d} \geq 1 \\ 1 - T_{h,d} & -23 \leq T_{h,d} < 1 \\ 1 - -23 & T_{h,d} < -23 \end{cases} \quad (2.6)$$

$$H_{2h,d} = \begin{cases} 0 & T_{h,d} \geq 13 \\ 13 - T_{h,d} & -23 \leq T_{h,d} < 13 \\ 13 - -23 & T_{h,d} < -23 \end{cases} \quad (2.7)$$

où $T_{h,d}$ est la température en °C.

L'ajout des autres variables exogènes telles que le vent et la radiation se fera de la même manière que la température, mais sans le processus de linéarisation. La formule prend la forme suivante avec l'ensemble V des variables exogènes, $X_{i,h,d}$, et les paramètres, γ_{hj} :

$$f_{h,d}^X = \sum_{i=i}^V \gamma_{h1} X_{i,h,d} + \gamma_{h2} X_{i,h,d-1} + \gamma_{h3} X_{i,h,d-7} \quad (2.8)$$

L'entraînement du modèle consiste à calculer les coefficients $\alpha, \beta, \gamma, \phi$ et θ . Pour ce faire, une méthode itérative proposée par Spliid est utilisée (Spliid, 1983). Les coefficients de départ sont d'abord calculés à l'aide d'une régression linéaire et ensuite appliqués au calcul des résiduels. L'erreur obtenue est alors ajoutée aux intrants de l'itération suivante et ce procédé est répété jusqu'à l'obtention de la convergence des coefficients ou jusqu'à un nombre d'itérations maximum, dans notre cas à 100 itérations.

2.1.1 Description des expériences

La série d'expériences effectuée dans cette section a pour but d'évaluer l'ajout de l'humidité relative, le vent et la radiation solaire dans les intrants du modèle. Afin de bien isoler l'impact de ces ajouts, les expériences sont lancées de manière indépendante et les intrants météorologiques choisis proviennent d'une seule et même position géographique, soit à l'aéroport YUL à Montréal. De plus, tous les paramètres d'entraînement sont fixés, ainsi que le paramètre d'inertie qui considère la journée de prévision, la veille et la semaine précédente (jour 0, -1, et -7) pour toutes les variables météorologiques, et ce pour chacune des expériences.

Les différentes expériences ont utilisé les variables exogènes suivantes :

- CTRL : Température réelle (expérience de contrôle)
- EXP1 : Température ressentie
- EXP2 : Température réelle et vitesse du vent
- EXP3 : Température réelle et température du point de rosée
- EXP4 : Température réelle et nébulosité
- EXP5 : Température réelle et radiation solaire
- EXP6 : Température réelle, température du point de rosée, vitesse du vent et radiation solaire

Les expériences utilisent une période de 5 ans soit de 2012-01-01 à 2017-01-01 pour l'entraînement. La prévision journalière a ensuite été effectuée sur une période de 7 ans soit de 2013-01-01 à 2019-12-30 permettant une évaluation distincte de la performance du modèle sur des données vues et des données nouvelles.

La performance du modèle est calculée à l'aide de l'erreur MAPE (Mean Absolute Percentage Error), dont le résultat est en pourcentage.

$$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{|y_t - \hat{y}_t|}{y_t} \quad (2.9)$$

où y_t est la valeur réelle et \hat{y}_t est la valeur prédite. Dans le cas, du MAPE journalier la valeur de n sera de 24, soit une valeur par heure.

TABLEAU 2.1 Résultats du score MAPE journalier des expériences sur des données vues 2013-01-01 à 2017-01-01.

Expérience	Variables	Moyenne	Écart-type	Minimum	Maximum
CTRL	T	1.777	1.119	0.310	7.487
EXP1	Tress	1.740	1.030	0.289	7.465
EXP2	T, Vent	1.728	1.045	0.291	7.668
EXP3	T, Td	1.635	1.009	0.238	7.395
EXP4	T, Neb	1.608	0.949	0.258	6.611
EXP5	T, Rad	1.720	1.068	0.308	8.404
EXP6	T, Vent, Td, Rad	1.552	0.904	0.259	7.141

TABLEAU 2.2 Résultats du score MAPE journalier des expériences sur des nouvelles données 2017-01-01 à 2019-12-31.

Expérience	Variables	Moyenne	Écart-type	Minimum	Maximum
CTRL	T	1.846	1.175	0.326	8.503
EXP1	Tress	1.784	1.025	0.263	9.289
EXP2	T, Vent	1.819	1.085	0.306	8.027
EXP3	T, Td	1.757	1.180	0.309	16.136
EXP4	T, Neb	1.676	1.004	0.359	9.334
EXP5	T, Rad	1.789	1.123	0.317	8.952
EXP6	T, Vent, Td, Rad	1.674	1.045	0.314	14.835

Les tableaux 2.1 et 2.2 présentent les statistiques d'erreur des différentes expériences sur des données vues et non vues avec le nombre de variables météorologiques ajoutées. On remarque d'abord en se rapportant à l'expérience 1 de référence, que chacune des expériences a permis une amélioration du pourcentage d'erreur moyen autant sur les données vues que non vues. Il semblerait donc que l'ajout d'informations supposées par nos hypothèses initiales soit pertinent.

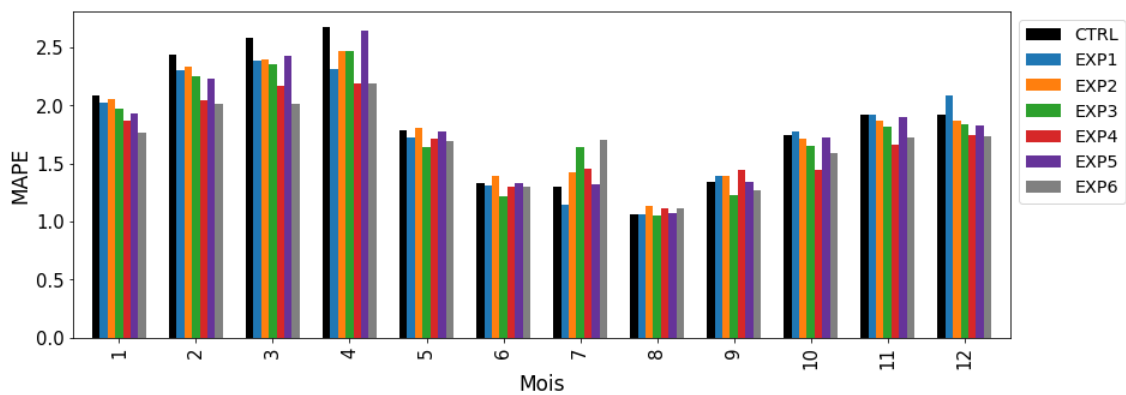


FIGURE 2.1 Score MAPE mensuel moyen pour chacune des expériences sur la période non-vue de 2017 à 2019.

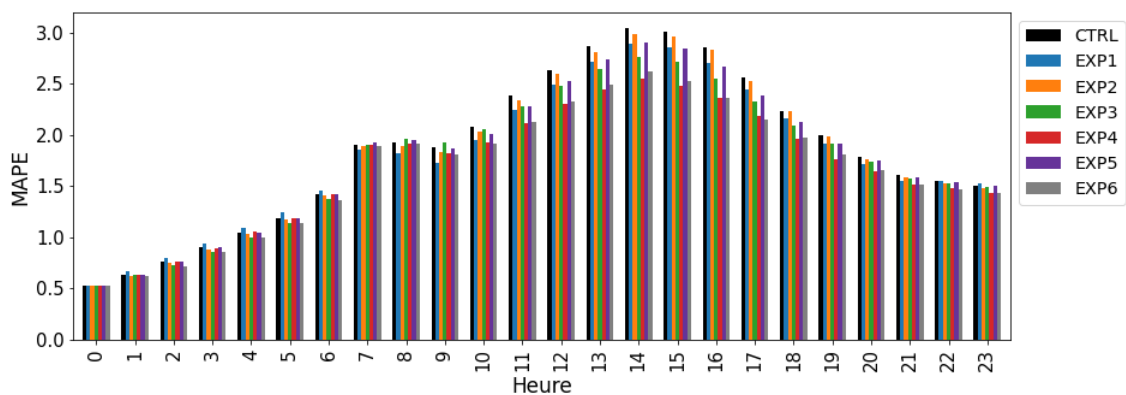


FIGURE 2.2 Score MAPE horaire moyen pour chacune des expériences sur la période non-vue de 2017 à 2019.

2.1.2 Analyse des expériences

Les résultats pour chacune des expériences sont détaillés dans la section suivante.

Expérience 1 : Impact de la température ressentie sur la prévision du modèle linéaire

L'expérience 1 contient la modification la moins coûteuse par rapport à la version initiale. Cette dernière n'ajoute aucune variable au modèle, il s'agit simplement d'un prétraitement apporté à la température. La température ressentie comprend l'effet de refroidissement éolien et de l'indice de chaleur. De cette manière, l'humidité relative est considérée lors des journées chaudes et le vent lors des journées froides. La combinaison de ces variables semble effectivement apporter un gain de performance soit 0.06% sur les données de test avec une faible variance.

Expérience 2 : Impact de la vitesse du vent sur la prévision du modèle linéaire

L'expérience 2 évalue l'apport direct de la vitesse du vent au modèle. La température réelle est toujours présente afin de conserver la cohérence avec le contrôle. Une étape de linéarisation est appliquée à la vitesse du vent avec un palier allant de 12 km/h à 39 km/h. On remarque une amélioration de la performance équivalente à l'expérience 1, mais inférieure sur les nouvelles données avec seulement 0.03%.

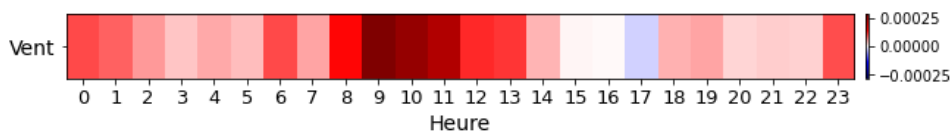


FIGURE 2.3 Coefficients calculés pour la vitesse du vent.

La figure 2.3 met en évidence la relation calculée par le modèle entre la vitesse du vent et le BQ. On voit un impact positif, soit une augmentation de la consommation avec le vent, et ce peu importe l'heure. On remarque également dans la figure 2.1 que cette relation amène de meilleures prévisions pour les mois froids, mais une détérioration pour les mois de mai à septembre. Ces résultats concordent avec un effet de refroidissement ressenti en présence de vent. La différence du score MAPE avec l'expérience précédente pourrait donc s'expliquer par le fait que le calcul du refroidissement éolien ne s'applique qu'aux températures inférieures à 10 °C. Il pourrait aussi s'agir d'une différence due à la présence de l'humidité relative dans la température ressentie, mais l'ajout de la vitesse du vent sans distinction de saisons semble insuffisant dans un modèle linéaire.

Expérience 3 : Impact de l'humidité sur la prévision du modèle linéaire

L'expérience 3 évalue l'apport de l'humidité sous forme de température de rosée. De la même façon que la température de l'air, la température de rosée est linéarisée selon les 4 paliers mentionnés précédemment. Toutefois, on remarque sur la figure 2.4 que la corrélation positive pour les journées chaudes n'est pas présente dans la température de rosée.

On remarque d'ailleurs une amélioration de précision par rapport au contrôle pour tous les mois de septembre à juin, mais une détérioration pour le mois de juillet et aucun impact en août. On peut en déduire que l'effet de refroidissement calculé à partir de la température de rosée est problématique, mais que l'effet de chauffage semble pertinent. Ce résultat s'oppose à l'hypothèse de départ selon laquelle la relation entre l'humidité relative et la consommation repose dans la sensation de chaleur en temps chaud.

Dans l'ensemble, cette expérience présente des résultats encourageants avec une

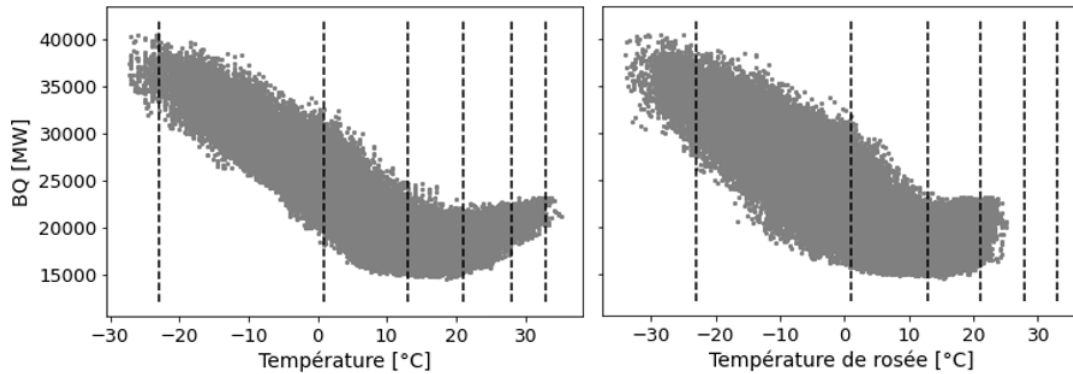


FIGURE 2.4 BQ en fonction de la température réelle à Montréal (gauche) et de la température du point de rosée (droite). Les lignes tiretées représentent les points de séparations des paliers de linéarisation.

amélioration du score MAPE moyen de 0.09%, mais elle présente une erreur maximale de 16.14% soit près de 2 fois l'erreur maximale du modèle de référence.

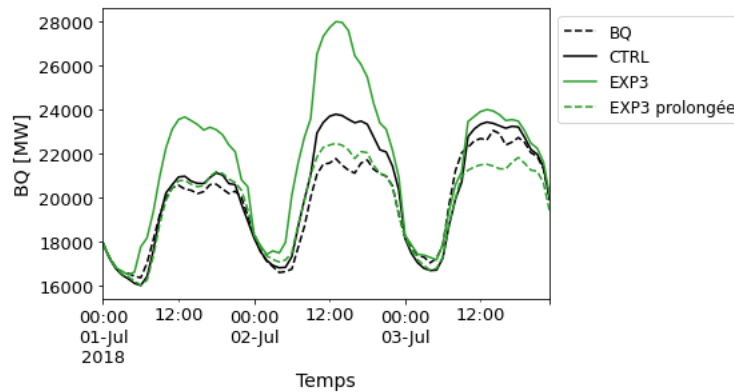


FIGURE 2.5 Valeurs de charge observées (noir tireté) et prévues par le modèle de contrôle (noir), avec la température de rosée (vert) et avec la température de rosée entraîné sur une période prolongée (vert tireté) pour les journées du 1er au 3 juillet 2018.

Cette erreur correspond à la journée du 2 juillet 2018. Si on regarde cette pé-

riode plus en détail sur la figure 2.5, on remarque qu'une erreur importante est déjà présente le 1er juillet 2018. On pourrait soupçonner une problématique due aux jours spéciaux, puisque le 1er juillet est un jour férié généralement dédié aux déménagements. Cela n'expliquerait toutefois pas l'écart avec le modèle de contrôle. La source du problème revient donc aux valeurs du nouvel intrant. Si on se concentre sur les conditions météorologiques à cette période, on constate que les températures sont largement au-dessus des normales (ECCC, 2023b) avec des températures supérieures à 35 °C et des températures de rosée de plus de 25 °C. Il s'agit d'ailleurs des valeurs maximales retrouvées dans notre échantillon de données pour la période complète, en entraînement et en validation. Il s'agit donc d'un cas extrême pour lequel le modèle n'a pas de comparatif. Une autre expérience a donc été effectuée toujours avec les mêmes intrants, mais en prolongeant l'entraînement sur la période d'été 2018. Lorsqu'on reproduit les prévisions pour ces premiers jours de juillet, on remarque alors une bien meilleure concordance des résultats. Cette expérience a bien sûr un net avantage puisque cette prévision est appliquée à des données connues, mais elle démontre tout de même qu'en élargissant la base de connaissances, il est possible de bien représenter ces situations avec la température de rosée en intrant.

Expérience 4 : Impact de la nébulosité sur la prévision du modèle linéaire

L'expérience 4 évalue l'apport de la nébulosité. La nébulosité ayant déjà fait ses preuves dans les modèles utilisés à Hydro-Québec, l'intérêt de cette expérience est principalement d'avoir un point de comparaison avec l'expérience 5 qui utilise la radiation solaire. Il est également intéressant de déterminer si l'impact semble lié à la luminosité ou à la température. Le processus de linéarisation de cette variable consiste à la diviser en 3 intervalles, soient [0-3], [3-10] et [9-10]. Les résultats sont meilleurs en général, mais on remarque que l'amélioration est maximale en

hiver et qu'une détérioration est présente pour les mois d'été. Cette distinction indique que l'information gagnée semble concerner davantage la température que la luminosité.

Si on regarde les coefficients calculés pour les paliers de nébulosité dans la figure 2.6, on peut y voir une relation positive lorsque la nébulosité est élevée (entre 9 et 10) pour les heures d'ensoleillement. Cette relation est aussi présente, mais moins prononcée pour la nébulosité moyenne (entre 3 et 10). Cette valeur indique que plus la couverture nuageuse est large, plus la consommation augmente, pendant le jour. À l'inverse, la relation est négative pour les heures de la nuit. Considérant l'écart de performance en fonction des saisons, cette logique semble cohérente avec l'impact sur la température. En effet, tel que décrit dans le chapitre 1, un couvert nuageux plus opaque laisse passer moins de rayonnement solaire, ce qui a pour effet de diminuer la température durant le jour et produit un effet de serre qui réchauffe la nuit. Ainsi la consommation augmenterait durant le jour en hiver et diminuerait pendant la nuit. Le comportement serait toutefois inverse pendant l'été, ce qui pourrait expliquer la baisse de performance du modèle pour les mois chauds. Une décomposition en fonction de la température serait donc pertinente dans le prétraitement de cette variable pour ce type de modèle.

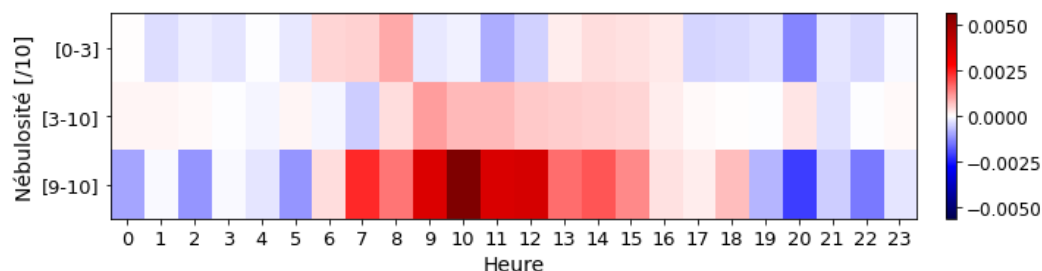


FIGURE 2.6 Coefficients calculés pour les trois paliers de nébulosité.

Expérience 5 : Impact de la radiation solaire incidente sur la prévision du modèle linéaire

L'expérience 5 évalue l'apport de la radiation solaire. La préanalyse de cette variable a permis de démontrer une corrélation inverse avec le BQ, mais uniquement à l'hiver. L'effet observable de cette variable s'explique probablement par le réchauffement des bâtiments en temps froid. En considérant cette information, un prétraitement a été appliqué aux données pour conserver seulement les valeurs lorsque la température est inférieure à 0 °C. De cette manière, l'effet de réchauffement est pris en compte uniquement dans l'effet de chauffage et ignoré dans l'effet de climatisation. Il est toutefois important de noter que contrairement à la couverture nuageuse qui peut persister, le rayonnement sera toujours nul pendant la nuit. On remarque donc l'amélioration des performances du modèle pour les heures d'ensoleillement seulement et les mois de novembre à avril. En comparaison avec l'expérience précédente, le gain de précision est moins important, mais aucune détérioration n'est présente pour les mois d'été.

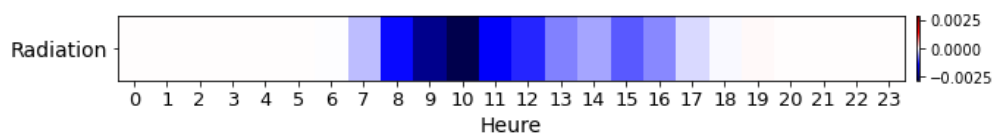


FIGURE 2.7 Coefficients calculés pour la radiation normalisés en fonction de la radiation solaire horaire moyenne.

Si on regarde les coefficients calculés par le modèle pour cette variable, une normalisation est nécessaire pour ne pas mal interpréter les forts coefficients présents au lever et au coucher du soleil. Ainsi en appliquant une normalisation avec les valeurs moyennes pour chaque heure, on obtient la relation présentée dans la fi-

gure 2.7, qui confirme la relation inverse attendue. En somme, on semble fournir les mêmes informations qu'à l'expérience précédente, mais la méthode de prétraitement doit être ajustée. On observe tout de même un gain de 0.06% avec l'ajout de cette nouvelle variable.

Expérience 6 : Impact des intrants météorologiques combinés sur la prévision du modèle linéaire

Pour compléter la première série, l'expérience 6 évalue l'impact de la combinaison des variables évaluées. Afin de combiner ces variables sans répéter l'information, la nébulosité et la température ressentie ont été ignorées. Ainsi, en combinant les expériences 2, 3 et 5 on obtient une amélioration du score MAPE de 0.22% sur les données vues et 0.17% sur de nouvelles données. Il est important de noter que l'exception rencontrée avec la température de rosée est à nouveau présente dans cette simulation puisque les périodes d'entraînement et de test sont inchangées. De plus, les erreurs produites pour les mois de juillet et août par l'ajout de la vitesse du vent et de la température de rosée sont reproduites. Toutefois, le gain important pour les autres mois permet de conclure que la combinaison de plusieurs variables peut être bénéfique.

2.2 Modèle neuronal

Pour compléter l'analyse des intrants météorologiques, une deuxième série d'expériences a été lancée avec un modèle neuronal. L'intérêt de cette deuxième approche est d'avoir la flexibilité d'apprendre des relations non linéaires. De plus, ce modèle a déjà fait ses preuves pour les prévisions opérationnelles de l'entreprise. Un inconvénient se pose toutefois dans la limite d'interprétabilité du réseau neuronal et sur ses capacités d'adaptation temporelle.

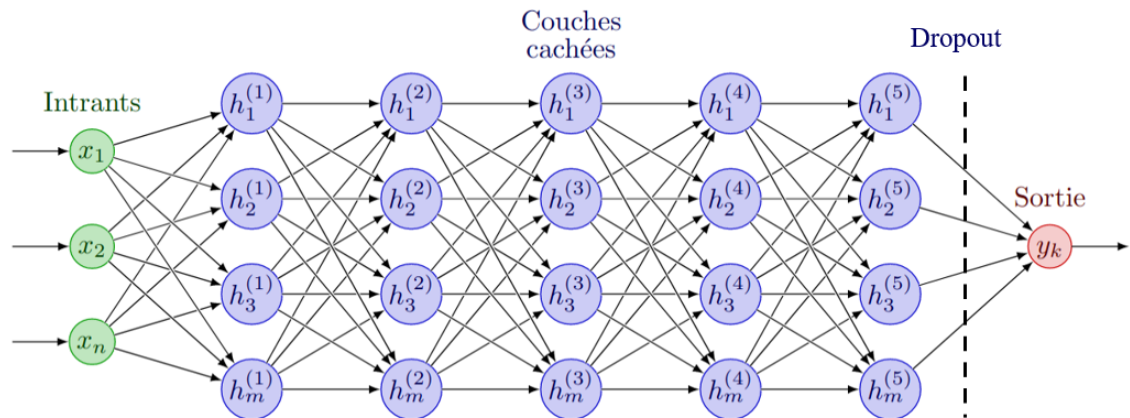


FIGURE 2.8 Architecture du modèle neuronal, avec n intrants (vert), m neurones par couche cachée (bleu) et k neurone en sortie (rouge).

L'architecture utilisée, telle que présentée dans la figure 2.8, est un perceptron multi-couches complètement connectés avec 5 couches cachées de 200 neurones et une couche de sortie composée d'un seul neurone. La prévision est faite pour chaque heure individuellement et utilise une inertie de 24 heures en plus de l'heure de prévision sur les intrants météorologiques. La fonction d'activation à l'intérieur du réseau est une tangente hyperbolique et une fonction linéaire pour le neurone de sortie. Le prétraitement appliqué aux données consiste à normaliser les données avec un encodage one-hot pour les variables catégorielles et un score Z pour les variables continues. L'encodage one-hot consiste à remplacer une variable catégorielle par n variables binaires où n correspond au nombre de catégories de la variable. Le score Z correspond à la différence entre une valeur et la moyenne de l'échantillon divisée par l'écart-type. Cette normalisation permet d'uniformiser les valeurs des différents intrants et de les centrer sur zéro pour faciliter l'entraînement du modèle.

En plus des intrants reliés à la météo, des variables temporelles sont ajoutées

telles que l'année, le mois, le jour, l'heure, le jour de la semaine et la période qui correspond à l'heure d'été ou l'heure avancée. Ainsi, avec l'encodage catégoriel, un total de 78 intrants non météorologiques est fourni au modèle (8 années + 12 mois + 31 jours + 24 heures + 2 périodes + 1 jour de la semaine).

L'entraînement se fait en lots de taille 32, c'est-à-dire que l'erreur utilisée dans la propagation arrière pour le calcul des poids est moyennée après l'évaluation de 32 échantillons différents. La fonction de perte utilisée pour ce calcul est l'erreur quadratique moyenne qui met l'accent sur les écarts plus larges produits par le modèle. Cette métrique est équivalente à celle utilisée dans l'entraînement du modèle linéaire présenté précédemment.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2 \quad (2.10)$$

où n est le nombre d'échantillons, \hat{y} la valeur prédite et y la valeur réelle.

L'ensemble des données d'entraînement est évalué à un maximum de 500 fois ou jusqu'à ce qu'il n'y ait plus d'amélioration pendant 50 itérations. La limite de 500 itérations a été fixée pour limiter le temps d'entraînement puisque le progrès obtenu au delà de ce nombre pour les tests préliminaires était négligeable. Des facteurs de régularisation sont ajoutés à chacune des couches cachées afin de recentrer les valeurs résultantes de celles-ci. Un filtre de type *dropout* est appliqué sur la dernière couche cachée avec un ratio de 0.4. Cette technique de régularisation consiste à "éteindre" certains neurones, c'est-à-dire de leur donner une valeur sortante nulle. Les valeurs résultantes de ceux-ci sont donc ignorées dans les calculs suivants pour un échantillon donné. Les neurones ignorés sont choisis aléatoirement à chaque itération. Cette pratique permet d'accélérer légèrement l'entraînement, mais surtout d'éviter le surentraînement. Pour finir, l'entraînement

est adapté par un optimiseur de type AdaGrad qui ajuste le taux d'apprentissage pour chaque paramètre en diminuant sa valeur avec le nombre d'itérations (Duchi *et al.*, 2011). Cette amélioration de la descente de gradient stochastique permet une convergence rapide en explorant l'espace des paramètres plus largement au départ puis en se spécifiant avec la progression de l'apprentissage.

Ce modèle est programmé dans le langage Python 3 et utilise les bibliothèques "scikit-learn" pour la normalisation des données et "keras" pour la construction et l'entraînement du RNA. L'entraînement prend environ 35 minutes pour les différentes expériences effectuées sur un processeur CPU 16 coeurs Intel Xeon Platinum 8375C @ 2.90GHz.

2.2.1 Description des expériences

En raison de la différente structure du modèle, l'expérience de référence comprend cette fois-ci les intrants catégoriels pour l'année, le mois, l'heure, le jour, le jour de la semaine et un indicatif pour l'heure d'été. Contrairement à l'approche statistique autorégressive, le besoin québécois des périodes passées n'est pas considéré comme intrant, ce qui retire la dépendance temporelle. Ces variables permettent au modèle d'apprendre les relations qui se rapportent au comportement humain habituel. Cependant, en incluant explicitement l'année dans les variables, les périodes d'entraînement (2012-2016) et de test (2017-2019) utilisées dans les expériences précédentes ne sont plus adaptées pour couvrir l'espace de possibilités lors de l'apprentissage. Pour cette raison, la séparation du jeu de donnée est faite aléatoirement dans l'intervalle complet avec une proportion de 60% pour l'entraînement, 20% pour la validation et 20% pour les tests. Cette séparation permet d'avoir des données distinctes pour entraîner le modèle, valider les résultats en cours d'entraînement et tester le modèle final. Pour avoir des comparaisons cohé-

TABLEAU 2.3 Scores MAPE journaliers moyens sur les données de test, de validation et d'entraînement pour les expériences réalisées avec le réseau de neurones.

Expérience	Variables	Test	Validation	Entraînement
CTRL	T	1.69	1.66	1.63
EXP1	Tress	1.75	1.76	1.7
EXP2	T, Vent	1.83	1.79	1.71
EXP3	T, Td	1.58	1.57	1.51
EXP4	T, Neb	1.56	1.55	1.47
EXP5	T, Rad	1.58	1.55	1.51
EXP6	T, Vent, Td, Rad	1.42	1.41	1.28

rentes, cette sélection est identique pour chacune des expériences effectuées dans cet ensemble. L'unique variation entre les expériences de cet ensemble est effectuée sur les variables météorologiques et celles-ci sont exactement les mêmes qu'à la série précédente.

Le tableau 2.3 présente le pourcentage d'erreur sur les jeux de données de test, de validation et d'entraînement. Les erreurs sont plus faibles pour les ensembles d'entraînement, avec des différences allant jusqu'à 0.14%, mais la diminution de la valeur de perte pour les données de validation jusqu'à la fin de l'entraînement ne semble pas suggérer de surentraînement.

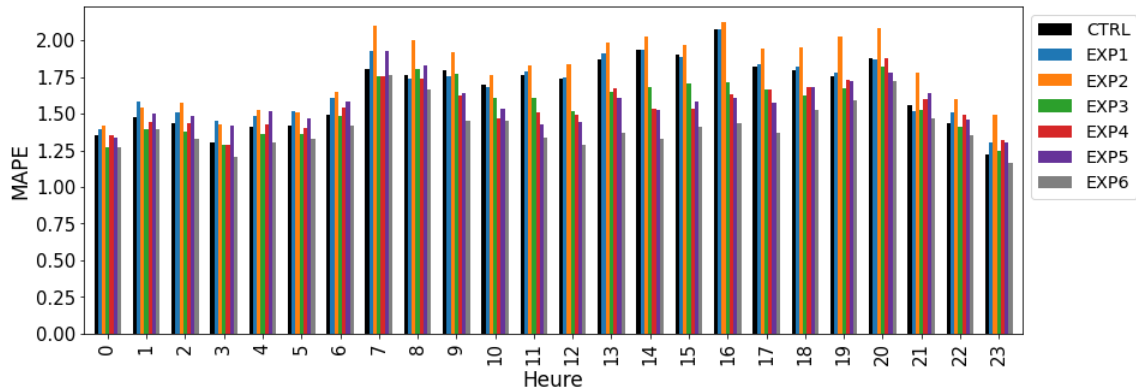


FIGURE 2.10 Score MAPE horaire moyen pour chacune des expériences sur les données de test de 2012 à 2019.

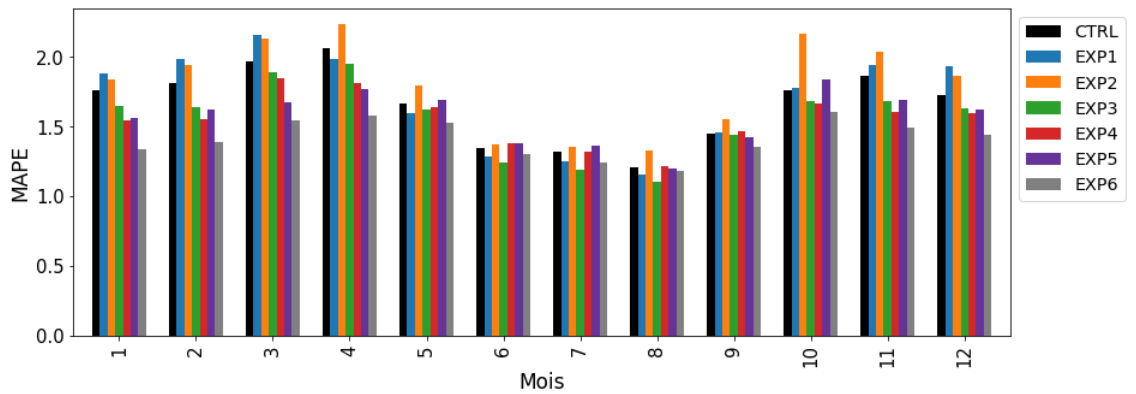


FIGURE 2.9 Score MAPE mensuel moyen pour chacune des expériences sur les données de test de 2012 à 2019.

2.2.2 Analyse des expériences

Pour interpréter les résultats obtenus pour ces expériences, une analyse supplémentaire est nécessaire pour relever les relations apprises par le modèle. Ainsi, les résultats et les valeurs de Shapley qui ont été calculées et comparées sont présentées dans la section suivante.

Valeurs de Shapley

Considérant la complexité du réseau de neurones multi-couches, une interprétation du comportement du modèle de prévision est plus complexe. Pour cette raison l'alternative proposée est de calculer les valeurs SHAP (SHapley Additive exPlanation) pour estimer au mieux les décisions prises par le modèle (Lundberg et Lee, 2017). Ces valeurs, développées dans la théorie des jeux, permettent d'évaluer la contribution des différents acteurs ou dans notre cas, des intrants dans l'atteinte d'un objectif. La méthode consiste à entraîner un modèle d'explication qui prend la forme d'une fonction linéaire de variables binaires. Le calcul est fait en évaluant les prévisions obtenues avec le modèle original avec et sans chacune des variables d'entrée et en moyennant les différences obtenues pour l'ensemble des combinaisons possibles. Cette approche assure une bonne représentation du modèle initial en respectant les propriétés de précision locale, d'absence et de cohérence.

1. Précision locale : le modèle explicatif doit être équivalent au modèle original pour un échantillon donné.
2. Absence : une variable absente dans les entrées originales ne doit avoir aucun impact dans l'explication.
3. Cohérence : si le résultat du modèle augmente avec une entrée donnée, l'attribution de cette entrée doit aussi augmenter.

Les valeurs SHAP, représentées par ϕ_i , pour chaque variable d'entrée i , sont calculées à l'aide de l'équation suivante :

$$\phi_i(f, x) = \sum_{z \subseteq x} \frac{|z|!(M - |z| - 1)!}{M!} [f_x(z) - f_x(z \setminus i)] \quad (2.11)$$

où

- x correspond à un échantillon de l'ensemble de données.
- $z \in 0, 1^M$ indique les combinaisons d'entrées considérées dans l'échantillon x . $z \setminus i$ équivaut à $z_i = 0$.
- $|z|$ le nombre d'entrées considérées (non nulles).
- M le nombre d'entrées par échantillon.
- f la fonction de prévision du modèle entraîné. $f_x(z)$ donne donc la prévision obtenue pour l'échantillon x , lorsque la combinaison z est activée.

Expérience 1 : Impact de la température ressentie sur la prévision du modèle neuronal

L'expérience 1 qui ajuste la température avec le facteur de refroidissement et l'indice de chaleur présente une détérioration du résultat par rapport au contrôle. Cet effet est présent autant sur les données vues que non vues. On remarque sur la figure 2.9 que l'erreur est plus importante pour les mois d'octobre à mars, mais une légère amélioration est présente sur les mois d'été. Comme le facteur de refroidissement est appliqué uniquement aux températures inférieures à 10 °C, il est majoritairement absent sur les mois d'été. La figure 2.11 présente les plus importantes sous-estimation et surestimation commises par le modèle. On remarque que l'ajustement par rapport au contrôle est dans la bonne direction, mais la correction est trop grande. Ainsi, avec un grand degré de liberté pour ajuster le refroidissement éolien, l'impact du vent est surestimé.

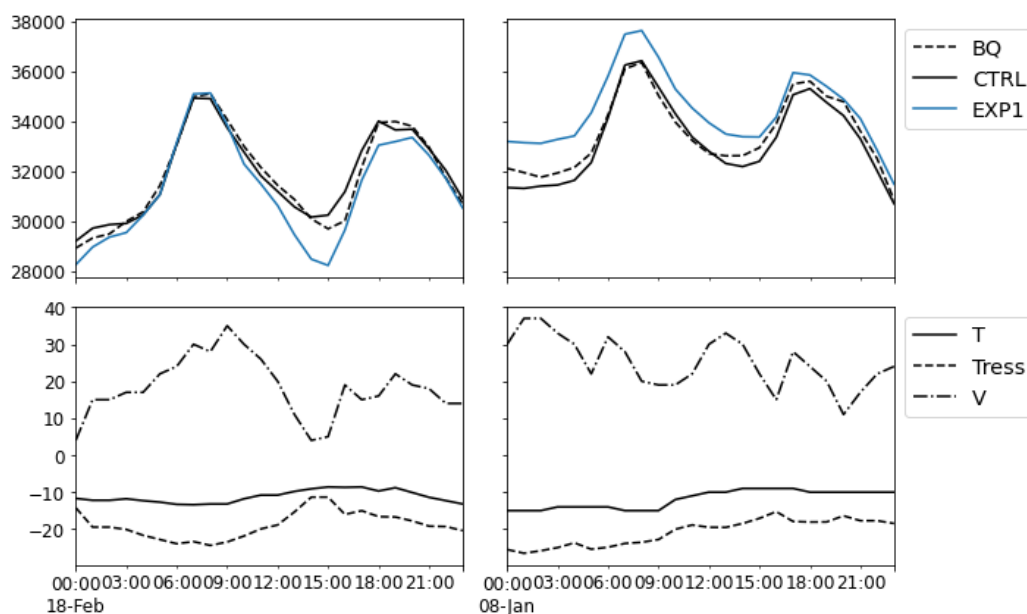


FIGURE 2.11 Consommation énergétique réelle (BQ) et prévue (CTRL,EXP1) dans les panneaux supérieurs et conditions météorologiques de température (T), température ressentie (Tress) et vitesse du vent (V) dans les panneaux inférieurs pour les journées du 18 février 2019 (gauche) et 8 janvier 2014 (droite)

On suppose donc que la problématique se trouve au niveau de ce facteur. Pour vérifier cette hypothèse, deux simulations ont été lancées pour évaluer séparément la température avec l'indice de chaleur et la température avec le refroidissement éolien. Les scores MAPE obtenus pour ces deux tentatives sont respectivement de 1.71% et 1.79%, ce qui est supérieur au contrôle dans les deux cas. On peut donc en déduire que les indices de température ressentie ne sont pas de meilleurs indicateurs du besoin énergétique.

Expérience 2 : Impact de la vitesse du vent sur la prévision du modèle neuronal

L'expérience 2 présente un score MAPE de 1.83%, soit 0.14% de détérioration. La vitesse du vent est une variable qui est déjà utilisée dans la version opérationnelle. Toutefois, lorsqu'elle est ajoutée à la température, la précision est diminuée. L'impact négatif se manifeste principalement durant les mois d'octobre et d'avril, mais est présent pour tous les mois et toutes les heures de la journée. Puisque l'erreur semble étendue à l'ensemble de la période, la source du problème pourrait se trouver dans les configurations des paramètres de la simulation. Cette étude visant à comparer l'impact des intrants météorologiques à architecture fixée, les paramètres d'entraînement se doivent de rester les mêmes, mais il est possible que la taille du réseau soit moins adaptée à l'apprentissage de l'impact du vent et que l'ajout de couches cachées puisse être bénéfique. Une autre hypothèse serait que l'inertie de 24h ne soit pas nécessaire pour cette variable. Il est vrai qu'une journée venteuse augmente les échanges convectifs avec les bâtiments et l'air extérieur, mais le temps de réponse des systèmes de climatisation ne devrait pas être considérable. Ainsi un test a été effectué en retirant l'inertie de la vitesse du vent uniquement. Le score MAPE des données de test obtenu est alors de 1.64% soit 0.19% mieux qu'avec l'inertie et 0.05% mieux que le contrôle. La figure 2.12 présente les plus importantes sous-estimation et surestimation commises par le modèle. On remarque que la version sans inertie n'aggrave pas la prévision, mais ne l'améliore pas non plus dans ces situations.

Si on s'intéresse à la contribution des variables météorologiques selon la valeur de Shapley, la figure 2.13 présente les entrées de température et de vitesse du vent avec leur inertie. La vitesse du vent, V , dans l'expérience 2 initiale, présente une plus faible participation avec une valeur de 0.0068 de contribution sur la valeur sortante normalisée du modèle. Les valeurs pour les pas de temps précédents consi-

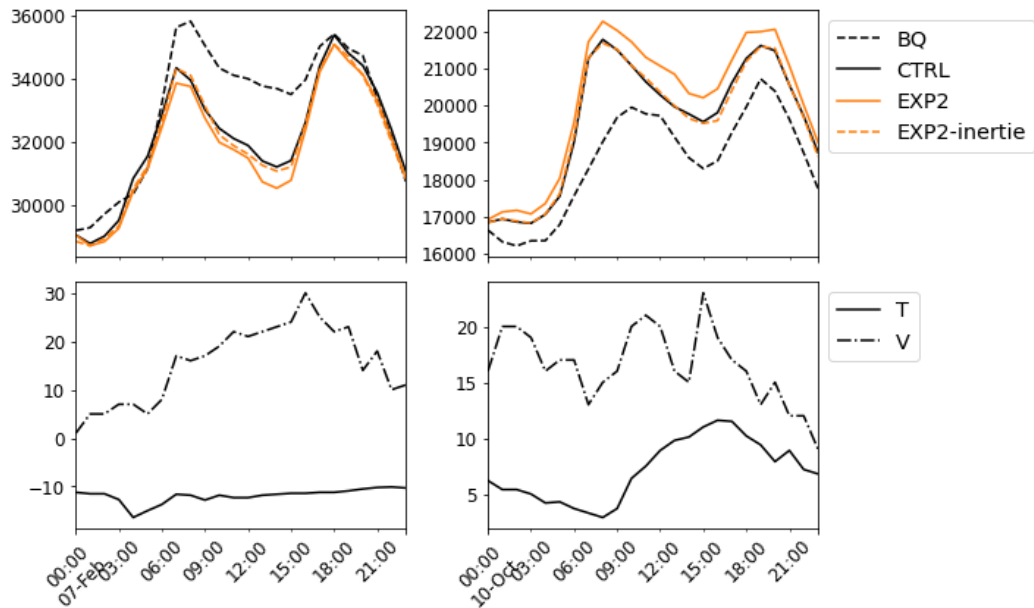


FIGURE 2.12 Consommation énergétique réelle, BQ, et prévue, CTRL et EXP2 avec et sans inertie du vent (EXP2-inertie), dans les panneaux supérieurs. Conditions météorologiques de température (T) et de vitesse du vent (V) dans les panneaux inférieurs pour les journées du 7 février 2018 (gauche) et 10 octobre 2016 (droite)

dérés dans l'inertie du vent sont moindres. En retirant l'inertie, on voit cependant l'impact de la valeur de vent restante augmenter jusqu'à 0.0160. On remarque également que la contribution des premiers pas de temps de la température est augmentée. Outre les 6 premières heures d'inertie, l'effet de la température appris par les deux versions de l'expérience est globalement similaire et met l'accent sur les pas de temps présent, de 1 heure et de 24 heures avant.

La figure 2.13 présente également les valeurs de Shapley de ces variables pour les deux journées problématiques mentionnées précédemment. La première situation est une journée d'hiver avec des vents modérés. L'ajustement positif avec le vent qui serait attendu est manqué par le modèle. On voit que la température a un

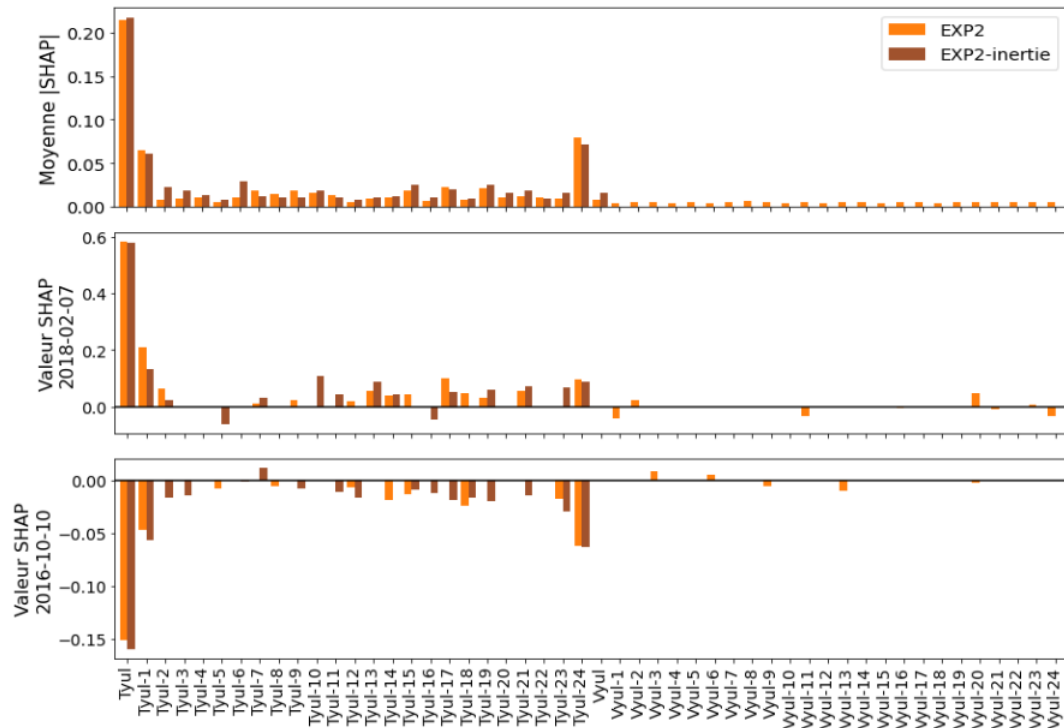


FIGURE 2.13 Moyenne de la valeur absolue des valeurs de Shapley sur 400 échantillons pour les entrées de température (Tyul) et de vitesse du vent (Vyul) au temps de prévision et au pas de temps d'inertie. Valeur de Shapley pour les journées du 7 février 2018 et 10 octobre 2016.

impact positif sur le résultat, avec une amplitude équivalente à la version sans inertie pour le temps de prévision. Les deux versions ne partagent pas le même impact de la part de l'inertie de température qui est opposée pour certains pas de temps. La deuxième situation est une journée d'automne avec des températures supérieures au point de congélation et des vents faibles à modérés. Dans ces conditions, le vent ne devrait pas jouer un rôle important, mais on voit que le modèle s'éloigne considérablement de la vraie valeur. Le graphique confirme une contribution globalement nulle de la part des entrées de vitesse du vent. La température en tout point indique plutôt une baisse de la consommation, mais la

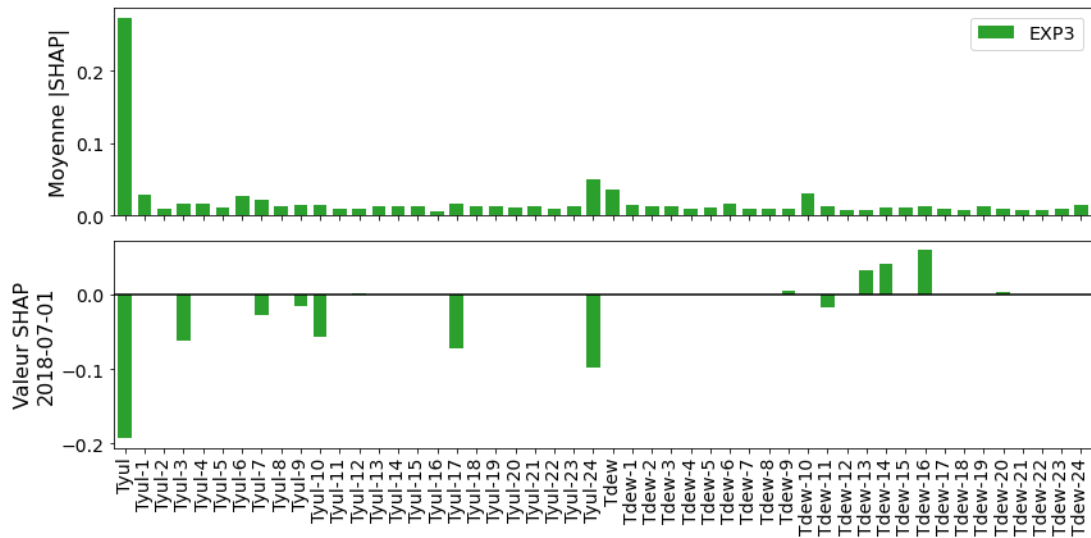


FIGURE 2.14 Moyenne de la valeur absolue des valeurs de Shapley sur 400 échantillons pour les entrées de température (Tyul) et de température de rosée (Tdew) au temps de prévision et au pas de temps d’inertie. Valeur de Shapley pour la journée du 1 juillet 2018.

contribution négative est plus faible que pour le modèle sans inertie de vent. On en déduit donc que le problème ne vient pas directement de la variable de vent, mais plutôt des paramètres appris pour les autres variables en présence du vent. L’ajustement apporté en retirant l’inertie semble permettre de retirer ce biais.

Expérience 3 : Impact de l’humidité sur la prévision du modèle neuronal

L’expérience incorporant la température du point de rosée présente une nette amélioration, soit 0.11% par rapport au point de référence, et ce autant pour les mois d’hiver que d’été. Ce gain est particulièrement remarquable pour les heures de l’après-midi et relativement uniforme sur le reste de la journée.

Les apports des variables météorologiques sont indiqués sur la figure 2.14. La

moyenne de la valeur absolue des valeurs de Shapley, sur le premier panneau, montre une contribution notable de la température du point de rosée, avec des amplitudes plus marquées, pour les pas de temps de 0 et 10 heures avant la prévision. Cette contribution est en moyenne positive, donc aura pour effet d'augmenter la consommation. Si on s'intéresse à la journée du 1er juillet 2018 qui était plus difficile avec le modèle linéaire en raison des conditions de chaleur et d'humidité les plus élevées de l'échantillon de donnée, on obtient cette fois-ci une erreur MAPE journalière de 0.96%. Les valeurs de Shapley pour cet événement, visible sur la figure 2.14, montrent un effet positif provenant de la température de rosée, mais négatif pour la température réelle. Cette direction étant relative à la valeur résultante de base concorde avec l'intuition puisque la climatisation consommera beaucoup moins que le chauffage. Cela dit, la distribution de l'impact dans les pas de temps d'inertie est difficilement interprétable. On retrouve tout de même la plus forte participation provenant de la température à 0 et 24 heures de décalage. Il est important de noter que les valeurs passées au modèle n'ont pas d'indice temporel selon ce dernier, mais dépendent les unes des autres. De plus, le modèle prend un total de 128 entrées avec l'encodage des variables catégoriques et l'ajout de l'inertie des variables continues. Ainsi l'espace de possibilité est beaucoup plus large que l'échantillon d'entraînement. Une hypothèse qui explique la distribution temporelle des valeurs de Shapley serait donc que l'apprentissage des poids accordés aux différents pas de temps soit biaisé par la limite d'échantillons distincts.

Un dernier test qui peut être effectué avec l'humidité est de fournir l'humidité relative plutôt que la température de rosée au modèle. Toujours en conservant l'inertie de 24h les erreurs obtenues sont cette fois-ci de 1.57%, 1.56% et 1.49% pour les données de test, validation et entraînement. Ces résultats sont presque équivalents à ceux obtenus avec la température de rosée, indiquant que le modèle

peut déduire la même information de ces 2 variables dépendantes.

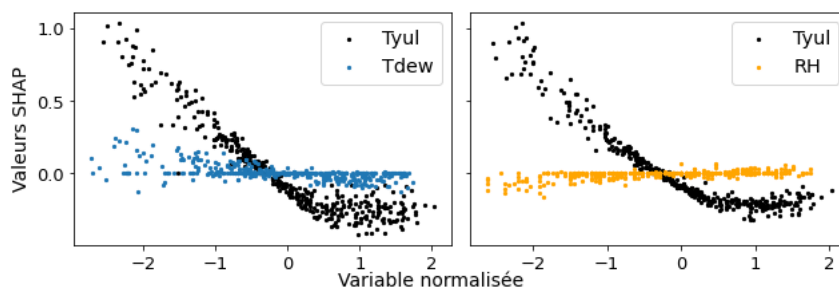


FIGURE 2.15 Contribution des intrants de température et d'humidité sur la prévision de 400 échantillons distribués aléatoirement sur l'ensemble de données. Les valeurs des intrants et de la sortie sont normalisées.

La figure 2.15 présente les valeurs SHAP associées aux intrants de température et d'humidité pour les deux expériences. On voit que le modèle apprend différemment, soit une relation négative pour la température de rosée avec une variance relativement prononcée, alors que l'humidité relative présente une relation linéaire faiblement positive dont la variance est plus limitée. On peut tout de même remarquer que les valeurs SHAP de la température réelle sont affectées par l'ajout de la variable d'humidité. La relation illustrée dans la figure 1.7 est bien représentée par le modèle dans les deux situations, mais la variance est plus faible pour la simulation avec l'humidité relative. En somme, on peut déduire que l'humidité relative est une bonne alternative et que l'impact appris pour cette dernière est une augmentation de la consommation avec l'humidité relative.

Expérience 4 : Impact de la nébulosité sur la prévision du modèle neuronal

L'évaluation de la nébulosité est très similaire aux résultats obtenus avec le modèle linéaire, soit une amélioration marquée pour les mois d'hiver et peu d'impact, voire une légère détérioration pour les mois d'été. La différence d'erreur est plus

importante dans les heures d'ensoleillement, ce qui concorde avec la période d'utilisation d'éclairage, mais aussi de réchauffement par radiation. L'effet nocturne de la nébulosité ne semble pas avoir été capté. Globalement, le score MAPE est réduit de 0.13% par rapport au contrôle grâce à l'ajout de cette variable. La distribution de l'impact moyen des variables est présentée sur la figure 2.16. La contribution de la nébulosité en soi est relativement faible, mais les poids des variables principales sont légèrement modifiés par l'ajout de cet intrant.

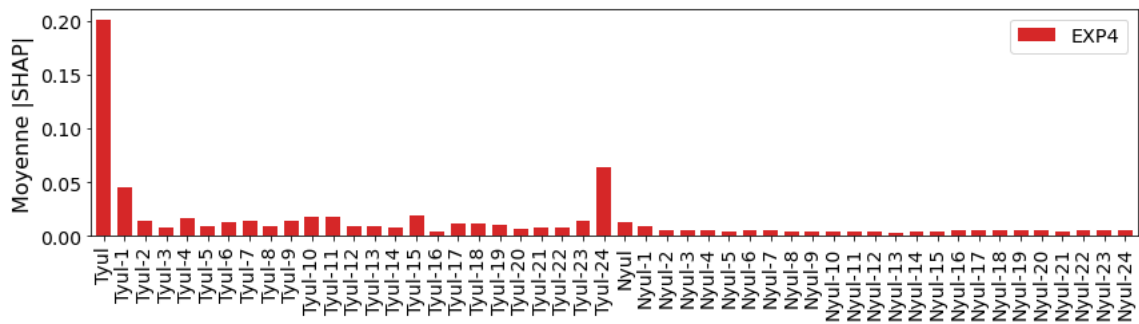


FIGURE 2.16 Moyenne de la valeur absolue des valeurs de Shapley pour les variables de température (Tyul) et de nébulosité (Nyul) au temps de prévision et au pas de temps d'inertie.

Expérience 5 : Impact de la radiation solaire incidente sur la prévision du modèle neuronal

La radiation, qui contient en principe les mêmes informations que la nébulosité durant le jour, semble se transmettre légèrement moins bien pour les mois moins froids et les heures matinales. Évidemment, la période nocturne ne reçoit aucune radiation solaire donc la valeur de l'intrant pour ces heures est de 0. Le fait que l'erreur augmente pour certaines de ces heures provient donc de l'apprentissage des poids pour les autres variables. Il aurait également été possible qu'elle soit due à l'inertie de la radiation, mais une tentative a été lancée en retirant l'inertie

de la radiation et les résultats sont comparables.

En regardant la contribution des variables selon la valeur de Shapley sur la figure 2.17, on voit que la radiation aux pas de temps de 0, 1, 13 et 14 heures est significativement considérée. Lorsqu'on compare aux valeurs calculées pour l'expérience de la nébulosité, on constate cependant que la contribution de la température réelle est rendue moins importante.

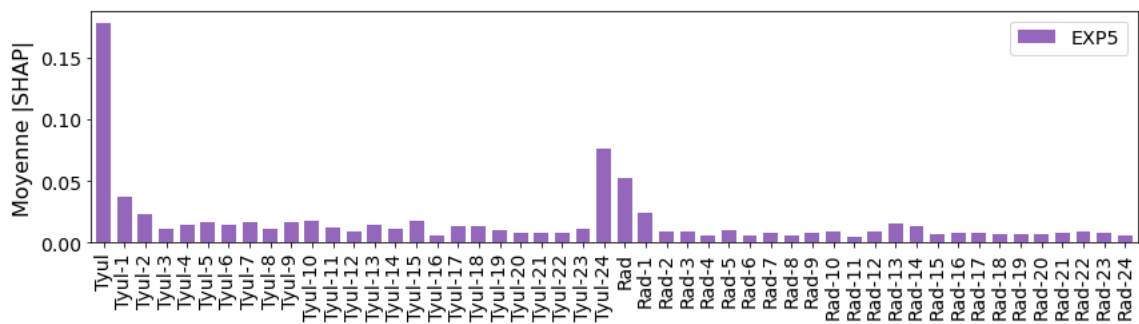


FIGURE 2.17 Moyenne de la valeur absolue des valeurs de Shapley pour les variables de température (Tyul) et de radiation (Rad) au temps de prévision et au pas de temps d'inertie.

Expérience 6 : Impact de la combinaison des intrants météorologiques sur la prévision du modèle neuronal

Considérant les résultats des précédentes expériences, la combinaison du vent, de la température de rosée et de la radiation avec inertie de 24h pourrait ne pas être optimale. Le score MAPE est toutefois grandement diminué avec une erreur moyenne de 1.42% sur les données de test et 1.28% sur celles d'entraînement. Le gain de précision est valide en toutes saisons, mais plus marqué sur les mois d'hiver. Le contraste d'erreur entre les saisons est ainsi diminué. Pour la répartition journalière, l'heure de pointe matinale est toujours difficile, mais les heures d'ensoleillement suivantes sont améliorées.

Pour compléter l'étude de la combinaison des variables, les variations suivantes ont été testées :

1. Retrait complet du vent
2. Retrait de l'inertie du vent
3. Remplacement de la température de rosée par l'humidité relative
4. Remplacement de la radiation par la nébulosité
5. Remplacement de la radiation et de la température de rosée par la nébulosité et l'humidité relative
6. Remplacement de la radiation et de la température de rosée par la nébulosité et l'humidité relative et retrait de l'inertie du vent

Chacune de ces tentatives obtient un score très similaire à la première combinaison, mais aucune ne réussit à le réduire sur l'échantillon de test. On voit tout de même que la combinaison qui néglige l'inertie du vent est pratiquement équivalente en utilisant 24 intrants (vent des 24 dernières heures) de moins. On peut donc gagner en performance sans perte de précision significative.

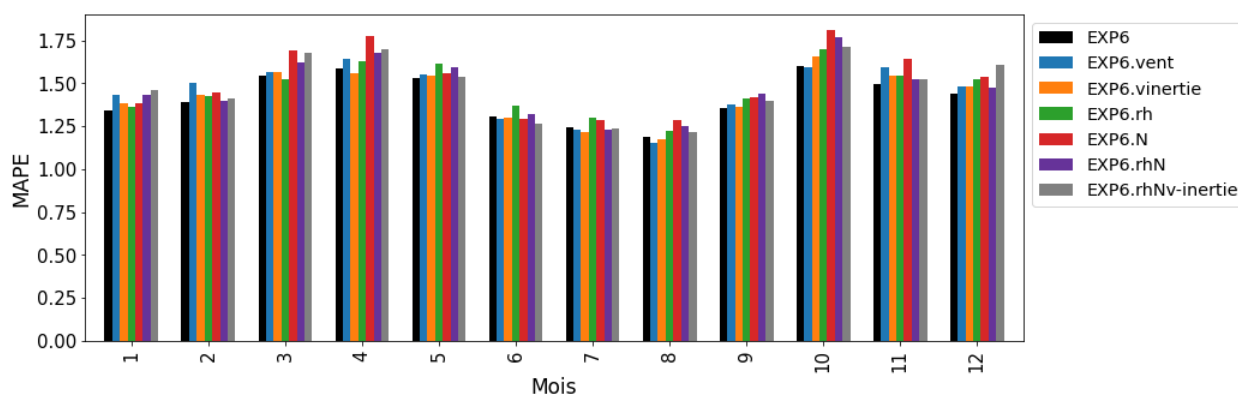


FIGURE 2.18 Score MAPE mensuel moyen des différentes combinaisons sur les données de test de 2012 à 2019. Description de la légende, EXP6 : combinaison de l'expérience 6 initiale, EXP6.vent : retrait du vent, EXP6.vinertie : retrait de l'inertie du vent, EXP6.rh : remplacement de la température de rosée par l'humidité relative, EXP6.N : remplacement de la radiation par la nébulosité, EXP6.rhN : remplacement de la radiation et de la température de rosée par la nébulosité et l'humidité relative et EXP6.rhNv-inertie : remplacement de la radiation et de la température de rosée par la nébulosité et l'humidité relative et retrait de l'inertie du vent

CHAPITRE III

ÉVALUATION DE LA RÉPARTITION SPATIALE DES INTRANTS

Les chapitres précédents ont évalué la corrélation entre les variables météorologiques à Montréal et le BQ. Toutefois, le BQ correspond à la quantité d'énergie nécessaire pour l'ensemble du Québec. La province qui couvre un large territoire a des conditions météorologiques qui peuvent grandement varier selon l'emplacement. L'est de la province sera plus affecté par la masse d'air maritime, qui est plus humide que la masse d'air continentale à l'ouest. De la même manière le sud reçoit les échanges avec la masse d'air tropicale plus chaude provenant des États-Unis alors que le nord interagit avec la masse arctique qui est beaucoup plus froide (Ahrens, 2015). Il est donc peu probable que la température à Montréal à elle seule suffise à représenter adéquatement la province en entier. Pour cette raison, les modèles actuels utilisent 5 stations météorologiques couvrant principalement les grandes villes qui ont été choisies en fonction de leur disponibilité et de l'intuition qu'elles contiendraient de l'information complémentaire. La performance du modèle a été évaluée lors de l'ajout de chacune de ces stations et une diminution de l'erreur leur accordait une place dans la liste des nouveaux intrants. Cette méthode a permis l'amélioration de la précision des résultats, mais a laissé peu de place à l'exploration de combinaisons potentiellement meilleures. Les 5 stations actuellement considérées par le modèle opérationnel sont Montréal, Québec, Baie-Comeau, Sherbrooke et Ottawa.

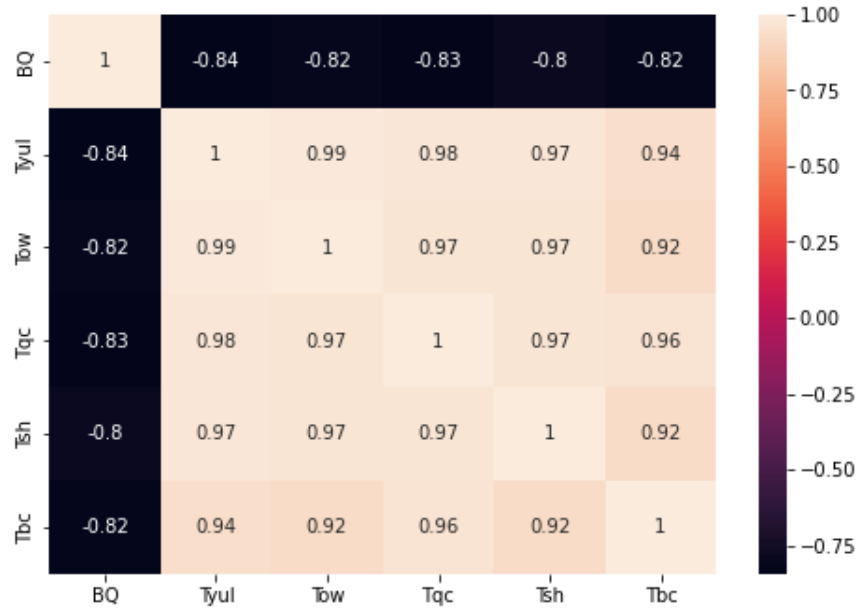


FIGURE 3.1 Matrice de corrélation du BQ et de la température aux 5 stations initiales, Montréal (yul), Ottawa (ow), Québec (qc), Sherbrooke (sh) et Baie-Comeau (bc).

Bien que chacune de ces stations présente des informations différentes, elles ne sont bien entendu pas indépendantes. On remarque tout de même que la corrélation avec le BQ est maximale avec la température de Montréal. Sachant que le lien entre la température et l'énergie consommée est fait lorsqu'il y a des bâtiments à tempérer, on peut associer une importance à la température mesurée en un endroit en fonction de la population de cette région. Ce raisonnement pourrait justifier les valeurs plus élevées pour Montréal dont la concentration de population est majoritaire, suivi par Québec.

3.1 Exploration des stations

En élargissant l'espace de possibilités pour l'exploration spatiale des différentes stations du Québec, on se retrouve rapidement contraint par l'aspect combinatoire du problème. Pour limiter le nombre de configurations, des hypothèses de départ sont posées. On suppose donc, comme mentionné précédemment, que la taille de la population aura un impact sur l'importance d'une station. On suppose également que les stations distantes entre elles seront moins dépendantes et plus complémentaires. Ainsi, pour vérifier ces hypothèses, un ensemble de 15 stations, présenté sur la figure 3.2, a été sélectionné pour couvrir la province et sa population.

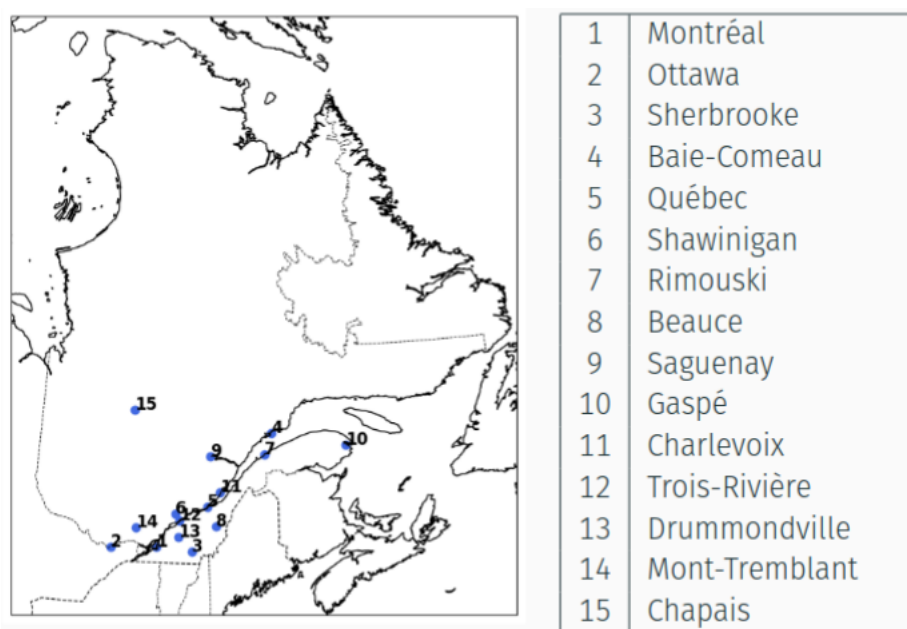


FIGURE 3.2 Positionnement géographique des 15 stations météorologiques sélectionnées pour l'étude combinatoire.

Pour connaître la pertinence d'une station, l'évaluer en présence de chacun des autres points de l'ensemble semble nécessaire puisque l'information peut être in-

complète ou redondante dans un cas, mais complémentaire dans un autre. Ainsi le nombre de combinaison d'intrants de température pour cet ensemble se calcule à l'aide du coefficient binomial.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (3.1)$$

où n correspond à la taille de l'ensemble et k la taille des sous-ensembles uniques tirés dans n . En suivant cette équation, le nombre de simulations à effectuer en évaluant des sous-ensembles de 4 stations monte à 1 365 et à 3 003 avec 5 stations. Considérant un temps d'entraînement et de simulation allant de 2 à 7 minutes en fonction du nombre de stations pour une prévision s'étalant sur 30 jours, une énumération exhaustive n'est pas envisageable. Ainsi pour réduire le temps de calcul, les tests ont été limités au mois de janvier pour l'année 2015 et 2017 totalisant 60 jours de simulation dont l'entraînement est ajusté à l'année. Le choix de cette période est justifié par de plus importantes erreurs produites par les modèles en hiver. On considère également uniquement les sous-ensembles de taille inférieure à 5, ce qui est suffisamment large pour répondre aux hypothèses de départ.

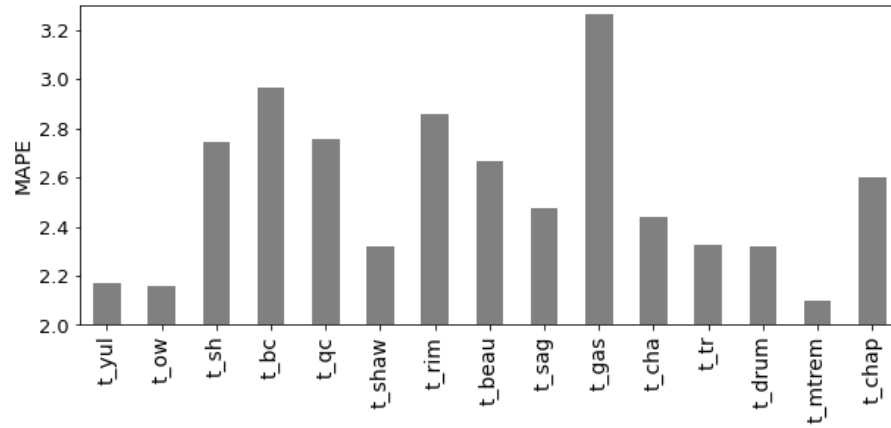


FIGURE 3.3 Score MAPE moyen pour la période de janvier 2015 et janvier 2017 obtenu avec chaque station météorologique individuelle.

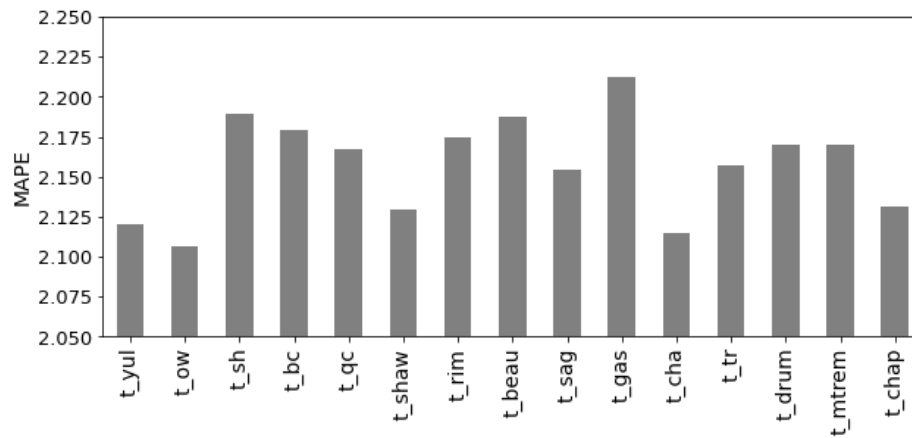


FIGURE 3.4 Score MAPE de chaque station météorologique moyenné pour toutes les combinaisons testées pour la période de janvier 2015 et janvier 2017.

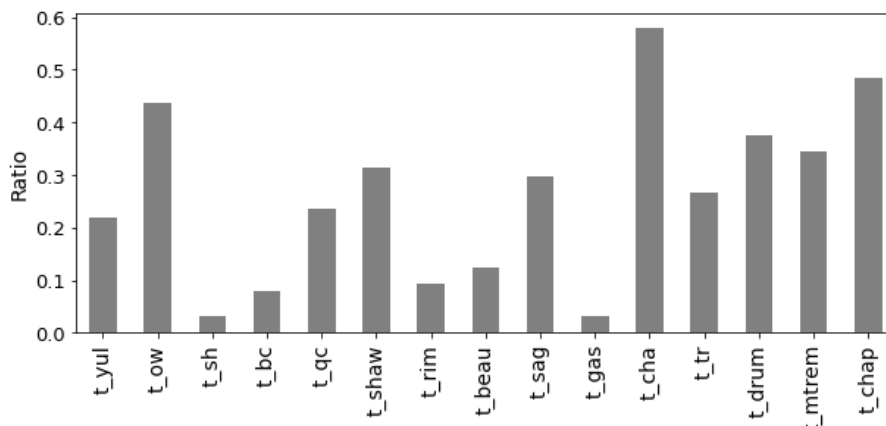


FIGURE 3.5 Ratio d'occurrence de chaque station parmi les combinaisons ayant obtenues un score MAPE inférieur à 1.95.

Les résultats de la série de simulations de chaque sous-ensemble sont illustrés sur les figures 3.3, 3.4 et 3.5. La figure 3.3 présente d'abord le score MAPE obtenu en présence d'une seule température pour chaque station de l'ensemble. On remarque qu'individuellement, les températures génèrent des erreurs très variables allant de 2.09% pour Mont-Tremblant à plus de 3.2% pour Gaspé. Les régions de Montréal et Ottawa obtiennent de bons résultats contrairement à Québec et Sherbrooke dont la concentration de population est tout de même importante. La figure 3.4 montre ensuite le score moyen pour l'ensemble des simulations dans lesquelles la station est présente. Ce score moyen permet d'avoir un aperçu de la performance générale d'une station. On y retrouve de meilleurs résultats pour les stations de Montréal, Ottawa, Shawinigan, Charlevoix et Chapais, mais Sherbrooke et Québec sont faibles à nouveau. Pour finir, la figure 3.5 montre le ratio d'occurrence de chaque station parmi les meilleurs sous-ensembles, soit ceux dont l'erreur MAPE obtenue est inférieure à 1.95%. La température de Québec se retrouve avec un ratio de 25%, ce qui est légèrement supérieur au pourcentage de Montréal indiquant que, lorsque bien combinée, cette variable peut améliorer la précision du modèle. Les

régions les plus fréquentes sont toutefois Ottawa, Charlevoix et Chapais malgré leur faible taux de population.

Les meilleures combinaisons résultantes de cette évaluation obtiennent un score MAPE moyen d'environ 1.87%.

- Ottawa, Shawinigan et Charlevoix
- Montréal, Saguenay, Charlevoix et Drummondville
- Charlevoix, Drummondville, Mont-Tremblant et Chapais

Ces résultats semblent indiquer que la taille de la population n'est pas à elle seule un facteur suffisant pour déterminer l'importance d'une station dans le calcul de prévision de la demande énergétique. La dispersion géographique semble toutefois être une condition pertinente, mais les régions à l'est de la province n'ont que très peu de poids.

3.2 Performances du modèle neuronal

L'étape suivante de l'analyse consiste à évaluer la performance du modèle neuronal avec les sous-ensembles de température déterminés précédemment. Pour ce faire, les paramètres utilisés dans la section 2.2 sont conservés, permettant une comparaison avec les dernières expérimentations. La seule modification apportée est le choix des intrants météorologiques qui, pour l'évaluation spatiale, est limité à la température. Une nouvelle expérience de contrôle est proposée pour identifier l'objectif de performance en utilisant les 5 températures actuellement présentes dans le modèle opérationnel.

TABLEAU 3.1 Scores MAPE journaliers des simulations sur les ensembles de test, validation et entraînement en fonction du sous-ensemble de température passé en intrant.

Stations	Nb stations	Test	Validation	Entraînement
1	1	1.69	1.66	1.63
1, 2, 3, 4, 5	5	1.54	1.53	1.47
2, 6, 11	3	1.55	1.55	1.49
1, 9, 11, 13	4	1.55	1.53	1.47
11, 13, 14, 15	4	1.59	1.59	1.50

La première entrée du tableau 3.1 est un rappel du résultat obtenu avec la station de Montréal seulement afin de pouvoir comparer. Le deuxième ligne du tableau indique que l'objectif de performance à atteindre ou dépasser pour cette étape est de 1.54% sur l'ensemble de test. Les 3 combinaisons testées obtiennent des scores supérieurs à ce contrôle, mais il est pertinent de voir un résultat presque équivalent pour les 2 premiers sous-ensembles (2,6,11 et 1,9,11,13) avec 1.55% sur les données de test malgré un nombre de stations inférieur. Avec une inertie de 24 heures, chaque station correspond à 25 valeurs entrantes pour le modèle. Ainsi, une diminution de 2 stations implique le retrait de 50 entrées seulement pour la température, représentant un gain d'efficacité. La combinaison d'Ottawa, Shawinigan et Charlevoix semble donc une alternative intéressante. Ces résultats suggèrent que l'utilisation de 5 stations n'est peut-être pas nécessaire et qu'il est possible d'optimiser l'information avec seulement 3 stations.

Un autre moyen de déterminer l'importance d'une station est d'observer les valeurs de Shapley après entraînement. Pour ce faire, une simulation a été lancée avec les 15 stations de l'ensemble pour entraîner le modèle. Les valeurs de Shapley moyennes pour chacune des stations ont ensuite été calculées à l'aide de

2000 échantillons de ce modèle. Les résultats présentés sur la figure 3.6 indiquent l'importance accordée par le modèle sur chacun de ces intrants. On peut d'abord s'intéresser aux 5 premières températures qui composent l'expérience de référence et constater que celles de Sherbrooke et Baie-Comeau ont un poids presque négligeable en comparaison aux autres régions. Avec cette nouvelle information, il est intéressant de tester la combinaison de départ en retirant ces 2 stations pour voir si la performance en est affectée.

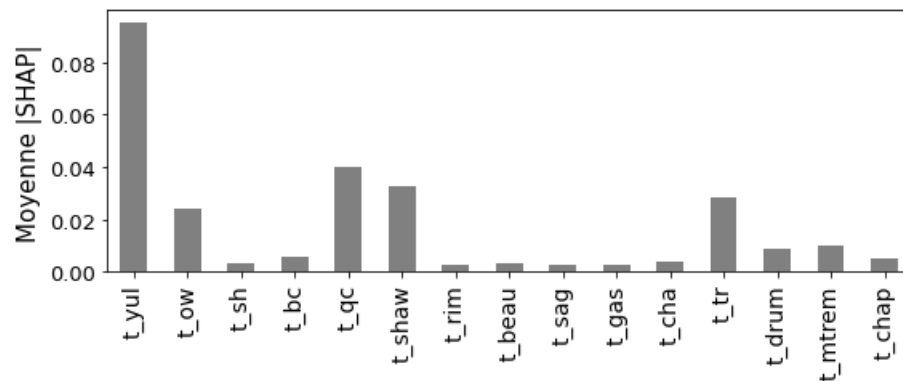


FIGURE 3.6 Valeurs de Shapley moyennes obtenues pour la simulation avec les 15 températures de l'ensemble.

Dans l'ensemble, on voit également que Shawinigan est la station la plus importante parmi les nouvelles stations observées. Trois-Rivières amène une contribution presque équivalente à Shawinigan, ce qui peut s'expliquer par leur proximité géographique. On peut donc délimiter nos recherches sur les régions les plus importantes, soit Montréal, Shawinigan, Québec et Ottawa.

TABLEAU 3.2 Score MAPE des simulations sur les ensembles de test, validation et entraînement en fonction du sous-ensemble de températures passé en intrant.

Station	Nb stations	Test	Validation	Entraînement
1, 2, 5	3	1.55	1.54	1.49
1, 5, 6	3	1.54	1.53	1.47
1, 2, 5, 6	4	1.52	1.50	1.45

Les résultats du tableau 3.2 confirment qu'un choix de 3 stations est suffisant pour maximiser les informations. En prenant les 4 stations les plus importantes selon la mesure de Shapley, il est possible de surpasser la performance ciblée, mais de manière non significative par rapport au résultat à 3 stations. Ainsi, le meilleur sous-ensemble permettant un résultat équivalent au contrôle est composé de Montréal, Québec et Shawinigan.

3.3 Pondération des stations

Dans l'optique d'affiner davantage le choix des intrants et éviter la redondance de l'information, une option est d'identifier une valeur moyenne pour représenter au mieux les conditions de la province. Pour ce faire, deux méthodes de pondération sont proposées : une pondération en fonction de la population et une pondération en fonction des valeurs de Shapley. Ces méthodes sont testées sur la combinaison de référence avec les 5 stations opérationnelles en partie pour l'aspect comparatif, mais également pour la facilité d'accès de cette information à l'interne.

La température globale est calculée à l'aide de l'équation de moyenne pondérée, où T_i est la température, X_i le poids associé et i la région correspondante. Les coefficients de pondération X_i utilisés sont définis dans le tableau 3.3.

$$T_{moy} = X_i T_i \quad (3.2)$$

TABLEAU 3.3 Coefficients de pondérations calculés en fonction de la population et de la valeur SHAP pour les 5 régions initiales.

Région	Population	SHAP
Montréal	0.630	0.477
Ottawa	0.104	0.151
Sherbrooke	0.062	0.035
Baie-Comeau	0.007	0.032
Québec	0.196	0.305

Les résultats des simulations effectuées avec les températures globales calculées par ces deux méthodes sont présentés dans le tableau 3.4.

TABLEAU 3.4 Scores MAPE journaliers des simulations sur les ensembles de test, validation et entraînement pour les températures globales calculées.

Pondération	Nb stations	Test	Validation	Entraînement
Montréal	1	1.69	1.66	1.63
Population	1	1.60	1.58	1.55
SHAP	1	1.58	1.57	1.53

L'avantage de la première alternative est que la population est une valeur objective, peu importe le choix du modèle de prévision. La deuxième option semble plus précise, mais devrait être testée sur d'autres modèles pour déterminer son potentiel de généralisation. Dans l'ensemble, on obtient tout de même une progression considérable par rapport à l'utilisation de la température de Montréal

uniquement. On ne rejoint pas la précision obtenue avec les 5 stations initiales, mais la différence de seulement 0.04% pourrait être un sacrifice envisageable pour le gain de ressources obtenu en retirant 200 valeurs d'entrée (4 stations avec 24 heures d'inertie). Il est toutefois important de noter que ces méthodes ont des limites puisque la relation entre la température et le BQ n'est pas linéaire. De plus, il n'est pas aussi facile de faire une moyenne pour d'autres variables, telles que l'humidité relative ou la nébulosité.

CHAPITRE IV

COMBINAISON DES STATIONS ET DES VARIABLES MÉTÉOROLOGIQUES

Pour compléter l'évaluation des intrants météorologiques, la dernière étape est de joindre les conclusions des derniers chapitres. On combine donc les variables de température, température de rosée, radiation et vitesse du vent pour les régions de Montréal, Québec et Shawinigan. L'inertie de la vitesse du vent est ignorée puisqu'elle ne semblait pas pertinente à la prévision. Pour que la simulation soit comparable avec les expériences précédentes, le reste des paramètres reste inchangé.

4.1 Évaluation du modèle de prévision résultant

Les expériences des derniers chapitres ont montré un gain de précision significatif des prévisions avec l'ajout de variables météorologiques et l'ajout de stations, soit respectivement de 0.27% et 0.15% par rapport au modèle neuronal de contrôle comprenant, comme unique variable météorologique, la température de Montréal. L'objectif est donc de combiner ces améliorations pour diminuer davantage l'erreur restante. Les détails du modèle résultant sont présentés dans le tableau 4.1.

Le nombre d'intrants météorologiques comprend les 4 variables pour les 3 stations, ce qui fait un total de 12 intrants. Globalement, la simulation obtient un score

MAPE de 1.32% sur les données de test, ce qui représente une amélioration de 0.37% par rapport à l'expérience de contrôle initiale. Ce résultat est encourageant, mais on peut voir que certaines erreurs importantes sont encore commises. La sous-estimation la plus importante est de 1750 MW et correspond à la journée du 25 novembre 2014. Les conditions météorologiques de cette journée sont des températures descendantes de 11 à 3 °C au cours de la journée et de forts vents. La sous-estimation du modèle pourrait être due à l'inertie de la température, c'est-à-dire que les températures plus élevées en début de journée auraient pu influencer la consommation prédite en fin de journée à la baisse. La surestimation maximale est de 3120 MW et correspond à la journée du 10 octobre 2016. Cette journée a été difficile à prévoir de manière précise pour l'ensemble des modèles testés dans cette étude. Les conditions météorologiques avec des températures variant entre 2 et 14 °C et des vents faibles à modérés ne suffisent pas à expliquer cette erreur. Il est donc possible qu'un biais soit présent dans le modèle ou qu'un facteur non climatologique soit en cause.

Malgré ces erreurs importantes, l'écart-type de la simulation est de 263 MW et 82.1% des résultats ont un écart inférieur à 500 MW, comparativement à 365 MW et 73% pour le contrôle initial, ce qui constitue une réduction considérable de l'erreur. Il reste toutefois 2.2% des cas qui sont supérieurs à 1000 MW, ce qui correspond à une erreur grave, mais ce pourcentage est grandement diminué par rapport au contrôle pour lequel ces erreurs représentent 6.7% des prévisions. Des améliorations pourraient être apportées en incluant les jours spéciaux et d'autres composantes non climatologiques.

La figure 4.1 présente la distribution des erreurs par rapport aux mois et aux heures. On remarque que l'erreur est plus grande à l'heure de pointe matinale et à la fin de la soirée. Il est possible que ces moments de la journée soient davantage influencés par le comportement humain que par la météorologie laissant une erreur

TABLEAU 4.1 Statistiques de précision du modèle de prévision sur les données de 2012 à 2019 avec séparation 60% entraînement, 20% validation et 20% test.

Nb intrants météo	12
MAPE entraînement	1.21 %
MAPE validation	1.31 %
MAPE test	1.32 %
Minimum test	-1750 MW
Maximum test	3120 MW
Écart-type test	263 MW
> 1000 MW test	2.2 %
> 500 MW test	17.9 %

inchangée par le travail effectué dans ce projet. Ensuite, la distribution mensuelle montre une erreur moyenne maximale pour les mois d'avril, mai, septembre et octobre. La période la plus difficile à prévoir est habituellement l'hiver en raison de la grande consommation d'électricité pour le chauffage. À l'inverse, les mois mentionnés ont habituellement des températures confortables durant lesquelles la climatisation et le chauffage ne sont pas nécessaires. Il est possible à nouveau que l'impact de la météorologie soit moins présent dans ces cas-là et que l'erreur résiduelle soit causée par un autre facteur. Finalement, le dernier panneau de la figure montre le nombre d'occurrence des erreurs importantes, supérieures à 1000 MW, pour chaque mois. On peut voir que ces erreurs arrivent tout de même encore souvent à l'hiver dans la période durant laquelle le BQ est le plus élevé. Ce résultat indique qu'un travail au niveau des intrants météorologiques en hiver pourrait encore être utile avec, par exemple, l'ajout des précipitations et tempêtes

hivernales.

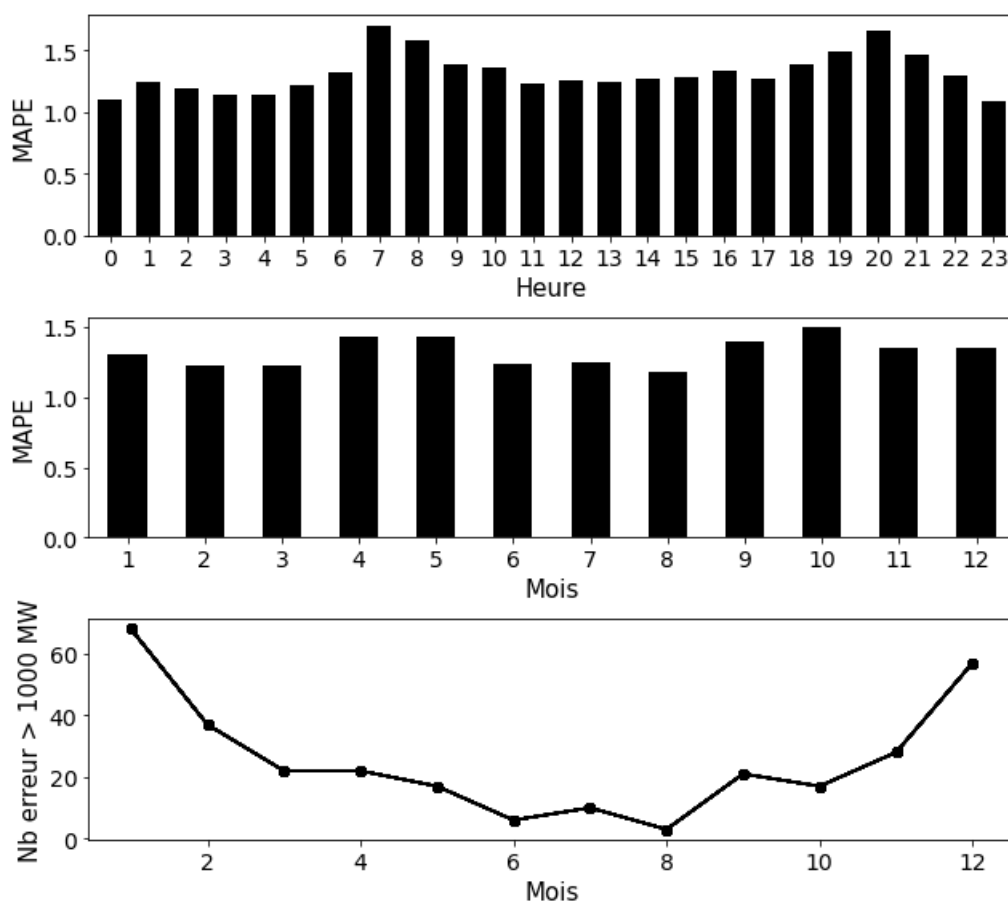


FIGURE 4.1 Scores MAPE journalier et mensuel moyens sur les données de test de 2012 à 2019. Occurrence des écarts de prévision supérieures à 1000 MW pour chaque mois.

Pour compléter cette analyse, la figure 4.2 présente les valeurs de Shapley des intrants météorologiques du modèle. La valeur moyenne des 24 pas de temps d'inertie est très faible pour chacune des variables, mais certaines heures, comme -1 et -24h, sont plus importantes. Ceci laisse penser qu'il pourrait être possible de retirer certains pas de temps et de conserver uniquement les intervalles de 6 ou 12h par exemple. La vitesse du vent est faible pour les 3 stations. Cela dit, il s'agit

d'une moyenne sur un ensemble de 400 échantillons et il est possible que l'impact du vent soit significatif seulement dans certaines situations. La radiation solaire incidente montre une participation importante pour la valeur à Montréal et diminue rapidement pour les deux autres stations. Pour la température du point rosée, la valeur de Shapley est supérieure pour Shawinigan, mais la moyenne d'inertie est plus importante pour Montréal. En somme, il est difficile de comprendre le raisonnement du modèle, mais on peut remarquer qu'il attribue la plus grande importance aux 3 températures, suivies par la radiation solaire incidente, la température de rosée et finalement la vitesse du vent. Il est possible que les variables de température de rosée à Québec et vitesse du vent à Shawinigan et Québec soient négligeables puisque leur apport n'a pas été démontré dans les expériences précédentes et que la valeur de Shapley calculée est très petite.

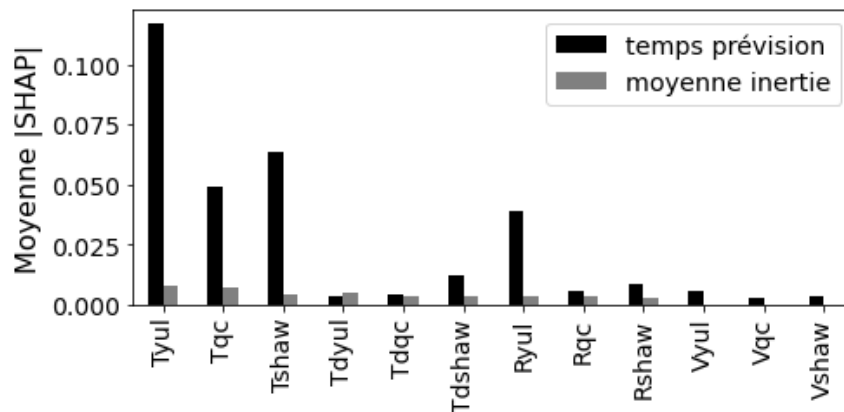


FIGURE 4.2 Moyenne de la valeur absolue des valeurs de Shapley sur 400 échantillons pour les intrants météorologiques au temps de prévision et pour la moyenne des 24 pas de temps d'inertie.

4.2 Comparaison de l'architecture

Un facteur important à considérer dans le choix d'un modèle, en plus de sa précision, est son utilisation de ressources. Le temps de calcul ainsi que l'espace mémoire utilisé peuvent présenter des contraintes dans les prévisions opérationnelles. De plus, un gain de vitesse d'entraînement pourrait se traduire en gain de précision s'il permet un réentraînement plus fréquent avec les données récentes.

Un autre aspect qu'il est possible d'explorer avec une architecture de modèle simple est la prévision d'ensemble. En ayant une architecture efficace et de petite taille, il est possible de combiner plusieurs modèles pour avoir la prévision moyenne de l'ensemble. Cette méthode est reconnue pour améliorer considérablement la précision des résultats (Rokach, 2010). De ce fait, l'objectif n'est pas uniquement d'améliorer la précision du modèle, mais également de trouver le bon équilibre entre précision et efficacité.

Pour évaluer la performance technique du modèle résultant, on compare avec une version utilisée par l'unité Prévision des apports et de la demande à Hydro-Québec. Le tableau 4.2 présente les paramètres de l'architecture, la taille et le temps d'entraînement de ce modèle (COSMOS), d'une version simplifiée de ce modèle (COSMOS_A), du modèle résultant de cette étude (RESULT) et du modèle de base utilisant uniquement la température de Montréal comme intrant météorologique (CTRL). Le temps indiqué est obtenu avec un processeur CPU 16 coeurs Intel Xeon Platinum 8375C @ 2.90GHz.

Il est important de noter que les paramètres non listés dans le tableau, tels que le nombre d'époques, le nombre de couches cachées ou la fonction d'activation, sont identiques pour tous les modèles présentés. On remarque d'abord que la taille des modèles est proportionnelle au nombre de paramètres entraînaibles et varie

TABLEAU 4.2 Architecture et utilisation des ressources des modèles de prévisions.

	COSMOS	COSMOS _A	RESULT	CTRL
Nb variables	4	4	4	1
Nb stations	5	5	3	1
Nb couches	5	5	5	5
Nb neurones	400	200	200	200
Inertie	72h	24h	24h	24h
Nb paramètres	1 236 801	276 801	222 401	181 801
Taille (Mo)	9.93	2.25	1.82	1.49
Temps (min)	73	52	42	30
MAPE entraînement	0.59	0.70	1.2	1.63
MAPE test	1.75	1.46	1.32	1.69

en fonction du nombre de neurones par couche cachée et du nombre d'entrées. Le temps d'entraînement est aussi influencé par ces facteurs. Avec un nombre de neurones par couche cachée équivalent, on voit une augmentation de 22 minutes pour l'entraînement de COSMOS_A avec 500 entrées météorologiques par rapport au contrôle avec 25 entrées météorologiques. On peut voir également que le nombre de pas de temps considérés dans l'inertie multiplie le nombre d'entrées météorologiques, ce qui impacte rapidement la taille et le temps d'entraînement du modèle. Une option pour réduire cet impact serait d'ajuster le temps d'inertie à chaque variable.

On observe également que la précision n'augmente pas forcément avec la complexité du modèle. À l'inverse, la version simplifiée de COSMOS permet une meilleure généralisation qui se traduit par une réduction de l'erreur MAPE sur les données de test. Le retrait des intrants peut toutefois dégrader la précision comme on peut voir pour le modèle de contrôle. Le modèle résultant de l'étude se

trouve donc à être un compromis entre taille, vitesse et précision.

Dans l'ensemble, il y a différentes manières de réduire la taille et le temps d'entraînement du modèle, mais il est important de conserver suffisamment d'espace et d'information pour bien apprendre le problème.

CONCLUSION

En conclusion, cette étude a évalué la contribution des données météorologiques dans deux modèles de complexité différente de prévision de la demande d'énergie pour l'ensemble de la province de Québec. Deux aspects des données météorologiques ont été pris en compte : le type d'information météorologique par l'utilisation de plusieurs variables telles que la température de l'air à 2m, l'humidité, la radiation solaire incidente, la nébulosité et la vitesse du vent à 10m ; et la source spatiale des données météorologiques par l'utilisation de 15 stations météorologiques réparties sur l'ensemble de la province de Québec.

L'évaluation est faite en trois parties, soit une comparaison statistique des données initiales, une analyse de performance d'un modèle statistique linéaire et une analyse de performance et de pertinence des intrants d'un réseau de neurones. Le modèle statistique utilisé est basé sur l'approche par équations multiples présentée par Clements et al. (Clements *et al.*, 2016) alors que le réseau de neurones est basé sur une architecture développée et optimisée par l'équipe de Prévision des apports et de la demande à Hydro-Québec. L'interprétation des résultats obtenus par le réseau de neurones a été possible avec le calcul des valeurs de Shapley (Lundberg et Lee, 2017). L'intégration des variables météorologiques a été testée sous différentes formes, dont l'effet de refroidissement éolien, l'index de chaleur et la température du point de rosée. Pour chaque expérience effectuée, le point de référence est une simulation du modèle qui utilise uniquement la température à Montréal comme donnée météorologique d'entrée.

Les résultats ont démontré une amélioration de la précision des modèles de pré-

vision avec l'ajout de chacune des variables météorologiques. En ajoutant une seule variable avec la température, la plus grande réduction de l'erreur MAPE est obtenue avec la nébulosité, tandis que la plus faible réduction est obtenue avec la vitesse du vent. La combinaison optimale de plusieurs variables comprend la température du point de rosée, la radiation solaire incidente et la vitesse du vent. Avec ces intrants, le MAPE est passé de 1.85% à 1.67% et de 1.69% à 1.42%, soit une amélioration de 0.18% et 0.27%, respectivement pour la précision du modèle statistique et celle du réseau de neurones.

L'analyse des modèles a démontré une relation faible, mais positive avec la consommation énergétique pour la vitesse du vent et l'humidité relative. La relation de la nébulosité avec le BQ est négative durant la nuit et positive pendant les heures d'ensoleillement. La radiation solaire incidente montre un impact négatif sur la consommation pendant les heures d'ensoleillement des journées froides.

L'analyse spatiale des intrants météorologiques a permis de démontrer que la combinaison de 3 stations, Montréal, Québec et Shawinigan, permet une amélioration considérable de la performance des modèles. Cette combinaison est cohérente avec nos hypothèses initiales concernant la répartition géographique et la représentation proportionnelle de la population. En combinant les données météorologiques des 3 stations, le score MAPE obtenu sur les données de test de 2012 à 2019 est de 1.32%, montrant ainsi une amélioration de 0.37% par rapport au modèle de référence (1.69%).

Les analyses présentées comportent tout de même une certaine incertitude en raison des valeurs manquantes dans les données météorologiques. L'absence de données représente jusqu'à 10% de la période de 8 ans pour certaines variables et régions. Le remplacement des valeurs manquantes par des valeurs climatologiques permet une cohérence, mais peut conduire à des erreurs significatives. Une

alternative pour les travaux futurs serait de remplacer ces valeurs par celles d'une réanalyse comme ERA5. De plus, ce mémoire présente des résultats obtenus avec des valeurs d'observation qui ne sont pas disponibles pour des projections futures. Ainsi, pour avoir un modèle de prévision applicable en temps réel, il aurait été pertinent de faire l'entraînement avec des valeurs issues d'un modèle de prévision météorologique.

L'étude pourrait être enrichie par d'autres avenues de recherche lors de travaux futurs. Une analyse détaillée de l'impact des précipitations comprenant le type, l'intensité et la quantité accumulée serait grandement pertinente pour comprendre leur influence sur les habitudes de consommation de la population et déterminer la meilleure façon de l'intégrer aux modèles de prévision. Une analyse ciblée sur les journées problématiques, dont la prévision est largement erronée, permettrait d'identifier les éléments manquants aux modèles de prévision actuels. Aussi, une étude de prévision de la demande séparée sur les différentes régions du Québec permettrait de mieux cerner les relations entre la météo de la région et la consommation énergétique. L'incertitude par rapport aux variations spatiales y serait grandement réduite. Finalement, le contexte opérationnel de l'étude présentée dans ce mémoire a motivé le choix de modèles simples, soutenables et bien connus. Il serait intéressant d'évaluer la performance de modèles de prévision temporelle plus élaborés en utilisant les intrants identifiés dans cette étude et éventuellement d'y intégrer des images satellitaires comme intrant.

RÉFÉRENCES

- Ahrens, C. D. (2015). *Meteorology today : an introduction to weather, climate, and the environment*. Cengage Learning Canada Inc.
- Aisyah, S., Simaremare, A. A., Adytia, D., Aditya, I. A. et Alamsyah, A. (2022). Exploratory weather data analysis for electricity load forecasting using svm and grnn, case study in bali, indonesia. *Energies*, *15*(10), 3566.
- Al-Musaylh, M. S., Deo, R. C., Adamowski, J. F. et Li, Y. (2018). Short-term electricity demand forecasting with mars, svr and arima models using aggregated demand data in queensland, australia. *Advanced Engineering Informatics*, *35*, 1–16.
- Amato, A. D., Ruth, M., Kirshen, P. et Horwitz, J. (2005). Regional Energy Demand Responses To Climate Change : Methodology And Application To The Commonwealth Of Massachusetts. *Climatic Change*, *71*(1), 175–201. <http://dx.doi.org/10.1007/s10584-005-5931-2>. Récupéré le 2022-05-10 de <https://doi.org/10.1007/s10584-005-5931-2>
- Apadula, F., Bassini, A., Elli, A. et Scapin, S. (2012). Relationships between meteorological variables and monthly electricity demand. *Applied Energy*, *98*, 346–356. <http://dx.doi.org/10.1016/j.apenergy.2012.03.053>. Récupéré le 2022-05-10 de <https://www.sciencedirect.com/science/article/pii/S0306261912002735>
- Bessoudo, M., Tzempelikos, A., Athienitis, A. et Zmeureanu, R. (2010). Indoor thermal environmental conditions near glazed facades with shading devices—part i : Experiments and building thermal model. *Building and environment*, *45*(11), 2506–2516.
- Blum, A. L. et Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, *97*(1-2), 245–271.
- Clements, A., Hurn, A. et Li, Z. (2016). Forecasting day-ahead electricity load using a multiple equation time series approach. *European Journal of Operational Research*, *251*(2), 522–530. <http://dx.doi.org/10.1016/j.ejor.2015.12.030>. Récupéré le 2022-05-24 de <https://linkinghub.elsevier.com/retrieve/pii/S0377221715011698>

- Cooley, J. W. et Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90), 297–301.
- David, C., Blondin Massé, A. et Zinflou, A. (2023). Fast short-term electrical load forecasting under high meteorological variability with a multiple equation time series approach. Dans *International Journal of Electrical and Computer Engineering*, volume 17. World Academy of Science, Engineering and Technology.
- Dickey, D. A. et Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a), 427–431.
- Duchi, J., Hazan, E. et Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).
- ECCC (2023a). Government of canada / gouvernement du canada. Récupéré de https://climate.weather.gc.ca/historical_data/search_historic_data_e.html
- ECCC (2023b). Government of canada / gouvernement du canada. Récupéré de https://climate.weather.gc.ca/climate_normals/results_1981_2010_e.html?searchType=stnName&txtStationName=montreal&searchMethod=contains&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=5415&dispBack=0
- Grenier, M. (2006). Short-term load forecasting at Hydro-Quebec TransEnergie. Dans *2006 IEEE Power Engineering Society General Meeting*, 5 pp.–. ISSN : 1932-5517, <http://dx.doi.org/10.1109/PES.2006.1709029>
- Gross, G. et Galiana, F. (1987). Short-term load forecasting. *Proceedings of the IEEE*, 75(12), 1558–1573. Conference Name : Proceedings of the IEEE, <http://dx.doi.org/10.1109/PROC.1987.13927>
- Hahn, H., Meyer-Nieberg, S. et Pickl, S. (2009). Electric load forecasting methods : Tools for decision making. *European journal of operational research*, 199(3), 902–907.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D. *et al.* (2020). The era5 global reanalysis. *Quarterly Journal of the Royal*

Meteorological Society, 146(730), 1999–2049.

Hobby, J. D. et Tucci, G. H. (2011). Analysis of the residential, commercial and industrial electricity consumption. Dans *2011 IEEE PES Innovative Smart Grid Technologies*, 1–7. IEEE.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D. et Zareipour, H. (2020). Energy forecasting : A review and outlook. *IEEE Open Access Journal of Power and Energy*, 7, 376–388.

Hydro-Québec (2023). Répartition de la consommation d’un ménage.

Accessed on July 3, 2023. Récupéré de

<https://www.hydroquebec.com/residentiel/espace-clients/consommation/consommation-electrique-sources.html>

IEA, I. (2019). World energy outlook 2019. *World Energy Outlook 2019*.

License : CC BY 4.0. Récupéré de

<https://www.iea.org/reports/world-energy-outlook-2019>

Ionescu, C., Baracu, T., Vlad, G.-E., Necula, H. et Badea, A. (2015). The historical evolution of the energy efficient buildings. *Renewable and Sustainable Energy Reviews*, 49, 243–253.

Lago, J., De Ridder, F. et De Schutter, B. (2018). Forecasting spot electricity prices : Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221, 386–405.

Lim, B., Arik, S. Ö., Loeff, N. et Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting.

International Journal of Forecasting, 37(4), 1748–1764.

Lundberg, S. M. et Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

MacKinnon, J. G. (2010). *Critical values for cointegration tests*. Rapport technique, Queen’s Economics Department Working Paper.

Makridakis, S., Spiliotis, E. et Assimakopoulos, V. (2022). M5 accuracy competition : Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364.

Mudelsee, M. (2019). Trend analysis of climate time series : A review of methods. *Earth-science reviews*, 190, 310–322.

Mukherjee, S. et Nateghi, R. (2017). Climate sensitivity of end-use electricity consumption in the built environment : an application to the state of florida,

united states. *Energy*, 128, 688–700.

Musbah, H., El-Hawary, M. et Aly, H. (2019). Identifying seasonality in time series by applying fast fourier transform. Dans *2019 IEEE Electrical Power and Energy Conference (EPEC)*, 1–4. IEEE.

Nateghi, R. et Mukherjee, S. (2017). A multi-paradigm framework to assess the impacts of climate change on end-use energy demand. *PLOS ONE*, 12(11), e0188033. Publisher : Public Library of Science, <http://dx.doi.org/10.1371/journal.pone.0188033>. Récupéré le 2022-05-10 de <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188033>

Olivares, K. G., Challu, C., Marcjasz, G., Weron, R. et Dubrawski, A. (2023). Neural basis expansion analysis with exogenous variables : Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2), 884–900.

Osczevski, R. et Bluestein, M. (2005). The new wind chill equivalent temperature chart. *Bulletin of the American Meteorological Society*, 86(10), 1453–1458.

Pantiskas, L., Verstoep, K. et Bal, H. (2020). Interpretable multivariate time series forecasting with temporal attention convolutional neural networks. Dans *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1687–1694. IEEE.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review*, 33, 1–39.

Rothfus, L. P. et Headquarters, N. S. R. (1990). The heat index equation (or, more than you ever wanted to know about heat index). *Fort Worth, Texas : National Oceanic and Atmospheric Administration, National Weather Service, Office of Meteorology*, 9023, 640.

Savić, S., Selakov, A. et Milošević, D. (2014). Cold and warm air temperature spells during the winter and summer seasons and their impact on energy consumption in urban areas. *Natural hazards*, 73, 373–387.

Song, B., Bai, L. et Yang, L. (2022). Analysis of the long-term effects of solar radiation on the indoor thermal comfort in office buildings. *Energy*, 247, 123499.

<http://dx.doi.org/https://doi.org/10.1016/j.energy.2022.123499>. Récupéré de <https://www.sciencedirect.com/science/article/pii/S0360544222004029>

Spliid, H. (1983). A fast estimation method for the vector autoregressive moving average model with exogenous variables. *Journal of the American Statistical Association*, 78(384), 843–849.

Zhao, K., Liu, X.-H., Zhang, T. et Jiang, Y. (2011). Performance of temperature and humidity independent control air-conditioning system in an office building. *Energy and Buildings*, 43(8), 1895–1903.