Titre: Title:	Prédiction d'économies d'énergie de bâtiments après rénovation par la modélisation basée sur les données
Auteur: Author:	Jonathan Kere
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Kere, J. (2021). Prédiction d'économies d'énergie de bâtiments après rénovation par la modélisation basée sur les données [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9741/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/9741/
Directeurs de recherche: Advisors:	Massimo Cimmino, & Marios-Eleftherios Fokaefs
Programme: Program:	Génie mécanique

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Prédiction d'économies d'énergie de bâtiments après rénovation par la modélisation basée sur les données

JONATHAN KERE

Département de génie mécanique

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès science appliquées*

Génie mécanique

Décembre 2021

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Prédiction d'économies d'énergie de bâtiments après rénovation par la modélisation basée sur les données

présenté par **Jonathan KERE**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Michaël KUMMERT, président
Massimo CIMMINO, membre et directeur de recherche
Marios-Eleftherios FOKAEFS, membre et codirecteur de recherche
Quentin CAPPART, membre

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner ma gratitude.

Je voudrais tout d'abord adresser ma reconnaissance à mes directeurs de ce mémoire, Massimo et Marios, pour leurs conseils toujours pertinents et précieux.

Je tiens à remercier spécialement Pierre-Luc, pour toujours avoir été à mon écoute et m'apporter l'aide dont j'avais besoin tout au long du projet, ainsi que pour m'avoir fait aimer mon temps chez Ecosystem.

Je désire aussi remercier les professeurs de Polytechnique Montréal et des Mines de Douai, qui m'ont fourni les outils nécessaires à la réussite de mes études universitaires.

J'aimerai également remercier Mitacs pour leur soutien financier. Cette recherche a reçu le soutien de Mitacs dans le cadre du programme Mitacs Accélération.

Je voudrais exprimer ma reconnaissance envers mes parents, Chantal et David, qui m'ont toujours soutenu dans mes choix de vie et qui mon toujours poussé à donner le maximum de moi-même, ainsi qu'envers ma sœur, Sarah, qui m'a appris dès le plus jeune âge à défendre mes convictions tout en restant à l'écoute des autres.

Enfin, je souhaite remercier Jessie-Lee, qui a transformé mon simple séjour au Canada en immigration, et qui m'a soutenu tout le long de mon mémoire.

RÉSUMÉ

Au Canada, le secteur du bâtiment est l'un des secteurs les plus énergivores (Gouvernement Canada, 2021). Sur le cycle de vie d'un bâtiment, la phase d'utilisation est responsable de plus de 80% de la consommation énergétique. Les travaux de rénovation énergétique permettent de réduire cet impact énergétique. Les entreprises de services énergétiques (ESE) ont pour objectif de réaliser des projets de travaux de rénovation énergétique pour leurs clients. Lors de la phase de design des travaux, il est important de pouvoir quantifier le potentiel d'amélioration de l'efficacité énergétique associé aux travaux. Par ailleurs, les domaines de l'apprentissage machine et de la science des données sont aussi de plus en plus présents dans le secteur du bâtiment. Ainsi, le présent mémoire vise à prédire le potentiel de réduction de la consommation d'énergie associé à un projet de rénovation énergétique en utilisant les données d'anciens projets de rénovation énergétique, et de suggérer des mesures d'économie d'énergie.

Ce mémoire se concentre sur les bâtiments de type éducatifs. La première partie porte uniquement sur les écoles primaires et vise à tester différents algorithmes d'apprentissage machine afin de prédire le potentiel d'économies d'énergie annuelles pour un nouveau projet de rénovation. L'étude utilise les paramètres du modèle « change-point » comme entrée du modèle et elle a permis d'identifier l'algorithme forêt aléatoire comme le plus efficace. Une seconde étude vise à améliorer et généraliser la méthode de prédiction des économies pour d'autres types de bâtiment éducatifs tels que des écoles secondaires, des universités et des centres de formations. On obtient finalement qu'avec les données d'anciens projets de travaux de rénovation énergétique de la base de données d'Ecosystem, la catégorie générale des écoles primaires et secondaires (K12) sont les bâtiments pour lesquels la méthode peut être validée, avec une erreur de prédiction *CV(RMSE)* de 42%. Parallèlement, une troisième étude est présentée. Elle cherche à recommander les mesures d'économie d'énergie à implémenter pour atteindre les économies d'énergies. Elle utilise la même source de données (K12) et un algorithme de voisinage kNN (« k-Nearest Neighbors ») est implémenté pour déterminer le regroupement des bâtiments voisins les plus similaires à partir du résultat intermédiaire de l'algorithme. Encore une fois, les paramètres du modèle « change-point » sont utilisés en entrée de l'algorithme. Une fois les voisins trouvés, les mesures d'économie d'énergie réalisées pour ces voisins sont récupérées et sont transférées en sortie de l'algorithme. Une recommandation avec un score F1 de 66% est alors atteinte pour la prédiction des mesures.

Mots-clés : Travaux de rénovation énergétique, Apprentissage machine, Forêt aléatoire, Modèle « change-point », Bâtiments éducatifs

ABSTRACT

In Canada, the construction sector is one of the most energy intensive (Gouvernement Canada, 2021). The operation phase is responsible for more than 80% of the energy consumption over the life cycle of a building. Energy retrofit projects present a solution to reduce this energy impact. The objective of energy service companies (ESCO) is to carry out energy retrofit projects for their clients. During the design phase of a project, it is important to be able to evaluate the potential energy savings. In parallel, the fields of machine learning and data science are increasingly present in the building sector. Thus, this master thesis aims to predict the energy savings potential of energy retrofit projects using a data-driven approach based on data from past projects, and to recommend measures to achieve the energy savings.

The project focuses on educational buildings as a case study. The first part focuses on primary schools only and aims to evaluate different machine learning algorithms for the prediction of the potential for annual energy savings of a new retrofit project. It uses the parameters of the change-point model as the input to the machine learning model and shows that the random forest algorithm is the most efficient one. The second part aims to improve and generalize the method of energy savings predictions for other types of educational buildings such as secondary schools, universities and training centers. It is shown that with the data from Ecosystem's database, the general category of primary and secondary schools (K12) are the buildings for which the method can be validated with a prediction error *CV(RMSE)* of 42%. The third party seeks to predict the energy saving measures to be implemented to achieve energy savings. It uses the same data source (K12) and a kNN algorithm is implemented to determine the most similar neighboring buildings using the intermediate result of the algorithm. Again, the parameters of the change-point model are used as input to the algorithm. Once the neighbors are found, the energy measures implemented for these neighbors are retrieved and transferred to the output of the algorithm. An F1 score of 66% is reached for the prediction of the measures.

Keywords: Energy retrofits, Machine learning, Random forest, Change-point model, Educational buildings

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ANN	« Artificial neural network »
ASHRAE	« American Society of Heating, Refrigerating and Air Conditioning Engineers »
BETTER	« Building Efficiency Targeting Tool for Energy Retrofits »
Ccp	« Cooling change point »
CP	« Change-point »
Csl	« Cooling slope »
CV	Coefficient de variation
CVCA	Chauffage, ventilation et conditionnement de l'air
DJ	Degrés-jour
DT	« Decision tree »
ESCO	« Energy service companies»
ESE	Entreprise de services énergétiques
EUI	« Energy use intensity »
Hcp	« Heating change point »
Hsl	« Heating slope »
ISD	« Integrated Surface Database »
K12	Écoles primaires et secondaires
kNN	« k Nearest Neighbors »
M&V	Mesures et vérifications
MAE	« Mean absolute error »
MAPE	« Mean absolute percentage error »

MBE	« Mean bias error »
MSE	« Mean square error »
NOAA	« National Oceanic and Atmospheric Administration »
R^2	Coefficient de détermination
RF	« Random forest »
RLM	Régression linéaire multivariable
RMSE	« Root mean square error »
RPG	Régression par processus Gaussien
SGD	« Stochastic gradient descent »
SVM	« Support vector machine »

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Le développement durable et de la réduction des émissions de carbone sont des enjeux cruciaux pour la génération actuelle et celles à venir. L'objectif est de viser à un monde meilleur, c'est-à-dire un environnement sain, une économie stable et durable, et une justice sociale, afin d'assurer le confort pour aujourd'hui comme pour demain. En faveur de cet objectif, le gouvernement canadien a lancé le « programme de développement durable à l'horizon 2030 » qui vise la réduction de plus de 37% des émissions de CO_2 du pays (Gouvernement Canada, 2021).

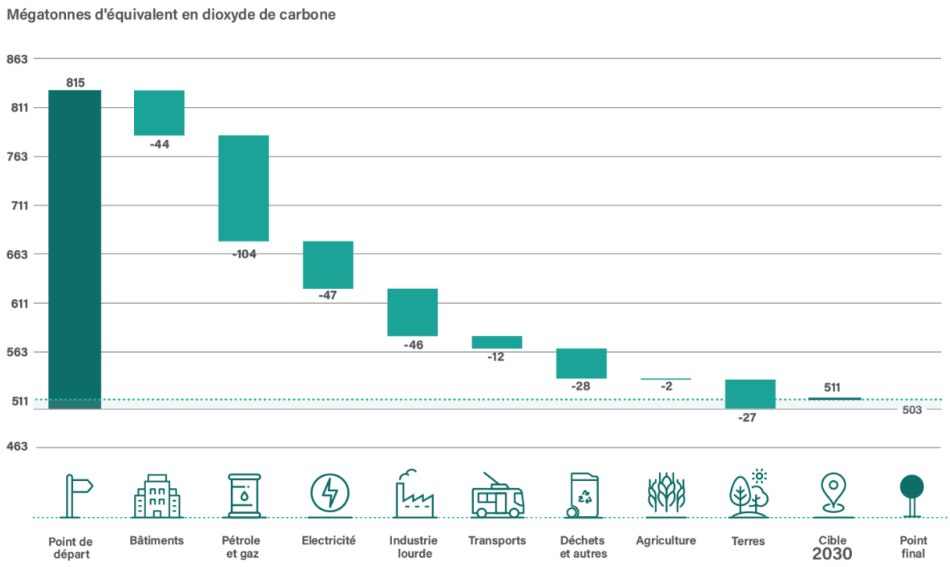

Figure 1.1. Sources de réduction des émissions contribuant à atteindre la cible de 2030, tirée de Gouvernement du Canada (2021) - reproduction avec autorisation

Le programme identifie un potentiel de réduction dans le secteur du bâtiment de 44 Mt CO_2 équivalent (15% de l'objectif total), comme le montre la Figure 1.1. Ce secteur est très énergivore : au Canada en 2019, il était à lui seul responsable de 31% de la consommation énergétique du pays et de 12% des émissions de gaz à effet de serre du Québec (Whitmore & Pineau, 2020). Ce scénario a attiré beaucoup d'attention sur la construction de bâtiments net zéro : un bâtiment ayant un rendement énergétique net zéro signifie qu'annuellement (ou sur une période de référence), le

bâtiment produit localement au moins autant d'énergie qu'il n'en consomme. Cependant, comme les bâtiments ont généralement une durée de vie d'environ cent ans, et 80% à 90% de leur consommation d'énergie (et par conséquent leurs émissions) totale a lieu durant la phase opérationnelle (Roth & Jain, 2018), la majorité des émissions de gaz à effet de serre provient du patrimoine déjà construit. L'impact énergétique des bâtiments net zéro nouvellement construits est donc moindre par rapport à l'impact d'une réduction de la consommation d'énergie des bâtiments déjà existants. Ainsi, la meilleure façon de réduire la consommation du secteur du bâtiment est de réaliser des rénovations améliorant l'efficacité énergétique des structures existantes.

Afin de répondre à cet objectif de réduction de la consommation énergétique dans le secteur du bâtiment, de plus en plus d'entreprises ont pour mission d'améliorer l'efficacité énergétique des bâtiments, elles portent l'étiquette d'Entreprise de Services Énergétiques (ESE). Une ESE est une firme d'ingénierie mandatée par un propriétaire d'un bâtiment pour concevoir des projets en efficacité énergétique. L'ESE peut aussi offrir le financement, réaliser la construction et même l'opération pour une période établie. Ces projets sont remboursés à même les économies d'énergie réalisées dans la période convenue au contrat à l'issue du projet. En outre, les infrastructures mises en place demeurent la propriété de l'utilisateur.

Ecosystem, l'entreprise partenaire de ce projet de maîtrise, est une ESE. Son siège social se trouve à Québec, et sa mission est la rénovation de bâtiments éducatifs, du domaine de la santé, ou d'autres écosystèmes énergétiques complexes situés au Canada et aux États-Unis. Au début d'un nouveau projet, les ingénieurs d'Ecosystem font une analyse du bâtiment à rénover et évaluent le potentiel d'économie d'énergie, c'est-à-dire les économies d'énergie qu'il est possible d'obtenir à la suite du déploiement de mesures d'efficacité énergétique. Les mesures d'efficacité énergétique correspondent aux travaux énergétiques qui sont réalisés afin de réaliser une économie d'énergie, par exemple un changement de chaudière. À la suite de l'analyse, les ingénieurs d'Ecosystem peuvent prédire les économies énergétiques annuelles, qui se traduisent par une économie financière. Les économies font généralement partie de l'entente contractuelle et de la garantie de résultat offerte par Ecosystem, puisque Ecosystem est payé par les économies.

À travers ce modèle d'affaires, on constate l'importance d'une bonne prédiction : si les économies sont incorrectement prédites, alors les économies d'énergie réelles seront différentes des économies d'énergie désirées. Les profits d'Ecosystem seront alors réduits ou mettront plus de

temps à être obtenus, ce qui peut engendrer des complications avec la planification interne des finances de l'entreprise. Le lien de confiance avec les clients pourrait également être détérioré si la prédiction d'économie surévalue la réalité. Par ailleurs, la garantie sur les économies s'accompagne d'une période de suivi des projets durant laquelle Ecosystem surveille les consommations énergétiques du projet pour s'assurer que le contrat est respecté ou parfois faire des ajustements (par exemple, des travaux supplémentaires ou l'optimisation des opérations). Au cours des années, Ecosystem a accumulé des données de consommation et d'économies réalisées sur des projets passés (environ 10 ans de données). Il y a donc une opportunité d'exploiter ces données qui n'ont jusqu'à maintenant pas été utilisées. Ainsi, Ecosystem désire affiner ses méthodes de prédictions, en se basant sur de nouveaux préceptes utilisant les données, c'est dans ce cadre que l'entreprise a commandité ce projet de recherche.

Le présent projet a donc pour objectif principal d'**utiliser les données obtenues sur d'anciens projets afin de prédire les économies énergétiques et les mesures d'efficacité énergétique requises sur de nouveaux projets**. Deux objectifs secondaires sont alors identifiés :

1. Déterminer le potentiel d'économie énergétique d'un bâtiment par une approche basée sur les données de projets antérieurs.
2. Recommander des mesures d'efficacité énergétique à implémenter afin d'atteindre les économies d'énergie par une approche basée sur les données de projets antérieurs.

Le présent mémoire commence par un chapitre d'introduction suivi par un chapitre de revue de littérature pertinente au projet. Le troisième chapitre présente la base de données d'Ecosystem. Ensuite, la quatrième chapitre traite de la prédiction des économies d'énergies seulement dans le cas d'écoles primaires afin de trouver et valider une méthode qui permet d'obtenir des résultats viables à partir des données disponibles. Le cinquième chapitre présente le développement de la méthode de prédiction des économies pour tous les types de bâtiments éducatifs. Le sixième chapitre concerne la recommandation des mesures d'économie d'énergie. Le septième chapitre est un chapitre de discussion. Enfin, le huitième et dernier chapitre est la conclusion, qui résume les méthodes développées pendant ce projet et les prochaines étapes qu'il serait possible de poursuivre.

CHAPITRE 2 REVUE DE LITTÉRATURE

L'objectif de ce chapitre est d'établir un cadre pertinent grâce aux travaux scientifiques en lien avec les objectifs du projet. Ce chapitre commence par l'examen des travaux concernant la prédiction des économies après rénovation pour les bâtiments. Les méthodes de prédiction basées sur des modèles physiques, ainsi que les méthodes basées sur les données et les outils d'apprentissage machine sont successivement présentées. Ensuite, les différents algorithmes d'apprentissage machine utilisés en science du bâtiment, ainsi que des exemples d'applications issus de la littérature, sont décrits.

2.1.1 Méthodes de prédiction de l'économie basées sur des modèles physiques

Coakley et al. (Coakley et al., 2014) définissent les modèles physiques comme des modèles qui appliquent un jeu de lois physiques (par exemple, la gravité, les transferts de chaleur et de masse, etc.) qui régissent un système, afin de prédire son comportement en fonction des propriétés et des conditions du système. Ces lois physiques sont représentées par des équations qui permettent de capturer une ou plusieurs caractéristiques du système en jeu.

En science du bâtiment, il existe deux approches principales utilisant des modèles physiques : l'utilisation de logiciels de simulation, ou l'utilisation d'équations physiques pour un phénomène physique particulier.

Logiciels de simulation

Il existe de nombreux logiciels de simulation énergétique du bâtiment : DOE-2 (Winkelmann et al., 1993), TRNSYS (University of Wisconsin--Madison. Solar Energy, 1975), EnergyPlus (Crawley et al., 2001), etc. Ces logiciels permettent le calcul détaillé de la consommation énergétique requise pour maintenir les critères de performance du bâtiment spécifiés (par exemple, la température et l'humidité des espaces), sous l'influence d'entrées externes, telles que les conditions météorologiques, l'occupation et l'infiltration, et d'entrées internes, telles les dimensions, les systèmes énergétiques installés, les propriétés thermiques de l'enveloppe. Ces

calculs sont réalisés en arrière-plan grâce à des systèmes d'équations physiques complexes sur des pas de temps discrets.

Pour la prédiction de l'économie, ces logiciels sont utilisés lors de la phase de design de la rénovation. Le bâtiment est alors simulé en considérant les mesures d'économie d'énergie à implémentées. Les mesures d'économie d'énergie correspondent aux travaux énergétiques qui sont à réaliser pour atteindre les objectifs énergétiques. Le logiciel est alors capable de fournir une prédiction de la consommation pour le bâtiment rénové, et donc une économie théorique. Il faut donc pour utiliser ces logiciels dans un contexte de prédiction de l'économie d'énergie avoir fait une étude préalable du bâtiment pour déterminer les entrées internes et externes du logiciel ainsi que les mesures d'économie d'énergie à mettre en place. Cependant, les objectifs du présent projet de recherche visent à prédire l'économie d'énergie sans avoir connaissance des entrées complexes d'un logiciel de simulation. L'utilisation des logiciels de simulation n'est donc pas une option. Pour cette raison, ils ne sont pas étudiés plus en détails dans cette revue de littérature et dans la suite du mémoire.

Modèles simplifiés pour l'estimation de la consommation énergétique globale

Il existe une grande quantité de modèles en science du bâtiment permettant de couvrir une grande quantité de situations physiques. Dans le cadre de la consommation globale d'un bâtiment, deux modèles sont retenus ici, les modèles degrés-jour en chauffage et en climatisation et le modèle « change-point ». Ils permettent de calculer la consommation d'énergie du bâtiment en fonction des conditions climatiques extérieures.

Les **modèles degrés-jour** reposent sur le calcul des degrés-jour pour le lieu concerné. Comme décrit dans l'handbook de l' « American Society of Heating & Air-Conditioning Engineers» (ASHRAE) (American Society of Heating Refrigerating Air-Conditioning Engineers, 2017), le degré-jour est une mesure de la fréquence et de combien de degrés la température moyenne quotidienne pour un emplacement est au-dessus (pour le refroidissement) ou au-dessous (pour le chauffage) d'une température d'équilibre. Par exemple, un jour où la température quotidienne moyenne est inférieure de 12 degrés à la température d'équilibre représenterait 12 degrés-jours de chauffage. Une température d'équilibre de 18°C pour le chauffage et la climatisation est communément acceptée comme la température la plus appropriée pour capturer l'impact du climat

sur la consommation des bâtiments commerciaux. Les degrés-jours sont principalement utilisés avec une résolution mensuelle (somme des degrés-jour sur un mois) ou annuelle (somme des degrés-jour sur une année) et permettent de caractériser les conditions thermiques du lieu géographique concerné. Le calcul de la consommation d'énergie pour un bâtiment peut alors être réalisé en appliquant :

$$Q_h = DJ_h * 24\left(\frac{h}{jour}\right) * \sum_i P_i \tag{2.1}$$

où Q_h est la consommation de chauffage en Wh, DJ_h sont les degrés-jour en chauffage pour la résolution de temps désirée, et P_i sont les pertes thermiques du bâtiment par l'enveloppe et l'infiltration en W/K, *i* représentant les sources de pertes. Pour la consommation en climatisation, l'équation devient :

$$Q_c = DJ_c * 24\left(\frac{h}{jour}\right) * \sum_i P_i \tag{2.2}$$

où Q_c est la consommation de climatisation en Wh, DJ_c sont les degrés-jour en climatisation pour la résolution de temps désirée, et P_i sont les gains thermiques du bâtiment par l'enveloppe et l'infiltration en W/K.

Le **modèle « change-point »** est un modèle intermédiaire entre un modèle physique et un modèle issu des données (Kissock et al., 2003). Ce type de modèle porte aussi le nom de modèle « boîte grise ». Ces modèles fonctionnent de telle sorte qu'une équation physique représente le phénomène physique, mais cette équation est calibrée grâce aux données historiques. Le modèle change-point est obtenu en faisant la régression des données mensuelles de consommation d'énergie par rapport aux températures moyennes de la période de facturation. Le modèle doit identifier les températures des points d'équilibre (ou points de changement) auxquels la consommation d'énergie passe d'un

comportement dépendant des conditions météorologiques à un comportement indépendant des conditions météorologiques (American Society of Heating Refrigerating Air-Conditioning Engineers, 2017). L'équation du modèle change-point est :

$$E = b_0 + b_1 * (b_3 - T)^+ + b_2 * (T - b_4)^+ \quad (2.3)$$

où E est la consommation énergétique quotidienne moyenne mensuelle, T est la moyenne mensuelle de la température extérieure, et les b_i sont les paramètres du modèle qui sont obtenus suite à la régression. b_0 représente la consommation d'énergie indépendante de la température, b_1 et b_2 représentent respectivement les pertes thermiques en chauffage et les gains thermiques en climatisation (en W/K), et b_3 et b_4 représentent respectivement les températures d'équilibre en chauffage et en climatisation. Le « + » en exposant signifie que la parenthèse n'est prise en compte que si l'intérieur est positif.

Les modèles degrés-jour et « change point » permettent de calculer la consommation énergétique d'un bâtiment en fonction des conditions météorologiques extérieures. Pour des conditions météo réelles ou virtuelles, il est possible de simuler le comportement énergétique du bâtiment grâce aux modèles. Par contre, ils ne permettent pas de calculer l'économie d'énergie potentielle d'un bâtiment, mais servent à créer une référence pour un scénario connu. Dans le cas de l'économie, ces modèles permettent pour un bâtiment rénové de calculer l'économie réalisée en prenant en compte les conditions météorologiques : le modèle avant travaux sert de référence, et la consommation après travaux est comparée avec la consommation obtenue grâce au modèle pour les températures après travaux. La différence de consommations représente l'économie d'énergie.

Peu importe le type de modélisation, que ce soit un modèle physique ou un modèle basé sur les données, Amasyali et El-Gohary (Amasyali & El-Gohary, 2018) rappellent qu'il est essentiel de pouvoir tester son modèle. Tester son modèle consiste à évaluer la prédiction en utilisant une mesure standard d'évaluation. Selon leur revue, en science du bâtiment, les mesures d'évaluation les plus communes dans la littérature sont le coefficient de variation (« coefficient of variation », *CV*), l'erreur absolue moyenne en pourcentage (« mean absolute percentage error », *MAPE*) et la racine de l'erreur quadratique moyenne (« root mean square error », *RMSE*). On retrouve

également d'autres mesures d'évaluation moins utilisées comme l'erreur absolue moyenne (« mean absolute error », *MAE*), l'erreur de biais moyenne (« mean bias error », *MBE*), l'erreur quadratique moyenne (« mean square error », *MSE*), le coefficient de détermination (R^2), ou le taux d'erreur (δ_i). Le *CV* est la mesure d'évaluation la plus couramment utilisée, et ce pour deux raisons. Premièrement, c'est l'une des mesures d'évaluation de la performance recommandées par l'ASHRAE pour évaluer les modèles de prédiction de la consommation d'énergie (American Society of Heating Refrigerating Air-Conditioning Engineers, 2017). Deuxièmement, le *CV(RMSE)* normalise l'erreur de prédiction par la consommation d'énergie moyenne et fournit une mesure sans unité qui est plus pratique à des fins de comparaison.

2.1.2 Méthodes de prédiction de l'économie basée sur les données et des outils d'apprentissage machine

L'intelligence artificielle est de plus en plus présente dans la vie de tous les jours grâce aux innovations technologiques et à l'avancée des outils numériques. Cette tendance apparaît également en science du bâtiment. L'apprentissage machine est un champ d'études de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données. Les modèles obtenus grâce à de l'apprentissage machine sont des algorithmes « boîtes noires », c'est-à-dire que le modèle exact n'est pas connu par l'opérateur et ne repose pas nécessairement sur un sens physique : l'opérateur fourni un jeu de données comprenant la ou les entrées du modèle et la ou les sorties du modèle, et l'algorithme génère automatiquement la loi mathématique liant l'entrée à la sortie.

Les applications de l'apprentissage machine en science du bâtiment sont diverses, que ce soit dans les algorithmes utilisés ou dans les objectifs des opérateurs. Cette sous-section se concentre sur les articles présents dans la littérature qui traite de la prédiction de l'économie d'énergie d'un système énergétique ou de l'ensemble du bâtiment en utilisant des outils d'apprentissage machine.

Ascione et al. (Ascione et al., 2017) présentent une méthode de prédiction de l'économie pour des édifices à bureaux selon les mesures d'économie d'énergie réalisées. Leur méthode repose sur l'utilisation d'un réseau de neurones artificiels. Les auteurs présentent deux réseaux de neurones

différents, un premier permettant le calcul de la consommation annuelle pour le chauffage, la climatisation, et le nombre d'heures d'inconfort thermique pour un bâtiment sans rénovations, et un second permettant de calculer les mêmes grandeurs ainsi que la quantité d'électricité produite par des panneaux photovoltaïques pour un bâtiment rénové (parmi une liste de mesures d'économie d'énergie définit par les auteurs). Le premier réseau permet donc d'établir une référence de la consommation énergétique. Ainsi, la différence entre les résultats de consommations obtenus avec le premier réseau et le second permet d'obtenir l'économie d'énergie qu'engendrent les rénovations. Les variables d'entraînement correspondent à des variables d'état (par exemple la hauteur du bâtiment, le type des fenêtres, etc.) et non à des variables temporelles comme la consommation énergétique. Ascione et al. (Ascione et al., 2017) comparent les résultats obtenus avec leur méthode aux résultats obtenus pour des scénarios similaires avec le logiciel de simulation EnergyPlus. L'erreur relative entre les deux méthodes est inférieure à 10% et les auteurs qualifient ces résultats de satisfaisants. Cette méthode basée sur les données est donc dans leur contexte aussi fiable qu'un logiciel complexe de simulation qui nécessite plus de données sur le bâtiment et plus de temps de calcul.

Re-Cecconi et al. (Re Cecconi et al., 2019) utilisent également une méthode de réseau de neurones artificiels pour la prédiction du potentiel d'économie. La différence avec le travail de Ascione et al. (Ascione et al., 2017) est qu'ici l'impact géographique est pris en compte comme entrée dans l'algorithme, ce qui n'était pas le cas pour l'étude précédente. Ce facteur se traduit par l'utilisation des degrés-jours annuels de chauffage en entrée du modèle. Les variables d'entraînement correspondent ici aussi à des variables d'état. L'étude se concentre uniquement sur des écoles. Re-Cecconi et al. (Re Cecconi et al., 2019) cherchent à prédire l'économie d'énergie en chauffage selon différents scénarios de rénovation. Bien que leurs résultats ne soient pas comparés à une méthode classique de prédiction de l'économie (en utilisant des outils de simulation), les auteurs rappellent que le principal avantage de l'approche proposée concerne la possibilité de calculer la consommation énergétique en chauffage post-rénovation des bâtiments sans inspection sur place et en utilisant des paramètres qui sont facilement récupérés ou calculés.

Deb et Lee (Deb & Lee, 2018) proposent une méthode utilisant des algorithmes de regroupement pour rassembler les bâtiments selon l'économie potentielle suite à des rénovations. En s'intéressant à 54 bâtiments de bureaux situés à Singapour, les auteurs cherchent à identifier le jeu de paramètres d'entrée parmi l'ensemble des paramètres disponibles qui permet le meilleur regroupement des

bâtiments dans des catégories d'économie. Ils cherchent donc les variables qui ont le plus grand impact sur l'économie. Ils utilisent un algorithme K-means et testent l'ensemble des combinaisons possibles des 14 paramètres disponibles afin de regrouper les bâtiments dans des groupes (« clusters ») selon leurs économies normalisées (kWh/m^2). Les paramètres disponibles sont soit des paramètres d'état, soit des paramètres temporels sur une résolution annuelle. Ils concluent que pour une séparation en deux ou trois groupes de potentiel d'économie, le meilleur jeu de paramètres est : la surface au sol, la consommation énergétique qui n'est pas associée à de la climatisation, l'efficacité moyenne du système de climatisation, et sa capacité. Ils obtiennent pour cette combinaison de paramètres et deux « clusters », tels que présentés à la Figure 2.1, où EUI est l'intensité de consommation énergétique (« energy use intensity »).

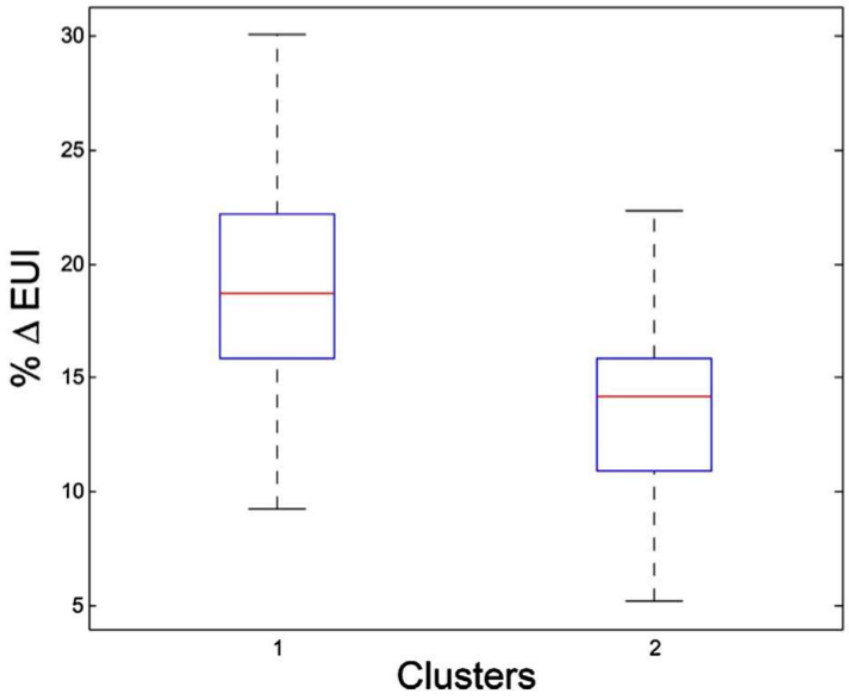

Figure 2.1. Diagramme « boîtes à moustaches » pour la prédiction du groupe d'économie (%) selon deux « clusters » (Deb & Lee, 2018) - reproduction avec autorisation

Li et al. (Li et al., 2019) présentent un outil reposant sur une méthode alternative hybride utilisant un modèle physique et un modèle basé sur les données pour prédire l'économie d'énergie globale

d'un bâtiment. La méthodologie derrière l'outil consiste à d'abord modéliser l'ensemble des bâtiments de la base de données selon un modèle « change-point ». Les données utilisées pour établir les modèles sont les consommations énergétiques avec une résolution mensuelle. Cela permet de caractériser l'ensemble des bâtiments selon cinq paramètres (b_i) et également d'obtenir un modèle pour un scénario de base sans rénovation. Une fois ces paramètres obtenus, les auteurs étalonnent chaque paramètre indépendamment et ils font pour l'ensemble des bâtiments une distribution des paramètres qui leur permet d'identifier une moyenne et une répartition statistique de chaque paramètre. Des limites sont alors fixées pour chaque paramètre pour définir un scénario du paramètre conservateur, moyen ou efficace (du moins bon au meilleur scénario). Un scénario global du bâtiment peut ensuite être défini, pour lequel tous les paramètres doivent être meilleurs ou égaux au seuil du scénario. Ce seuil est alors choisi pour chaque paramètre (si ce dernier est moins bon) pour virtuellement obtenir le nouveau modèle « change-point » du bâtiment rénové et ainsi obtenir une consommation énergétique du bâtiment si ces paramètres étaient améliorés suite à une rénovation. Enfin, la consommation avec le scénario de base et la consommation avec le scénario final peuvent être comparées afin d'obtenir l'économie d'énergie. Des mesures d'économie d'énergie sont également associées à chaque amélioration des paramètres du modèle pour pouvoir proposer des pistes de rénovations aux utilisateurs.

La prédiction du potentiel de l'économie avant le projet n'est pas la seule application de l'apprentissage machine en science du bâtiment lié à l'économie énergétique. En effet, il est également important de pouvoir faire le suivi après projet et de vérifier que les économies vendues aux contrats sont atteintes. Cela correspond à la phase de mesures et vérification (M&V). Afin de pouvoir vérifier les économies, il faut donc avoir une consommation de référence « si les travaux n'avaient pas eu lieu » qui permet de comparer la consommation énergétique sur des conditions similaires. Yeonsook et al. (Yeonsook & Zavala, 2012) présentent une méthode de calcul de cette consommation de référence si les travaux n'avaient pas eu lieu. Leur méthode repose sur un modèle de procédé gaussien (PG) selon un modèle bayésien en utilisant des données avec une résolution horaire. Les auteurs montrent que ce modèle est plus fiable que des modèles obtenus via régression linéaire, car les PG permettent de capturer des phénomènes complexes et non linéaires entre les différentes variables en jeu. Yeonsook et al. (Yeonsook & Zavala, 2012) utilisent la méthode développée pour calculer la consommation horaire d'un climatiseur. Le profil horaire de la consommation du climatiseur comparée au modèle comme présenté à la Figure 2.2 permet de

valider l'utilisation de PG pour la prédiction de la consommation d'un climatiseur et donc l'utilisation d'un PG dans la phase de M&V.

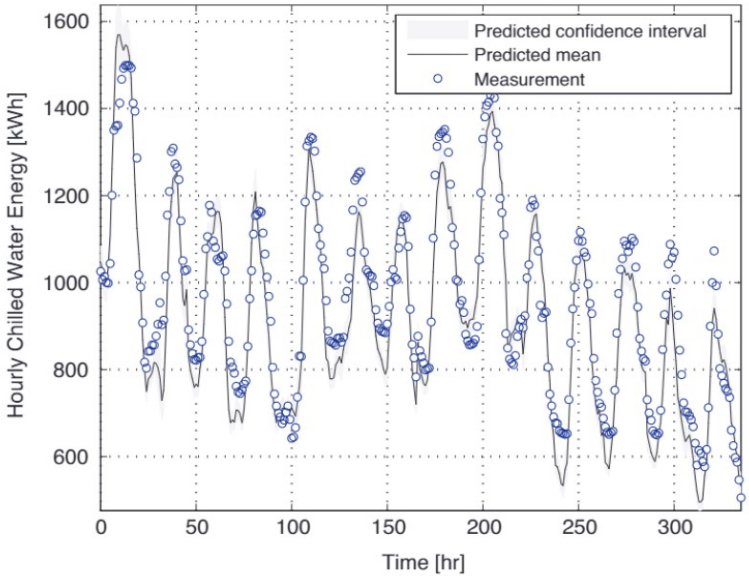

Figure 2.2. Comparaison horaire entre les mesures et le modèle développé par Yeonsook et al. (Yeonsook & Zavala, 2012) - reproduction avec autorisation

D'autres méthodes de modélisation existent pour obtenir la consommation de référence. Zhang et al. (Zhang et al., 2015) comparent quatre types de modélisation différente : un modèle « change-point », un modèle de mélange gaussien, un modèle de procédé gaussien, et un modèle en réseau de neurones artificiels. Ces modèles sont appliqués à un bureau pour prédire la consommation d'énergie pour l'eau chaude servant au chauffage, à la ventilation et au conditionnement de l'air (CVCA). Les auteurs ont réalisé pour chaque modèle une prédiction horaire et une prédiction journalière de cette consommation afin de valider quel modèle fonctionne le mieux et dans quel contexte, en utilisant pour l'entraînement respectivement des données horaires et journalière. En évaluant le R^2, le *RMSE*, le *CV* et le *MBE*, ils concluent que dans leur contexte, le réseau de neurones artificiels génère les moins bons résultats, ce qui est dû à la quantité insuffisante de données d'entraînement, et le modèle ayant les meilleurs résultats est le modèle de mélange gaussien. Le modèle change-point est quand même celui qui est défini par les auteurs comme le

plus approprié dans le contexte de leur étude, car en termes d'effort de modélisation contre la précision, il constitue le modèle le plus fiable et le plus simple à implémenter.

2.2 Algorithmes d'apprentissage machine utilisés en science du bâtiment

L'apprentissage machine est une méthode rapide pour la prédiction puisque l'opérateur n'a pas à générer ou à connaître le modèle complet car la machine le fait automatiquement. Différents algorithmes d'apprentissage machine existent et c'est à l'opérateur de choisir lequel utiliser selon le contexte : quantité de données, objectifs, erreur tolérée, etc. Dans cette section, les différents types d'algorithmes d'apprentissage machine utilisés dans la littérature en science du bâtiment sont présentés afin de faire une rapide présentation du fonctionnement de l'algorithme et des applications associées. Les réseaux de neurones artificiels, les modèles par processus gaussien, les modèles linéaires multivariables, les modèles en structure arborescente et les modèles par partitionnement sont successivement discutés.

Réseau de neurones artificiels

Les réseaux de neurones artificiels (« artificial neural networks », ANN), est un type de modélisation imitant les connexions neuroniques d'un cerveau humain, d'où le nom du modèle. Le principe de fonctionnement d'un tel algorithme est que chaque variable d'entrée est mise en lien avec toutes les autres variables. Par la suite, il y a une couche (« layer ») d'entrée où les variables d'entrées sont fournies. Suite à cela, des couches cachées vont réaliser des opérations sur les variables afin de capturer la complexité des relations entre elles. Enfin, une couche de sortie va permettre d'obtenir la ou les variables de sorties désirées. Les différentes couches sont liées entre elles par une relation mathématique (linéaire, exponentielle, etc.) mettant en relation l'ensemble des éléments de la couche $n-1$ avec les éléments de la couche n. Ces éléments de couche sont appelés neurones. Le schéma de fonctionnement d'un tel algorithme est présenté à la Figure 2.3.

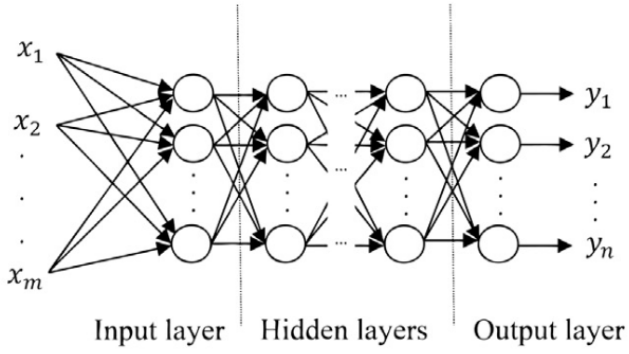

Figure 2.3. Schéma de fonctionnement d'un modèle ANN (Yixuan et al., 2018) - reproduction avec autorisation

Ce modèle est dit implicite, car l'ensemble des couches centrales sont cachées, ce sont des « hidden layers ». L'opérateur ne connaît pas directement les équations mathématiques liant toutes les couches. Cependant, c'est à l'opérateur de définir le nombre de couches et les types de relations mathématiques devant lier chacune des couches. Le nombre de couches optimal est à déterminer expérimentalement, notamment par une validation du modèle avec les données de test. Plus le nombre de couches augmente, plus il y a de risque de trop coller aux données d'entraînement (surapprentissage des données) et le modèle ne se généralise ensuite pas bien (Ascione et al., 2017). Il faut alors réaliser une réitération du modèle en faisant varier le nombre de couches pour déterminer au mieux ce nombre. Ce type de modèle nécessite cependant une grande quantité de données pour pouvoir générer des résultats fiables. En effet, Conraud (Conraud-Bianchi, 2008) rappelle que pour un ANN utilisé dans le domaine du bâtiment, il faut un minimum d'échantillon d'entrainement tel que :

$$N_{echantillon} > 5 * V_{entrée} * V_{sortie} \quad (2.4)$$

où $N_{echantillon}$ est le nombre d'échantillon, $V_{entrée}$ est le nombre de variables d'entrée et V_{sortie} est le nombre de variables de sortie. Si l'on ajoute la séparation des échantillons disponibles en un

jeu d'entraînement et un jeu de test, il faut que le jeu d'échantillons disponible initialement soit très important.

Comme présenté dans la **section 2.1.2**, il existe de nombreuses applications scientifiques en science du bâtiment utilisant ce type de modélisation. En plus des travaux cités précédemment, on peut également noter les travaux de Amasyali et El-Gohary (Amasyali & El-Gohary, 2018) qui font la revue des travaux scientifiques traitant d'apprentissage machine dans le bâtiment. Ils obtiennent sur l'ensemble des documents traités que 47% utilisent une modélisation ANN, pour des raisons multiples et variées. Cela permet d'affirmer que l'utilisation des réseaux de neurones artificiels est démocratisée en science du bâtiment.

Modèle par processus gaussien

Le modèle de régression par processus Gaussien (RPG) consiste à déterminer la ou les sorties grâce à une densité de probabilité établie avec les données d'entraînement et les variables d'entrée. Ce modèle a une implémentation qui est généralement très rapide pour des résultats précis. La création d'un tel modèle repose sur la définition de deux fonctions : une fonction « moyenne », et une fonction « covariance »:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{2.5}$$

où $m(x)$ est la fonction moyenne, et $k(x, x')$ est la fonction covariance. La plupart des outils numériques permettent d'automatiser une partie de ce processus, ainsi il suffit de choisir la forme des équations selon le type de données, et l'algorithme pourra trouver automatiquement les paramètres de l'équation. En général, une constante ou une fonction polynomiale de degré simple peut être utilisée pour la fonction moyenne : la forme finale de l'équation est définie par le jeu de donnée, une fonction moyenne simple n'est ainsi pas trop restrictive. L'enjeu principal dans un modèle RPG est la détermination de la forme de la fonction de covariance : selon le type de donnée, il faut combiner différents types d'équations pour obtenir une forme convenable d'équation finale.

Rajagopal et Young (Rajagopal & Hae Young, 2013) utilisent ce type de modélisation pour la prédiction à long terme (quelques jours) de la consommation horaire en éclairage et pour les systèmes CVCA de l'université de Stanford. Ils arrivent à obtenir une prédiction avec un coefficient de détermination de 0,987 et un intervalle de confiance de 95% contenant l'ensemble des données. Ils concluent que cette méthode de prévision peut améliorer les performances des systèmes de réseaux intelligents en permettant un approvisionnement énergétique fiable grâce à une planification précise. De plus, elle peut être utilisée comme méthode de diagnostic et de contrôle pour détecter les dysfonctionnements du système et améliorer l'efficacité énergétique.

Prakash et al. (Prakash et al., 2018) utilisent également ce type de modélisation pour la prédiction des charges énergétiques de trois campus américains. Une prédiction à court terme (jusqu'à 10 minutes) et à long terme (1 à 5 jours) sont réalisées pour prédire la consommation horaire. Le modèle des auteurs a donné une précision de prédiction allant jusqu'à 94,38 % et 99,26 % pour les prévisions à long et à court terme, respectivement.

Modèle linaire multivariable

Le modèle par régression linéaire multivariable (RLM) est un modèle qui est mathématiquement simple. Pour obtenir un tel modèle, il s'agit de faire une régression linéaire entre les variables d'entrées et la variable de sortie. Il est à noter qu'un modèle RLM ne permet de fournir qu'une seule variable de sortie, contrairement aux autres algorithmes présentés précédemment. Le modèle RLM s'exprime mathématiquement par :

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_n * x_n + \varepsilon \qquad (2.6)$$

où Y est la variable de sortie modélisée, β_0 est l'ordonnée à l'origine, les β_i sont les coefficients estimés du modèle RLM, les x_i sont les variables d'entrée et ε est l'erreur du modèle. Bien que ce modèle soit simple, il faut cependant que les variables d'entrée soient choisies de manière pertinente pour la représentation du phénomène physique désiré. Dans le contexte des bâtiments, si un tel modèle cherche à prédire la consommation énergétique, ce type de modèle ne fonctionne

que si et seulement si les variables d'entrée incluses dans le modèle affectent la consommation énergétique finale du bâtiment (Capozzoli et al., 2015).

Capozzoli et al. (Capozzoli et al., 2015) utilisent cette modélisation pour évaluer la consommation énergétique en chauffage pour des écoles italiennes. Avec un jeu de neuf paramètres sur les caractéristiques et les opérations des bâtiments, ils arrivent à prédire la consommation annuelle des écoles avec une erreur relative moyenne de 15%. Les auteurs concluent sur l'efficacité du modèle RLM dans leur contexte, mais expliquent cependant que certaines variables nécessaires à la construction du modèle sont parfois difficiles à obtenir comme la conductivité thermique des mur/fenêtres, il faut donc s'assurer de pouvoir obtenir un nombre important de variables sur l'ensemble des bâtiments.

Chung et al. (Chung et al., 2006) utilisent le même type de modélisation, mais pour des bâtiments commerciaux. Ils cherchent à développer un modèle RLM permettant d'obtenir l'intensité de la consommation d'énergie normalisée, c'est-à-dire où l'ensemble des variables d'entrées sont standardisées selon un Z-score. Une fois le modèle développé, ils sont capables de faire une analyse comparative au niveau des EUI afin d'établir des limites de percentiles pour les consommations des bâtiments de leur base et de l'ensemble des bâtiments commerciaux grâce à une méthode par amorçage.

Modèle en structure arborescente

Les modèles en structure arborescente sont les modèles d'arbre de décision ou de forêt aléatoire. L'adjectif arborescent définit la représentation graphique de ces modèles qui prend la forme d'un arbre où chaque embranchement correspond à une décision basée sur les variables d'entrée. L'arbre de décision (« decision tree », DT) est la version simple de ce type de modèle. On cherche ici soit à faire une régression, c'est-à-dire prédire une valeur continue, ou à faire de la classification, c'est-à-dire prédire une classe ou une valeur discrète. Le DT est construit en divisant un nœud parent en plusieurs nœuds enfants à plusieurs reprises, en commençant par le nœud racine qui contient l'ensemble de l'échantillon d'apprentissage. Chaque nœud enfant contient un sous-échantillon respectant une condition d'embranchement précise. Les derniers nœuds (les nœuds sans nœuds enfants) sont les résultantes de l'arbre. La forêt aléatoire (« random forest », RF) est un algorithme d'ensemble, c'est-à-dire qu'il est composé de plusieurs algorithmes plus simples, en l'occurrence

ici il est composé de plusieurs DT. Chaque DT composant un RF est légèrement différent : les variables d'entrées n'ont pas toujours le même poids et les échantillons d'entrées ne sont pas toujours les mêmes. Ce processus est réalisé automatiquement par la machine. Cette composition d'ensemble permet d'avoir un modèle plus robuste face au biais induit par l'échantillon fixe de bâtiments d'entraînement, et de mieux généraliser le modèle que dans le cas d'un simple DT. Cependant, c'est également un modèle plus lourd et complexe, il faut donc selon l'objectif et le contexte définir si un algorithme DT simple suffit ou si un RF est nécessaire.

La modélisation par DT est utilisée par Capozzoli et al. (Capozzoli et al., 2015) dans la même étude et le même contexte que pour la modélisation RLM cité précédemment. Les auteurs comparent ces deux méthodes dans la prédiction de la consommation énergétique en chauffage pour des écoles italiennes. La méthode par DT permet d'obtenir une erreur relative moyenne de 14% (contre 15% pour RLM) avec un jeu de seulement cinq variables d'entrée pertinentes (contre 9 variables d'entrée pertinentes pour l'algorithme RLM). Bien que les résultats soient meilleurs pour le modèle DT, les auteurs concluent en affirmant que les deux modélisations sont complémentaires et présentent des forces et des faiblesses différentes. Le plus grand avantage du DT est que la sortie consiste en un ensemble de règles de décision pratiques que l'opérateur peut rapidement utiliser. Le plus grand avantage du RLM est qu'il fournit une équation mathématique avec l'ensemble des variables pertinentes représentées.

Prakash et al. (Prakash et al., 2018) utilisent la modélisation RF pour comparer la prédiction de la consommation à court et long terme obtenu grâce à leur modélisation RPG. Les auteurs définissent le modèle RF comme l'une des méthodes de prévision de pointe et donc ce modèle leur sert de référence. Bien que le cœur de leur étude ne soit pas cette modélisation, il apparaît ici qu'une modélisation RF peut servir de référence par rapport à des méthodes plus expérimentales.

Modèle par partitionnement

Un modèle par partionnement, ou « cluster » en anglais, est une méthode de regroupement non supervisée. Différents groupes ayant au sein du même groupe des propriétés similaires peuvent être obtenus, ce qui permet ensuite une utilisation ou une interprétation de ces différents groupes. Cette méthode peut être utilisée notamment pour transformer une régression en classification. Un des algorithmes les plus populaires pour un modèle par partitionnement est l'algorithme « K-means ».

C'est un algorithme simple d'implémentation et efficace. Il est donc souvent privilégié pour générer automatiquement des groupes. Cet algorithme repose sur une minimisation de la distance euclidienne au sein d'un même « cluster ». Ainsi, tous les éléments d'un même groupe sont proches (en termes de distance mathématique) du centre du groupe, appelé « centroïde ». Cette distance est évaluée selon l'équation :

$$d_{AB}^2 = \sum_{j=1}^{p} (a_j - b_j)^2 \tag{2.7}$$

où d_{AB} est la distance entre les éléments A et B, a est la position de l'élément A dans l'espace (a= (a_1, a_2, …, a_p)), b est la position de l'élément B, et p est la dimension de l'espace, soit le nombre de paramètres utilisés pour réaliser le regroupement. Cependant, afin de pouvoir exécuter l'algorithme, il faut également définir dès le début le nombre de groupe pour que l'algorithme puisse ensuite automatiquement regrouper les données selon la bonne résolution de groupe. La méthode utilisée pour déterminer le nombre de groupe optimal est appelée la méthode du coude : il faut faire varier le nombre de groupe pas à pas, et évaluer pour chaque point la somme des erreurs au carré entre les données et leur centroïde. On obtient ainsi une courbe ayant un coude (transition rapide de la pente de la courbe), et ce point correspond au nombre de groupe optimal. D'autres algorithmes de regroupement existent, comme l'algorithme « DBscan », mais ils ne seront pas abordés ici par manque de pertinence avec les objectifs de l'étude.

Kazaki et Papadopoulos (Kazaki & Papadopoulos, 2018) utilisent l'algorithme « K-means » dans leur étude pour regrouper les profils de charges mensuelles d'une université grecque. Ils arrivent ainsi en décomposant les données de charges en fréquence par une transformée de Fourrier, à créer quatre groupes de profil de charges distinct, chaque groupe représentant un comportement énergétique différent. Ces différents groupes peuvent ensuite être sujets de différentes études séparées selon le besoin, bien que les auteurs ne réalisent pas eux-mêmes ces études séparées dans leur article.

CHAPITRE 3 PRÉSENTATION DE LA BASE DE DONNÉES

La base de données d'Ecosystem contient des données pour 540 bâtiments situés au Canada sur lesquels l'entreprise est intervenue lors de projets de travaux de rénovation énergétique entre 2006 et 2020. On y retrouve pour chaque bâtiment des métadonnées (ou données générales), des données mensuelles issues des factures énergétiques, et des données calculées qui ont été ajoutées à la base de données pour chaque projet.

3.1 Métadonnées

Les métadonnées correspondent aux données constantes dans le temps pour un bâtiment. Ces données permettent d'identifier le bâtiment et d'obtenir un portrait général du bâtiment sans prendre en compte sa consommation énergétique. Les attributs de métadonnées sont présentés dans le Tableau 3.1. Il y a pour chaque bâtiment de la base de données 11 attributs de métadonnées permettant de caractériser le bâtiment.

Tableau 3.1. Attributs de métadonnées

Nom du champ	**Type**	**Description**
building_id	Nombre entier	Numéro d'identification unique assigné pour chaque bâtiment
building_name	Chaîne de caractères	Nom du bâtiment
project	Chaîne de caractères	Nom du projet
vertical	Chaîne de caractères	Type de bâtiment (ex : Éducation, Industriel, …)

Tableau 3.1. Attributs de métadonnées (suite)

Nom du champ	**Type**	**Description**
region	Chaîne de caractères	Région administrative du Québec ou la province
measures	Chaîne de caractères	Liste des mesures d'économie d'énergie implémentées
area	Nombre entier	Surface du bâtiment en m^2
address	Chaîne de caractères	Adresse du bâtiment
users	Chaîne de caractères	Liste des employés d'Ecosystem impliqués dans le projet
gj_savings_sold	Nombre entier	Économie d'énergie en GJ vendue au contrat
dollars_savings_sold	Nombre entier	Économie en dollars canadien vendue au contrat
climate_station_name	Chaîne de caractères	Nom de la station climatique associée au bâtiment et aux données climatiques

3.2 Données mensuelles

Les données mensuelles correspondent aux données de facturation sur une résolution mensuelle. Les attributs de données mensuelles sont présentés dans le Tableau 3.2. Il y a 34 attributs distincts pour les données mensuelles. Les données dans la base sont représentées sous la forme d'un tableau

où les colonnes sont les attributs, et les lignes sont les données pour une facture d'un type d'énergie spécifique pour un mois de facturation. Ainsi, pour un bâtiment et un mois donné, on retrouve autant de lignes que de factures énergétiques pour ce mois. Si le bâtiment a deux compteurs électriques et un compteur de gaz naturel, il y aura donc trois lignes de données pour chaque mois.

Tableau 3.2. Attributs mensuels

Nom du champ	**Type**	**Description**
year_start	Date	Date représentant le premier jour du suivi de 12 mois
month	Nombre entier	Nombre correspondant au mois du calendrier, associé à la facture d'énergie
days	Nombre entier	Nombre de jours couverts par la facture
energy	Chaine de caractère	Type d'énergie associé à la mesure « *usage* »
unit	Chaine de caractère	Unité associée à « *energy* ».
meter_index	Nombre entier	Numéro séquentiel associé à chaque facture de même type d'énergie (1, 2, 3, …)
usage	Nombre décimal	Consommation d' « *energy* » en « *unit* » pour le mois correspondant

Tableau 3.2. Attributs mensuels (suite)

Nom du champ	**Type**	**Description**
reference_usage	Nombre décimal	Consommation calculée d' « *energy* » en *unit* pour le mois correspondant si le projet n'avait pas eu lieu
cost	Nombre décimal	Facture en dollars canadien associée à « *usage* »
reference_cost	Nombre décimal	Facture en dollars canadien associée à « *reference_usage* »
peak	Nombre décimal	Demande maximale associée à « *energy* » (exemple : kW crête pour une consommation électrique)
reference_peak	Nombre décimal	Demande maximale calculée/ajustée associée à « *energy* » si le projet n'avait pas eu lieu
is_pre_follow_up	Nombre entier	Un indicateur de la période des données. -1 pour pré-travaux, 1 pour pré-suivi, 0 pour suivi

Tableau 3.2. Attributs mensuels (suite)

Nom du champ	**Type**	**Description**
year_index	Nombre entier	Numéro séquentiel indiquant l'année par rapport à une période où 0 est la première année (ex : si « *is_pre_follow_up* » = 0 et « *year_index* » =0, cela représente la première année de suivi)
year_id	Nombre entier	Numéro d'identification unique associé à chaque année en relation avec la modélisation météo
heating_is_adjusted	Booléen	« *True* » si un modèle de chauffage est associé aux données de consommation
heatind_dd	Nombre décimal	Degrés-jours de chauffage pour le mois
heating_base_temp_celsius	Nombre entier	Température de référence pour le calcul des degrés-jours de chauffage en celsius

Tableau 3.2. Attributs mensuels (suite)

Nom du champ	**Type**	**Description**
heating_base_units	Nombre décimal	Ordonnée à l'origine pour un modèle degrés-jours de chauffage calculé sur une année avec l' « *unit* » correspondante
heating_units_per_dd	Nombre décimal	Pente du modèle degrés-jours de chauffage calculé sur une année avec l' « *unit* » correspondante
heating_r2	Nombre décimal	Coefficient de corrélation entre le modèle degrés-jours de chauffage et les valeurs mesurées
cooling_is_adjusted	Booléen	« *True* » si un modèle de climatisation est associé aux données de consommation
cooling_dd	Nombre décimal	Degrés-jours de climatisation pour le mois
cooling_base_temp_celsius	Nombre entier	Température de référence pour le calcul des degrés-jours de climatisation

Tableau 3.2. Attributs mensuels (suite)

Nom du champ	**Type**	**Description**
cooling_base_units	Nombre décimal	Ordonnée à l'origine pour un modèle degrés-jours de climatisation calculé sur une année avec l' « *unit* » correspondante
cooling_units_per_dd	Nombre décimal	Pente du modèle degrés-jours de climatisation calculé sur une année avec l' « *unit* » correspondante
cooling_r2	Nombre décimal	Coefficient de corrélation entre le modèle degrés-jours de climatisation et les valeurs mesurées
climate_station_year	Nombre entier	Année du premier mois pour la requête météo
climate_station_start_month	Nombre entier	Premier des 12 mois associés à la requête météo
usage, GJ	Nombre décimal	La consommation d'« *usage* » convertie en GJ
reference_usage, GJ	Nombre décimal	La consommation de « *reference_usage* » convertie en GJ

Tableau 3.2. Attributs mensuels (suite)

Nom du champ	**Type**	**Description**
EUI, GJ/m2	Nombre décimal	Intensité de consommation, « *usage* », *GJ* divisé par la surface du bâtiment en m^2
EUIref, GJ/m2	Nombre décimal	Intensité de consommation de référence, « *reference_usage* », *GJ* divisée par la surface du bâtiment en m^2

Pour chaque bâtiment, on retrouve des données sur trois périodes par rapport à l'avancée des travaux énergétiques : les données prétravaux, les données durant les travaux, et les données post-travaux. Les données prétravaux correspondent aux factures du client avant l'intervention d'Ecosystem. Le nombre d'années pour les données prétravaux est variable selon le projet et dépend principalement des informations que transmet le client en début de projet. Le nombre d'années prétravaux varie d'un à trois ans de données selon le bâtiment. Les données durant les travaux, aussi appelées données présuivi, sont des données intermédiaires, car elles correspondent à un régime transitoire entre un bâtiment non rénové et un bâtiment rénové. Le nombre d'années de travaux est variable selon l'envergure du projet. Le nombre d'années présuivi varie d'un à cinq ans de données selon le bâtiment. Les données post-travaux, aussi appelées données de suivi, sont les données les plus nombreuses, car Ecosystem fait le suivi des projets pour s'assurer que les économies d'énergie vendues aux contrats sont atteintes. Ces données ne sont pas toutes équivalentes à un régime permanent, car lors de la phase de suivi, des nouveaux systèmes ou mesures de contrôle sont parfois implémentés, car les travaux peuvent ne pas être aussi performants que prévu initialement. De plus, des changements opérationnels, de vocations ou même l'agrandissement du bâtiment peuvent survenir et affecter la consommation d'énergie durant la période de suivi. Le nombre d'années de suivi varie d'un à huit ans selon le bâtiment.

Parmi les données mensuelles, deux sous-catégories d'attributs se distinguent : les données directement issues des factures, et les données issues d'analyse énergétique. Les attributs issus de l'analyse énergétique correspondent aux attributs relatifs au modèle degrés-jour, et aux attributs étant identifiés par la mention « reference » ou « ref » dans leur nom, ce sont les attributs qui représentent des données si les projets de travaux énergétiques n'avaient pas eu lieu.

Les modèles degrés-jour de chauffage et de climatisation sont calculés pour chaque bâtiment à partir d'une année de référence, généralement la première année précédant les travaux ou une moyenne de quelques années précédant les travaux. Les degrés-jours de chauffage (DJ de chauffage) correspondent à l'écart entre une température de référence de 18 °C et la température moyenne journalière si cette dernière est inférieure à la température de référence (Atlas Climatique du Canada, 2021). Les degrés-jours de climatisation (DJ de climatisation) correspondent à l'écart entre une température de référence de 18 °C et la température moyenne journalière si cette dernière est supérieure à la température de référence. Une fois ces degrés-jour journaliers calculés, on peut les additionner sur un mois pour obtenir des degrés-jours mensuels. Pour un bâtiment donné, le modèle est obtenu avec une régression linéaire entre la consommation mensuelle globale (c'est-à-dire tout type d'énergie confondu) et les degrés-jours en chauffage et en climatisation mensuels. On obtient :

$$\hat{E}_{month,i} = b_0 + b_1 * DJ_{chauffage} + b_2 * DJ_{climatisation} \quad (3.1)$$

où $\hat{E}_{month,i}$ est la consommation énergétique mensuelle calculée, les b_j sont les facteurs associés au modèle, et $DJ_{chauffage}$ et $DJ_{climatisation}$ sont les degrés-jours mensuels qui proviennent des données climatiques. Les informations sur le modèle sont stockées dans la base de données dans les différents attributs associés au modèle.

Le modèle degrés-jours est ensuite utilisé pour calculer les attributs de consommation de « reference », c'est-à-dire une consommation s'il n'y avait pas eu de travaux de rénovation énergétique. Pour calculer ces consommations pour un bâtiment spécifique, l'ingénieur responsable du projet utilise le modèle degrés-jour obtenu sur l'année de référence pour calculer la consommation d'énergie sur une base mensuelle. Cette consommation peut ensuite être ajustée

selon les événements particuliers du mois (interventions spécifiques, événements particuliers, pandémie, etc.). Cependant, l'information sur la pondération appliquée n'est pas disponible dans la base de données et il est donc impossible de déterminer le modèle exact qui a été appliqué pour obtenir les consommations de référence.

3.3 Données ajoutées

Les données ajoutées correspondent à des attributs intermédiaires qui ont été récupérés et/ou calculés pour lesquels l'obtention est simple, mais qui seront par la suite utiles pour la méthodologie employée dans les chapitres suivants. Les attributs ajoutés à la base de données sont présentés au Tableau 3.3. La température extérieure moyenne mensuelle, $\bar{T}^{(n)}_{month,i}$, est récupérée pour chaque bâtiment et chaque mois. Ces températures ont été obtenues en utilisant la base de données de la « National Oceanic and Atmospheric Administration » (NOAA) et leur base de données « Integrated Surface Database (ISD) » (NOAA - National Centers for Environmental Information). L'ISD fournit des données météorologiques horaires associées à une station météo. Les données accessibles ne sont pas similaires entre chaque station météo, les équipements de mesures varient, et les données sur certains pas de temps peuvent être manquantes. Ainsi, pour obtenir la température moyenne mensuelle pour chaque bâtiment, l'adresse du bâtiment est associée aux cinq stations météo les plus proches via sa latitude et longitude, et la distance euclidienne avec les stations de la base de la NOAA. Une fois ces cinq stations obtenues, les températures moyennes horaires sont récupérées pour la station la plus proche. Pour un an de données météo, s'il y a moins de 15% de données manquantes pour cette station, alors les températures moyennes mensuelles sont calculées à partir des moyennes des données horaires. Sinon, s'il y a plus de 15% de données manquantes pour un an et pour la station la plus proche, on passe à la seconde station la plus proche pour obtenir le jeu de données météo horaire d'une année, et ainsi de suite si la condition des 15% n'est toujours pas respectée, jusqu'à obtenir moins de 15% de données manquantes par année. Une fois les températures moyennes mensuelles obtenues pour chaque année désirée, elles sont ajoutées à la base de données.

Tableau 3.3. Attributs ajoutés

Nomenclature	**Type**	**Description**
$\bar{T}_{month,i}^{(n)}$	Nombre décimal	Température extérieure moyenne mensuelle pour l'année n et le mois i
type	Chaine de caractères	Version détaillée de l'attribut « vertical » (ex : Pour « vertical » = « Education », on retrouve « primaire », « secondaire », etc
$E_{month,i}^{''(n)}$	Nombre décimal	Consommation énergétique mensuelle globale par mètre carré pour l'année n et le mois i
$Y_{month,i}^{(n)}$	Nombre décimal	Économie énergétique mensuelle en GJ pour l'année n et le mois i
$Y_{year}^{(n)}$	Nombre décimal	Économie énergétique annuelle en GJ pour l'année n
$Y_{year}^{''(n)}$	Nombre décimal	Économie énergétique annuelle en GJ /m^2 pour l'année n
$\bar{Y}_{year}^{''}$	Nombre décimal	Économie énergétique annuelle moyenne en GJ /m^2 sur les années de suivi

Les différentes économies sont calculées selon :

$$Y_{month,i}^{(n)} = \hat{E}_{month,i}^{*(n)} - E_{month,i}^{(n)} \tag{3.2}$$

où $\hat{E}_{month,i}^{*(n)}$ est la consommation énergétique mensuelle globale de référence pour le mois i de l'année n (calculé avec le modèle DJ) si les travaux n'avaient pas eu lieu. Le symbole « * » indique la consommation d'énergie d'un bâtiment pré-travaux. L'accent « ˆ » indique une valeur calculée à partir d'un modèle du bâtiment.

Les économies annuelles sont alors :

$$Y_{year}^{(n)} = \sum_{i=1}^{12} Y_{month,i}^{(n)} \tag{3.3}$$

$$Y_{year}^{\prime\prime(n)} = \frac{Y_{year}^{(n)}}{A} \tag{3.4}$$

$$\bar{Y}_{year}^{\prime\prime} = \frac{1}{N}\sum_{n=1}^{N} Y_{year}^{\prime\prime(n)} \tag{3.5}$$

où A est la surface du bâtiment en m^2 et N est le nombre d'années de suivi.

La Figure 3.1 représente schématiquement l'économie d'énergie par rapport à la chronologie d'un projet de rénovation énergétique. Les étoiles vertes sur la figure représentent les données de facturation avant et après travaux $E_{month,i}^{(n)}$, et les étoiles bleues représentent la consommation projetée si les travaux n'avaient pas eu lieu $\hat{E}_{month,i}^{*(n)}$, obtenue en utilisant le modèle théorique du bâtiment avant-travaux.

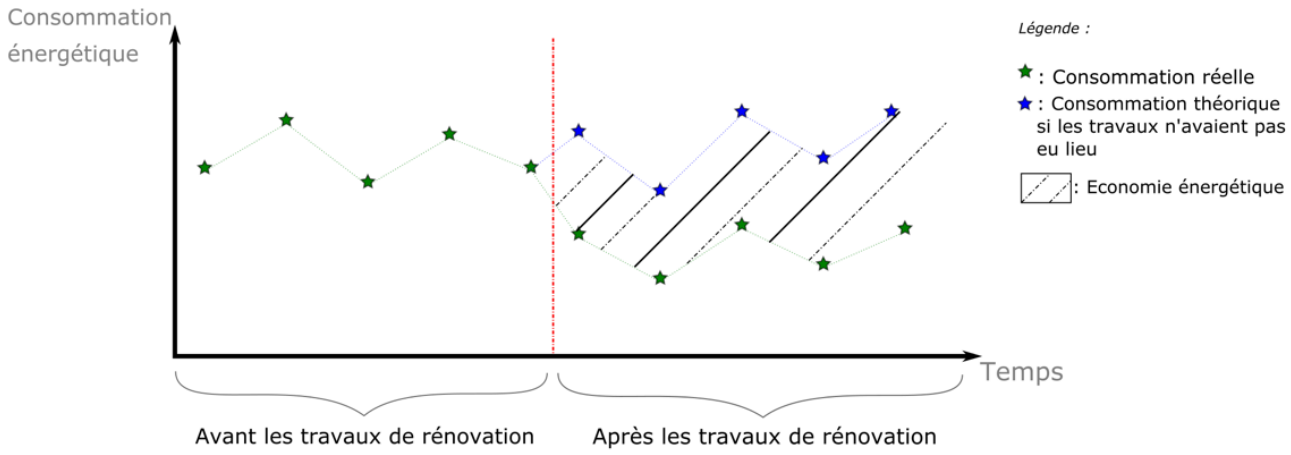

Figure 3.1. Schématisation de l'économie par rapport à la chronologie des travaux de rénovation énergétique

3.4 Statistiques sur la base de données

3.4.1 Ensemble des données

Le nombre de bâtiments, la consommation d'énergie moyenne après le projet de rénovation ainsi que l'économie moyenne sont donnés pour chaque vertical de marché dans le Tableau 3.4. On constate que les bâtiments éducatifs représentent plus de 88% des projets réalisés, suivis ensuite par les bâtiments de type santé (environ 10%), puis enfin les autres types de bâtiments qui sont minoritaires. Dans les bâtiments éducatifs, l'économie moyenne est de 18%, cela signifie qu'en moyenne, le bâtiment consomme 18% moins d'énergie après rénovation qu'avant rénovation.

Tableau 3.4. Statistiques de rénovation par vertical de marché

Vertical	**Nombre de bâtiments**	**Consommation moyenne après projet (GJ/m².an)**	**Économie moyenne réalisée (GJ/m².an)**	**Économie moyenne réalisée (%)**
Éducation	478	0,58	0,14	18%
Santé	52	1,3	1,27	22%
Municipal	4	0,81	0,19	16%
Société immobilière	3	0,77	0,50	38%
Industriel	1	1,27	0,38	23%
Tour à bureaux	1	0,74	1,17	64%

D'un point de vue géographique, les projets de la base de données sont tous situés au Canada. Si l'on s'intéresse plus particulièrement aux projets visant les bâtiments éducatifs, la Figure 3.2 donne la localisation géographique des projets et leurs consommations énergétiques respectives. La plupart des projets sont regroupés dans l'est du pays, dans la province du Québec, et un seul projet se situe en Alberta. Le contexte climatique de l'étude est donc restreint géographiquement puisque les seuls climats représentés par les données sont le climat subarctique et le climat continental humide pour le Québec, et le climat continental sec pour l'Alberta.

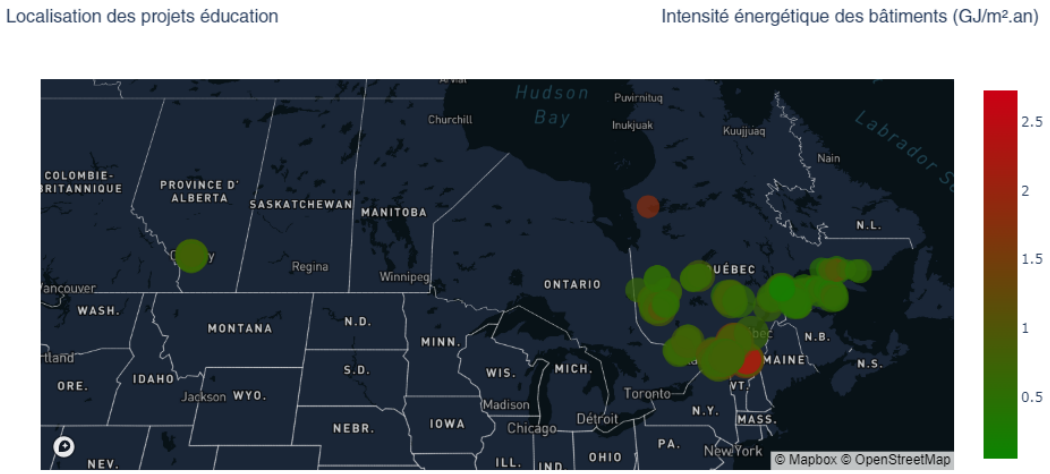

Figure 3.2. Carte des projets de bâtiments éducatifs

3.4.2 Bâtiments éducatifs

La base de données contient des données pour des bâtiments éducatifs, des bâtiments du milieu de la santé et des bâtiments industriels, mais la présente étude ne se concentre que sur les bâtiments éducatifs qui représentent la majorité des bâtiments de la base de données (88% des bâtiments).

On retrouve en tout cinq types de bâtiments éducatifs différents :

- Les écoles primaires : des écoles accueillant des élèves de 6 à 11 ans, entre la 1ère et la 6e année primaire. Ces écoles primaires comprennent parfois des classes de maternelles pour les élèves de 4 à 6 ans, mais ce n'est pas le cas pour toutes les écoles primaires.
- Les écoles secondaires : des écoles accueillant des élèves de 11 à 16 ans, entre la 1ère et la 5e année secondaire. Elles forment avec les écoles primaires un groupe appelé K12 (« Kindergarten to grade 12 »), soit les écoles accueillant des élèves entre la maternelle et la 5e année secondaire.
- Les CEGEP : l'acronyme de « Collège d'enseignement général et professionnel ». Ce sont des établissements scolaires uniques à la région du Québec, où le premier niveau de l'enseignement supérieur est dispensé.
- Les universités.

- Les centres de formation professionnelle.

Le Tableau 3.5 présente pour chaque type de bâtiment le nombre de bâtiment ayant des données pré-travaux et de suivi complètes disponibles dans la base de données.

Tableau 3.5. Nombre de bâtiments par type

Type de bâtiment	Primaire	Secondaire	Cegep	Université	Centre de formation
Nombre d'échantillons	272	147	6	3	50

Ces types de bâtiments éducatifs peuvent être regroupés en catégories de bâtiments. Ces regroupements consistent à combiner différents types de bâtiments dans un même groupe selon des critères logiques afin d'obtenir des tailles d'échantillons plus grandes. Ces différents groupes sont : les K12, les K12 clés, les bâtiments d'éducation supérieure, et l'ensemble des bâtiments éducatifs.

K12

Les K12 sont le regroupement des écoles primaires et secondaires. Elles forment un groupe par la ressemblance au niveau des activités qui s'y déroulent. Il s'agit d'élèves de 4 à 16 ans qui assistent à des cours en salle, sans activité particulière énergétiquement parlant. Ce sont des bâtiments d'envergure petite ou moyenne (en moyenne 7 500 m^2 de surface dans la base de données).

K12 clés

Pour des missions annexes d'Ecosystem, certains bâtiments ont été ciblés comme étant des projets représentant bien les activités de l'entreprise, dont des K12 qui font donc partie des K12 clés. Un traitement particulier a été réalisé sur les données liées à ces projets pour s'assurer d'obtenir sur ce

jeu une base de données propre et complète. Ce traitement consiste à s'assurer que les données de consommation soient correctement entrées et complètes dans la base de données.

Bâtiments d'éducation supérieure

Les bâtiments pour l'éducation supérieure sont le regroupement des cégeps, des universités et des centres de formation professionnelle. Ce sont principalement des élèves de 17 ans et plus qui fréquentent ces établissements et les activités qui y sont réalisées sont très diversifiées et dépendent de la spécialité de l'établissement. Ce sont des bâtiments de grande envergure qui accueillent un grand nombre d'étudiants (en moyenne 43 000 m^2 dans la base de données).

Ensemble des bâtiments éducatifs

L'ensemble des bâtiments éducatifs est le regroupement de l'ensemble des bâtiments éducatifs. Le point commun de ce groupe est le caractère éducatif de l'établissement. C'est le groupe le plus diversifié en termes d'activités et de taille.

Le Tableau 3.6 présente pour chaque groupe de type de bâtiment le nombre de bâtiments ayant des données d'avant travaux et de suivi complètes disponibles dans la base de données.

Tableau 3.6. Nombre de bâtiments par regroupement

Groupe de bâtiment	K12	K12 clés	Éducation supérieure	Bâtiments éducatifs
Nombre d'échantillons	419	157	50	479

CHAPITRE 4 SÉLECTION DE LA MÉTHODE DE PRÉDICTION DE L'ÉCONOMIE D'ÉNERGIE

L'objectif de ce chapitre est de trouver et valider une méthode pour la prédiction de l'économie d'énergie. Afin d'évaluer différentes méthodes, un échantillon simple de la base de données doit également être défini afin de pouvoir valider une méthode dans un cadre simplifié.

4.1 Sélection de l'échantillon

Puisque Ecosystem est principalement sollicité pour des bâtiments scolaires comme présenté dans le Tableau 3.4, un échantillon des bâtiments éducatifs est utilisé afin de choisir la méthode pour la prédiction de l'économie. En effet, les bâtiments éducatifs représentent 479 bâtiments dans la base de données, soit 88% des projets. C'est donc ce groupe de bâtiments qui a été retenu pour cette étude et c'est dans ce groupe qu'un échantillon doit être obtenu.

Il existe au sein des bâtiments éducatifs différents types de bâtiments, comme présenté dans le Tableau 3.5. Les comportements énergétiques similaires des bâtiments du même échantillon permettent de restreindre le cadre de la modélisation et donc de faciliter la prédiction. Des comportements similaires signifient des activités intérieures proches, des taux de fréquentation qui à l'année évoluent de manière similaire, et qui ont tendance à utiliser l'énergie de la même façon. Pour obtenir des bâtiments similaires, seules les écoles dont l'attribut « type » est primaire (écoles primaires pour le Québec et « elementary school » pour les autres provinces) ont été sélectionnées, soit 272 écoles. Le choix de cet échantillon est justifié, car c'est l'échantillon des bâtiments éducatifs le plus important (57% des bâtiments éducatifs sont des écoles primaires) et les activités entre les différentes écoles primaires sont très peu diversifiées, peu importe le lieu. Cependant, dans cet échantillon, certaines écoles n'avaient pas de données prétravaux, certaines n'avaient pas encore de données de suivi, et certaines avaient des données incomplètes ou erronées. Ces écoles ont donc été exclues de l'échantillon de données. Suite à cette exclusion, il reste 175 écoles primaires.

Le cas simplifié permettant de valider ou rejeter les différentes méthodes est réalisé sur un échantillon de 175 bâtiments composé uniquement d'écoles primaires canadiennes. Les méthodes validées pourront ensuite être développées et appliquées à d'autres types de bâtiments.

4.2 Modèle « change-point »

Bien que l'échantillon de données ne regroupe que des écoles primaires, chaque école est unique et les caractéristiques de chaque école sont variables selon l'école : la taille, le nombre d'élèves inscrits, la localisation, les systèmes énergétiques installés, etc. Pour utiliser un outil d'apprentissage machine basé sur les données de plusieurs écoles différentes, il faut d'abord obtenir des mesures ou grandeurs comparables entre les écoles qui permettent la prédiction de la consommation d'énergie (ou bien de l'économie d'énergie) et qui serviront d'entrées à l'outil d'apprentissage machine. Dans le cadre d'un outil énergétique, l'ASHRAE conseille d'utiliser un modèle en régime permanent pour caractériser des bâtiments lorsque les données sont mensuelles (American Society of Heating Refrigerating Air-Conditioning Engineers, 2017) (en opposition avec un modèle dynamique). En s'appuyant sur les données disponibles et la revue de littérature, notamment les travaux ayant permis de développer l'outil « Building Efficiency Targeting Tool for Energy Retrofits » (BETTER) issu d'une collaboration entre le Lawrence Berkeley National Laboratory (Berkeley Lab) et Johnson Controls (Johnson Controls, 2021), ou encore l'outil Wattscale (Iyengar et al., 2020), une modélisation via un modèle « change-point » pour caractériser le comportement énergétique des bâtiments est la plus indiquée, car la base de données dispose de consommation énergétique avec une résolution mensuelle.

Le modèle « change-point » (ou modèle CP) est un modèle à une seule variable et à cinq paramètres, obtenu en faisant la régression de la consommation en fonction de la température moyenne sur la période de facturation. Le modèle identifie les températures des points d'équilibres (ou les points de changement) auxquels la consommation d'énergie passe d'un comportement dépendant des conditions météorologiques à un comportement indépendant des conditions météorologiques. On obtient un modèle linéaire par partie en deux ou trois parties selon le comportement énergétique du bâtiment comme présenté à la Figure 4.1 et à la Figure 4.2. Le

modèle en deux parties, appelé modèle trois paramètres (3P) correspond aux bâtiments ayant une consommation énergétique saisonnière sensible aux conditions météorologiques, mais seulement pour un seul type de condition de température. Cela revient à considérer des bâtiments qui ont uniquement des besoins énergétiques en chauffage (Figure 4.1) ou uniquement des besoins énergétiques en climatisation. Le modèle en trois parties, appelé modèle cinq paramètres (5P) correspond aux bâtiments ayant une consommation énergétique saisonnière sensible aux conditions météorologiques, et cela pour des conditions froides et chaudes. Cela revient à considérer des bâtiments qui ont des besoins énergétiques en chauffage et des besoins énergétiques en climatisation. Cette modélisation a été introduite dans une publication de l'ASHRAE par Kissock et al. suite à leurs travaux sur la modélisation inverse (Kissock et al., 2003).

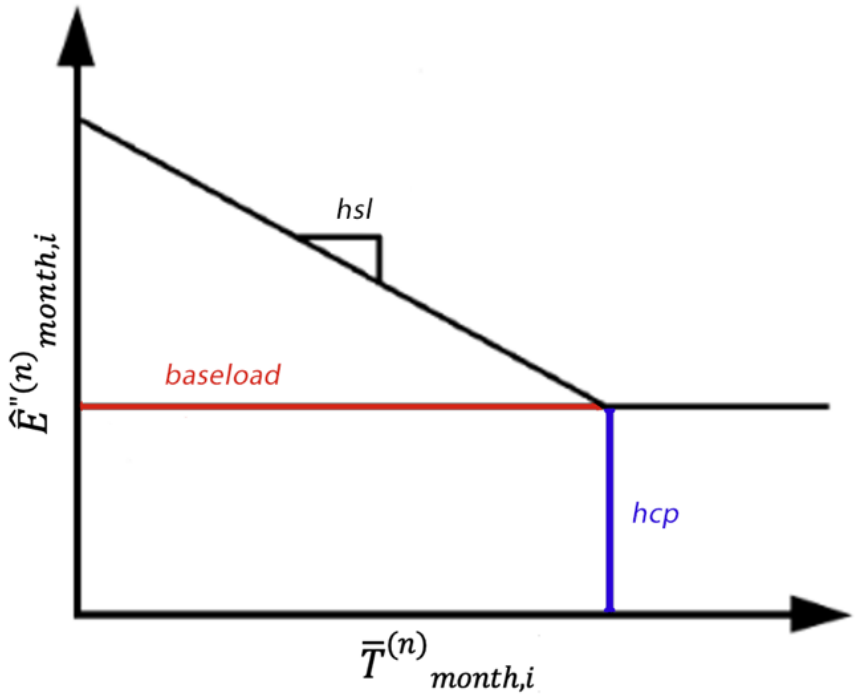

Figure 4.1. Modèle change-point 3P

chauffage

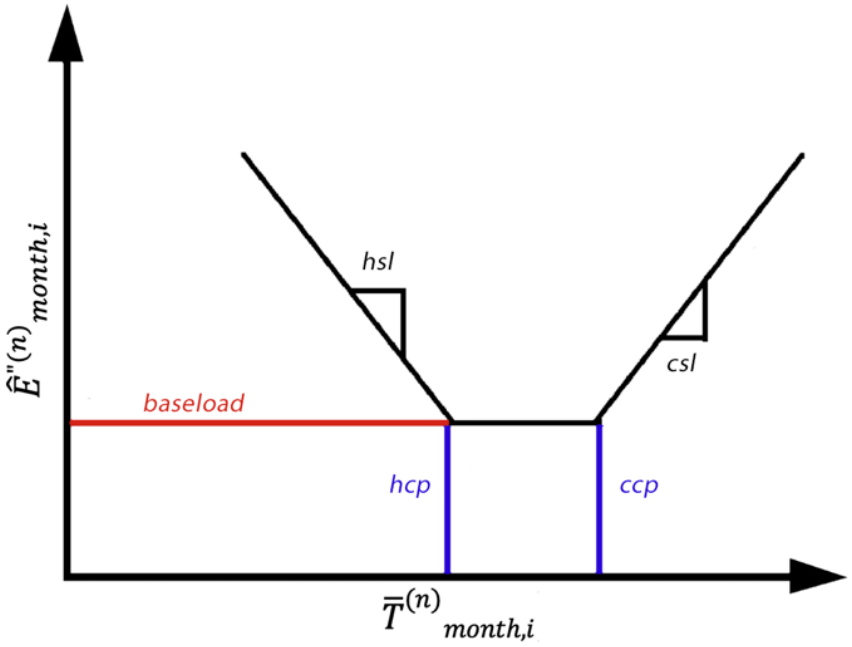

Figure 4.2. Modèle change-point 5P

L'équation du modèle CP à cinq paramètres est :

$$\hat{E}''^{(n)}_{month,i} = baseload + hsl * \left(hcp - \bar{T}^{(n)}_{month,i}\right)^{+} + csl * \left(\bar{T}^{(n)}_{month,i} - ccp\right)^{+} \qquad (4.1)$$

où $baseload$ est la consommation indépendante des conditions météorologiques (consommation électrique pour l'éclairage, matériel électronique, etc.), hsl (ou « heating slope ») est la pente de consommation pour le chauffage en fonction de la température extérieure, csl (ou « cooling slope ») est la pente de consommation pour la climatisation en fonction de la température extérieure, hcp (ou « heating change-point ») est la température d'équilibre à partir de laquelle le système passe d'une indépendance météorologique à une dépendance pour le chauffage et ccp (ou « cooling change-point ») est la température d'équilibre à partir de laquelle le système passe d'une indépendance météorologique à une dépendance pour la climatisation. La température moyenne

mensuelle $\bar{T}_{month,i}^{(n)}$ est la seule variable de l'équation. Les parenthèses ne sont prises en compte que si l'intérieur de ces dernières est positif.

Une régression est basée sur l'équation 4.1 afin d'identifier les paramètres qui minimisent l'erreur quadratique entre les données prétravaux et le modèle. La régression est effectuée en utilisant une adaptation de l'outil BETTER (Johnson Controls, 2021) qui est l'implémentation en logiciel libre de la modélisation inverse de Kissock (Kissock et al., 2003). On obtient ainsi pour chaque bâtiment de l'échantillon les cinq paramètres du modèle change-point ainsi que le coefficient de détermination R^2. Ce coefficient est un indicateur de la qualité de la régression et est défini par :

$$R^2 = \frac{\sum_{n,i} (E''^{*(n)}_{month,i} - \hat{E}''^{*(n)}_{month,i})^2}{\sum_{n,i} (E''^{*(n)}_{month,i} - \bar{E}''^{*}_{month})^2} \quad (4.2)$$

où $E''^{*(n)}_{month,i}$ est la consommation normalisée par la surface mesurée pour le mois i et l'année n prétravaux, $\hat{E}''^{*(n)}_{month,i}$ est la valeur prédite correspondante, et $\bar{E}''^{*}_{month}$ est la moyenne des consommations normalisées prétravaux.

Dans le sous-échantillon de données, plus de 90% des bâtiments ont un coefficient R^2 supérieur à 0,80, ce qui signifie que cette modélisation est adéquate. Un seuil minimum a été fixé à 0,70 et les bâtiments ayant un coefficient R^2 inférieur à ce seuil sont exclus de l'échantillon. Cette sélection basée sur le seuil du R^2 est réalisée lors de la sélection de bâtiments expliquée dans la **section 4.1**, ainsi les 175 écoles primaires conservées ont toutes un R^2 supérieur à 0,70. La Figure 4.3 présente le nuage de points de consommation d'un bâtiment autour du modèle pour un R^2=0,94 et la Figure 4.4 présente le nuage de points de consommation d'un bâtiment autour du modèle pour un R^2=0,45. Les points pour le cas R^2=0,45 sont très dispersés autour de la courbe, le modèle ne reflète pas correctement la consommation énergétique du bâtiment. Ce bâtiment ne sera donc pas conservé dans le jeu de données, contrairement au bâtiment pour lequel R^2=0,94.

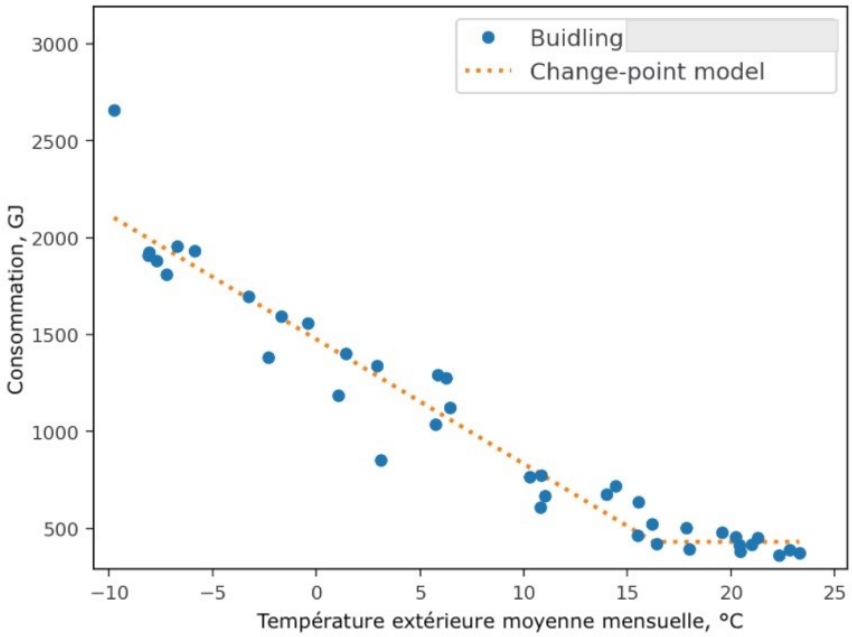

Figure 4.3. Modèle "change-point" d'un bâtiment pour R^2=0,94

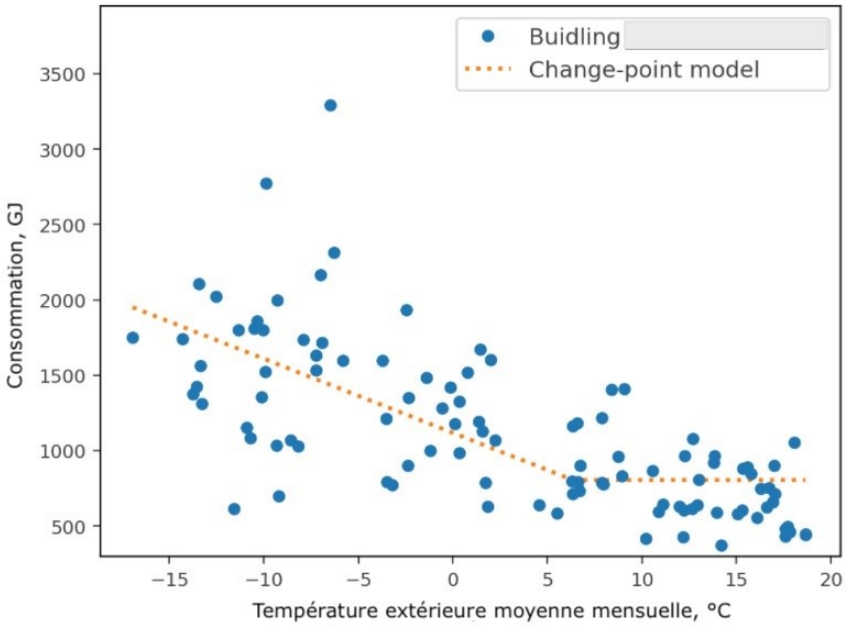

Figure 4.4. Modèle "change-point" d'un bâtiment pour R^2=0,45

Une fois l'ensemble des facteurs CP et du R^2 obtenus pour chaque école primaire de l'échantillon de données, les facteurs sont standardisés en Z-score. Cette démarche consiste à mettre à la même échelle les variables d'entrées pour qu'elles aient toutes le même poids initial (Kreyszig, 1979). Cette standardisation s'effectue de manière indépendante pour chaque facteur en effectuant l'opération :

$$Z_i = \frac{x_i - M}{\sigma} \tag{4.3}$$

Où Z_i est le facteur standardisé pour un bâtiment i, x_i est la valeur avant standardisation du facteur, M est la moyenne du facteur sur l'ensemble des bâtiments et σ est l'écart-type du facteur sur l'ensemble des bâtiments. Pour un facteur donné, cette opération correspond à une distribution normale centrée réduite de l'ensemble de l'échantillon de bâtiment.

Bien que le modèle CP soit un modèle à cinq paramètres, il est possible que certains paramètres soient nuls ou non-définis selon le comportement énergétique du bâtiment. Le $baseload$ est toujours défini car une consommation indépendante des conditions météorologiques peut toujours être définie même si elle est nulle. A contrario, s'il n'y a pas de consommation dépendante des conditions froides, c'est-à-dire pas de chauffage, alors le hsl est nul (pas de pente de chauffage) et le hcp est non défini, et similairement s'il n'y a pas de climatisation pour le csl et ccp. Dans le cas des écoles primaires de la base de données, la climatisation n'existe que rarement : ce sont des bâtiments petits ou moyens, où les occupants principaux sont des enfants de bas âge, où les activités réalisées sont peu énergivores, et ce sont des bâtiments qui ne sont pas occupés pour les deux mois de la période d'été (juillet et août). De plus, les climats rencontrés ne nécessitent pas de climatisation pendant la période active des bâtiments. Ce constat dans les systèmes physiques se traduit dans la modélisation par un csl nul et un ccp non défini pour l'ensemble des écoles primaires de la base de données. Ainsi, seul le $baseload$, hsl et hcp feront réellement partie des paramètres d'entrée pour les écoles primaires.

4.3 Méthodes de prédiction de l'économie d'énergie

Une fois l'ensemble des paramètres définis, la prédiction de l'économie peut être réalisée. Les entrées du problème sont les cinq facteurs CP standardisés et la sortie est $\bar{Y}''_{year}$, soit l'économie annuelle moyenne sur les années de suivi comme schématisé à la Figure 4.5. Deux types d'algorithmes existent dans les problèmes d'apprentissage machine : les algorithmes de classification, qui cherchent à classer les échantillons d'entrée dans des groupes selon des critères continus ou discontinus, et les algorithmes de régression qui cherchent à prédire une valeur continue à partir des différents paramètres d'entrée. Si l'on cherche à développer un outil capable de placer un bâtiment dans un groupe d'économie potentielle, des groupes comme « peu d'économies réalisables » ou « importantes économies réalisables », il s'agit alors d'un problème de classification. Si par contre on cherche à développer un outil capable de prédire l'économie réalisable de manière chiffrée, c'est-à-dire prédire une valeur continue de l'économie, alors c'est un algorithme de régression qui devra être utilisé. Comme ces deux types d'algorithmes sont valides sur le principe de fonctionnement, ils seront présentés en détail et la performance de chaque algorithme sera évaluée afin de déterminer quel algorithme permet d'obtenir les meilleurs résultats et la meilleure performance par rapport à l'objectif de l'outil.

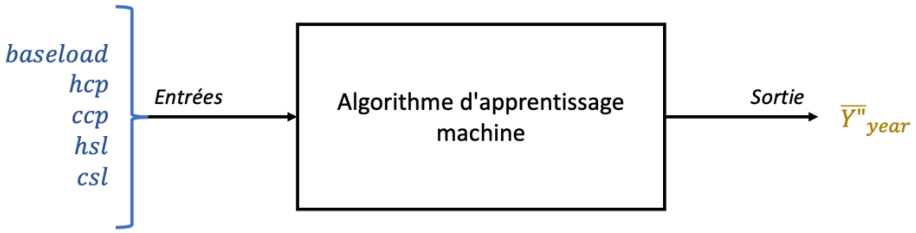

Figure 4.5. Schéma du principe de fonctionnement de l'algorithme

4.3.1 Méthodes de classification

La première méthode étudiée pour déterminer si un bâtiment à plus ou moins de potentiel à se faire rénover est la méthode de classification. L'objectif derrière cette méthode est d'obtenir différents groupes représentant chacun un degré de potentiel d'économie d'énergie plus ou moins important.

Le premier travail est de définir ces différents groupes, ou classes. On commence par choisir le nombre de classes. Un petit nombre de classes simplifie la détermination dans chaque classe, mais la quantité d'information est plus faible. Pour le cas minimum de deux classes, cela signifierait une classe « faible potentiel d'économie » et une classe « fort potentiel d'économie », ce qui sépare l'échantillon de manière brutale et il n'y a pas de contraste entre deux bâtiments d'une même classe même s'ils sont aux extrémités opposées de la classe. À l'inverse, un nombre de classes élevé permet de faire ressortir les contrastes et d'obtenir une information plus précise sur chaque classe, mais cela au détriment d'une complexité plus élevée pour l'algorithme.

Dans le cadre d'une étude de validation de la méthode, l'approche avec un faible nombre de classes a été choisie. Une méthode simple pour un problème à deux classes (binaire) et un à trois classes ont été implémentés. Dans ces deux scénarios, chacune des classes a été créée en séparant uniformément le jeu de données selon $\bar{Y}''_{year}$. Pour le cas à deux classes, une classe A équivalente aux bâtiments à faible économie et une classe B équivalente aux bâtiments à forte économie ont été créées. Les deux classes sont de même taille (même nombre d'échantillons dans les deux classes). Similairement, dans le cas à trois classes, les bâtiments sont répartis uniformément en faible économie, économie moyenne, et forte économie. Les limites d'économies des différentes classes sont présentées au Tableau 4.1 mais ces limites représentent un cas arbitraire de séparation uniforme basée sur les données et non pas à une méthode de séparation existante.

Tableau 4.1. Limites des économies pour les classes

Scénario	**Limites des classes [GJ/(m^2.an)]**
Deux classes	- Économie faible : [0 ; 0,133] - Économie élevée :]0,133 ; 0,565]
Trois classes	- Économie faible : [0 ; 0,099] - Économie moyenne :]0,099 ; 0,140] - Économie élevée :]0,140 ; 0,565]

Les algorithmes de classification, ou classifieurs, sont nombreux. Leur objectif est de classer dans des groupes similaires les échantillons d'entrée qui ont des propriétés similaires. Basé sur l'étude général des algorithmes existants et des outils disponibles au développement (modules et progiciels Python), quatre algorithmes différents ont été retenus: l'arbre de décision, le classifieur à gradient stochastique, le « boosting gradient », et le « benchmark ». Dans tous les cas, l'objectif de la classification est de prédire la classe du bâtiment à partir des cinq facteurs CP.

Arbre de décision

L'arbre de décision, ou DT, est un classifieur tirant son nom de sa représentation. Chaque donnée d'entrée passe à travers des « feuilles » ayant chacune un critère simple de classification sur un seul paramètre. Les différentes classes possibles sont situées aux extrémités des branches (les « feuilles » finales de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape. L'arbre est généré grâce aux données d'entraînement. Il est obtenu de telle sorte que les échantillons d'entraînement soient classés de manière optimale, c'est-à-dire que le plus d'échantillons soient dans la bonne classe. À chaque feuille de l'arbre, l'algorithme évalue la condition d'embranchement afin d'optimiser le classement, et ceci à chaque feuille. C'est donc un algorithme récursif. L'optimisation du classement se fait en calculant l'indice $GINI$ de la condition d'embranchement pour toutes les conditions possibles et en choisissant la condition minimisant l'indice $GINI_{split}$. Cet indice est calculé selon :

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} * GINI(i) \quad (4.4)$$

$$GINI(i) = 1 - \sum_{j}^{m} {i_j}^2 \quad (4.5)$$

où k est le nombre d'embranchements, n_i est le nombre d'échantillons dans la feuille enfant, n est le nombre d'échantillons dans la feuille parent, m est le nombre de classes, i est une condition d'embranchements, et i_j est la fraction des éléments de la feuille enfant avec la classe j. Ce calcul est réalisé sans supervision par l'algorithme. Un hyperparamétrage est cependant nécessaire pour aider l'algorithme à générer l'arbre optimal, notamment pour le choix de la profondeur de l'arbre (nombre d'étages) et du nombre minimum d'échantillons par feuille. Cet hyperparamétrage est fait grâce à une recherche en grille des différents hyperparamètres : des hyperparamètres initiaux sont fixés, et l'on procède ensuite à une variation d'abord avec un pas large puis avec un pas de plus en plus réduit des hyperparamètres jusqu'à converger vers une performance optimale de l'outil. Une fois l'arbre entraîné, il est ensuite simple de faire la prédiction pour de nouveaux échantillons, car il suffit de « suivre le chemin » des conditions fixées par l'arbre. Ce type d'algorithme est apprécié pour le fait que l'arbre puisse être vu et analysé. L'arbre obtenu dans notre cas est présenté à la Figure 4.6, obtenu en utilisant la fonction DecisionTreeClassifier du progiciel Sklearn (Pedregosa et al., 2011). Les hyperparamètres obtenus pour cet algorithme grâce à la méthode expliquée sont ceux proposés par défaut par la fonction, sauf pour la profondeur maximale (« max_depth ») qui est fixée à 4 et le nombre minimum d'échantillons par feuille (« min_samples_split ») fixé à 2.

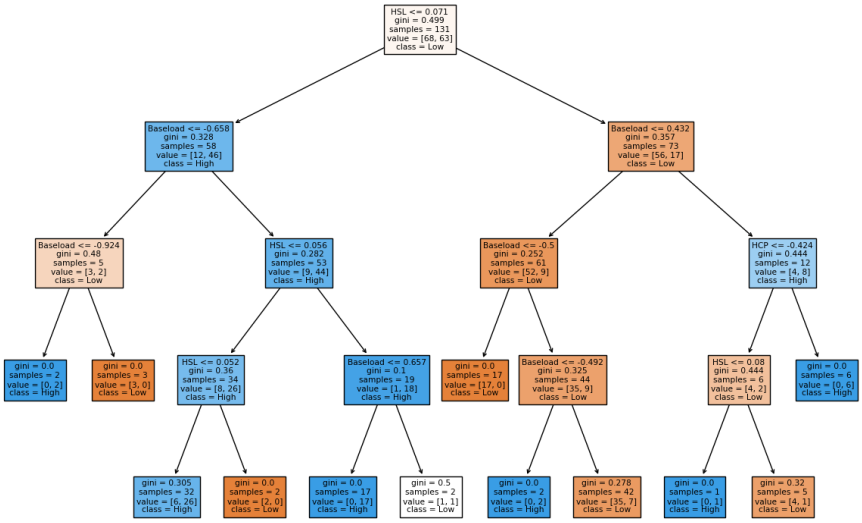

Figure 4.6. Schéma de l'arbre de décision pour une classification à deux classes

Classifieur à gradient stochastique

Le classifieur à gradient stochastique descendant (classifieur SGD), est un classifieur linéaire « Support Vector Machine » (SVM) optimisé par SGD. Il cherche à séparer les données par une séparation linéaire, séparation étant déterminée en optimisant la minimisation de l'erreur pas à pas sur chaque échantillon et non pas sur l'ensemble des données, contrairement à un classifieur SVM sans SGD. Un classifieur linéaire signifie que la séparation des échantillons selon la classe est linéaire. Ainsi, pour un problème 2D, la séparation entre deux classes serait représentée sur un graphique par une droite. D'un point de vue mathématique, un classifieur SGD cherche à résoudre :

$$\min_{w,b,\zeta} \frac{1}{2} * w^T w + c \sum_{i=1}^{n} \zeta_i \tag{4.6}$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i \tag{4.7}$$

$$w_{i+1} = w_i - \alpha \frac{d}{dw}(erreur_i) \tag{4.8}$$

où i représente le $i^{ème}$ échantillon, ζ est la distance avec la marge, c est le terme de régularisation, w est le paramètre de l'algorithme, $w^T w$ est le vecteur normal à la séparation, $\phi(x_i)$ représente vecteur d'entrée, b est le biais, y est le vecteur de sortie, et α est le pas d'apprentissage. La différence entre une méthode SVM classique et une méthode SGD réside dans l'équation (4.8) où le paramètre n'est optimisé qu'avec le i^{ème} élément et non l'ensemble des échantillons. Le classifieur SGD permet de converger bien plus rapidement vers l'erreur minimum, et avec des résultats similaires à une optimisation sur l'ensemble des données dans la plupart des problèmes. L'algorithme a été implémenté en utilisant la fonction « SGD-Classifier » développée par Sklearn (Pedregosa et al., 2011). Les hyperparamètres obtenus pour cet algorithme sont ceux proposés par défaut par la fonction, sauf pour la tolérance de l'erreur finale (« tol ») qui est fixée à 10^{-4} et le terme de régularisation (« alpha ») qui est fixé à 10^{-5}.

« Boosting gradient »

Le « boosting gradient » est un modèle d'ensemble, ce qui signifie qu'il est issu d'une combinaison de N classifieurs plus simples. Le principe d'un tel classifieur est qu'il va générer N classifieurs linéaires simples en réévaluant à chaque pas le poids des entrées mal classées. Au pas suivant, l'emphase est mise sur les échantillons mal classés et le prochain classifieur fonctionne de telle sorte à ce que ces échantillons mal classés au pas précédent deviennent bien classés. Ceci peut générer de nouveaux échantillons mal classés, ce qui crée une boucle itérative. Ainsi, chaque classifieur simple est efficace pour un jeu de l'échantillon, et l'objectif final est de combiner l'ensemble de ces classifieurs pour rassembler la connaissance de chacun et d'obtenir un classifieur global qui sera efficace sur l'ensemble des échantillons. La Figure 4.7 montre un schéma de fonctionnement illustrant le fonctionnement d'un tel algorithme pour un problème à deux classes 2D et où trois classifieurs simples sont générés pour obtenir le classifieur final. On observe la réévaluation du poids de chaque échantillon selon leur classement : s'ils sont bien classés, ils rapetissent (le poids devient plus faible sur ces échantillons), et s'ils sont mal classés, ils s'agrandissent (le poids devient plus grand sur ces échantillons). L'implémentation repose sur la fonction Catboost-Classifier développée du progiciel Catboost (Anna et al., 2017). Les

hyperparamètres sont ceux proposés par défaut par la fonction, sauf pour le nombre de classifieur simple (« iteration ») fixé à 25.

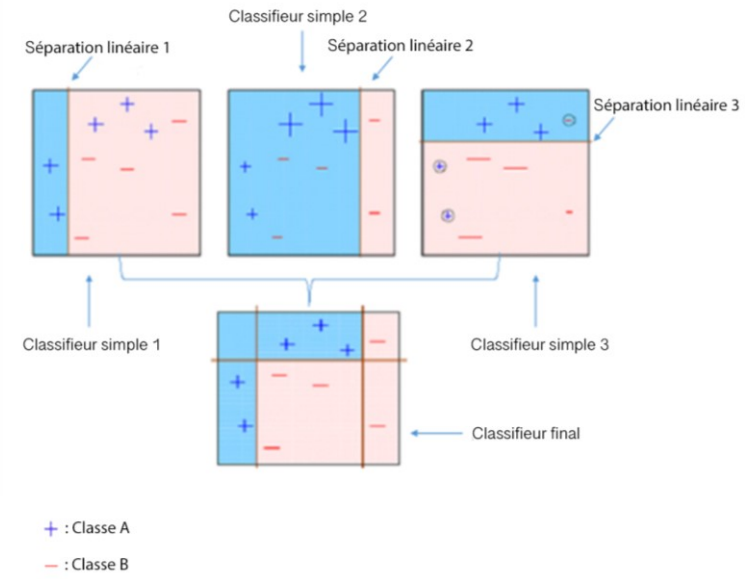

Figure 4.7. Schéma de fonctionnement d'un algorithme boosting de gradient pour un problème à deux classes

« Benchmark »

Le « benchmark » consiste à noter les paramètres d'entrée et d'effectuer la classification à partir de ces notes. Il faut donc attribuer une note aux facteurs CP qui sont les paramètres d'entrée. De manière analogue à la définition des classes, deux types de notation ont été créés, une notation binaire et une notation en trois notes. Ces notes ont été définies en ordonnant de manière croissante chacun des facteurs et en séparant de manière uniforme dans deux notes (0 ou 1) pour la notation binaire, et en trois notes (-1, 0 ou 1) pour la notation en trois notes. Le Tableau 4.2 présente les limites des facteurs pour chaque note pour les deux types de notation, ces limites étant en valeurs standardisées Z-score. Ainsi, le problème est simplifié, les entrées ne sont plus les valeurs continues des facteurs, mais des classes. À partir de ces entrées, un classifieur « boosting gradient » a été utilisé pour obtenir la classification finale.

Tableau 4.2. Limites des paramètres CP en Z-score pour le « benchmark »

Type de notation	**Notes**	**Limites *baseload***	**Limites *hsl***	**Limites *hcp***
Binaire	0	] -∞ ; -0,170]	] -∞ ; 0,078]	] -∞ ; 0,150]
	1	] -0,170 ; +∞]	] 0,078 ; +∞]	] 0,150 ; +∞]
Trois notes	-1	] -∞ ; -0,355]	] -∞ ; 0,061]	] -∞ ; -0,477]
	0	] -0,355 ; 0,135]	] 0,061 ; 0,093]	] -0,477; 0,605]
	1	] 0,135 ; +∞]	] 0,093 ; +∞]	] 0,605 ; +∞]

Validation croisée

L'efficacité des algorithmes d'apprentissage machine est évaluée par validation croisée (Trevor et al., 2009). Il s'agit de séparer son échantillon de données en deux sous-groupes, un groupe d'entraînement et un de test. L'algorithme est alors entraîné avec les données d'entraînement et l'évaluation de l'efficacité de l'algorithme est faite sur les données de test que l'algorithme ne connaît pas. Cette opération est alors réalisée *X* fois sur des séparations de l'échantillon aléatoires pour ne pas induire de biais de séparation, et la métrique d'efficacité est alors la moyenne de l'efficacité sur l'échantillon test sur ces *X* séparations. Pour une validation croisée, il faut donc définir le pourcentage de répartition des échantillons dans chacun des deux groupes, ainsi que le nombre d'itérations de la séparation. Une séparation de 75% de données d'entraînement et 25% de données de test a été utilisée afin d'avoir un nombre important de données d'entraînement tout en conservant une partie non négligeable des données pour le test. Cette séparation est réalisée sur 100 itérations aléatoires. On définit cette appellation comme $cv_{100}(75/25)$, l'indice représentant le nombre d'itérations, et la parenthèse la répartition en pourcentage. Une fois cette validation croisée introduite, la métrique pour évaluer l'efficacité sur le jeu de test a pu être définie.

La mesure retenue pour évaluer l'efficacité du classement est la précision P_{test}. Pour calculer cette dernière, il faut déjà calculer $P_{i_{test}}$ qui se définit par :

$$P_{i_{test}} = \frac{Nombre\ d'échantillons\ correctement\ classés}{Nombre\ total\ d'échantillons}\% \quad (4.9)$$

où $test$ fait référence au sous-groupe de données de test et i correspond à une occurrence de séparation des données aléatoires.

Ainsi, la précision sur l'ensemble du $cv_{100}(75/25)$ est définie par :

$$P_{test} = \frac{1}{100}\sum_{i=1}^{100} P_{i_{test}} \quad (4.10)$$

4.3.2 Méthodes de régression

La seconde méthode étudiée pour déterminer si un bâtiment présente plus ou moins de potentiel à se faire rénover repose sur les algorithmes de régression. L'objectif derrière ces méthodes est de prédire la valeur continue d'économie annuelle moyenne réalisable avec des travaux énergétiques. Le résultat fournit plus d'information qu'avec un algorithme de classification, car une valeur continue est obtenue, cependant cela implique que l'erreur de prédiction n'est plus contrastée par une classe. L'autre différence majeure de cette méthode par rapport à la classification est qu'il n'y a pas de supervision nécessaire pour créer les classes. En effet, la création des différentes classes était arbitraire et dépendait de l'échantillon de données, car elles avaient été créées pour séparer uniformément les données de la base de données. Cela pouvait donc générer un biais par rapport au choix du nombre de classes, un biais par rapport au type de séparation (uniforme ou pas), et un biais par rapport aux données initiales. Les algorithmes de régression permettent donc d'éviter ces biais.

Comme pour les classifieurs, il existe plusieurs algorithmes de régression (ou régresseurs). Basé sur l'étude générale des algorithmes existants et des outils disponibles au développement (modules et progiciels Python), seul l'algorithme de forêt aléatoire a été conservé. D'autres algorithmes auraient théoriquement pu être testés, mais le nombre de bâtiments disponibles et les résultats recherchés encouragent à analyser seulement l'algorithme de forêt aléatoire.

Forêt aléatoire

La forêt aléatoire, ou « random forest » en anglais (RF), est une méthode d'ensemble qui repose sur une méthode d'agrégation (« bagging » en anglais) : *N* arbres de décision sont créés et sont entraînés avec des échantillons de données légèrement différents pour chaque arbre, et la valeur finale prédite est la moyenne de la prédiction de tous les arbres. Le fonctionnement de la méthode est illustré à la Figure 4.8. Cette méthode permet d'éviter les phénomènes de surapprentissage (quand l'algorithme est trop fidèle aux données d'entraînement) et également de limiter l'erreur en diminuant la variance par rapport à un arbre simple. De plus, cet algorithme opère automatiquement une sélection des paramètres d'entrées. Cela signifie que chaque arbre, en plus d'avoir un sous-échantillon d'entraînement différent, a également un jeu de paramètres d'entrée différent et des poids des paramètres d'entrée différents. Cette opération permet sur l'ensemble de donner un poids plus important aux paramètres ayant le plus d'impact sur la prédiction. Pour d'autres algorithmes, cette opération est généralement à réaliser par l'utilisateur en amont. Ce type d'algorithme permet donc de limiter également la supervision et d'utiliser la combinaison des poids des facteurs CP optimaux à la prédiction de l'économie. Pour implémenter cet algorithme, la fonction « RandomForestRegressor » du progiciel Sklearn a été utilisée (Pedregosa et al., 2011).

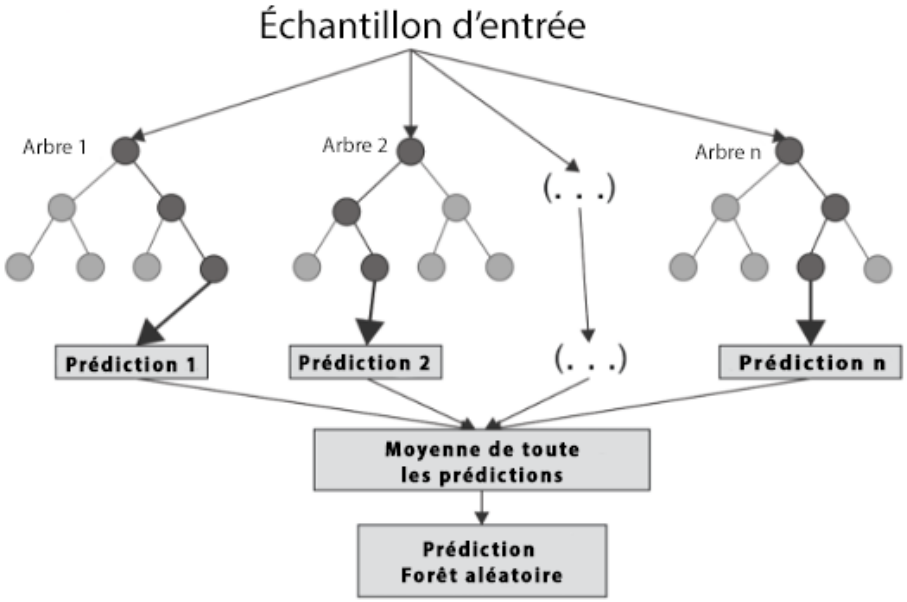

Figure 4.8. Schéma de fonction d'une forêt aléatoire

Les hyperparamètres de cet algorithme sont ceux proposés par défaut par la fonction, sauf pour le nombre d'arbres générés (« n_estimators ») qui est fixée à 1000, et la profondeur maximale de l'arbre (« max_depth ») qui est fixée à 7, le nombre minimum d'échantillon par feuille (« min_samples_leaf ») qui est fixé à 2, et le nombre minimum d'échantillon par séparation d'arborescence (« min_samples_split ») qui est fixé à 5.

Validation croisée

Pour évaluer l'efficacité de l'outil prédictif, une validation croisée $cv_{100}(75/25)$ a été utilisée comme pour le classement. Cependant, pour la régression, la métrique ne peut plus être la précision, car il n'y a pas de classes. Une autre métrique a dû être utilisée. Pour chaque séparation de la validation croisée, l'erreur quadratique moyenne sur l'échantillon test a été calculée, ou $RMSE_i$, et le coefficient de variation de l'erreur quadratique moyenne, ou $CV(RMSE)_i$, tel que :

$$RMSE_i = \sqrt{\frac{\sum_{k=1}^{N}\left(\hat{\bar{Y}}''_{year,k} - \bar{Y}''_{year,k}\right)^2}{\kappa}} \tag{4.11}$$

$$CV(RMSE)_i = \frac{RMSE_i}{\bar{\bar{Y}}''_{year,i}} \tag{4.12}$$

$$\bar{\bar{Y}}''_{year,i} = \frac{\sum_{k=1}^{N} \bar{Y}''_{year,k}}{N} \tag{4.13}$$

où i représente la $i^{ème}$ séparation, N le nombre d'écoles dans l'échantillon i, $\hat{\bar{Y}}''_{year,k}$ est la prédiction de l'économie annuelle $\bar{Y}''_{year,k}$ pour l'école k et $\bar{\bar{Y}}''_{year,i}$ est l'économie moyenne réelle sur l'ensemble des bâtiments de l'échantillon i. L'erreur $RMSE_i$ est dans l'unité de $\bar{Y}''_{year,k}$, soit des GJ/m^2, et le $CV(RMSE)_i$ est en pourcentage.

On obtient pour l'ensemble des cross-validation $CV(RMSE)$ tel que :

$$CV(RMSE) = \frac{\sum_{i=1}^{100} CV(RMSE)_i}{100} \tag{4.14}$$

4.4 Résultats

4.4.1 Méthodes de classification

En apprentissage machine, tous les paramètres d'entrée n'ont pas le même poids et/ou la même représentativité du comportement qu'ils cherchent à prédire. La combinaison optimale correspond donc aux paramètres qui capturent le mieux le comportement du bâtiment par rapport à la prédiction et pour un algorithme précis. Ici, les paramètres d'entrée sont les cinq facteurs CP car ils représentent un ensemble, mais ces cinq facteurs capturent tous un aspect différent du

comportement énergétique, et tous les aspects ne sont pas forcément utiles pour traiter le problème de cette étude.

Comme expliqué à la section 4.2, la climatisation est inexistante dans cet échantillon de bâtiments et les facteurs csl et hcp sont non définis, ils ne peuvent donc pas faire partie de la combinaison optimale. Les possibilités pour les paramètres d'entrée sont donc réduites au *baseload*, hsl et hcp. Cependant, même dans cet échantillon de paramètre d'entrée, tous n'ont pas la même représentativité : la combinaison optimale des paramètres maximisant la précision n'est pas la même pour chaque méthode. L'ensemble des combinaisons de paramètres est testé pour chaque méthode afin de déterminer la combinaison permettant de maximiser P_{test}.

Tableau 4.3. Résultats pour la méthode de classification à 2 classes pour les écoles primaires

Algorithme	**Précision (P_{test})**	**Combinaison de paramètres d'entrée optimale**
« Gradient boosting »	73,3%	baseload – hsl
Classifieur SGD	67,3%	baseload – hsl
Arbre de décision	66,1%	baseload – hsl
Benchmark à 2 notes	71,3%	hsl
Benchmark à 3 notes	70,6%	baseload - hsl
Classe majoritaire	50,0%	-

Tableau 4.4. Résultats pour la méthode de classification à 3 classes pour les écoles primaires

Algorithme	**Précision (P_{test})**	**Combinaison de paramètres d'entrée**
Gradient boosting	58,5%	baseload – hsl - hcp
SGD Classifier	44,2%	baseload – hsl
Arbre de décision	55,0%	baseload – hsl
Benchmark à 2 notes	51,0%	hsl
Benchmark à 3 notes	44,0%	baseload – hsl
Classe majoritaire	33,3%	-

Le Tableau 4.3 présente les résultats pour un classement en deux classes des données de test, et le Tableau 4.4 pour un classement en trois classes des données de test. Les résultats sont comparés à un classifieur classe majoritaire. Ce classifieur fonctionne en prédisant automatiquement la classe qui est majoritaire dans ses données d'entraînement. Si un tel classifieur est appliqué avec une distribution uniforme dans chaque classe pour les données d'entraînement et les données de test, ce qui est le cas présentement, on obtient $P_{test} = 50\%$ pour une classification binaire et $P_{test} = 33{,}3\%$ pour une classification selon trois classes. Des valeurs maximales de précision de 73,3% pour le classement en deux classes et 58,5% pour le classement en trois classes en utilisant la méthode « boosting gradient » sont observables. Dans tous les cas, la précision des algorithmes de classification est supérieure à celle obtenue par la classe majoritaire. L'utilisation des facteurs CP comme entrée du modèle est donc une méthode judicieuse et correcte. De plus, on note que hsl est le facteur faisant toujours parti de la combinaison optimale des paramètres d'entrée, suivi à 80% du temps par le $baseload$. Ces deux facteurs sont donc ceux qui capturent le mieux un comportement associé à l'économie potentielle d'énergie pour les écoles primaires parmi les cinq facteurs CP.

L'algorithme « boosting gradient » permet d'obtenir pour chacune des classifications la précision maximale pour la classification binaire et pour la classification en trois classes. C'est l'algorithme qui fonctionne le mieux avec les données. Le gain de performance sur la précision par rapport aux cas de référence est de 23,3% pour la classification binaire et de 25,2% pour la classification en trois classes. C'est un gain non négligeable qui justifie le choix d'un tel algorithme dans ce contexte.

4.4.2 Méthodes de régression

Dans le cadre de la régression, une seule méthode prédictive a été implémentée, à savoir l'algorithme forêt aléatoire. Pour avoir une référence à laquelle comparer et valider les résultats, un algorithme de référence non intelligent a également été implémenté. Cet algorithme de référence doit jouer pour la régression le même rôle que l'algorithme « classe majoritaire » pour la classification. Un algorithme simple qui prédit toujours l'économie moyenne du jeu de données d'entraînement a été utilisé comme référence. Ainsi, pour la validation croisée i, la prédiction est telle que :

$$\hat{\bar{Y}}''_{year,k} = \bar{\bar{Y}}''_{year,i} \tag{4.15}$$

où $\hat{\bar{Y}}''_{year,k}$ est la prédiction d'économie pour l'école k et ce quel que soit le k, et $\bar{\bar{Y}}''_{year,i}$ est la valeur moyenne d'économie mesurée dans le jeu d'entraînement pour cette validation croisée. Cet algorithme est non intelligent, car il n'apprend pas réellement des données. C'est un algorithme simple et il est justifié de l'utiliser comme cas de référence, car les économies des futurs bâtiments ont des chances d'être aux alentours de l'économie moyenne des anciens bâtiments. Cet algorithme sera nommé dans les résultats comme « référence ».

La Tableau 4.5 présente les résultats obtenus pour l'application de la méthode avec un algorithme forêt aléatoire et l'algorithme de référence. On notera qu'ici, la combinaison des facteurs CP

optimale n'est pas signalée, car l'algorithme de forêt aléatoire opère de manière interne de la sélection des paramètres et définit de manière autonome et implicite le poids attribué à chaque paramètre d'entrée. Cela veut dire que même si le jeu de paramètres d'entrées est exhaustif, l'algorithme arrive seul à déterminer quels sont les paramètres principaux par rapport au phénomène modélisé, et opère la prédiction sans supervision des paramètres d'entrées de l'opérateur.

Tableau 4.5. Résultats pour la méthode de régression classes pour les écoles primaires

Algorithme	**CV(RMSE)**
Référence	68%
Forêt aléatoire	46 %

On remarque que l'algorithme de forêt aléatoire permet d'obtenir de meilleurs résultats qu'avec l'algorithme de référence. L'erreur *CV(RMSE)* passe de 68% dans le cas de référence à 46% d'erreur avec l'algorithme forêt aléatoire, soit une diminution du *CV(RMSE)* de plus de 20% entre un algorithme non intelligent et l'algorithme entraîné. Cela signifie que les choix des paramètres d'entrée sont judicieux également dans le cas de la régression. Les facteurs CP comme entrée de l'algorithme permettent de capturer le comportement énergétique et de générer une prédiction plus précise qu'un algorithme non intelligent qui ne prend pas en compte les facteurs.

Ensuite, on peut noter que l'algorithme RF permet en l'état d'obtenir une erreur de prédiction moyenne proche des 45%, cela signifie donc que l'erreur de prédiction est tout de même importante. Cependant cette erreur est à mettre en contraste avec l'objectif de l'outil. L'objectif n'est pas de faire une prédiction avec une erreur nulle, c'est même impossible en apprentissage machine, mais de développer un outil capable d'aider rapidement l'ingénieur et de le guider vers un potentiel d'économie en début de projet. En prenant cet aspect en considération, ces résultats sont bien plus encourageants.

4.5. Conclusion

À partir des résultats sur la méthode par algorithmes de classification, cette méthode est à remettre en question. En effet, si les algorithmes de classification avaient permis d'obtenir une prédiction dont la précision dépasse un seuil arbitraire de 80%, l'objectif aurait été considéré comme atteint. Mais dans le cas étudié ici qui est un cas simplifié, au moins 25% des données sont mal classifiées, alors que les différentes classes ont des intervalles larges par rapport à l'économie minimale et maximale au sein d'une même classe. La précision de la prédiction est donc faible et peu d'informations précises quant à l'économie possible sont obtenues même quand la prédiction est correcte. De plus, en augmentant le nombre de classes, ce qui permettrait d'obtenir plus d'information sur l'économie potentielle, l'erreur augmente également. Enfin, les classes subissent un biais de par leur population. La définition de la limite « bonne/mauvaise économie » n'est pas empirique, mais dépend du jeu de bâtiments que l'on utilise pour définir cette limite. Cette limite est de plus variable dans le temps dépendamment du contexte économique/énergétique et de programmes de subventions en vigueur. C'est un choix qui est fait individuellement pour chaque projet. Il ne fait pas de sens de réentraîner l'outil à chaque fois que la limite change.

Passer par une méthode de classification a permis de valider que les facteurs CP pouvaient fournir une information sur les économies possibles, mais c'est aussi une méthode qui atteint rapidement ses limites dans notre cas : les résultats ne sont pas mauvais en soi, mais ils ne fournissent pas assez d'informations pour être exploités et réellement utiles dans le cadre du projet.

La méthode par régression est plus encourageante que la méthode par classification, car elle permet d'obtenir plus d'informations dans le résultat produit et avec moins de supervision et une meilleure précision. En plus de cette simplicité à implémenter, une méthode par régression permet d'éviter un biais en séparant les données en différentes classes de façon arbitraire.

En conclusion, la méthode privilégiée dans le cadre d'un cas simplifié est celle par algorithme de régression forêt aléatoire. C'est donc cette méthode qui a été conservée pour développer la méthode dans un cas plus général non restreint aux écoles primaires.

CHAPITRE 5 MÉTHODE DE RÉGRESSION VIA FORÊT ALÉATOIRE POUR LA PRÉDICTION DE L'ÉCONOMIE D'ÉNERGIE

La démarche dans un cas simple a permis de valider la méthode par régression via forêt aléatoire pour la prédiction de l'économie d'énergie. Le développement de la méthode dans un cas plus général permet de s'intéresser à l'optimisation de la méthode à différentes étapes du processus, ainsi que de redéfinir les échantillons de données sur lesquels la méthode s'applique.

5.1 Correction de la définition d'économie annuelle moyenne

L'économie annuelle moyenne est la valeur qui doit être prédite. Elle était définie dans la section 3.3 comme étant pour un bâtiment la moyenne de l'économie d'énergie annuelle sur les années de suivi (équation (3.5)).

Cependant, ce calcul pose un problème, car cette somme prend en compte l'ensemble des années de suivi et notamment celles pour lesquelles les objectifs d'économie fixés par Ecosystem ne sont pas atteints. Certaines années peuvent donc avoir une économie faible, mais il est surtout possible d'avoir des années avec des économies négatives, ce qui impacte fortement le calcul de l'économie moyenne annuelle. Une économie négative correspond à une année où la consommation énergétique est supérieure à la consommation de référence calculée grâce au modèle degrés-jour de l'année de référence. Ce phénomène arrive majoritairement lors des premières années de suivi. En effet, pendant les premières années de suivi, le comportement du bâtiment rénové n'est pas encore parfaitement connu et des ajustements sont parfois nécessaires, autant dans les systèmes que dans les contrôles, afin d'assurer les économies vendues. Certains bâtiments ont donc une économie négative lors de leurs premières années de suivi, mais une fois les ajustements faits, l'économie devient conforme au contrat les années suivantes.

Pour améliorer le calcul de l'économie moyenne, seules les années pour lesquelles l'économie est positive sont conservées, et ce d'un point de vue énergétique global c'est-à-dire tous types d'énergie confondus. Ainsi, l'équation (3.5) devient :

$$\bar{Y}''_{year} = \frac{1}{N^+} * \sum_{n^+=1}^{N^+} Y''^{(n^+)}_{year} \tag{5.2}$$

où N^+ est le nombre d'années où $Y''^{(n)}_{year}$ est positif.

Cette définition n'est pas parfaite, car toutes les années en régime transitoire ne sont pas ignorées, mais elle permet d'ignorer les années ayant le plus d'impact négatif sur le calcul de l'économie. À l'échelle globale des 479 bâtiments éducatifs, il y a 2196 années de suivi tous bâtiments confondus, et ce calcul permet d'exclure 302 années de suivi pour lesquelles l'économie est négative, soit 13,75% des années de suivi. De plus, si l'on s'intéresse à l'impact de ce changement sur les valeurs du Tableau 3.4, cela revient à une économie moyenne sur l'ensemble des 479 bâtiments de 0,16 GJ/m^2.an au lieu de 0,14 GJ/m^2.an, soit une économie moyenne de 21,5% au lieu de 18%.

Cette nouvelle définition de l'économie est utilisée dans la suite du développement de la méthode, car elle a un sens physique plus juste que l'ancienne définition par rapport à l'objectif du projet.

5.2 Correction du modèle « change-point »

Le modèle CP présenté en section 4.2 génère une prédiction mensuelle de consommation, mais la durée des mois n'est pas fixe et varie entre 28 à 31 jours, il y a donc un biais de quelques jours dans la prédiction de la consommation du modèle. Il faut donc corriger la définition du modèle CP pour limiter ce biais.

La méthode utilisée pour pallier ce problème consiste à ne plus utiliser un modèle CP en consommation mensuelle, mais en consommation journalière moyenne pour le mois, c'est-à-dire de normaliser la consommation pour chaque mois par le nombre de jours au mois. L'équation (4.1) devient alors :

$$\hat{\bar{E}}''^{(n)}_{daily,i} = baseload + hsl * \left(hcp - \bar{T}^{(n)}_{month,i}\right)^{+} + csl * \left(\bar{T}^{(n)}_{month,i} - ccp\right)^{+} \tag{5.3}$$

$$\hat{E}''^{(n)}_{month,i} = D^{(n)}_{i} * \hat{\bar{E}}''^{(n)}_{daily,i} \tag{5.4}$$

où $\hat{\bar{E}}''^{(n)}_{daily,i}$ est la consommation moyenne journalière pour le mois i de l'année n, $D^{(n)}_{i}$ est le nombre de jours pour le mois i de l'année n, et $\hat{E}''^{(n)}_{month,i}$ est donc le produit de ces deux derniers. Ce nouveau modèle CP est obtenu en faisant la régression entre $\bar{E}''^{*(n)}_{daily,i}$ et $\bar{T}^{(n)}_{month,i}$ tel que :

$$\bar{E}''^{*(n)}_{daily,i} = \frac{E''^{*(n)}_{month,i}}{D^{(n)}_{i}} \tag{5.5}$$

Cette nouvelle définition du modèle CP permet d'éviter de fournir une consommation mensuelle sans prendre en compte la variation du nombre de jours par mois. Il est cependant à noter que cette nouvelle méthode pour le calcul du modèle CP à un impact direct sur la valeur des facteurs. En effet, les facteurs subiront un effet non négligeable puisque ce n'est plus une consommation mensuelle qui est la résultante du modèle, mais une consommation journalière moyenne pour le mois. Ainsi pour l'ensemble des bâtiments, le $baseload$, hsl et csl se verront diviser d'un facteur d'environ 30, facteur correspondant au nombre de jours moyen dans un mois.

L'ASHRAE déconseille d'utiliser cette modélisation pour une résolution inférieure au mois (American Society of Heating Refrigerating Air-Conditioning Engineers, 2017) mais dans notre cas, l'équation (5.3) ne correspond pas à une résolution journalière puisque $\hat{\bar{E}}''^{(n)}_{daily,i}$ est une consommation moyenne journalière pour le mois. Le cadre défini par l'ASHRAE est donc respecté.

5.3 Résultats du développement de la méthode

Le développement de la méthode a permis d'améliorer la définition de l'économie annuelle moyenne pour se rapprocher d'un calcul réaliste, et a permis de développer une méthode plus précise pour calculer le modèle CP d'un bâtiment. Les résultats présentés cherchent à comparer l'implémentation de la méthode et de ces améliorations aux différents jeux de données afin de valider pour quels types ou groupes de bâtiments la méthode est valable. La méthode est appliquée aux types et groupes de bâtiments présentés dans la section 3.4. Cependant, seuls les bâtiments ayant un R^2 supérieur à 0,70 sont conservés, et la méthode est testée sur seulement les types et groupes ayant une taille d'échantillon assez grande. Ainsi le Tableau 5.1 et Tableau 5.2 présentent la taille des échantillons selon le seuil du R^2 pour respectivement les types et les regroupements, et indiquent si la méthode est testée.

Tableau 5.1. Nombre de bâtiments conservés dans la méthode par type

Type de bâtiment	Primaire	Secondaire	Cegep	Université	Centre de formation
Nombre d'échantillons	272	147	6	3	50
Nombre d'échantillons R^2>0,70	161	90	4	3	39
Méthode testée	Oui	Oui	Non	Non	Oui

Tableau 5.2. Nombre de bâtiments conservés dans la méthode par regroupement

Groupe de bâtiment	K12	K12 clés	Éducation supérieure	Bâtiments éducatifs
Nombre d'échantillons	419	157	50	479
Nombre d'échantillons R^2>0,70	251	157	46	297
Méthode testée	Oui	Oui	Oui	Oui

On s'intéresse toujours au *CV(RMSE)* sur une validation croisée $cv_{100}(75/25)$, en ajoutant ici l'erreur absolue moyenne sur la validation croisée *(MAE)* définie par :

$$MAE_i = \frac{\sum_{n=1}^{\kappa} \left| \hat{\bar{Y}}''_{year} - \bar{Y}''_{year} \right|}{\kappa} \tag{5.6}$$

$$MAE = \frac{\sum_i MAE_i}{100} \tag{5.7}$$

où MAE_i est l'erreur absolue moyenne sur le i[ème] cross validation et κ est le nombre de bâtiments dans l'échantillon test. L'unité de la *MAE* est l'unité pour l'économie, soit des GJ/m^2.

Tableau 5.3. Résultats pour chaque type

Type de bâtiments sélectionné	**Taille de l'échantillon**	**CV(RMSE)**	**MAE [MJ/m^2]**	$\bar{\bar{Y}}''_{year}$ **[MJ/m^2]**
Primaire	161	48%	59	157
Secondaire	90	40%	72	224
Centre	39	63%	95	195

Tableau 5.4. Résultats pour chaque groupe

Groupe de type de bâtiment	**Taille de l'échantillon**	**CV(RMSE)**	**MAE [MJ/m^2]**	$\overline{\overline{Y}}''_{year}$ **[MJ/m^2]**
K12	251	45%	64	181
K12 clés	157	42%	61	179
Éducation supérieure	46	63%	101	210
Bâtiments éducatifs	297	57%	85	185

Les résultats sont présentés au Tableau 5.3 et au Tableau 5.4 où l'on retrouve respectivement les résultats pour une séparation en chaque type de bâtiment éducatif, et les résultats pour les regroupements de types. Les résultats du Tableau 5.3 montrent que la méthode performe mieux pour les écoles primaires et secondaires que pour les centres de formation. Si l'on s'intéresse d'abord aux résultats pour les écoles primaires, on constate que les modifications réalisées lors du développement de la méthode n'ont pas permis d'améliorer le *CV(RMSE)* par rapport aux résultats présentés dans la section 4.4.2, il y a au contraire une légère perte de performance de l'algorithme (environ 2%). Pour les écoles secondaires, le *CV(RMSE)* est inférieur à celui des écoles primaires, mais la *MAE* est cependant plus élevée que celle des écoles primaires. Ce comportement s'explique par le fait que les écoles secondaires sont des bâtiments généralement plus grands en surface et en nombre d'élèves que les écoles primaires. La consommation aura donc tendance à être plus élevée et l'erreur absolue de prédiction va donc suivre cette tendance. Cette consommation plus élevée pour les écoles secondaires peut se lire dans la dernière colonne du tableau qui correspond à l'économie moyenne annuelle pour l'échantillon de bâtiments. La méthode développée ne peut cependant pas être validée pour les centres de formation. On constate des erreurs *CV(RMSE)* et *MAE* bien plus élevées que pour les autres types de bâtiments. Si l'on prend comme référence de performance l'erreur de l'algorithme non intelligent de la section 4.4.2, l'erreur CV(RMSE*)* pour

les centres est seulement 4% plus basse que dans le cas de référence, contre 19% pour les écoles primaires et 27% pour les écoles secondaires. Ce comportement s'explique par la diversité des activités entre les différents centres professionnels, une diversité bien plus importante qu'entre deux écoles primaires ou secondaires. En effet, les centres de formation professionnels visent chacun des domaines de formation très précis et n'ont donc pas les mêmes comportements. De plus, la taille de l'échantillon pour les centres est également plus petite que celle des écoles secondaires et primaires, cela augmente également intrinsèquement l'erreur.

Les résultats du Tableau 5.4 permettent de justifier le choix de regrouper les bâtiments du type K12. En effet, les erreurs pour l'ensemble des K12 sont dans la gamme des erreurs pour les écoles primaires et secondaires séparément. L'erreur *CV(RMSE)* pour l'ensemble des K12 est à moins de 3% de l'erreur *CV(RMSE)* pour les écoles primaires, et l'erreur *MAE* des K12 est presque égale à la moyenne de la *MAE* des écoles primaires et secondaires. La méthode marche encore mieux dans le cas des K12 clés. Ce sous-groupe de K12 permet d'obtenir une erreur *CV(RMSE)* à 2% de différence de l'erreur *CV(RMSE)* minimale obtenue tous sous-groupes confondus, et une erreur *MAE* à 1,6 MJ/m^2 de différence de l'erreur *MAE* minimale obtenue tous sous-groupes confondus. L'algorithme pour les K12 clés est également plus robuste, car il est entrainé avec plus d'échantillons que pour individuellement les écoles primaires et les écoles secondaires. C'est le groupe pour lequel la méthode fonctionne le mieux. Pour les bâtiments de l'éducation supérieure ou l'ensemble des bâtiments éducatifs, la méthode ne peut pas être validée. L'erreur *CV(RMSE)* pour le groupe de bâtiments d'éducation supérieure est très élevée, et l'échantillon est de petite taille ce qui ne favorise pas l'utilisation des données. L'algorithme est donc peu efficace et peu robuste pour ce regroupement. Pour le groupe de l'ensemble des bâtiments éducatifs, l'algorithme est altéré par le mauvais fonctionnement pour les bâtiments d'éducation supérieure et le bon comportement pour les K12, ce qui donne un *CV(RMSE)* de 57%. La Figure 5.1 présente les résultats de la méthode de prédiction pour un échantillon aléatoire de K12 issues d'une validation croisée aléatoire. Cette figure permet de constater que la méthode développée permet de prédire fidèlement les économies mesurées.

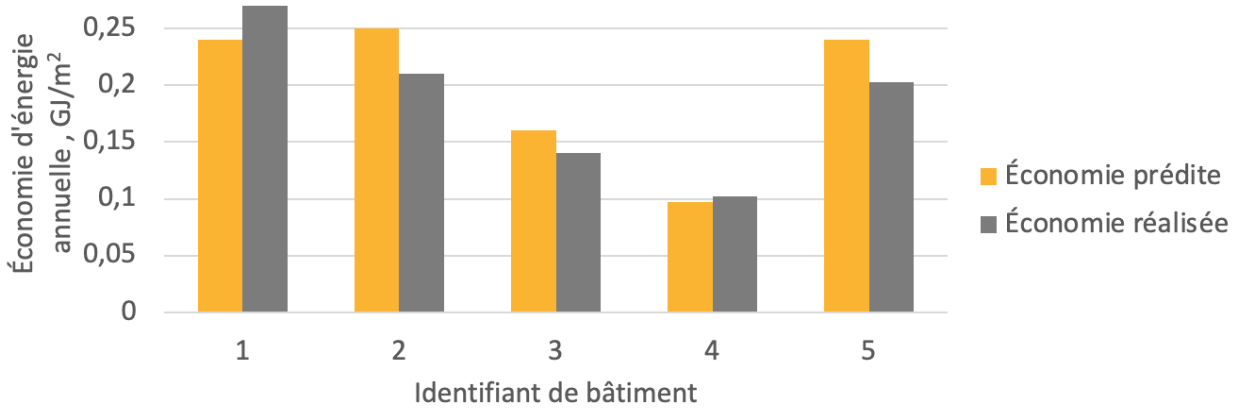

Figure 5.1. Prédiction des économies pour un échantillon de 5 bâtiments aléatoires

La méthode de prédiction de l'économie d'énergie grâce à un algorithme de régression basé sur les paramètres du modèle change-point est donc validée dans le cadre des écoles primaires et secondaires. La méthode est également validée dans le cadre d'un regroupement de ces deux types sous l'étiquette bâtiment K12. Plus particulièrement, pour le groupe des K12 clés, on obtient une erreur *CV(RMSE)* moyenne d'environ 42% et une erreur *MAE* moyenne de 61 MJ/m^2, qui sont des erreurs acceptables dans le cadre du projet. Pour les centres de formation professionnelle, une conclusion tranchée sur la méthode ne peut pas être faite : les erreurs sont élevées, mais il y a peu d'échantillons de centres dans la base comparativement aux écoles primaires et secondaires. Avec plus d'échantillons, il serait possible de conclure quant à la méthode pour ce type de bâtiment. Pour les autres regroupements, la méthode est rejetée, car les erreurs de prédiction sont trop élevées. Cependant, ce sont des types de bâtiment qui sont minoritaires parmi le nombre de projets réalisés, ils ne sont donc pas prioritaires. La méthode développée permet donc de traiter la majorité des bâtiments éducatifs sur lesquels interviendra Ecosystem dans le futur si les types de sollicitations restent constants.

CHAPITRE 6 RECOMMANDATION DES MESURES D'ÉCONOMIE D'ÉNERGIE

Ce chapitre présente la méthode et les résultats de la recommandation des mesures d'économies d'énergie associées à la prédiction du potentiel d'économies d'énergies présentée au chapitre précédent. Cette méthode utilise les mêmes données que celle utilisées dans le chapitre 5, puisque dans le flux de travail de l'outil final, la prédiction de l'économie est réalisée en premier, suivie par la recommandation de mesures.

6.1 Données disponibles

Les économies d'énergies d'un bâtiment sont toujours associées à une ou plusieurs mesures d'économie d'énergie. Ces mesures sont les interventions qui sont réalisées par le maître d'œuvre du projet, en l'occurrence ici l'entreprise Ecosystem. Différents types de mesures d'économie d'énergie existent : amélioration de l'enveloppe du bâtiment, remplacement/installation de systèmes énergétiques, optimisation des opérations et des contrôles, gestion des tarifs (American Society of Heating Refrigerating Air-Conditioning Engineers). Les mesures à réaliser pour les travaux de rénovation énergétique sont définies lors de l'étude détaillée, suite à une analyse complète du bâtiment. Pour un bâtiment donné, seules les mesures pertinentes qui permettent une diminution de la consommation énergétique ou une amélioration du confort des occupants sont considérées. Certaines interventions sur le bâtiment peuvent aussi être associées à du maintien d'actif.

Dans la base de données, on retrouve l'attribut « measures » comme présenté dans le Tableau 3.1 qui présente l'ensemble des attributs de la base de données. La base de données contient alors l'ensemble des mesures qui ont été implémentées pour chacun des 479 bâtiments. À l'amorce du projet de recherche, il n'existait pas de terminologie définie pour nommer les mesures d'économie d'énergie au moment de la saisie des données : chaque responsable de projet a rempli cet attribut selon son propre arbitre et donc la terminologie varie d'un projet à l'autre. Deux problèmes se

posent : un problème de compréhension humaine et un problème d'exploitation automatique des données. Le problème de compréhension correspond à des noms de mesure qui sont trop peu détaillés (par exemple « changement des systèmes énergétiques »). Ce problème empêche la compréhension et l'utilisation des données par l'humain, il n'est pas simple de réinterpréter ces mesures pour leur donner un nom standard. Le problème d'exploitation automatique correspond aux données qui ne peuvent pas être analysées grâce à un outil informatique car elles ne sont pas standardisées. Une même mesure d'économie d'énergie peut être entrée sous plusieurs noms différents et il est donc difficile d'implémenter un traitement automatique des données. Face à ces problèmes, il est impossible de traiter les mesures d'économie d'énergie pour ces projets de la base de données, et donc impossible d'utiliser l'ensemble des données pour la prédiction des mesures d'économie d'énergie grâce à un modèle basé sur les données.

Pour pallier ce problème, un traitement manuel pour un sous échantillon de la base de données a été réalisé. Ce traitement consiste à définir une liste de noms de mesures standards couvrant l'ensemble des mesures pouvant être implémentées par Ecosystem, et de redéfinir un nouvel attribut « mesures standards » où la liste des mesures implémentées dans leur nom standard est entrée. Une liste de 85 mesures standards a ainsi été obtenue suite à l'analyse manuelle de la base de données, permettant de couvrir l'ensemble des mesures possibles. Le sous échantillon choisit pour réaliser ce traitement est le groupe de bâtiment K12 clés définis section 3.3.2. L'attribut « mesures standards » a été ajouté à ce groupe de 157 bâtiments. Ce groupe forme donc le jeu de données disponibles pour développer un modèle basé sur les données pour recommander les mesures d'économies d'énergie.

6.2 Facteurs « change-point » comme indicateur des performances énergétiques du bâtiment

L'objectif est de développer un algorithme reposant sur de l'apprentissage machine capable de recommander les mesures d'économie d'énergie à implémenter dans un nouveau bâtiment. La sortie du modèle est donc une liste de mesures avec une notion de probabilité pour chaque mesure, mais il faut également définir l'entrée afin de pouvoir entraîner puis utiliser l'algorithme. Il faut

pouvoir définir des entrées qui auront un impact sur les mesures d'économie d'énergie, et qu'il est possible d'avoir pour l'ensemble des bâtiments.

Iyengar et al. (Iyengar et al., 2020) présentent une analyse des paramètres du modèle CP afin d'identifier les défauts d'un bâtiment. Ils montrent que chacun des cinq paramètres est interprétable physiquement et permet de capturer certaines capacités thermiques et énergétiques d'un bâtiment. La pente de chauffage, hsl, et la température d'équilibre de chauffage, hcp, sont les deux paramètres qui permettent d'interpréter les inefficacités en chauffage d'un bâtiment. Les bâtiments avec un hsl élevé ont un taux de déperdition de chaleur élevé ce qui affecte directement les systèmes de chauffage qui consomment plus pour compenser ces déperditions. Un taux de déperdition de chaleur élevé peut être attribué à une mauvaise isolation de l'enveloppe du bâtiment, à des fuites d'air importantes ou à des systèmes de chauffage inefficaces. Séparément, hcp indique également des inefficacités dans le système de chauffage d'un bâtiment. Un hcp élevé peut signifier deux choses : une température de consigne du thermostat en chauffage élevée ou une mauvaise isolation du bâtiment. Si la température de consigne est élevée en hiver, les systèmes de chauffage fonctionnent plus fréquemment pour maintenir la température intérieure au point de consigne. En revanche, si l'isolation du bâtiment est mauvaise, plus de chaleur est perdue à travers l'enveloppe du bâtiment. Encore une fois, les systèmes de chauffage s'allument fréquemment et « plus tôt » pour maintenir la température de consigne en chauffage. On peut interpréter la pente en climatisation csl et la température d'équilibre en climatisation ccp de manière analogue. Un csl élevé peut correspondre à une mauvaise isolation de l'enveloppe du bâtiment, à des fuites d'air importantes ou à des systèmes de climatisation inefficaces. Un ccp faible peut signifier deux choses : une température de consigne du thermostat en climatisation basse ou une mauvaise isolation du bâtiment. Les bâtiments avec un $baseload$ élevé indiquent une utilisation élevée des appareils électriques, ou des appareils électriques inefficaces. Ces bâtiments peuvent bénéficier du remplacement des anciens appareils (éclairage, escalier mécanique, portes automatiques, etc.) afin de réduire la consommation énergétique, ou d'une optimisation de l'utilisation (détecteur de mouvement, gestion des allumages selon un horaire, etc.).

Les paramètres du modèle CP permettent donc de capturer les performances énergétiques du bâtiment, ce qui implique directement que les paramètres du modèle CP sont capables de capturer les voies d'améliorations énergétiques d'un bâtiment afin de contrer les défauts. Pour la recommandation des mesures d'économie d'énergie comme pour la prédiction de l'économie, il

est donc pertinent de prendre comme entrée les cinq paramètres du modèle CP. Ces derniers ont un lien direct avec les mesures qui peuvent être implémentées. Ces données sont accessibles facilement puisque les paramètres CP de l'ensemble des bâtiments de la base de données ont déjà été calculés pour les travaux des chapitres 4 et 5. Les paramètres sont standardisés sous leur Z-score comme défini dans la section 4.2 afin que tous les paramètres aient un poids unitaire commun au début du procédé.

6.3 Sélection de l'algorithme

L'algorithme doit prendre en entrée les cinq paramètres du modèle CP et fournir en sortie une liste de mesures d'économie d'énergie à implémenter comme schématisé sur la Figure 6.1. La sortie n'est pas une valeur continue : utiliser un algorithme de régression n'est donc pas pertinent. La sortie n'est pas non plus une valeur discontinue unique ou une classe unique, car la liste des mesures en sortie n'a pas forcément une longueur fixe et le nombre de combinaisons des mesures d'économie d'énergie possibles dans la liste est trop grand pour pouvoir faire une classe pour chaque combinaison. Il n'est donc pas pertinent d'utiliser directement un algorithme de classification. L'algorithme doit permettre de fournir les mesures d'économie d'énergie grâce aux données d'anciens projets, c'est-à-dire être capable de trouver le ou les anciens projets les plus similaires au nouveau bâtiment afin de pouvoir se baser sur les connaissances de ces anciens projets pour faire la prédiction. De ces constats, un algorithme en deux étapes a été implémenté pour répondre aux différents problèmes : la première étape consiste à trouver les bâtiments similaires au nouveau bâtiment, et la seconde étape consiste à récupérer les mesures d'économie d'énergie implémentées sur ces bâtiments. Ce sont ces mesures qui composent la recommandation pour le nouveau bâtiment.

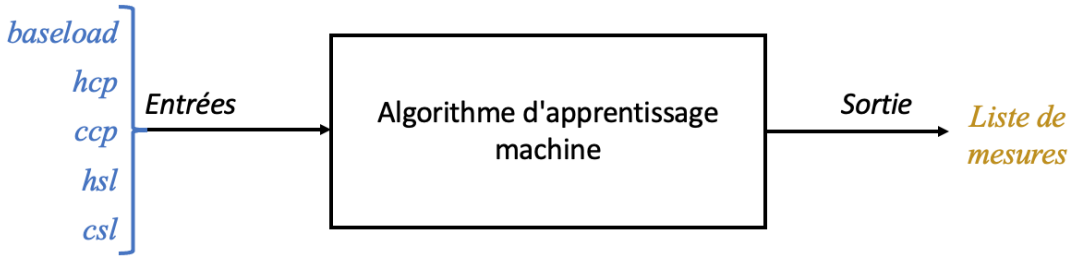

Figure 6.1. Schéma de fonctionnement de l'algorithme d'apprentissage machine pour la prédiction des mesures d'économie d'énergie

Pour réaliser ces deux étapes, nous avons implémenté un algorithme de voisinage kNN (« k Nearest Neighbors »), également appelé méthode des *k* voisins les plus proches, qui permet de retourner les mesures d'économie d'énergie implémentées aux bâtiments similaires. Cet algorithme est usuellement un algorithme de classification qui permet de classer en fonction de la classe majoritaire du voisinage. Cependant, l'algorithme est utilisé non pas pour du classement, mais pour le résultat intermédiaire sur les voisins. Il est donc utilisé pour déterminer un groupe (« cluster ») autour du nouveau bâtiment. Un algorithme de regroupement n'est pas utilisé puisqu'un tel algorithme crée plusieurs groupes dans la base de données, ce qui n'est pas ce qui est recherché ici. Puisque seulement un groupe autour du nouveau bâtiment doit être obtenu et que le nouveau bâtiment n'est pas ajouté à la base de données, utiliser le résultat intermédiaire d'un algorithme kNN est justifié. L'algorithme fonctionne de telle sorte que pour une nouvelle entrée, dans notre cas un nouveau bâtiment, les *k* échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée sont utilisés pour la recommandation, selon une distance à définir. Le nombre de voisins est défini selon *k* et fait partie des hyperparamètres du modèle. La distance fait également partie des hyperparamètres. Il existe plusieurs distances mathématiques, mais la plus commune est la distance euclidienne qui définit la distance entre deux points tels que :

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{6.1}$$

où a et b sont les deux points concernés et n est la dimension de l'espace de a et b. La distance euclidienne est la mesure de la vraie distance en ligne droite entre deux points dans l'espace euclidien. Dans un espace euclidien à deux dimensions, un problème de voisinage peut se représenter comme sur la Figure 6.2 qui illustre un voisinage pour $k = 3$ et $k = 6$.

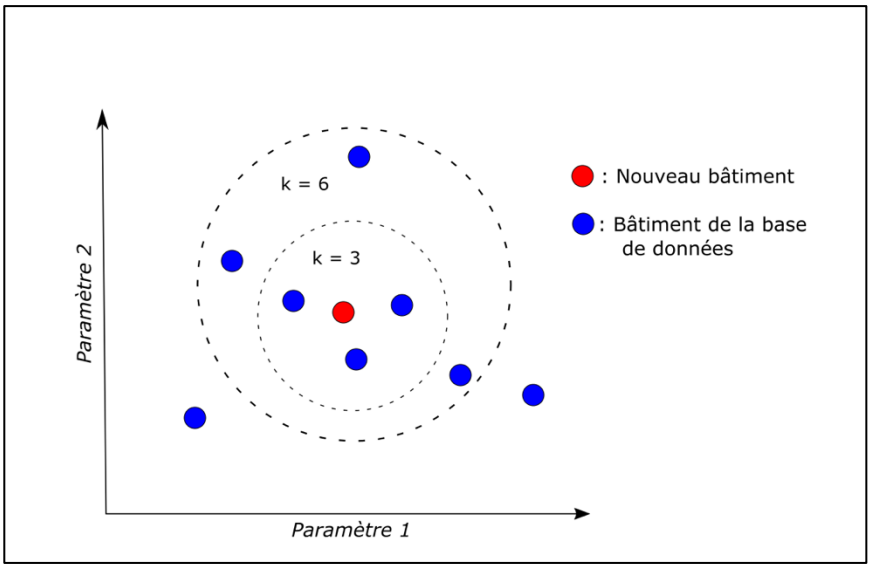

Figure 6.2. Schématisation d'un voisinage pour k=3 et k=6

Pour déterminer le nombre de voisins le plus adéquat, une validation de l'algorithme grâce aux données de test doit être réalisée. Le nombre de voisins est varié entre 2 et 10 afin de déterminer pour quel nombre de voisins les résultats sont les meilleurs.

Une fois les bâtiments similaires obtenus grâce au résultat intermédiaire de l'algorithme kNN, les mesures d'économie d'énergie pour ces bâtiments sont récupérées. Elles sont mises sous la forme d'un tableau à deux colonnes où la première colonne correspond à une mesure énergétique, et la seconde colonne correspond à la récurrence de la mesure parmi l'ensemble du voisinage. Le

nombre de récurrences est majoré par le nombre de voisins. En effet, comme une mesure ne peut être recensée qu'une seule fois par bâtiment, pour un voisinage de *k* bâtiments, la mesure ne pourra être apparaitre que *k* fois ou moins. La prédiction des mesures se fait donc avec un aspect probabiliste, car plus la récurrence de la mesure est élevée parmi les voisins, plus il y a de chance que la mesure prédite soit en effet à implémenter. Pour chaque mesure prédite, il est possible de définir une probabilité de prédiction telle que :

$$\varphi_{x,i} = \frac{r_{x,i}}{\sum_{j=1}^{p} r_{x,j}} * 100\% \quad (6.2)$$

où $\varphi_{x,i}$ est la probabilité de prédiction pour la mesure *i* du bâtiment x et les $r_{x,j}$ sont les récurrences pour chaque mesure prédite j, et p est le nombre de mesures prédites. Le Tableau 6.1 montre en exemple le calcul des mesures à recommander pour lequel les $k = 3$ plus proches voisins sont les bâtiments A, B et C.

Tableau 6.1. Exemple de recommandation de mesures pour un bâtiment quelconque

Mesures implémentées bâtiment A	**Mesures implémentées bâtiment B**	**Mesures implémentées bâtiment C**	**Mesures recommandées pour le nouveau bâtiment et $\varphi_{x,i}$**
-« Efficient lighting » -« Installation of efficient boiler » -« Air-source heat pumps »	-« Installation of efficient boiler » -« Centralized control »	-« Efficient lighting » -« Installation of efficient boiler » -« Air-source heat pumps »	-« Efficient lighting » : 25% -« Installation of efficient boiler » : 37,5% -« Air-source heat pumps » : 25% -« Centralized control » : 12,5%

Validation croisée et métriques

Afin de pouvoir évaluer l'efficacité de la recommandation, il faut définir des métriques d'évaluation. Une méthode par validation croisée est utilisée pour obtenir des résultats sans biais sur la séparation entraînement/test. Les métriques obtenues sont la moyenne des métriques sur l'ensemble des bâtiments test de la validation croisée. Un $cv_{100}(75/25)$ est utilisé. La recommandation fournit une liste de mesures, soit une liste de valeurs discontinues ou de classes. Il est donc possible d'utiliser comme métrique d'évaluation des métriques utilisées en classification. Les deux premières métriques considérées sont la précision et le rappel. La précision est une mesure de la capacité d'un modèle de classification à identifier uniquement les points de données pertinents, tandis que le rappel est une mesure de la capacité d'un modèle à trouver tous les cas pertinents dans un ensemble de données. Dans le contexte de la recommandation des mesures, la précision peut s'expliquer par « parmi toutes les mesures recommandées, combien sont réellement réalisées », et le rappel par « parmi toutes les mesures réellement réalisées, combien font partie des recommandations ». Mathématiquement, cela se traduit par :

$$P_{i,x} = \frac{TP_x}{TP_x + FP_x} \tag{6.3}$$

$$R_{i,x} = \frac{TP_x}{TP_x + FN_x} \tag{6.4}$$

où $P_{i,x}$ et $R_{i,x}$ sont respectivement la précision et le rappel pour le bâtiment x de la validation croisée i. TP_x est le nombre de « vrais positifs » pour le bâtiment x, c'est-à-dire les mesures recommandées réellement implémentées. FP_x est le nombre de « faux positifs » pour le bâtiment x, c'est-à-dire le nombre de mesures recommandées non implémentées. Finalement FN_x est le nombre de « faux négatif » pour le bâtiment x, c'est-à-dire le nombre de mesures non recommandées mais qui sont implémentées. Ces deux métriques sont ambivalentes et complémentaires, et permettent d'évaluer deux aspects de la classification différents. Selon le résultat désiré, il est parfois intéressant d'optimiser la précision et parfois le rappel, mais optimiser l'un implique généralement de négliger l'autre. Si l'on cherche à optimiser la précision, cela revient à chercher à ce que l'ensemble des mesures recommandées soient effectivement implémentées, même si certaines mesures implémentées ne sont pas recommandées. À l'inverse, si l'on cherche à optimiser le rappel, cela revient à chercher à ce que l'ensemble des mesures implémentées soient recommandées, même si certaines mesures recommandées ne sont pas implémentées. Un compromis entre ces deux métriques est recherché, alors le score F1 est introduit. Le score F1 est une moyenne harmonique entre la précision et le rappel et se définit mathématiquement par :

$$F1_{i,x} = \frac{P_{i,x} + R_{i,x}}{2 * P_{i,x} * R_{i,x}} \tag{6.5}$$

où $F1_{i,x}$ est le score F1 pour le bâtiment x de la validation croisée i. Optimiser le score F1 revient à chercher un équilibre entre la précision et le rappel.

Pour une validation croisée, les métriques obtenues sont :

$$P_i = \frac{1}{N} * \sum_{j=1}^{N} P_{i,j} \tag{6.6}$$

$$R_i = \frac{1}{N} * \sum_{j=1}^{N} R_{i,j} \tag{6.7}$$

$$F1_i = \frac{1}{N} * \sum_{j=1}^{N} F1_{i,j} \tag{6.8}$$

où *N* est le nombre de bâtiments test.

Finalement, pour l'ensemble du $cv_{100}(75/25)$:

$$P = \frac{1}{100} * \sum_{j=1}^{100} P_j \tag{6.9}$$

$$R = \frac{1}{100} * \sum_{j=1}^{100} R_j \tag{6.10}$$

$$F1 = \frac{1}{100} * \sum_{j=1}^{100} F1_j \tag{6.11}$$

Il est à noter que ces trois métriques ne prennent pas en compte la récurrence des mesures dans le voisinage. Il est donc intéressant d'également utiliser une métrique prenant en compte cet aspect. Une métrique est alors définie qui évalue la distribution correcte de la recommandation pour un bâtiment φ_i tel que :

$$\varphi_x = \sum_{j=1}^{c} \varphi_{x,j} \ \% \tag{6.12}$$

où φ_x est donc la distribution correcte de la recommandation pour le bâtiment x, et c est le nombre de mesures recommandées réellement implémentées. Les éléments de la somme ne correspondent donc seulement qu'à ces mesures. On peut introduire pour une cross validation Φ_i tel que :

$$\Phi_i = \frac{1}{N} * \sum_{j=1}^{N} \varphi_j \tag{6.13}$$

où *N* est le nombre de bâtiments test.

Finalement, pour l'ensemble du $cv_{100}(75/25)$:

$$\Phi = \frac{1}{100} * \sum_{j=1}^{100} \Phi_j \tag{6.14}$$

Pour un $cv_{100}(75/25)$, les métriques d'évaluation sont la précision, le rappel, le score F1, et la distribution correcte de la recommandation. Il est important de maximiser le score F1 et la distribution correcte de la recommandation alors que le rappel et la précision nous permettent d'avoir des informations complémentaires sur l'efficacité du modèle dans un contexte plus précis.

6.4 Optimisation des poids des paramètres d'entrées

Les paramètres CP forment l'ensemble des paramètres d'entrée de l'algorithme. Comme les paramètres sont standardisés en Z-score, chaque paramètre à un poids égal unitaire, et l'espace créé par les entrées est donc un espace à cinq dimensions égales. Cependant, choisir des poids unitaires égaux pour tous les paramètres n'est pas forcément la bonne solution. En effet, bien que l'ensemble des paramètres permettent de capturer un ou plusieurs phénomènes physiques liés au comportement énergétique du bâtiment, ils n'ont pas forcément le même impact sur la prédiction finale. Il faut donc faire une recherche des poids de chacun des paramètres d'entrée afin de déterminer la combinaison optimale. La variation du poids se fait simplement en multipliant le paramètre unitaire par un scalaire qui représente le nouveau poids. Par exemple, un poids de 10 correspond à une multiplication du paramètre en Z-score par 10.

Afin de déterminer les poids optimaux des paramètres, une recherche par grille est effectuée. C'est une méthode de recherche exhaustive, c'est-à-dire qu'elle repose sur de l'itération en changeant à chaque itération les paramètres. Pour réaliser une telle recherche, il faut fixer un intervalle de variation et un pas de variation pour chaque paramètre. L'ensemble des combinaisons possibles pour ces paramètres est testé. La combinaison ayant permis d'obtenir les meilleurs résultats est le résultat de cette méthode. Pour un problème à cinq paramètres, il faut faire attention, car si pour chaque paramètre, cinq valeurs sont testées, ce qui revient à tester 3125 combinaisons (5^5). La puissance de calcul de la machine peut donc vite être saturée si le nombre de valeurs testées augmente. Ainsi, pour la première étape de recherche, l'ensemble des paramètres sont testés avec des poids de : 0, 1, 10, 20, 30, 40. Cela correspond à 7776 possibilités. On obtient ainsi le vecteur de poids ω_1 , avec pour chaque paramètre $\omega_{1,k}$ où k représente le paramètre, qui est le poids optimal obtenu suite à la première recherche. Lors de la seconde phase de recherche, chaque paramètre est testé en prenant comme poids $\omega_{1,k}$, $\omega_{1,k} + 5$, et $\omega_{1,k} - 5$ (0 si négatif), cela représente 343 possibilités. On obtient ainsi pour chaque paramètre $\omega_{2,k}$. Ensuite, pour la troisième phase chaque paramètre est testé en prenant comme poids $\omega_{2,k}$, $\omega_{2,k} + 3$, et $\omega_{2,k} - 3$ (0 si négatif), ce qui représente 343 possibilités. On obtient ainsi pour chaque paramètre $\omega_{3,k}$. Enfin, pour la dernière phase chaque paramètre est testé en prenant comme poids $\omega_{3,k}$, $\omega_{3,k} + 1$, et $\omega_{3,k} - 1$ (0 si négatif), cela représente 343 possibilités. On obtient ainsi pour chaque paramètre $\omega_{4,k}$. Cette recherche de

plus en plus ciblée permet de trouver la combinaison optimale en détaillant toujours plus le poids optimal précédent, et ce sans saturer les capacités de calcul qui aurait été causé par une variation entre 0 et 50 avec un pas de 1 pour l'ensemble des paramètres (345 x 10^6 combinaisons).

Cette recherche des poids optimaux est réalisée après la recherche du nombre optimal de voisins qui est, elle, réalisée à un poids unitaire pour tous les paramètres. Ce choix est justifié par le coût calculatoire de cette méthode. Il vaut donc mieux la réaliser à la fin une fois que l'ensemble des autres hyperparamètres du modèle sont optimisés.

6.5 Résultats

La Figure 6.3 présentant les métriques en prenant comme entrée de l'algorithme les paramètres avec un poids unitaire, et le nombre *k* de voisins variables entre 2 et 10. La Figure 6.3 permet de constater que plus le nombre de voisins augmente, plus le score F1 et Φ diminue de manière quasi linéaire. On retrouve un maximum du score F1 et de Φ pour 2 voisins à respectivement 55,3% et 58,0%. Si l'on s'intéresse au rappel et à la précision, on constate leur caractère ambivalent, car le premier augmente et le second diminue avec l'augmentation du nombre de voisins. Ce comportement est logique d'après la définition du rappel et de la précision : plus le nombre de voisins est grand, plus le bruit des données est important, donc la précision diminue, mais il y a également plus de mesures recommandées et donc plus de chance que les mesures implémentées soient parmi le lot de recommandation. Si on cherche simplement à maximiser le score F1 et Φ, il faudrait choisir 2 voisins comme nombre optimal de voisins. Cependant, il faut également prendre en compte l'utilité de l'algorithme. On cherche ici à présenter à l'utilisateur un ensemble de mesures qu'il pourrait implémenter dans son bâtiment. Même si les bâtiments voisins ont de fortes probabilités d'avoir les mêmes mesures à implémenter, prendre un si faible nombre de voisins peut induire un biais si un des deux voisins en question n'est pas commun. Il vaut donc mieux recommander plus de mesures au risque d'émettre une erreur légèrement plus élevée, si cette erreur permet d'éviter un biais trop grand. Ainsi, $k = 3$ a été défini comme nombre de voisins optimal. Cela permet d'avoir un score F1 et un Φ de respectivement 52,4% et 55,5%, ce qui ne diminue pas non plus trop l'efficacité de l'algorithme par rapport à 2 voisins.

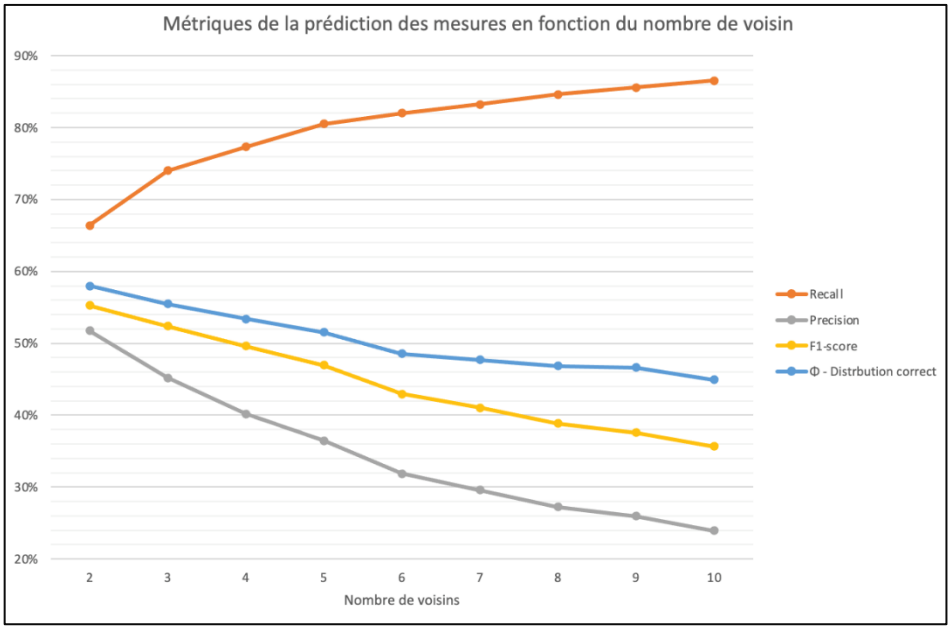

Figure 6.3. Évolution des métriques sur la prédiction des mesures en fonction du nombre de voisins

Une fois le nombre de voisins optimal obtenu, la recherche en grille donne les résultats du Tableau 6.2**,** qui présente les poids optimaux à chaque étape de la recherche ainsi que les métriques obtenues pour chaque étape correspondante pour $k = 3$. Le point optimal est atteint dès la deuxième étape et correspond à la combinaison ω_2, tel que $\omega_{2_{baseload}} = 1$, $\omega_{2_{hsl}} = 1$, $\omega_{2_{hcp}} = 1$, $\omega_{2_{csl}} = 35$, $\omega_{2_{ccp}} = 0$. On constate en premier lieu que le ccp à un poids de 0 et n'est donc virtuellement pas pris comme entrée de l'algorithme pour la prédiction. Le second paramètre lié à la climatisation csl à lui un poids de 35, ce qui ne signifie pas qu'il est 35 fois plus important que les autres, mais qu'au contraire son espace mathématique est dilaté par 35 pour lui donner de l'importance, cela signifie que csl avec un poids unitaire a moins d'importance que les trois autres paramètres avec un poids unitaire. Ce comportement est explicable grâce à l'analyse sur les bâtiments faites à la section 4.2 expliquant que les écoles primaires, et par extension une grande partie des K12, n'ont pas de climatisation. Ainsi les paramètres liés à la climatisation sont forcément « moins importants » d'un point de vue global. Cependant, cette analyse et ces résultats pour les poids ne sont vrais que car les bâtiments sont situés au Québec, une région du monde où les températures

sont relativement basses et où les besoins en climatisation sont donc limités. Pour des bâtiments situés dans des climats plus chauds, il faut répéter le travail de recherche par grille afin de trouver les poids adaptés aux données. La combinaison finale permet d'obtenir un score F1 et un Φ supérieur qu'avec des poids unitaires pour tous les paramètres, avec une augmentation de respectivement 13,5% et 8,7%.

Tableau 6.2. Efficacité de la recommandation des mesures d'économie d'énergie en fonction des poids des paramètres d'entrées

Combinaison	ω_1	ω_2	ω_3	ω_4
Poids (dans l'ordre *[baseload, hsl, hcp, csl, ccp])*	[1;1;1;40;0]	[1;1;1;35;0]	[1;1;1;35;0]	[1;1;1;35;0]
Rappel	76,0%	76,3%	76,3%	76,3%
Précision	57,9%	59,6%	59,6%	59,6%
Score F1	64,3%	65,9%	65,9%	65,9%
Φ	62,2%	64,2%	64,2%	64,2%

Un modèle kNN a été développé pour la recommandation des mesures d'économie d'énergie d'un bâtiment en optimisant les hyperparamètres du modèle, à savoir le nombre de voisins et le poids des entrées. Ce modèle permet d'avoir un score F1 final de 65,9% et une distribution correcte prenant en compte les récurrences parmi les voisins Φ de 64,2%. Au niveau de la recommandation pour l'utilisateur, l'algorithme fournit dans un premier temps les noms des trois bâtiments voisins afin que l'utilisateur puisse récupérer des informations sur ces bâtiments s'il le désire. Dans un second temps, l'algorithme fournit les mesures recommandées qui peuvent être représentées de deux façons. Les mesures recommandées peuvent être représentées par un histogramme en barre, où l'abscisse représente les différentes mesures et l'ordonnées la récurrence de ces mesures parmi les voisins comme illustré sur la Figure 6.4. Cette représentation est compréhensive puisque

l'ensemble des informations est présenté simplement pour que le lecteur interprète rapidement les résultats. Les mesures recommandées peuvent aussi être représentées par un diagramme polaire (ou toile d'araignée), où les angles représentent les différentes mesures et le rayon la récurrence de ces mesures parmi les voisins comme illustré sur la Figure 6.5. Cette représentation est graphique puisque l'aspect global de la figure est plus esthétique au profit de la rapidité de lecture. C'est cette dernière représentation qui a été utilisée dans l'outil final, dont l'interface graphique est présentée à l'Annexe B, où l'on retrouve les trois bâtiments voisins ainsi que la recommandation des mesures par diagramme polaire.

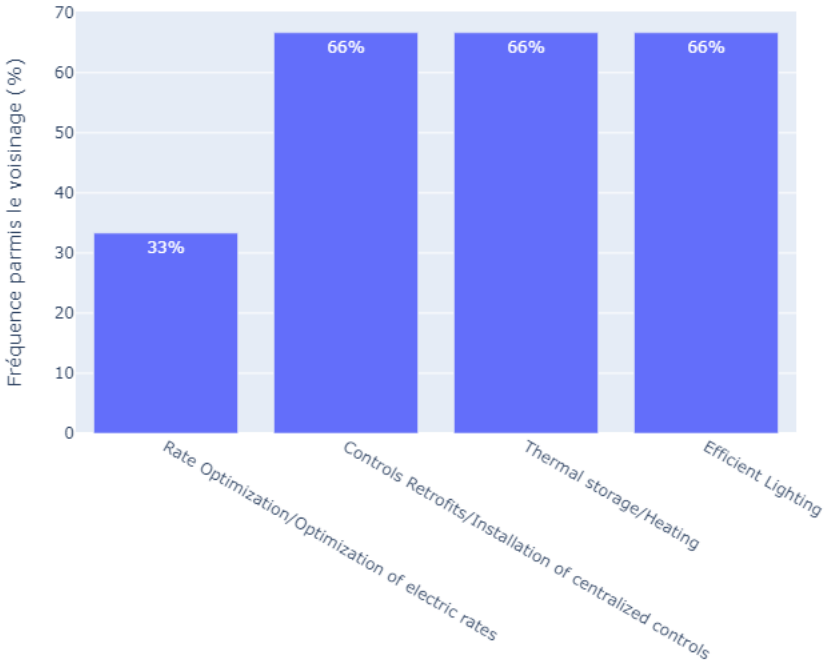

Figure 6.4. Histogramme en barre pour la recommandation des mesures d'économie d'énergie

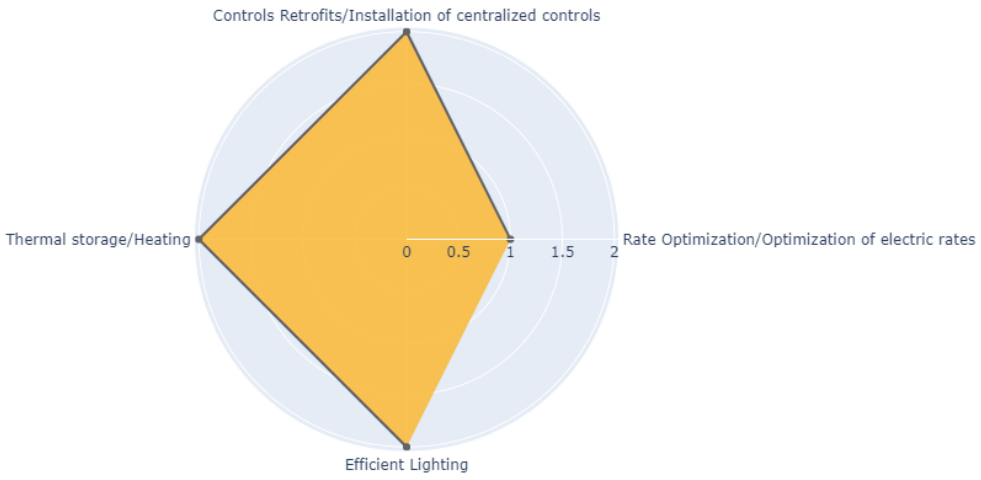

Figure 6.5. Diagramme polaire pour la recommandation des mesures d'économie d'énergie

CHAPITRE 7 DISCUSSION

7.1 Prédiction de l'économie d'énergie

Les chapitres 4 et 5 présentent le développement et la validation d'une méthode de prédiction du potentiel d'économie annuelle pour des écoles primaires et secondaires. Les données de consommation mensuelle et la température mensuelle moyenne extérieure sont les seules données nécessaires pour appliquer la méthode à un bâtiment. Le modèle « change-point » de Kissock (Kissock et al., 2003) d'un bâtiment est généré à partir de ces données. Les paramètres du modèle « change-point » sont utilisés comme entrée de l'algorithme d'apprentissage machine. Ces paramètres permettent de tenir compte du comportement énergétique global du bâtiment sans faire une étude détaillée de l'enveloppe et des systèmes énergétiques du bâtiment. Un algorithme de forêt aléatoire est utilisé pour la prédiction, suggéré par Prakash et al. (Prakash et al., 2018) comme un algorithme prédictif de pointe. L'utilisation de ce type d'algorithme pour la prédiction du potentiel d'économie n'est cependant pas présente dans la littérature. L'algorithme entraîné avec les données de la base de données est implémenté en Python (Van Rossum, 2020). Ainsi, la méthode représente un nouvel apport à l'état de l'art des méthodes de prédiction de l'économie basée sur les données, puisqu'aucune méthode utilisant les mêmes données ou le même algorithme n'est recensée dans la littérature.

7.2 Recommandation des mesures d'économie d'énergie

Le chapitre 6 présente le développement et la validation d'une méthode de recommandation des mesures d'économie d'énergie à implémenter afin d'atteindre les économies d'énergie. Comme pour la prédiction des économies, les paramètres du modèle « change-point » de Kissock (Kissock et al., 2003) sont utilisés comme les données d'entrée. Pour l'algorithme servant à la prédiction, un algorithme kNN est utilisé, non pas pour de la classification mais pour le résultat intermédiaire de l'algorithme donnant les voisins les plus proches. Cet algorithme permet de trouver les trois bâtiments de la base de données les plus proche du bâtiment à traiter. Cette proximité est évaluée

selon la distance euclidienne de l'espace mathématique créé par les cinq paramètres du modèle. Le poids optimal de chaque dimension est déterminé afin d'obtenir le meilleur résultat de prédiction. Les mesures d'économie d'énergie implémentées aux trois bâtiments « voisins » sont celles recommandées pour le bâtiment traité. Dans la littérature, les mesures d'économie d'énergie peuvent être utilisées comme des données d'entrée pour la prédiction des économies, mais aucun article ne présente une méthode où les sorties sont les mesures d'économie d'énergie. Ainsi, cette méthode présente un apport à l'état de l'art puisqu'aucune méthode de recommandation des mesures d'économie d'énergie basée sur les données n'est présente dans la littérature.

CHAPITRE 8 CONCLUSION

8.1 Synthèse des travaux

La présente étude a permis de développer deux méthodes de prédiction pour les travaux de rénovation énergétique des bâtiments éducatifs basée sur les données.

La première méthode concerne la prédiction du potentiel des économies d'énergies annuelles pour un bâtiment associés aux travaux d'amélioration énergétique. Dans un premier temps, les données des écoles primaires ont été utilisées pour sélectionner une méthode de prédiction. Une méthode par algorithme de forêt aléatoire a été sélectionnée. La méthode génère les facteurs CP standardisés à partir des factures et les utilise en entrée de l'algorithme pour prédire en sortie l'économie annuelle. Cette méthode a dans un second temps été améliorée et généralisée pour d'autres types de bâtiments. L'amélioration consiste en une redéfinition de l'économie pour les données d'entraînement et un calcul optimisé du modèle CP. La généralisation de la méthode à d'autres types de bâtiments consiste à tester et valider ou rejeter la méthode pour les autres types de bâtiments éducatifs ou pour des regroupements de types de bâtiments. Finalement, la méthode a été validée pour les écoles primaires et les écoles secondaires, et également pour le regroupement de ces deux types (le groupe K12). Le groupe K12 permet d'avoir un grand nombre de bâtiments pour l'entraînement avec une erreur sur la prédiction de l'économie d'énergie de 42% pour le *CV(RMSE)* et de 64 MJ/m^2 pour la *MAE*.

La deuxième méthode concerne la recommandation des mesures d'économie d'énergie à implémenter afin de réaliser le potentiel d'économies d'énergies. Cette méthode utilise les même données (K12) car ce sont les seuls bâtiments de la base de données qui ont de l'information sur les mesures pour l'entraînement. Un algorithme de voisinage kNN est utilisé, prenant en entrée les paramètres du modèle CP avec des poids optimisés pour l'algorithme, et fournit en sortie la liste des trois bâtiments voisins les plus proches (selon les paramètres du modèle CP). Une liste des mesures est obtenue à partir de ces voisins. La méthode permet d'obtenir un score F1 de 66% sur la recommandation des mesures.

Les deux méthodes développées ont permis de répondre aux objectifs du projet de recherche : Déterminer le potentiel d'économie énergétique d'un bâtiment, et recommander des mesures d'économie d'énergie à implémenter afin d'atteindre les économies d'énergies, ces deux objectifs reposant sur une approche basée sur les données de projets antérieurs. Les méthodes ont été implémentées dans des outils numériques afin de faciliter la distribution et l'utilisation. Les outils permettent d'obtenir les résultats des prédictions via une interface web. Ainsi, trois outils numériques ont été développés dans le cadre du projet. Un premier outil sert à visualiser les données de chaque bâtiment dans la base de données, à comparer entre différents bâtiments la consommation énergétique ou monétaire, et à superposer des modèles (degrés-jour, « change-point », etc) sur les données selon des modèles. Un deuxième outil sert à la prédiction à partir de données de consommation pour un nouveau bâtiment. Il fournit ensuite la modélisation graphique par modèle CP, la prédiction de l'économie d'énergie annuelle et la prédiction des mesures énergétiques. Cet outil est l'implémentation des différentes méthodes développées dans ce projet. Enfin, un troisième et dernier outil sert à la gestion des données. Cet outil permet d'ajouter des bâtiments à la base de données, ou à mettre à jour les modèles utilisés dans l'outil de prédiction. Ce dernier outil est à destination des développeurs mais permet de faciliter la gestion des données via une interface graphique. L'ensemble de ces outils ont été développés par code Python (Van Rossum, 2020) en utilisant le progiciel Streamlit (*Streamlit*, 2020) pour l'interface graphique. L'interface graphique pour l'outil de prédiction, qui est l'outil principal, est présenté à l'ANNEXE B.

8.2 Limitations

La quantité et la qualité des données sont des enjeux principaux en apprentissage machine (Halevy et al., 2009). La quantité correspond à la taille de l'échantillon d'entraînement et aux attributs disponibles pour chaque échantillon. Dans le contexte de ce projet de recherche, l'échantillon d'entraînement correspond aux bâtiments de la base de données. La qualité des données correspond à l'adéquation d'un ensemble de données et de son objectif spécifique. Elle est basée sur des caractéristiques qualitatives telles que l'exactitude, l'exhaustivité, la cohérence, la validité et l'unicité. Les données constituent l'élément principal d'un algorithme d'apprentissage machine, car

l'algorithme utilise ces données pour construire un modèle cohérent par rapport à l'objectif. De manière générale, plus la quantité des données est importante, meilleur est le résultat, car l'algorithme est entraîné avec un large spectre de scénarios. Il est donc apte à faire une prédiction sur un éventail de situations différentes. Similairement, plus les données sont complètes et sans bruit, meilleur est l'algorithme, car le modèle construit a une plus faible erreur de bruit. Cependant, il est impossible de définir un nombre minimum de données d'entraînement ou un seuil minimum de qualité de données de manière générale : chaque problème est unique et donc la qualité et la quantité des données nécessaires varient pour chaque contexte. Ici, le choix a été fait de conserver uniquement les bâtiments avec un R^2 du modèle CP supérieur à 0,7, ce qui est la mesure de qualité, et ce choix a permis de conserver 157 bâtiments pour la catégorie des bâtiments K12 clés. Les résultats finaux montrent que le seuil de qualité et le nombre de bâtiments sont suffisants par rapport aux objectifs.

Un second enjeu important dans un problème d'apprentissage machine concerne les entrées disponibles pour l'algorithme. La base de données d'Ecosystem a permis d'utiliser les paramètres CP comme entrée du modèle mais ce choix a été fait par rapport aux données disponibles. En effet, selon la littérature, il apparaît que d'autres méthodes basées sur d'autres algorithmes et d'autres entrées ont permis d'obtenir des résultats similaires à cette étude, comme notamment la méthode développée par Re-cecconi et al. (Re Cecconi et al., 2019) ou la méthode développé par Ascione et al.(Ascione et al., 2017). Par contre, les types de données disponibles dans la littérature n'étaient pas toutes disponibles dans la base de données d'Ecosystem. Il est donc important de se questionner sur les entrées disponibles et l'exploitation qu'il est possible d'en faire en rapport avec l'objectif. Par exemple, il aurait pu être envisagée d'ajouter en entrée à l'algorithme un paramètre caractérisant la localisation du bâtiment, le modèle final aurait donc pris en compte un facteur environnemental en plus des facteurs énergétiques issus du modèle CP. Ainsi dans un problème d'apprentissage machine, la méthode de résolution est directement dépendante des entrées disponibles. Il faut donc parfois chercher à obtenir de nouvelles entrées si celles initiales ne permettent pas de développer une méthode adéquate.

8.3 Pistes d'amélioration

Une première piste d'amélioration serait la prédiction de l'économie pour chaque énergie (électricité ou carburants). Cela permettrait de raffiner les résultats de l'outil et d'améliorer ses fonctionnalités. La prédiction de l'économie d'énergie est actuellement réalisée sur la consommation totale pour un bâtiment, toutes énergies confondues. Cependant, dans le cas de travaux de rénovation énergétique, il n'est pas rare qu'un remplacement de combustible ait lieu, c'est-à-dire qu'une des énergies utilisées avant travaux est remplacée par des alternatives plus propres et économiques. Dans ce cas, il est possible d'obtenir une économie énergétique globale mais qu'il y ait une augmentation de la consommation pour une des sources énergétiques et une diminution pour une autre. Avoir une méthode de prédiction qui permet de faire la prédiction de l'économie pour chaque source d'énergie permettrait une vision détaillée sur l'évolution des consommations pour chaque énergie. De plus, une séparation en sources d'énergie permettrait également d'estimer une économie monétaire par source d'énergie, et donc de simuler les factures après projet. L'outil permettrait donc de simplifier encore plus l'élaboration de la proposition aux clients et de leurs fournir encore plus d'information.

Une deuxième piste de développement concerne l'analyse des données aberrantes. Le R^2 est la seule mesure de qualité utilisée dans ce projet pour accepter ou rejeter un bâtiment du jeu de données. Cependant, il est possible que même au sein d'un seul bâtiment ayant un R^2 acceptable, certaines données de consommations soient aberrantes. Il est aussi possible que les données d'un bâtiment le rendent intraitable par l'outil sans que l'utilisateur s'en rende compte en visualisant les données. La détection automatique de ces données erronées est l'enjeu de la détection de données aberrantes. Dans cette étude, une analyse des données aberrantes sur les facteurs CP a été réalisée mais les résultats n'étant pas concluants et ils n'ont pas été présentés. Des outils plus complet et complexe d'apprentissage machine existent pour la détection de données aberrantes (test de Chauvenet, test de Grubbs, etc.). Il serait alors possible de faire une étude détaillée sur la détection pour pouvoir éliminer les données aberrantes et également mieux définir les bâtiments pour lesquels l'outil peut être utilisé.

Enfin, une dernière piste de développement du projet serait d'améliorer les interfaces utilisateurs et les fonctionnalités des outils numériques développés. En effet, bien que l'ensemble des outils développés soient fonctionnels et répondent aux besoins pour lesquels ils ont été créés, il serait possible de les améliorer en optimisant l'interface utilisateur, notamment en maximisant l'utilisation de code HTML plutôt que le cadre limitant qu'impose le progiciel Streamlit (*Streamlit*, 2020). De plus, certaines fonctionnalités pourraient également être implémentées, comme la génération automatique d'un rapport de la prédiction mis en forme, ce qui simplifierait le partage à l'interne et même à l'externe des résultats générés par les outils.

RÉFÉRENCES

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews, 81*, 1192-1205. https://doi.org/10.1016/j.rser.2017.04.095

American Society of Heating Refrigerating Air-Conditioning Engineers. 36.7 ENERGY-EFFICIENCY MEASURES. Dans *2015 ASHRAE Handbook*. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. (ASHRAE).

American Society of Heating Refrigerating Air-Conditioning Engineers. (2017). *2017 ASHRAE handbook*.

Anna, V., Vasily, E., & Andrey, G. (2017). *CatBoost: gradient boosting with categorical features support* NIPS 2017.

Ascione, F., Bianco, N., De Stasio, C., Mauro, G. M., & Vanoli, G. P. (2017). Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. *Energy, 118*, 999-1017. https://doi.org/10.1016/j.energy.2016.10.126

Atlas Climatique du Canada. (2021). *Variables climatiques*. https://atlasclimatique.ca/variables

Capozzoli, A., Grassi, D., & Causone, F. (2015). Estimation models of heating energy consumption in schools for local authorities planning. *Energy and Buildings, 105*, 302-313. https://doi.org/10.1016/j.enbuild.2015.07.024

Chung, W., Hui, Y. V., & Lam, Y. M. (2006). Benchmarking the energy efficiency of commercial buildings. *Applied Energy, 83*(1), 1-14. https://doi.org/10.1016/j.apenergy.2004.11.003

Coakley, D., Raftery, P., & Keane, M. (2014). A review of methods to match building energy simulation models to measured data. *Renewable and Sustainable Energy Reviews, 37*, 123-141. https://doi.org/10.1016/j.rser.2014.05.007

Conraud-Bianchi, J. (2008). *A Methodology for the Optimization of Building Energy, Thermal, and Visual Performance.*

Crawley, D. B., Lawrie, L. K., Winkelmann, F. C., Buhl, W. F., Huang, Y. J., Pedersen, C. O., . . . Glazer, J. (2001). EnergyPlus: creating a new-generation building energy simulation program. *Energy and Buildings, 33*(4), 319-331. https://doi.org/https://doi.org/10.1016/S0378-7788(00)00114-6

Deb, C., & Lee, S. E. (2018). Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data. *Energy and Buildings, 159*, 228-245. https://doi.org/10.1016/j.enbuild.2017.11.007

Gouvernement Canada. (2021). *Environnement et ressources naturelles*. https://www.canada.ca/fr/services/environnement.html

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems, 24*(2), 8-12. https://doi.org/10.1109/MIS.2009.36

Iyengar, S., Lee, S., Irwin, D., Shenoy, P., & Weil, B. (2020). *WattScale: A Data-driven Approach for Energy Efficiency Analytics of Buildings at Scale*.

Johnson Controls. (2021). *Building Efficiency Targeting Tool for Energy Retrofits (BETTER)*. Johnson control. https://github.com/LBNL-JCI-ICF/better

Kazaki, A. G., & Papadopoulos, T. A. (2018). *Cluster analysis of university campus smart meter data.* 2018 53rd International Universities Power Engineering Conference (UPEC), 4-7 Sept. 2018, Piscataway, NJ, USA (p. 6 pp.). https://doi.org/10.1109/UPEC.2018.8541941

Kissock, J., Haberl, J., & Claridge, D. (2003). Inverse modeling toolkit: Numerical algorithms. *ASHRAE Transactions, 109*, 425-434.

Kreyszig, E. (1979). *Advanced engineering mathematics*. John Wiley & Sons.

Li, H., Szum, C., Lisauskas, S., Bekhit, A., Nesler, C., & Snyder, S. C. (2019). *Targeting building energy efficiency opportunities: An Open-source Analytical Benchmarking Tool.* ASHRAE Winter Conference,, Atlanta, GA, United states (vol. 125, p. 470-478).

NOAA - National Centers for Environmental Information. https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825--2830.

Prakash, A. K., Susu, X., Rajagopal, R., & Noh, Y. (2018). Robust Building Energy Load Forecasting Using Physically-based Kernel Models. *Energies, 11*(4), 862 (821 pp.). https://doi.org/10.3390/en11040862

Rajagopal, R., & Hae Young, N. (2013). *Data-driven Forecasting Algorithms for Building Energy Consumption.* Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2013, 10-14 March 2013, USA (vol. 8692, p. 86920T (86928 pp.)). https://doi.org/10.1117/12.2009894

Re Cecconi, F., Moretti, N., & Tagliabue, L. C. (2019). Application of artificial neutral network and geographic information system to evaluate retrofit potential in public school buildings. *Renewable & Sustainable Energy Reviews, 110*, 266-277. https://doi.org/10.1016/j.rser.2019.04.073

Roth, J., & Jain, R. (2018). *Data-driven, Multi-metric, and Time-varying (DMT) Building Energy Benchmarking Using Smart Meter Data*.

Streamlit. (2020). https://streamlit.io/

Trevor, H., Robert, T., & Jerome, F. (2009). *The Elements of Statistical Learning*. Springer.

University of Wisconsin--Madison. Solar Energy, L. (1975). *TRNSYS, a transient simulation program*. Madison, Wis. : The Laboratory, 1975.

Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.

Whitmore, J., & Pineau, P.-O. (2020). Etat de l'énergie au Québec. *HEC Montreal*.

Winkelmann, F. C., Birdsall, B. E., Buhl, W. F., Ellington, K. L., Erdem, A. E., Hirsch, J. J., & Gates, S. (1993). *DOE-2 supplement: Version 2.1E*. https://www.osti.gov/biblio/10147851

https://www.osti.gov/servlets/purl/10147851

Yeonsook, H., & Zavala, V. M. (2012). Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings, 53*, 7-18. https://doi.org/10.1016/j.enbuild.2012.06.024

Yixuan, W., Xingxing, Z., Yong, S., Liang, X., Song, P., Jinshun, W., . . . Xiaoyun, Z. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable & Sustainable Energy Reviews, 82*, 1027-1047. https://doi.org/10.1016/j.rser.2017.09.108

Zhang, Y., O'Neill, Z., Dong, B., & Augenbroe, G. (2015). Comparisons of inverse modeling approaches for predicting building energy performance. *Building and Environment, 86*, 177-190. https://doi.org/10.1016/j.buildenv.2014.12.023

ANNEXE A - ASYNCHRONICITÉ DES FACTURES

Dans la base de données, toutes les factures pour les différents bâtiments sont mensuelles, et la mensualité correspond aux douze mois annuels. Cette répartition des factures permet de faciliter le calcul du modèle CP décrit en section 5.2, car le modèle est calculé en cumulant la consommation de toutes les énergies au mois, il est donc simple de sommer les factures des différentes énergies sur les mêmes périodes. Cependant, ce cas de figure n'est pas universel. Les périodes des factures des différentes énergies ne sont pas nécessairement fixées sur les mois de l'année et ne coïncident pas sur les mêmes périodes entre-elles, il est donc impossible de sommer les différentes factures directement car les périodes temporelles ne correspondent pas. Ce problème doit être résolu car même si les bâtiments de la base de données peuvent être traités, l'outil a pour ambition de traiter des nouveaux bâtiments pour lesquels les factures ne coïncident pas forcément.

Pour pallier les problèmes des factures asynchrones, la méthode développée consiste à calculer le modèle CP pour chaque énergie indépendamment, puis de calculer grâce à ces modèles une consommation journalière virtuelle pour tous les jours de l'année; cette fausse consommation est ensuite utilisée pour évaluer la distribution journalière de la vraie facture d'énergie. Une fois les vraies factures distribuées journalièrement pour chaque énergie, il suffit de sommer selon les mois les consommations, puis de calculer le modèle CP global sur ces mois et ces consommations. La distribution journalière des factures ne se fait pas de manière uniforme mais est dépendante de la température moyenne journalière. De plus, la somme des consommations utilisées pour calculer le modèle CP est la même que la somme des consommations aux factures, car les modèles CP des énergies servent à évaluer une distribution et non à estimer une consommation réelle. Il est également possible ici que pour certains bâtiments, une ou des énergies ne se modélisent pas correctement, c'est-à-dire que le coefficient de détermination R^2 soit inférieur au seuil fixé à 0,7. Dans ce cas, le modèle change-point de cette énergie n'est défini que par le $baseload$ (tous les autres facteurs sont fixés à zéro) et le $baseload$ est égal à la moyenne des consommations mensuelles. D'un point de vue mathématique, en s'intéressant à un bâtiment cette méthode revient à réaliser :

$$E^{k}_{virtuel,day\ j} = CP^{k}(\bar{T}^{(n)}_{day\ j}) \tag{A.1}$$

$$CP^{k}\left(\bar{T}^{(n)}_{day\ j}\right) = baseload^{k} + hsl^{k} * \left(hcp^{k} - \bar{T}^{(n)}_{day\ j}\right)^{+} + csl^{k} * \left(\bar{T}^{(n)}_{day\ j} - ccp^{k}\right)^{+} \tag{A.2}$$

où k fait référence à une source d'énergie, $\bar{T}^{(n)}_{day\ j}$ est la température moyenne journalière pour le jour j et $E^{k}_{virtuel,day\ j}$ est une consommation virtuelle pour l'énergie k et le jour j. L'équation (A.2) correspond au modèle CP mais est différente de l'équation 5.3 car il s'agit ici de calculer une consommation journalière avec une température moyenne journalière. Ensuite :

$$E^{k}_{dist,day\ j} = E^{k}_{bill\ x} * \frac{E^{k}_{virtuel,day\ j}}{\sum_{\eta=1}^{F} E^{k}_{virtuel,day\ \eta}} \tag{A.3}$$

$$E_{month,i} = \sum_{k} \sum_{\gamma=1}^{D_i} E^{k}_{dist,day\ \gamma} \tag{A.4}$$

où $E^{k}_{dist,day\ j}$ est la consommation distribuée pour le jour j et l'énergie k, $E^{k}_{bill\ x}$ est la consommation inscrite à la facture pour la période x pour l'énergie k, F est le nombre de jours couverts par la facture, et $E_{month,i}$ est la consommation pour un mois i, et D_i est le nombre de jours dans ce mois.

Une fois $E_{month,i}$ calculée pour chaque mois, il est ensuite possible de calculer le modèle CP basé sur ces consommations mensuelles et sur les températures moyennes mensuelles correspondantes, comme dans la méthode décrite plus tôt dans cette section. Cependant dans le calcul de la régression permettant d'obtenir le modèle, les mois incomplets ne sont pas pris en compte. Il s'agit soit du premier, soit du dernier mois apparaissant aux factures. En effet, il est possible avec cette méthode que pour le premier et le dernier mois, les factures ne couvrent pas l'entièreté du mois. Les mois incomplets sont donc retirés pour le calcul du modèle.

ANNEXE B - INTERFACE GRAPHIQUE DE L'OUTIL DE PRÉDICTION

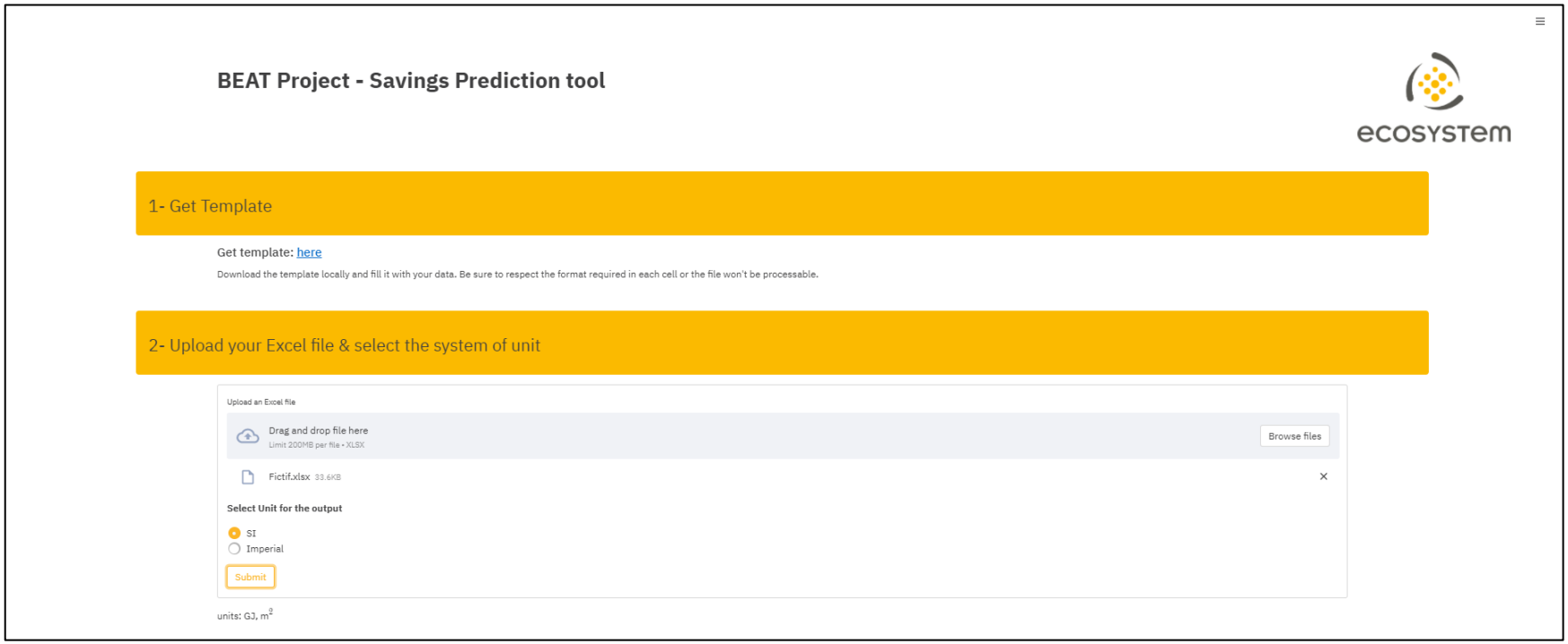

Figure B.1. Interface graphique de l'outil de prédiction

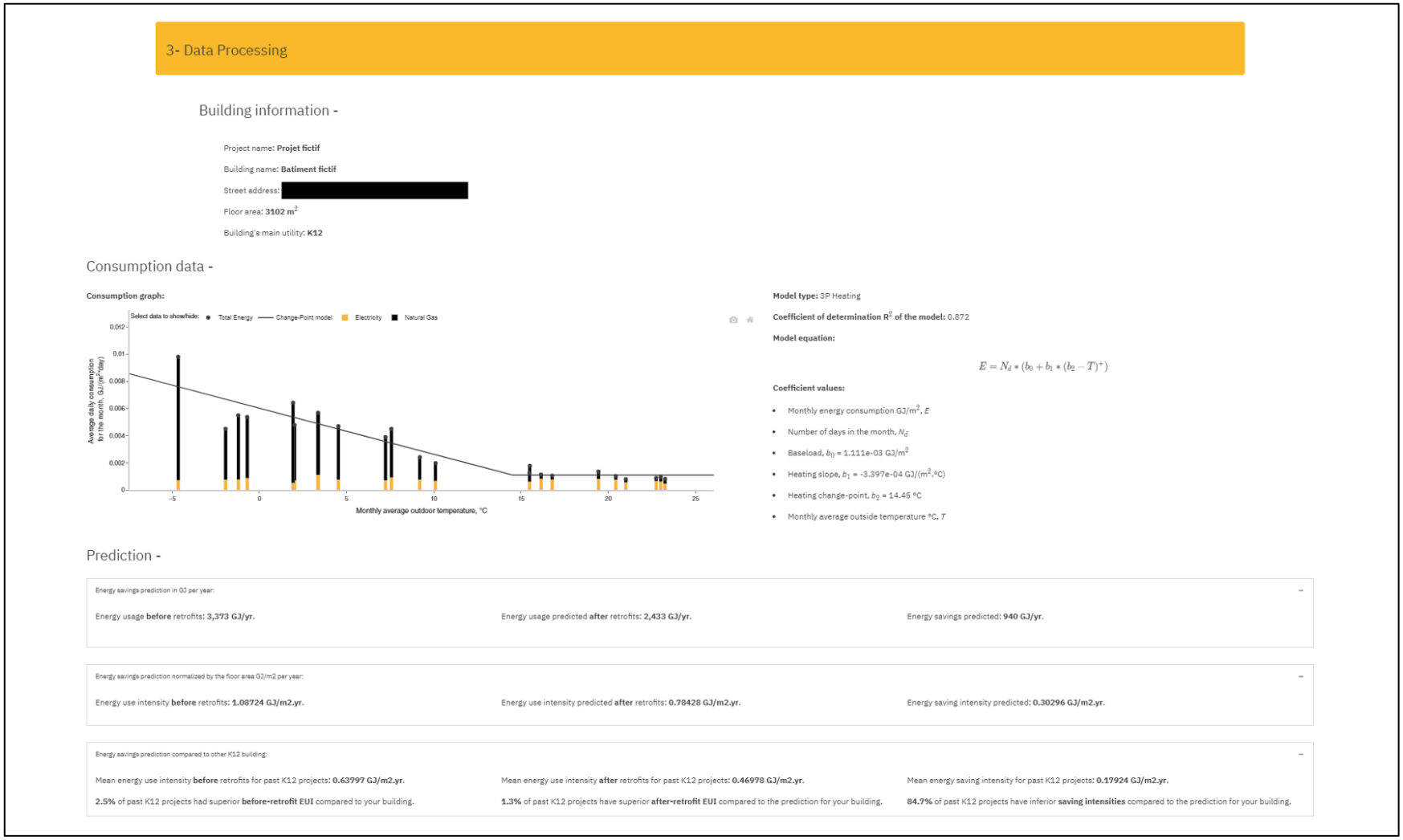

Figure B.1. Interface graphique de l'outil de prédiction (suite)

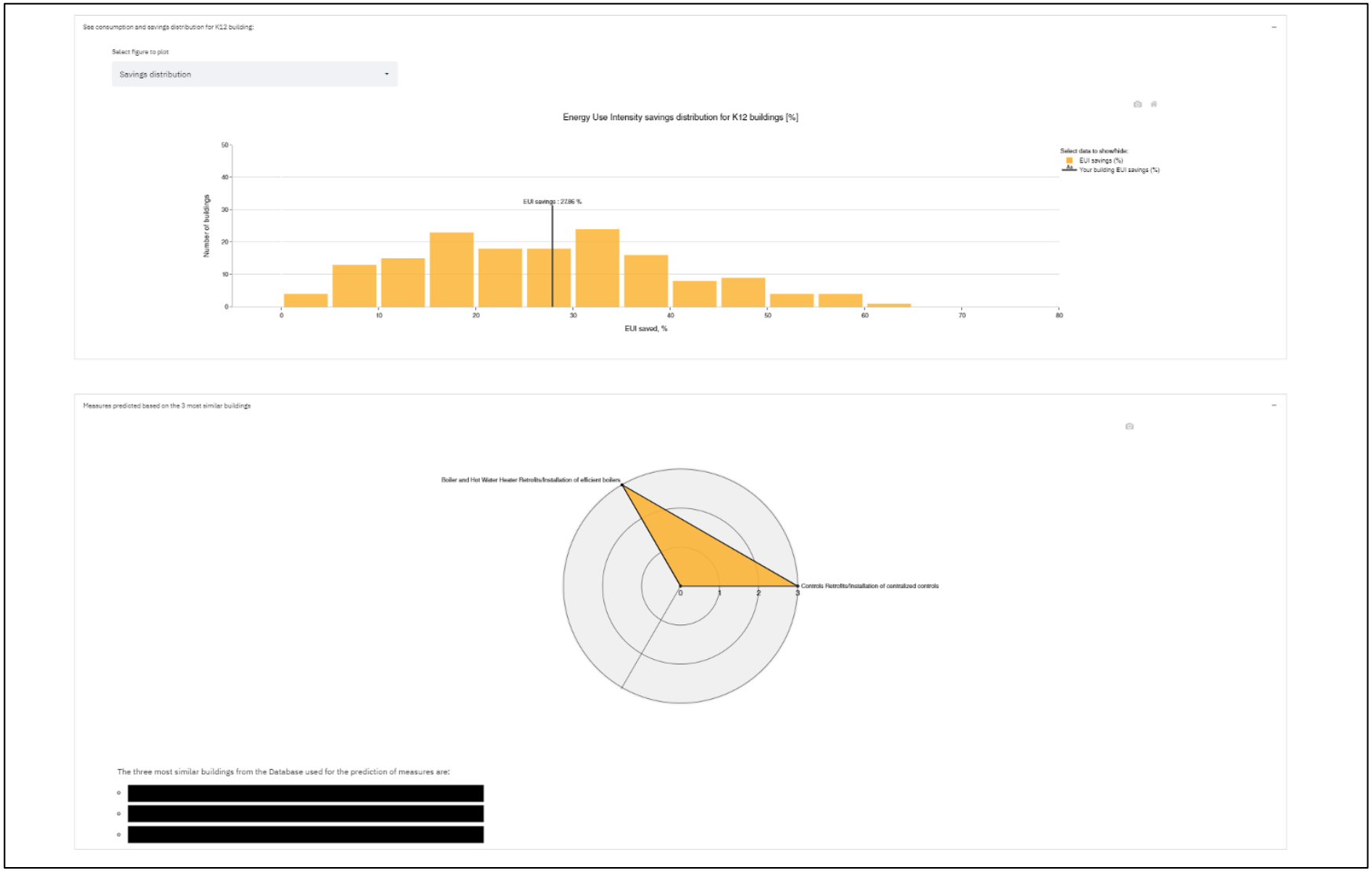

Figure B.1. Interface graphique de l'outil de prédiction (suite)