Titre: Title:	Regroupement par algorithme hiérarchique et k-means des profils d'usage des bornes de recharge publiques de véhicules électrique et analyse des facteurs d'usage liés à leur environnement
Auteur: Author:	Ismail Zejli
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Zejli, I. (2021). Regroupement par algorithme hiérarchique et k-means des profils d'usage des bornes de recharge publiques de véhicules électrique et analyse des facteurs d'usage liés à leur environnement [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9914/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Regroupement par algorithme hiérarchique et k-means des profils d'usage des bornes de recharge publiques de véhicules électrique et analyse des facteurs d'usage liés à leur environnement

ISMAIL ZEJLI
Département de mathématiques et de génie industriel

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

Décembre 2021

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Regroupement par algorithme hiérarchique et k-means des profils d'usage des bornes de recharge publiques de véhicules électrique et analyse des facteurs d'usage liés à leur environnement

présenté par **Ismail ZEJLI**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Luc-Désiré ADJENGUE, président
Hanane DAGDOUGUI, membre et directrice de recherche
Martin TRÉPANIER, membre et codirecteur de recherche
Jean Luc DUPRÉ, membre

DÉDICACE

À ma famille et mes amis

REMERCIEMENTS

Je tiens tout d'abord à remercier très sincèrement les professeurs Dr. Hanane Dagdougui et Dr. Martin Tréapnier pour leur soutien continu, leur patience et leurs précieux conseils.

Je tiens aussi à remercier Mr. Karim Er-Rafia de Jalon MTL pour son support et sa disponibilité dans l'accompagnement du projet.

Que Hydro-Québec trouve ici ma sincère reconnaissance pour avoir mis à ma disposition les données sur l'utilisation des bornes de recharge au Québec.

Je tiens à remercier les membres du jury qui m'ont fait l'honneur d'évaluer mon travail.

Mes remerciements vont également à mes parents, ma sœur et mon beau-frère qui m'ont toujours soutenu, encouragé et aidé. Enfin, merci à mes camarades de Polytechnique Montréal, particulièrement Amine Bellahsen et mes amis d'enfance, qui m'ont guidé dans mes débuts et m'ont encouragé tout au long de la complétion de mon mémoire.

RÉSUMÉ

Le réchauffement climatique représentant un danger important pour la protection des générations à venir, plusieurs initiatives ont été lancées pour réduire l'émission de gaz à effet de serre. L'une de ces initiatives a été la promotion de l'adoption des véhicules éclectiques (VEs) par la population puisque le transport terrestre représente plus de 14% des émissions de CO_2 dans le monde. L'introduction et le développement rapide des véhicules électriques a cependant introduit plusieurs défis. Parmi ces défis figure le problème du déploiement et de gestion des bornes de recharge de véhicules électriques (BRVEs) dans nos routes. En effet, les BRVEs représentent non seulement une barrière pour la vente exponentielle de VEs mais aussi un aspect important à considérer dans la gestion du réseau électrique. En premier lieu, l'angoisse liée à l'autonomie des VEs étant une préoccupation majeure chez les propriétaires de VEs, le manque ou l'expansion inefficace des BRVEs pourrait mener à un ralentissement des ventes de VEs. Cette angoisse est non seulement liée à la crainte de ne pas trouver de BRVEs à proximité où recharger, mais aussi qu'elles soient toutes occupé ou que l'attente soit trop longue pour qu'une devienne disponible. En deuxième lieu, les BRVEs présentent un défi majeur chez les opérateurs qui doivent assurer une demande constante pour maintenir un profit étant donné leur investissement important lors de l'installation et la maintenance des BRVEs. Enfin, les distributeurs d'électricité doivent prendre en compte l'utilisation des BRVEs installées et potentielles pour non seulement mieux planifier l'expansion et l'investissement sur le réseau électrique mais aussi pour prévenir des charges de pointe. Plusieurs études ont déjà été menées pour non seulement optimiser la planification des BRVEs mais aussi leur opération. Peu d'articles ont cependant exploité les données pour développer leur solution en raison du manque de base de données libres d'accès. Cette recherche se sert donc de données récupérées d'Hydro-Québec pour développer une approche à deux étapes qui a pour but de non seulement analyser l'utilisation des BRVEs installées mais aussi d'étudier les facteurs externes pertinents qui affectent l'utilisation des BRVEs. Cette approche vise principalement à conseiller les opérateurs sur l'utilisation des bornes déjà installées et potentielles.

L'approche utilisée se démarque d'autres articles par sa structure et ses méthodes employées. En premier lieu, chaque BRVE est représentée par tous ses profils de connexion journalier datant de 2019 à 2020. Ces profils de connexion sont représentatifs de l'utilisation des BRVEs dans une période de 24 heures où une valeur de 1 est représentative d'une utilisation de la borne et 0, de la disponibilité de la borne. Ensuite, les profils de connexion journalier de chaque borne sont regroupés grâce à un algorithme de regroupement hiérarchique en utilisant

la mesure de distance Dynamic Time Warping (DTW). Le regroupement est exécuté de façon à préserver la granularité des données et donc de sorte à généraliser le moins possible les profils de connexion des bornes. Cette étape sert principalement à étudier la diversité des profils de connexion des différentes stations ainsi que de réduire le nombre de profils de connexion à étudier pour la prochaine étape. En effet, les centroïdes générés par cette étape sont ensuite utilisés pour pouvoir comparer les profils de connexion de toutes les BRVEs. Pour compléter cette étape, un regroupement hiérarchique utilisant le DTW est utilisé pour regrouper les centroïdes de manière à produire le moins de regroupements possible. Cette étape permet non seulement de comparer les stations mais aussi d'offrir une analyse complète sur les profils de connexion types observés dans les BRVEs de l'île de Montréal. Enfin, ces résultats sont utilisés pour le développement d'un modèle KNN de classification permettant de comprendre la pertinence d'attributs sur l'utilisation des BRVEs à partir d'une analyse SHAP. Plus précisément, des attributs tels que les points d'intérêts avoisinant les BRVEs et le profil socio-économique de la population autour des BRVEs sont évaluées.

Plusieurs conclusions ont pu être tirées de ces étapes. En premier lieu, le regroupement des profils de connexion de chaque borne a permis de mieux comprendre la diversité de l'usage des bornes. D'un côté, l'hypothèse que l'âge de la borne pourrait avoir un impact sur la variété des profils a été rejeté, cependant il existe une légère corrélation entre le nombre de jour où la borne a été utilisée et la variété de ses profils de connexion. En deuxième lieu, l'étude des profils de connexion de toutes les bornes a présenté plusieurs résultats sur le profil de connexion général des différentes bornes. D'un côté, plus de 41% des profils de toutes les bornes considérées représente une utilisation nulle durant toute la journée. D'un autre côté, les autres profils généraux de connexion journaliers les plus populaires sont ceux pour lesquels l'utilisation est répandue durant la journée (entre 8h et 15h), et l'après-midi et le soir. Une étude plus accrue sur les proportions des profils généraux de connexion des différentes bornes a ensuite révélé que certains profils de connexion sont plus populaires dans certaines régions, tels que l'aéroport, que d'autres. Enfin, le modèle de classification a démontré que les attributs les plus important pour l'explication des profils de connexion d'une borne sont la proximité de la borne à des services financiers, des stations de métro ou de location de bicyclette, des endroits religieux et des endroits de divertissement et d'arts, parmi plusieurs autres attributs. L'analyse SHAP a aussi démontré que l'importance des attributs diffère selon le profil d'usage de la borne.

ABSTRACT

Concerns over the increasing levels of CO_2 emissions has led countries to meet every year for the Conference of the Parties (COP) to discuss problems and solutions to save our planet. While agreements have been signed to push countries to reduce their carbon footprint such as the Paris agreement, countries' approach to abide by them have differed. A similar approach followed by many has however been to encourage the adoption of Electric Vehicles (EVs) by their population as the transportation sector represents 14% of the global greenhouse emissions. While in many countries, government incentives, technological advancement, and the growing offer of EVs have led to a spike in EV sales, their growing adoption has introduced many challenges that need to be tackled. Of the many challenges is the efficient planning and operation of Electric Vehicle Charging Stations (EVCSs). Indeed, the inconsideration of the infrastructure for EV charging represents not only a barrier to the consistent growth of EV sales but also a concern for electricity providers and EVCS operators. On one hand, lack of an optimal deployment of EVCSs in the roads can lead to a spike in EV users' range anxiety and fear of finding no available EVCS or having to wait long hours for one to become available. This in turn can dissuade many from switching to the use of an EV. On the other hand, operators need to be aware of the demand of installed and potential EVCSs to ensure profit from their installation and maintenance investments. Similarly, electricity providers' access to this information can be crucial for better investment on the power grid and to prevent peak loads through services like vehicle-to-grid (V2G). Many articles have already been presented in the literature covering these challenges and offering different approaches to tackle them. However, only a few have used historical data to better represent the demand for EVCSs, whereas many have sought to simulate the demand through vehicle traffic. The present research seeks to use historical data on the use of EVCSs in the city of Montreal to better understand the demand and offer more reliable insight on their use. This data is leveraged in a two-stage approach used to not only analyze the use of EVCSs but also understand external factors most relevant in predicting their demand profiles. The results are then used to better advise operators and electricity providers on the demand profile of currently installed EVCSs and of different potential locations.

The two-stage approach sets itself apart from other articles not just by the analysis done but also by the methods it uses. First, each EVCS is represented by a set of its daily connection profiles from all its ports' data between 2019 to 2020. These are 24-hour sequences for which a 1 is set to a specific hour if a vehicle is connected to the EVCS and 0 otherwise. This data is then first used in a hierarchical clustering algorithm using the Dynamic Time

Warping (DTW) distance to not only better understand how diverse the connection profiles of the different EVCSs are but also to reduce the number of profiles analyzed in the next step. Using the centroids produced in the previous section, another hierarchical clustering algorithm using the DTW distance is executed to compare the profiles of the different EVCSs. Finally, a KNN classification model is used to predict the profiles of the different EVCSs given input data such as points of interests at most 500 meters away from the station and socio-economic data on the population living close to the EVCS.

Various conclusions were made through the different steps followed. First, EVCSs' connection profiles are very diverse leading to the idea that their use is very stochastic and can be thought of as hard to predict. This observation was made through the clustering of individual profiles which resulted in a very high number of optimal clusters in certain cases. Moreover, clustering profiles all together showed that there exists a high proportion (41%) of daily connection profiles for which no use is recorded during the day. Other clusters on the other hand, particularly those for which a connection is recorded between 8am and 3pm (19%), and in the afternoon and at night have very high proportions amongst other profiles. Associating the different clusters to EVCSs' daily profiles showed that unpopular clusters were prominent in isolated stations located in the airport for instance. Finally, the SHAP analysis completed on the k-NN classification model showed that points of interests were most relevant in predicting the connection profiles of and EVCS. More specifically, locations offering financial, transportation and educational services as well as places of worship and arts and entertainment venues had the greatest impact on the final model. The SHAP analysis however also showed that attributes' relevance differed depending on the connection profile of the EVCS.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

BRVE	Borne de Recharge de Véhicules Électrique
EVCS	*Electric Vehicle Charging Station*
VE	Vehicule Electrique
EV	*Electric Vehicle*
V2G	*Vehicule-to-Grid*
G2V	*Grid-to-Vehicule*
SIG	Système d'Information Géographique
PoI	Points d'intérêt
API	Interfaces de Programmation d'Applications
DTW	*Dynamic Time Warping*
ED	Distance Euclidienne
PCA	*Principal Component Analysis*
GMM	*Gaussian Mixture Model*
KNN	*K-Nearest Neighbors*
SHAP	*SHapley Additive exPlanations*
PAM	*Partition Around Medoids*
DBA	*DTW Barycenter Averaging*
CVI	*Cluster Validation Indices*

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Le réchauffement climatique est un problème majeur qui est constamment évoqué par de nombreux scientifiques à mesure qu'il devient de plus en plus dur de le limiter. Depuis la révolution industrielle, le niveau de dioxyde de carbone dans l'atmosphère a augmenté de manière alarmante, dépassant les niveaux les plus élevés vus dans le passé. Aujourd'hui, il existe plus de 410 parties par million de CO2 dans l'atmosphère, une augmentation de 110 (37%) parties par million du niveau le plus haut jamais enregistré dans l'histoire, il y a plus de 300 000 ans [1]. Cette augmentation est due notamment à notre dépendance croissante aux produits industriels dans tous les secteurs et d'une société de plus en plus matérialiste [2]. D'une météo de plus en plus instable provoquant diverses catastrophes naturelles à une population de plus en plus vulnérable aux maladies et à la malnutrition [3], les conséquences sont désastreuses. En effet, malgré les projections climatiques futures, de nombreux chiffres alarmants tels que la hausse des températures ont déjà étaient enregistrés dans certaines parties du globe [1]. Pour contrer ces effets, plusieurs mesures ont été prises. À travers les Nations Unies par exemple, plusieurs conférences et organismes ont été créés comme la Conférence des Nations Unies sur les Changements climatiques (COP) pour discuter du problème et des mesures à entreprendre par les différents pays pour limiter les conséquences du réchauffement climatique. Ces conférences ont donné naissance à de nombreuses ententes telles que l'accord de Paris en 2015 qui a pour objectif d'inciter les pays du monde entier à réduire leurs émissions de gaz à effet de serre pour réduire le réchauffement climatique à 1,5 degré Celsius [4]. Par conséquent, plusieurs pays comme le Canada, l'Angleterre, le Maroc, les États-Unis et les pays de l'Union européenne ont optés pour des investissements accrus dans le domaine des énergies renouvelables pour la production d'électricité qui fut majoritairement générée à partir de charbon. Pour accélérer la transition énergétique, les entreprises compte à elle se penche aussi sur des opérations plus durables en utilisant des matériaux moins polluants, en mieux consommant l'énergie ou en intégrant des ressources renouvelables décentralisées dans leurs bureaux et usines parmi d'autres stratégies.

Bien que le changement soit entraîné par les entreprises et les gouvernements, il est aussi important d'inciter la population à changer ses habitudes pour contribuer à la réduction de l'empreinte carbone du pays. De l'installation de panneaux solaires dans le toit pour le chauffage de l'eau ou la production d'électricité au recyclage et contrôle de l'utilisation d'appareils à forte consommation, il existe plusieurs solutions. L'une des plus importantes reste cependant de repenser nos habitudes de transport. Comme il est conseillé d'utiliser le co-voiturage ou les bus plus souvent, il est aussi fortement conseillé de favoriser l'achat

de véhicules électriques (VEs) plutôt que les véhicules à combustion, puisque le secteur du transport terrestre représente plus de 14% des émissions de gaz à effet de serre dans le monde [5].

Les Véhicules électriques (VEs) ont connu une importante croissance en popularité durant les dernières années. En effet, les ventes des VEs dans le monde ont connu une croissance exponentielle depuis 2010, allant de 0,02 million de véhicules vendus durant cette année à plus de 1,18 million vendus en 2016 et 4,79 millions vendus en 2019 [6]. Au Québec, les ventes de VEs ont représenté plus de 48% des ventes totales dans le Canada pour plus de 17 000 unités vendues en 2020 [7]. Ces chiffres peuvent être attribués à plusieurs facteurs, en particulier aux mesures prises par plusieurs gouvernements pour inciter la population à acquérir des VEs et à l'augmentation de l'offre de VEs dans le marché.

En effet, des pays tels que le Canada, les États-Unis d'Amérique ou même les pays scandinaves, parmi d'autre, proposent des programmes incitatifs qui réduisent le coût d'achat de ces véhicules, les rendant plus abordables pour certains clients potentiels [8]. Au Québec, ces rabais peuvent atteindre 8 000 $CAD pour certains véhicules pour lesquels le prix suggéré du fabricant est inférieur à 60 000 $CAD [9]. D'autre part, les constructeurs automobiles continuent d'innover dans ce secteur où certains des plus réputés, tel que Volkswagen AG, Mercedes-Daimler et Jaguar et Land Rover, se sont joints à Tesla pour proposer des alternatifs aux véhicules à combustion qui sont 100% électrique en plus de leurs offres hybride. En 2020, il existait plus de 350 modèles de véhicules électriques proposés dans le marché mondial, une augmentation importante de moins de 100 que ceux proposés en 2015 [8]. Ces chiffres continuent d'augmenter à mesure que les constructeurs introduisent des choix de plus en plus variés de VEs répondant aux attentes de tous les types de consommateurs. Malgré ces avancées technologiques, le développement rapide de cette filière à tout de même introduit de nombreux défis. Parmi ces défis, le problème lié au développement de l'infrastructure est abordé dans cette recherche. Plus précisément, une étude de l'utilisation des Bornes de Recharges de Véhicules électriques (BRVEs) est réalisée pour mieux comprendre le comportement des utilisateurs dans différentes stations. La section 4.1 offre plus de détails sur la situation des VEs dans le monde et au Québec.

Les BRVEs représentent l'un des plus grands défis dans l'augmentation de taux de pénétration des VEs, ainsi qu'à l'instauration des flottes de VEs dans le monde. Alors que les propriétaires de VEs sont généralement fournis d'un système de recharge de leurs véhicules à l'achat, il reste important d'assurer l'offre d'un service public. D'un point de vue technique, l'autonomie et le temps de recharge constituent les principales barrières à la pénétration de véhicules électriques dans le marché [10]. Assurer un réseau de BRVEs permettrait aux usagers d'utiliser leurs

véhicules plus souvent sans avoir à s'inquiéter de l'autonomie du véhicule. D'un point de vue psychologique, [11] explique que la présence d'un réseau de BRVEs important va contribuer à améliorer l'index 'comfortable range' qui se traduit par l'autonomie de confort et donc diminuerait l'anxiété associée à l'autonomie.

Aujourd'hui alors, de plus en plus de villes investissent dans le déploiement de ce service au sein d'environnements interurbains et intra-urbains sous la forme de bornes de recharge rapides et standards. Cet investissement présente cependant de nombreuses contraintes et défis allant au-delà du confort des usagers et doit donc être entamé de manière réfléchi. En effet, bien que le confort des usagers soit important, il faut aussi prendre en considération les répercussions de tels investissements sur les opérateurs, notamment sur le besoin de garantir un profit positif et de protéger le réseau électrique des périodes de pointe de demande de puissance.

Pour cela, plusieurs recherches présentent des solutions visant à optimiser d'une part le déploiement optimal des BRVEs et d'une autre part, la gestion de l'opération des BRVEs tenant compte de diverses variables et de différentes approches. Un aspect important considéré dans ces recherches est la définition de la demande pour les différents BRVEs. Ceci dit, peu d'articles exploitent les données massives qui deviennent de plus en plus répandues pour définir la demande des BRVEs existantes et potentielles. Ces articles utilisent ces données pour non seulement expliquer la popularité des BRVEs, mais aussi le comportement des usagers. De même, cette recherche utilise des données sur l'utilisation des BRVEs dans l'île de Montréal et ceux de points d'intérêts et de recensement pour offrir une analyse détaillée sur l'utilisation des bornes à Montréal. Plus précisément, l'étude est complétée selon deux axes principaux :

1. Analyse de l'utilisation des BRVEs en considérant les temps d'arrivée et de départ ainsi que le temps de connexion des usagers à travers une approche utilisant la distance DTW (Dynamic Time Warping) appliquée à un algorithme de regroupement hiérarchique des données.
2. Évaluation de l'impact des facteurs externes tels que les points d'intérêts et les données de recensement sur l'utilisation des bornes à partir d'un modèle de classification K-Nearest Neighbors.

Pour réitérer l'information présentée dans la section 4.1 détaillant les objectifs de la recherche, les résultats présentés offrent plusieurs conclusions pour améliorer la planification et l'opération du réseau des BRVEs.

Le premier axe permet non seulement de définir les profils type de chaque station évaluée, mais aussi d'offrir une analyse comparative de l'utilisation des différentes stations de l'île de Montréal. Cet aspect est important pour mieux identifier les BRVEs ayant des profils d'usage

similaires et permet donc d'exécuter le deuxième axe.

Le deuxième axe permet de mieux identifier les différences d'utilisation des BRVEs selon leurs situations géographiques sur l'île de Montréal. Cette analyse est primordiale pour la planification court-terme et long-terme des BRVEs dans le contexte urbain. En effet, elle offre une évaluation de l'impact des différents facteurs sur le profil d'usage des BRVEs. Cette approche permettra aux opérateurs d'être conscients de l'usage type attendu de certaines stations, et par conséquent d'optimiser les coûts d'investissement et de connexion au réseau électrique.

Enfin, les modèles de cette étude peuvent potentiellement être appliqués en outre dans d'autres villes.

CHAPITRE 2 REVUE DE LITTÉRATURE

Avant d'aborder et de développer la méthodologie de l'étude, une revue de la littérature a été réalisée en prenant en compte plusieurs aspects importants abordés dans la recherche. Pour cela, un résumé est tout d'abord présenté sur les études traitant la planification des BRVEs avant de traiter ceux s'intéressant sur leurs opérations.

En outre, ces analyses sont détaillées sous deux différents angles. Dans un premier lieu, plusieurs articles détaillent leurs solutions à partir de modèles d'optimisation. Tandis que plus récemment, plusieurs auteurs se sont penchés sur l'utilisation de données historiques. Ces données sont ensuite appliquées à différentes approches supervisées et non-supervisées. D'un côté, plusieurs articles suivent l'approche supervisée pour le développement de modèles de classification ou de régression pour prédire la popularité ou l'utilisation des BRVEs parmi d'autres attributs. D'un autre côté, certains exploitent des méthodes d'apprentissage non-supervisé, notamment des méthodes de regroupement pour regrouper les utilisateurs ou les stations parmi d'autres aspects selon divers attributs.

2.1 Problème de Planification

Le problème de planification des BRVEs a beaucoup été étudié au cours des dernières années avec un intérêt croissant pour plusieurs différentes approches. La plus populaire a été celle traitant le problème à travers divers modèles d'optimisation pour lesquels la demande des BRVEs est estimée sur base de la circulation routière, les points, ou les cellules ou polygones d'une carte géographique. Chaque modèle prend en compte différentes contraintes et paramètres dans la sélection des emplacements les plus appropriés. Toutefois, plus récemment, l'étude d'indicateurs de popularité des BRVEs à travers des approches de prédiction a également suscité un intérêt croissant pour remédier au manque de données historique dans les modèles d'optimisation présentée.

La section 2.1.1 présente les articles traitant ce défi à travers différents modèles d'optimisation tandis que la section 2.1.2 s'intéresse sur la résolution du problème à travers les méthodes d'apprentissage statistique.

2.1.1 Modèles d'optimisation

Malgré sa complexité, l'approche se basant sur la circulation routière est largement utilisée dans la littérature. Elle considère le comportement des conducteurs potentiels en considérant

également l'autonomie des véhicules pour définir un ensemble d'emplacements stratégiques pour les futurs BRVEs. Dans les environnements interurbains, de nombreux articles utilisent cette approche pour le dimensionnement et la planification des BRVEs [12, 13]. À propos de l'environnement intra-urbain, [14] prennent en compte l'autonomie, le choix de l'itinéraire et le temps de recharge, parmi d'autres variables, pour suggérer des emplacements de BRVEs optimaux au sein d'un réseau d'itinéraire simplifié. L'approche utilise un modèle à deux niveaux qui est reformulé comme un programme à un seul niveau pour le rendre plus efficace. D'autre part, [15] maximise les activités des conducteurs de VEs en plaçant de manière optimale les BRVEs, prenant en compte l'autonomie des véhicules. Il utilise des données sur le comportement de conduite des habitants de Pékin pour comprendre leurs activités et créer ainsi un modèle d'optimisation résolu grâce aux algorithmes génétiques. La solution est cependant limitée à recommander des régions idéales plutôt que des emplacements spécifiques. Enfin, [16] évalue les aspects stochastiques et dynamiques de la demande de recharge des VEs pour résoudre un modèle de programmation stochastique en nombres entiers à phases multiples. Une séquence de scénarios est construite pour représenter l'incertitude du flux de VEs. Malgré de bon résultats présentés par ces articles, la représentation du problème de planification à partir de données sur la circulation dispose de nombreuses contraintes. En effet, la simulation de la demande à partir de données sur la circulation peut être plus complexe à représenter efficacement en milieu urbain qu'en milieu interurbain, notamment en raison de la complexité de la représentation du réseau routier intra-urbain [17]. Une autre limitation découlant de cette approche est le manque de prise en compte de la demande réelle des BRVEs, menant à des résultats potentiellement biaisés. Il existe cependant d'autres approches prenant en considération certaines contraintes.

Récemment, un intérêt a été porté sur l'utilisation d'une approche basée sur les cellules qui vise à diviser la carte en plusieurs polygones représentant les emplacements potentiels de futur BRVEs. [18] utilise par exemple un problème linéaire avec nombres entiers à bi-critères pour minimiser le coût et maximiser la qualité du service. Les auteurs considèrent les données sur les trajectoires GPS de milliers de véhicules pour construire le modèle. Cette solution permet de mieux comprendre la portée et la distribution des BRVEs sur la carte en considérant celles déjà installées. D'autre part, [19] propose un modèle de somme pondérée à deux niveaux en analysant des aspects macros et micros. Au niveau macro, une analyse de la demande des BRVEs est effectuée au niveau territorial. Cette demande est estimée à partir de données sur le flux de véhicules locaux et touristiques. Au niveau micro, le terrain est subdivisé en hexagones résumant la présence de services comme le stationnement, les bureaux et les supermarchés pour représenter la demande potentielle des BRVEs. Les auteurs concluent que les emplacements très peuplés à proximité de plusieurs services et de places de stationnements

sont les mieux adaptés pour installer des BRVEs.

Enfin, parmi les approches les plus simples et les plus considérées est la méthode de modélisation de la demande basée sur des points géographiques. Cette approche vise à maximiser la demande dans différents emplacements spécifiques dans la carte. Elle ouvre essentiellement la voie à l'utilisation d'un plus large éventail de variables. Les auteurs de [20] utilisent des données sur les points d'intérêt (PoI), des informations de recensement sur la population dans chaque zone, la circulation automobile et la popularité des zones en fonction de l'afflux de personnes et des informations sur les réseaux sociaux pour localiser de manière optimale les BRVEs à partir d'un algorithme génétique. Un algorithme de croisement impliquant des pavages de Voronoi et une triangulation de Delaunay est également utilisé pour représenter les emplacements de BRVEs potentiels sur la carte. De nombreux articles envisagent également une perspective hybride dans laquelle la circulation et la demande pour les BRVEs sont pris en compte. Dans [21], le problème de planification est présenté à travers une architecture à deux couches prenant en considération la circulation, la sécurité du réseau électrique, la satisfaction des conducteurs de VEs et l'intérêt économique des opérateurs. Tandis que [22] utilise un modèle de programme linéaire avec nombres entiers basé sur le système d'information géographique (SIG) pour choisir les points optimaux qui maximiseraient les profits. Comme beaucoup d'articles, les auteurs utilisent la circulation routière pour estimer la demande des BRVEs et l'affectation du sol pour définir la pertinence d'un emplacement.

Plusieurs contraintes sont toutefois présentes dans les articles susmentionnés. En premier lieu, l'absence dans la littérature de données historiques sur l'utilisation des BRVEs pourrait limiter la validité de la demande des BRVEs qui est estimée à partir de facteurs externes associés aux activités de la population générale. Bien que cela soit approprié pour comprendre la demande potentielle à long terme des BRVEs une fois l'adoption des VE répandue [18], cela n'est pas représentatif de la demande actuelle des propriétaires de VEs et potentielle au cours terme. De plus, contraindre les emplacements potentiels sur la base des profits des opérateurs peut limiter la planification stratégique de la solution. Des méthodes de prédiction ont donc été développées pour remédier à certaines contraintes en prenant en compte des données historiques détaillées dans la section qui suit.

2.1.2 Apprentissage Statistique

Ces dernières années, plusieurs études ont valorisé les données historiques sur les BRVEs en étudiant l'importance de facteurs externes sur leur utilisation. Ces facteurs servent principalement comme outil d'aide à la décision pour les opérateurs sur les régions où l'installation de nouvelles stations est la plus appropriée. [23] évalue par exemple la popularité d'une borne

de recharge sur la base de données économiques, sociales et géographiques. Les auteurs utilisent des ensembles de données SIG pour construire leurs données d'entrées et des données sur l'utilisation des bornes pour définir la sortie. Plusieurs données d'entrées sont considérées comme l'utilisation des terres, l'énergie utilisée dans la zone, l'indice d'habitabilité de la zone, la circulation, la densité de population et les points d'intérêt. La popularité des bornes de recharge d'autre part, est représentée par le nombre unique d'utilisateurs qui ont utilisé les bornes de recharge en 2015 et est mesurée sans prendre en compte le nombre de bornes par stations. Ces données sont extraites de la base de données d'EVnetNL qui est un réseau de BRVEs installées en Hollande. L'article compare en outre la précision de trois méthodes d'apprentissage automatique pour la prédiction de la popularité des EVCS. En évaluant les résultats, les auteurs concluent que les attributs les plus importants sont l'affectation du sol ainsi que le nombre de magasins, de restaurants et de centres sportifs. En utilisant la même base de données, [24] étudie les activités, les fonctions et les caractéristiques de l'environnement qui ont le plus important impact sur la distribution de la consommation en énergie des différents BRVEs. En premier lieu, la consommation en énergie des différents BRVEs est modélisée à partir de différentes distributions tels que Weibull, Beta et Gamma avant d'utiliser un modèle Lasso pour évaluer l'importance de facteurs externes sur la consommation des différentes BRVEs. Ces facteurs externes incluent les données sur la circulation routière, les points d'intérêts, l'affectation du sol, la densité de la population durant la journée et l'index d'habitabilité. L'article conclut que le nombre de sessions d'opération des BRVEs est l'attribut le plus important dans la prédiction de l'énergie consommée par BRVE. De plus, les BRVEs situées dans des zones où les résidents ou les commerces ont un revenu plus élevé et où les maisons sont récemment construites sont plus susceptibles d'avoir une consommation d'énergie plus élevée.

Malgré l'utilisation de données historiques, un aspect important qui n'est pas considéré dans les articles susmentionnés est l'aspect temporel de l'utilisation des BRVEs. En effet, les valeurs prédites à partir des modèles sont des moyennes estimant le comportement général des usagers. Or, il est aussi important de considérer les périodes d'utilisation de BRVEs pour répondre à diverses préoccupations des opérateurs tels que les périodes à forte utilisation.

2.2 Opération

Étant donnée la complexité du comportement d'usage des BRVEs, plusieurs articles de recherche s'intéressent sur le comportement de leurs utilisateurs. Dans le cas du développement de modèles d'optimisation, cet aspect renferme généralement le contrôle de la recharge des VEs pour assurer un service V2G et G2V. Tandis que les modèles d'apprentissage statistiques

sont généralement utilisés pour analyser les comportements des usagers ou des BRVEs pour en déduire des observations et conseiller les opérateurs à exécuter certaines décisions plus variées pouvant aussi porter sur la planification des BRVEs.

2.2.1 Modèles d'optimisation

Les BRVEs présentent un potentiel important dans la distribution d'électricité, particulièrement dans des zones susceptibles de subir des charges de pointe. Pour cette raison, plusieurs articles utilisent des modèles d'optimisation pour évaluer la possibilité de contrôler la recharge de VEs à travers un service Vehicle-to-Grid et Grid-to-Vehicle. [25] développe un algorithme de planification de la recharge intelligente (ISCA) à l'aide du 'Henry gas solubility optimization', qui est un algorithme méta-heuristique récemment introduit s'inspirant de la physique [26]. Plusieurs variables sont considérées comme l'état de charge à l'instant d'arrivée et de départ du véhicule ainsi que la disponibilité de places de stationnement parmi d'autres.

Les simulations démontrent que le modèle développé est robuste, il propose une alternance entre les modes V2G, G2V et inactif nettement plus profitable qu'un service s'appuyant seulement sur le G2V. De la même manière, [27] utilise l'état de charge parmi d'autres variables et paramètres pour développer un modèle d'optimisation visant à prévenir une charge de pointe en alternant entre le transfert d'énergie du réseau au véhicule et vice-versa. Le résultat à conduit à la réduction de la variance entre les charges creuses et de pointe de 5 MW à 1,5 MW tout en optimisant la recharge des VEs.

2.2.2 Apprentissage Statistique

Contrairement au problème de planification, de nombreux articles utilisent des modèles d'apprentissage non-supervisés pour évaluer l'utilisation des BRVEs, notamment les modèles de regroupement. [28] utilise le K-Means et la distance euclidienne (ED) pour regrouper les propriétaires des VEs sur la base du temps moyen passé à recharger leurs véhicules et à être connectés à la borne ainsi que l'écart type de ces valeurs pour résoudre le problème de la planification centralisée du réseau. Les regroupements sont ensuite utilisés pour classer les nouveaux utilisateurs dans le regroupement approprié à l'aide d'un modèle k-Nearest Neighbors. Ce classement permet aux opérateurs du réseau électrique d'intégrer les technologies V2G et G2V de manière efficace au réseau des BRVEs. De même, les auteurs de [29] regroupent différentes caractéristiques du comportement des utilisateurs tels que le temps de charge, les heures d'arrivée et de départ et les heures entre les charges à partir du Gaussian Mixture Model (GMM). L'article conclut l'analyse en offrant un résumé des différents

groupes d'utilisateurs et de leur comportement de charge. Dans [30], les auteurs utilisent le regroupement pour définir les profils de demande de puissance journalière de la recharge des VEs au Royaume-Uni. Les partitions sont générées en utilisant un modèle K-Means pour lequel le nombre de k regroupement est validé à partir de l'index Davies-Bouldin. Les regroupements sont ensuite utilisés pour définir l'impact de la demande d'énergie des VEs sur le réseau électrique. [31] évalue l'état de charge (SoC) des véhicules électriques pour regrouper les conducteurs ayant un comportement similaire. Les auteurs utilisent également la distance euclidienne et le regroupement hiérarchique pour définir les différents groupes, en précisant l'incapacité du K-Means de traiter efficacement les valeurs aberrantes. Les regroupements résultants sont ensuite utilisés pour améliorer la précision des modèles prédictifs utilisés pour la prévision du comportement des utilisateurs les jours à venir. Différentes mesures de distance, en particulier la distance euclidienne (ED), la distance euclidienne modifiée (MED) et le Dynamic Time Wrapping (DTW), sont comparées pour regrouper la queue de la courbe représentant le courant des différentes sessions de recharge dans [32]. La précision de la distance euclidienne modifiée s'avère cependant supérieure à celle des autres mesures. Enfin, les auteurs de [33] et [34] passent en revue les différentes approches utilisées dans la littérature pour la prévision et le regroupement des données des bornes de recharge. [33] détaille plusieurs articles qui utilisent des modèles de regroupement pour définir des groupes d'éléments, principalement ceux représentant les utilisateurs, en fonction de leurs profils d'usage des BRVEs. L'article met l'accent sur l'importance d'utiliser les modèles de regroupement dans le problème de planification des BRVEs et de la partition des BRVEs ayant une utilisation similaire. [34] présente également différents articles appliquant le regroupement d'utilisateurs des BRVEs en considérant leur comportement de recharge à l'aide de méthodes telles que le GMM, K-Means et Kernel Density Estimators (KDE).

Cependant, plusieurs limitations peuvent être relevées des études susmentionnées. La majorité des travaux se sont par exemple concentrés sur le regroupement du comportement de charge des utilisateurs. Hors, le manque de considération des profils d'utilisation ou de consommation d'énergie des BRVEs restreint l'implication de l'utilisation des BRVEs pour la planification opérationnelle du réseau électrique. De plus, les attributs, comme le comportement de recharge des utilisateurs, sont fréquemment généralisés, limitant l'interprétabilité des résultats comme dans [28]. Enfin, Alors que la distance euclidienne est largement utilisée et appropriée pour le regroupement de données individuelles, elle reste nettement moins fiable pour l'analyse de séries chronologiques, car elle ne parvient pas à saisir les aspects temporels des données [35]. Cet aspect est cependant extrêmement essentiel, compte tenu notamment du caractère aléatoire et bruyant des séries temporelles évaluées dans cette recherche.

CHAPITRE 3 DÉMARCHE SUR L'ENSEMBLE DU TRAVAIL DE RECHERCHE ET ORGANISATION GÉNÉRALE DU DOCUMENT

Le développement d'une solution à la problématique a nécessité diverses étapes clés. En premier lieu, les données sur l'utilisation et la situation géographique des BRVEs ont été collectées et analysées. Ensuite, plusieurs méthodes ont été utilisées et comparées pour le développement de solutions pour les deux axes de la recherche, à savoir le partitionnement des données temporelles et l'analyse d'influence des facteurs externes sur le profil de connexion des BRVEs. La figure 3.1 résume la séquence suivie tandis que les sous-sections qui suivent détaillent tout d'abord les données utilisées avant de s'attarder sur les algorithmes considérés pour le développement de la solution. Puis, la section 4 présente l'article soumis au journal international 'Trasportation Research Part C : Emerging Technologies', tandis que la section 6 conclut le document en présentant un résumé de la recherche et des résultats présentés dans l'article.

3.1 Collecte et analyse de données

Plusieurs données ont dû être collectées pour la réalisation de ce projet. D'une part, les données se rapportant aux BRVEs ont servi au développement de modèles nous permettant de mieux comprendre leurs utilisations. D'autre part, des données socio-économiques et géographiques nous ont permis de mettre en relation l'utilisation des bornes à leur situation géographique. Différentes méthodes et sources ont été utilisées pour la collecte de ces données avant d'être incorporées à une base de données MySQL. Les sections qui suivent commencent par introduire la structure de la base de données utilisée avant de présenter les entités la définissant.

3.1.1 Développement de la base de données

Le développement de la base de données a été une étape primordiale pour optimiser l'extraction de données requises pour les différentes analyses. En effet, en raison de l'usage de données variées de différentes sources, le développement d'une base de données relationnelles a permis de connecter tous les attributs comme indiqué dans le modèle entité-association de l'annexe A. Pour réaliser cette étape, deux programmes écrits en Java à l'aide des librairies *Hibernate* et *Spring Framework* ainsi qu'une base de données MySQL ont été créés. Un programme Python a ensuite était développé à l'aide de la libraire *SQLAlchemy* pour permettre

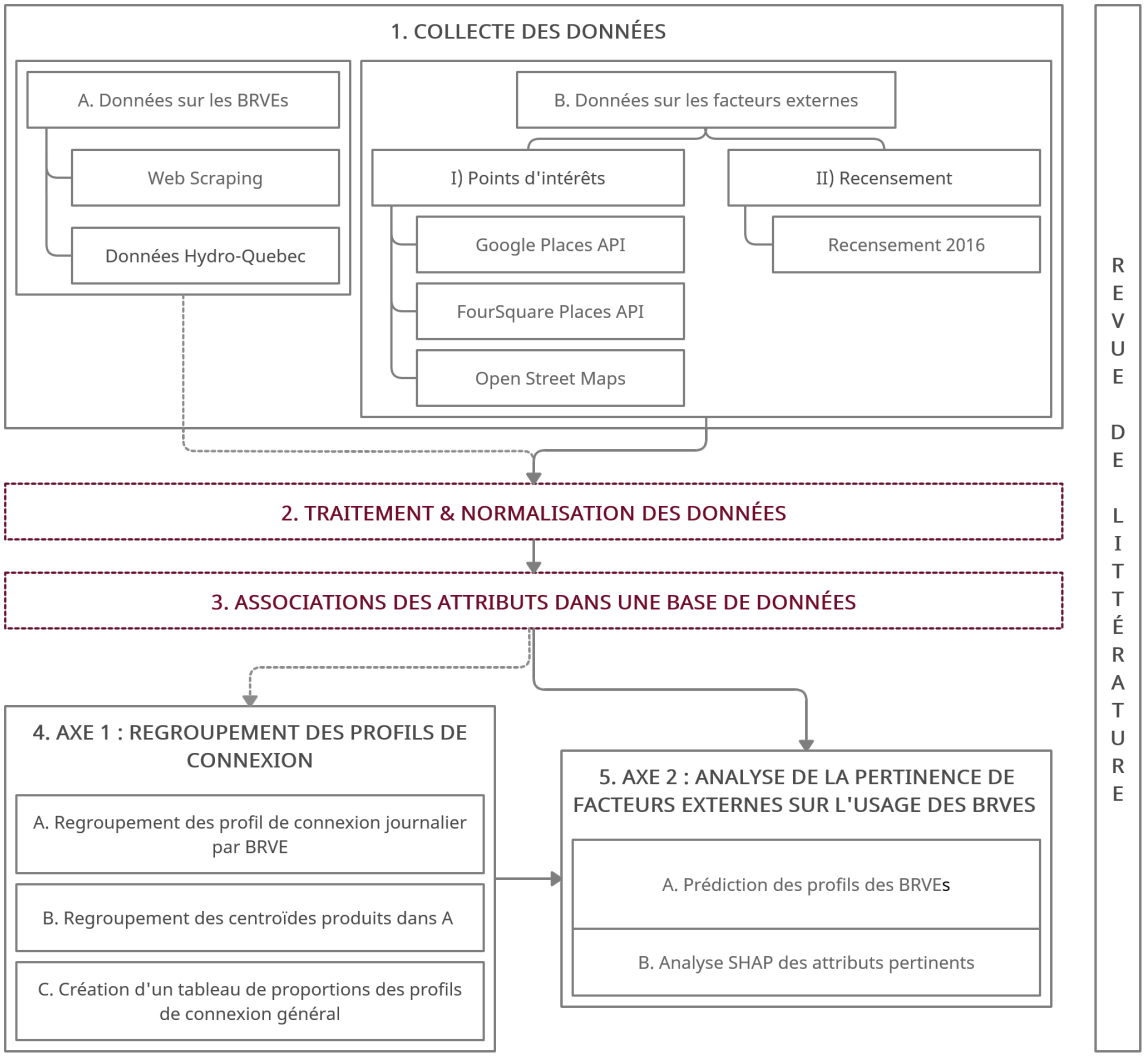


Figure 3.1 Méthodologie suivie pour la complétion de la recherche.

l'extraction des données pour toute analyse subséquente. Les sections qui suivent détaillent le traitement des données et leur incorporation à la base de données à l'aide de programmes écrits en Java.

3.1.2 Données des BRVEs

Les données sur l'utilisation des BRVEs sont très limitées. Il existe cependant un répertoire grandissant de base de données libre d'accès permettant de conduire des recherches sur l'utilisation des stations. Bien que dans la revue de littérature, certains articles se penchent sur l'utilisation de service comme EVnetNL pour la capture de données sur le comportement des utilisateurs [23, 24], l'article [36] fait état de plus de 60 bases de données complètes libres

d'accès utile pour ce type de recherche. Ceci dit, il est important de prendre en compte que le comportement des conducteurs diffère de pays en pays. En raison du cas d'étude visant à évaluer l'utilisation des BRVEs dans l'île de Montréal, il était donc important de prendre en compte des données locales. Deux approches ont été suivies pour assurer cet aspect :

1. Le web scraping ;
2. Réception de données d'Hydro-Québec.

Web Scraping

Dû au manque de données libre d'accès sur les BRVEs installées au Canada, l'interface du fournisseur le plus important de point de recharge au Québec et dans l'est de l'Ontario, Le Circuit Électrique, a en premier lieu été utilisé pour extraire des données sur les BRVEs installées à Montréal, notamment leurs utilisations. Celui-ci est inauguré en 2012 en tant que premier service de réseau de bornes de recharge au Canada et compte plus de 3000 stations dans son réseau, dont 400 sont des points de recharges rapides [37]. Dans le cadre de la recherche cependant, seules les données sur 703 stations sont extraites. Ces stations se situent toutes au sein d'une surface représentée dans la figure 3.2. Parmi ces stations, 450 sont prises en compte étant donné leur situation géographique les positionnant au sein de l'île de Montréal. Pour extraire l'information pertinente sur ces stations, un programme de collecte de données est développé et incorporé sur le service Amazon Elastic Compute Cloud (EC2) pour assurer une collecte régulière des données avant de les analyser. L'annexe B présente la structure des données JSON recueillies détaillant les attributs pris en considération.

Le web scraping présente cependant plusieurs problèmes, principalement du fait de l'approche utilisée pour définir l'utilisation des bornes. Chaque observation extraite à chaque intervalle de 5 minutes considère les valeurs suivantes pour définir le statut d'utilisation :

- Available : Disponible ;
- InUse : En cours d'utilisation ;
- Unknown : Inconnu ;
- OutOfService : Hors service ;

Ce format présente plusieurs obstacles. En premier lieu, il présente une approximation de l'heure d'arrivée et de départ de chaque véhicule puisque ces valeurs sont définies par la date et l'heure de l'extraction de l'information. De plus, une borne qui vient de se libérer pourrait être utilisée aussitôt par un autre utilisateur durant l'intervalle considérée d'extraction des données, ce qui biaiserait l'heure d'arrivée et de départ de certains utilisateurs tout comme la durée d'utilisation d'une borne par différents utilisateurs. Cet obstacle peut cependant être résolu à l'aide de l'attribut *Last Status Update* qui servirait à communiquer la date et l'heure

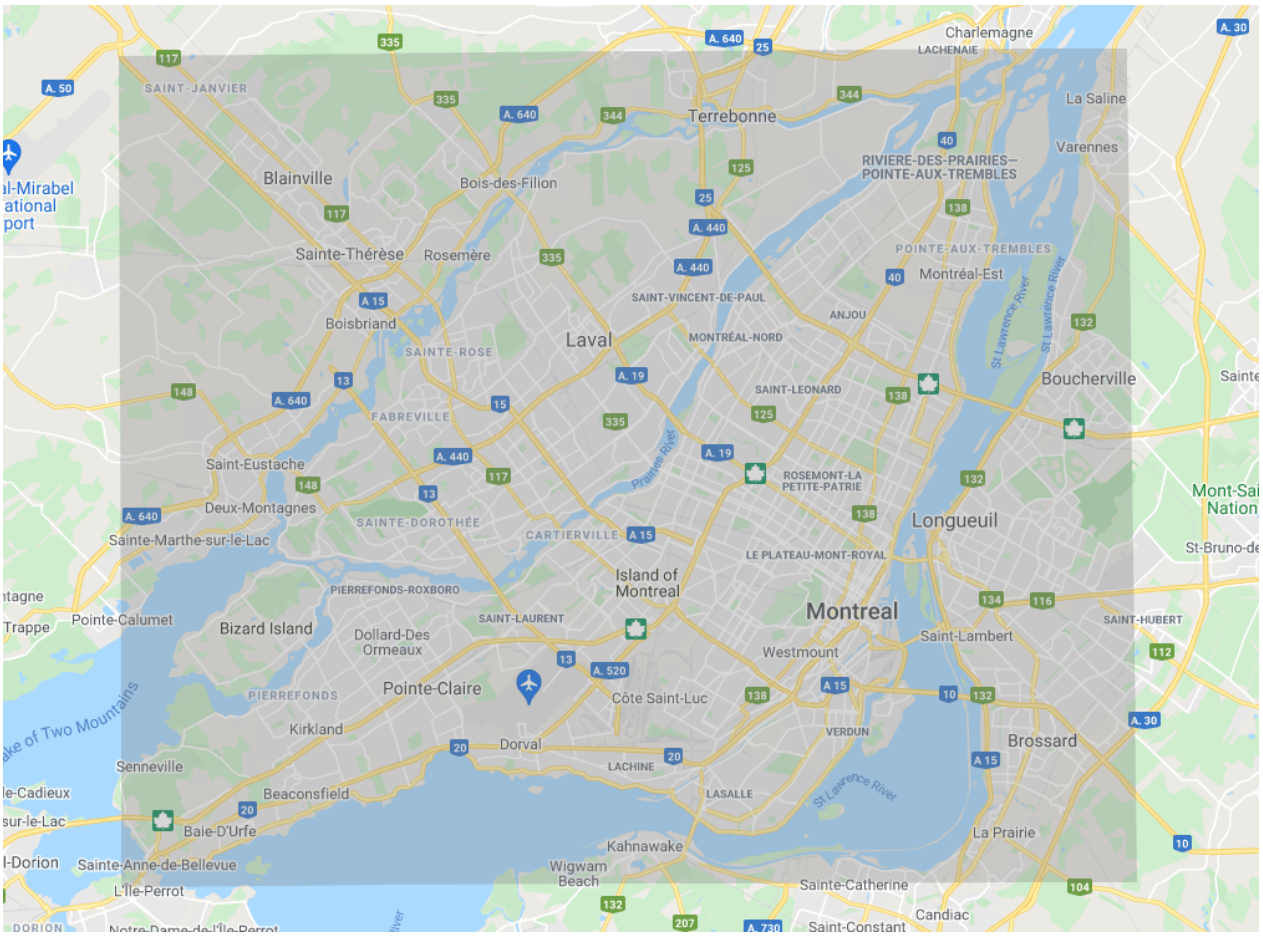


Figure 3.2 Région considérée pour l'extraction de données sur les BRVEs

de la dernière communication établie entre la borne et le serveur du circuit-électrique. En deuxième lieu, il existe plusieurs cas où la valeur de l'état d'utilisation est considéré inconnue, un aspect moins évident à résoudre.

Enfin, à la suite d'une mise à jour effectuée sur le site le 1^{er} juin 2020, limitant par ailleurs l'accès aux données, il n'était plus possible de les extraire. Malgré l'obtention de données sur 7 mois d'utilisation de BRVEs dans l'île de Montréal, seules les données de novembre 2019 à mars 2020 sont pertinentes dû à l'impact que le COVID-19 a eu sur l'utilisation des bornes. Cette solution a donc été délaissée et des données fournies par Hydro-Québec ont été privilégiées.

Données Hydro-Québec

À la suite des défis rencontrés dans l'extraction de données du circuit-électrique, l'avancement de la recherche dépendait de l'obtention de données plus structurées et complètes. La collecte de données d'Hydro-Québec nous a donc permis de regrouper des données appropriées pour une analyse plus approfondie de l'utilisation des bornes. Hydro-Québec est une société publique québécoise qui produit, transporte et distribue l'électricité dans le Québec [38]. C'est le plus large producteur d'électricité au Canada [39] et l'un des plus importants acteurs dans la production d'hydro-électricité dans le monde [38]. C'est aussi la société qui a mis en place le-circuit-électrique en partenariat avec d'autres acteurs clés. Les données fournies représentent donc une solide base sur l'utilisation historique des BRVEs installées depuis 2012.

Les données ont été fournies en format Power BI et regroupent de l'information variée sur chaque borne, leurs utilisations et les utilisateurs abonnés au service. Les données sur plus de 1 890 stations ainsi que 2 879 bornes au Québec et l'Ontario sont disponibles. De plus, plus de 2 millions de sessions d'utilisations détaillant les heures de début, de fin ainsi que les temps de connexions et de recharges sont inclus. Enfin, l'information sur plus de 30 000 utilisateurs est aussi couverte détaillant même les véhicules utilisés par certains utilisateurs. Le tableau C.6 résume l'entité la plus importante extraite de l'outil Power BI, à savoir le tableau détaillant toutes les sessions enregistrées, tandis que le reste des tableaux dans l'annexe C résume les entités restantes. Ces tableaux excluent certains attributs générés à partir de la division d'attributs, particulièrement ceux résumant les attributs de type *Date & Heure*. Au total, six entités sont présentées regroupant divers attributs importants :

- Parks ;
- Owners ;
- ChargingSessions ;
- Stations ;
- Members ;
- Vehicles.

Sur ces tableaux, la colonne *Description* décrit l'utilité de la variable, *Format* le type de données, et la colonne *Traitement* décrit les étapes suivies pour le pré-traitement des données utilisées s'il y a eu lieu. Le traitement des données représente, mais n'a pas été limité à l'évaluation de données aberrantes, l'analyse de données manquantes, le traitement de variable de type chaîne de caractères et la combinaison de plus de deux variables. Les valeurs de type chaîne de caractères sont principalement rendues minuscules ou corrigées en cas d'erreur d'orthographe. Ce processus a été important particulièrement pour les données décrivant les villes des clients. La seule contrainte rencontrée dans l'exploration des données est le manque de

l'attribut *Indications* fourni dans les données collectées à partir du site du circuit-électrique. Ces données sont essentielles pour la compréhension des restrictions liées au stationnement associé aux différentes bornes, telles que les heures d'ouverture et les durées maximales de stationnement.

Parmi ces données, seules certaines pertinentes sur les utilisateurs, les sessions et les stations ont été extraites et utilisées cependant. Ces attributs sont représentés en gras dans les tableaux de l'annexe C. De plus, seules les données en relation avec les stations se situant à l'intérieur de la surface décrite dans la figure 3.2 sont extraites. À partir de la figure 4.1 présentée dans la section 4.3.1, il est apparent que les BRVEs de cette régions sont généralement situés au centre de l'île de Montréal dans des zones à forte activité économique. Ces observations sont appuyées par les points d'intérêts représentés dans la section qui suit.

La section 4.3.1 décrit plus en détail le profil des stations et des sessions prises en compte dans la recherche.

3.1.3 Points d'intérêt et données de recensement

Au-delà de l'analyse de l'utilisation des BRVEs, l'objectif principal de la recherche étant d'analyser la corrélation entre l'utilisation des bornes et des facteurs externes variés, plusieurs données socio-économiques et géographiques ont été pris en compte. Plus précisément, les données sur les points d'intérêts de l'île de Montréal et de recensement de l'année 2016 du Canada ont été employées. La section qui suit détail les différentes sources utilisées.

Points d'intérêts

Les points d'intérêt (PoI) représentent un aspect important à prendre en compte dans l'analyse des facteurs qui influencent l'utilisation des BRVEs. En effet, ces données sont essentielles dans plusieurs études de planification de BRVEs [19, 20, 23, 24]. Ce sont aussi des données volumineuses pour lesquelles la pertinence dans des modèles dépend fortement de leurs traitements. Plusieurs étapes ont donc été suivies pour les organiser.

La première étape a été de choisir un des nombreux services d'extraction de PoIs en prenant en compte les critères suivantes :

1. Classification des PoIs sous des catégories disparates ;
2. Accès non-restreints et complets aux PoIs des régions spécifiées ;
3. Qualité des résultats en terme de : (a) Intégralité des données ; (b) Absence de données redondantes.

Il existe plusieurs interfaces de programmation d'applications (API) disponibles pour la collecte de ces données, notamment OpenStreetMap, Google Places API et FourSquare Places API. Ces services font partie des APIs les plus populaires dans le marché et fonctionnent relativement de la même manière. En effet, d'un côté, le Google Places API et FourSquare Places API permettent à n'importe quel utilisateur de créer un compte pour avoir accès à une clé secrète. Cette clé permet ensuite de lancer des requêtes *GET* pour extraire une multitude d'informations sur les PoIs. Chaque requête *GET* nécessite la définition de certains paramètres clés cependant, notamment le point autour duquel les données sont voulues et le rayon maximal à considérer. Le code Java ci-dessous présente un exemple d'une requête Four Square API comportant les paramètres requis :

```
uri = new URIBuilder()
          .setScheme("https")
          .setHost("api.foursquare.com")
          .setPath("/v2/venues/explore")
          .setParameter("client_id", client_id)
          .setParameter("client_secret", client_secret)
          .setParameter("v", "20180323")
          .setParameter("ll", latitude + "," + longitude)
          .setParameter("radius", 1500)
          .setParameter("time", "any")
          .setParameter("day", "day")
          .setParameter("openNow", false.toString())
          .setParameter("offset", offset.toString())
          .setParameter("limit", limit.toString())
          .build();
```

Malgré la simplicité de leur utilisation, ces services présentent de nombreuses contraintes. Tout d'abord, le service Google est payant et le service FourSquare ne permet l'exécution gratuite quotidienne que de 900 requêtes. De plus, chaque requête ne peut renvoyer que 20 PoIs dans le cas du Google Places API et 50 dans le cas du FourSquare Places API, ce qui augmente le nombre de requêtes requises pour extraire l'information complète sur les PoIs avoisinant chaque point. De même, alors que l'organisation des catégories des PoIs extraite à partir du FourSquare est structurée, chaque PoI de l'API de Google peut appartenir à plus d'une catégorie, ce qui rend leur analyse plus complexe. Enfin, les services ne renvoient qu'une proportion des PoIs se situant dans chaque zone, ce qui sous-représente certaines catégories et pourrait entraîner à des analyses biaisées. Une solution existe pour le service FourSquare,

mais ces requêtes renvoient plusieurs données dupliquées qui ne peuvent être traitées dû au manque de similarité entre leurs attributs. Ci-dessous, un exemple de deux PoIs renvoyés par une requête pour laquelle les coordonnées du point sont celles d'une des BRVEs considérées dans la recherche, le rayon est 500 mètres et la catégorie est *Outdoors & Recreation* :


```
"reasons":{ ... },
"venue":{"id":"523110c711d2dc8b3a86ce76",
         "name":"Yoga Nat",
         "contact":{ ... },
         "location":{"address":"Cote Des Neiges",
                     "lat":45.49651890210206,
                     "lng":-73.62397329749199,
                     "labeledLatLngs":[{ ... }],
                     "distance":433,
                     "cc":"CA",
                     "country":"Canada",
                     "formattedAddress":[ ... ]},
         "categories":[{ ... }],
          ... },
"referralId":"****"

"reasons":{ ... },
"venue":{"id":"53f9fc35498ee6361e578f95",
         "name":"Studio YogaNat",
         "contact":{ ... },
         "location":{"address":"5450 Chemin de la Côte-des-Neiges #400",
                     "crossStreet":"Edouard Montpetit",
                     "lat":45.49705253513732,
                     "lng":-73.62395524978638,
                     "labeledLatLngs":[{ ... }],
                     "distance":380,
                     "postalCode":"H3T 1Y5",
                     "cc":"CA",
                     "city":"Montréal",
                     "state":"QC",
                     "country":"Canada",
                     "formattedAddress":[ ... ]},
```

```
          "categories":[{...}],
           ... },
"referralId":"****"
```

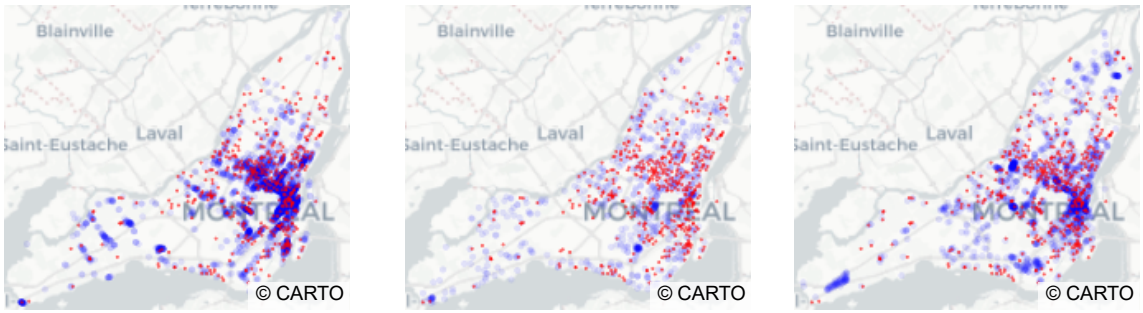
Ces deux résultats réfèrent au même endroit, mais pour lesquelles ni l'adresse, ni les coordonnées géographiques, ni le nom sont similaires ce qui rend la tâche d'élimination de données redondantes compliquée. Par conséquent, le nombre de PoIs autour de chaque borne peut être biaisé. En revanche, malgré une organisation plus complexe, OpenStreetMap offre plusieurs avantages. Tous d'abord, les PoIs d'OpenStreetMap peuvent être simplement extraits de la librairie *overpy* sur Python qui assure l'accès à l'API Overpass. Cette librairie est libre d'accès, contrairement aux options alternatives et permet à n'importe quel utilisateur d'extraire les PoIs d'une région compte tenu des catégories voulues comme suit :

```
area[name="Montréal (06)"][admin_level=5][place=region]->.searchArea;
(
    node["amenity"="restaurant"](area.searchArea);
    way["amenity"="restaurant"](area.searchArea);
    relation["amenity"="restaurant"](area.searchArea);
);
out center;
```

De plus, ces requêtes renvoient les données sur tous les PoIs mentionnés dans la région voulue. Ceci dit, contrairement aux services susmentionnés, il est important de définir les catégories à considérer dû à la complexité de la base de données d'OpenStreeMap. Inspiré par la structure utilisée par FourSquare pour l'organisation des catégories de PoIs [40], le tableau 4.1 décrit l'organisation suivie des catégories de PoIs extraites d'OpenStreetMap. De plus, cette complexité a aussi mis en avant un aspect important concernant la définition d'un point d'intérêt. En effet, comme est le cas de nombreux endroits tels que les universités ou même les industries, il peut exister plusieurs structures représentant une même institution. Ces structures peuvent aussi être très loin l'une de l'autre et représenter une surface très large. Il a donc été décidé de définir chaque établissement, notamment les universités, par les nombreux bâtiments qui les constituent, plutôt que la surface qui les représente.

En raison des nombreux avantages d'OpenStreetMap, particulièrement de l'extraction de données complètes par zone, le service a été utilisé pour la collecte de données pertinentes sur les PoIs de l'île de Montréal. La distribution de certains points d'intérêt est représentée dans la figure 3.3 alors que l'annexe D résume la distribution de tous les PoIs considérés. En

raison de leur représentation à partir des bâtiments qui les composent, la figure 3.3 détaille une forte concentration d'établissement d'enseignement dans la carte, particulièrement dans le cas où les PoIs sont représentatifs d'établissements universitaires. Il est notamment évident que le centre-ville regroupe une concentration importante de divers restaurants et cafés ainsi que de lieux de travail. Les concentrations de lieux de travail autour de la ville sont cependant représentatives de lieux industriels.



(a) Restaurants et cafés (b) Établissements d'enseignement (c) Lieux de travail

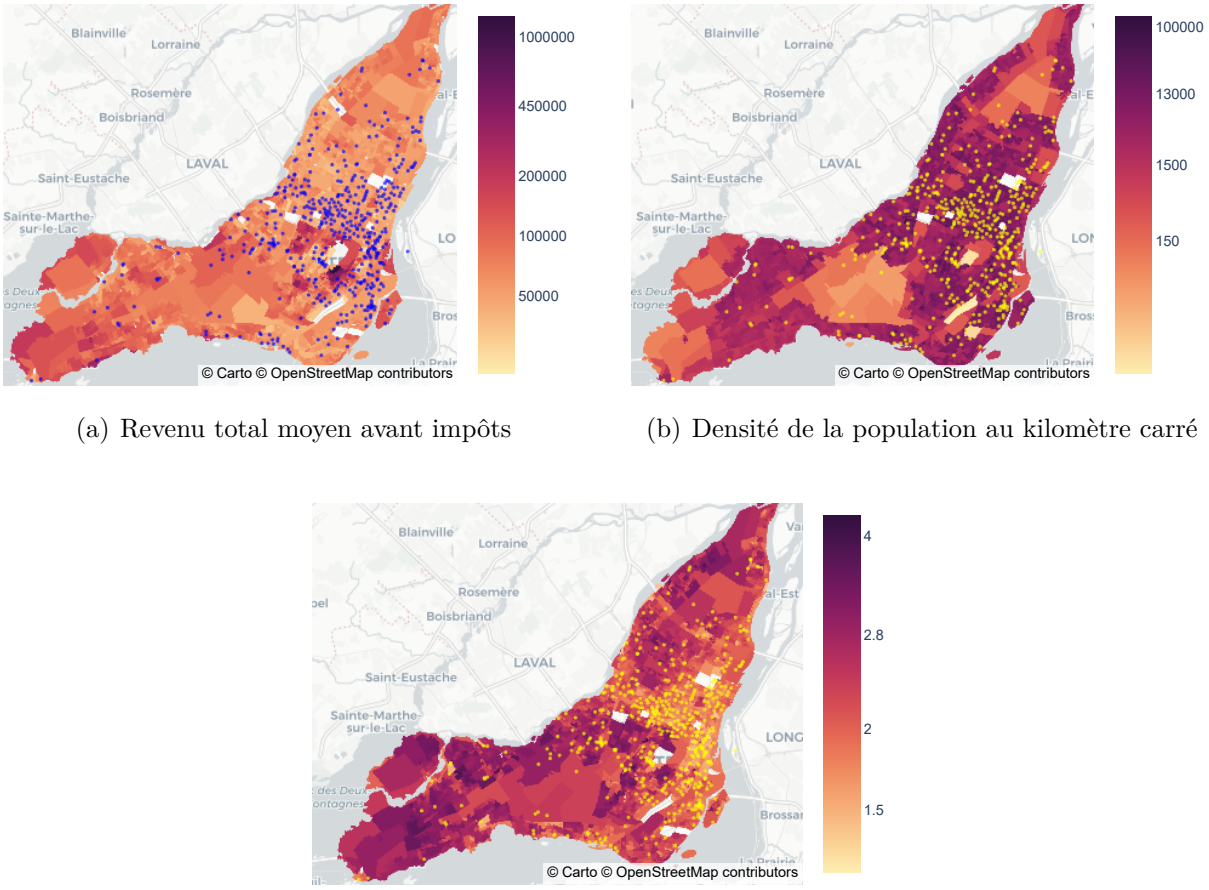
Figure 3.3 Distribution des PoIs clés en bleus et des stations publics en rouge.

La section 4.3.2 offre plus de détails sur les données collectées et utilisées.

Données de recensement

Les données de recensement représentent une information importante sur la population habitant au sein des différentes aires de diffusion de la ville de Montréal. A date de l'écriture de cette mémoire, le rapport de recensement le plus récent date de 2016 et a donc été considéré dans la recherche. Contrairement aux PoIs, ces données sont utiles pour la représentation de l'utilisation des BRVEs par des résidents. Les données de recensement contiennent diverses informations telles que l'âge, les origines, le type le logement et le revenu des ménages parmi d'autres aspects essentiels. Au total, plus de treize aspects sur la démographie et le profil socio-économique de la population sont évalués. Ces données sont proposées sous plusieurs formes spatiales. Dans le cadre de la recherche, les données des aires de diffusions qui constituent l'île de Montréal sont utilisées. Une Aire de Diffusion (AD) est une unité géographique représentant la plus petite surface géographique pour laquelle les données de recensement sont disponibles. La représentation du profil démographique et socio-économique de la population avoisinant les BRVEs est donc très précise. Ceci dit, seul un sous-ensemble de ces données a été pris en compte selon l'analyse faite sur celles considérées dans des articles de

recherche similaire [20,23,24]. La figure 3.4 détaille les distributions de certains attributs clés tandis que l'annexe E présente tous les attributs considérés dans la recherche.



(a) Revenu total moyen avant impôts

(b) Densité de la population au kilomètre carré

(c) Taille moyenne des ménages privés

Figure 3.4 Représentation de données de recensement clés de chaque aire de diffusion avec en bleu ou en jaune les stations publiques.

Pour mieux identifier le profil socio-économique de chaque station, les données de recensement des aires de diffusion pas plus loin de 500 mètres de chaque station sont combinées. Pour mieux représenter cette combinaison, la moyenne de la densité de la population, du revenu moyen, de la taille des ménages ainsi que de l'âge est calculée en prenant en compte le nombre de personnes considérées dans la collecte de ces données pour chaque AD. Tandis que la moyenne du nombre des différents types de logement est calculée en prenant en compte la surface terrestre des aires de diffusions considérées.

3.2 Problème d'apprentissage non-supervisé

Le partitionnement ou *clustering* des données est une méthode d'apprentissage non supervisé qui a pour objectif de regrouper les éléments d'une base de données en groupes homogènes. La recherche se sert de cette méthode pour non seulement analyser les données temporelles détaillant l'utilisation des BRVEs, mais aussi pour mieux les différencier par leurs attributs géographique et socio-économique. Les sections qui suivent présentent les différents algorithmes et méthodes considérés.

3.2.1 Algorithmes de regroupement

Plusieurs algorithmes de regroupement peuvent être considérés selon la distribution et la forme d'un ensemble X de données de taille N. Le K-Means et l'algorithme hiérarchique font partie des méthodes évaluées et considérées pour les différents objectifs de la recherche en raison de leurs approches divergentes. Pour utiliser ces algorithmes, la librairie *scikit-learn* disponible sur Python a été employée.

K-Means

La simplicité du K-Means en fait l'algorithme le plus populaire. En effet, seul le paramètre k détaillant le nombre de regroupements à prendre en compte doit être ajusté. L'algorithme fonctionne en générant k points distincts représentant les candidats potentiels de centroïdes. Bien que l'initialisation de ces points soit normalement faite de manière aléatoire, le *K-Means++* est un algorithme qui sert à optimiser cette étape pour réduire le temps de convergence du K-Means et donc améliorer sa performance [41]. L'algorithme sert à répartir les points au sein de l'espace de données pour couvrir le plus de regroupement possible. Pour cela, l'algorithme suit les étapes suivantes :

1. Initialiser aléatoirement le premier centroïde c_1 à partir de l'ensemble X de points ;
2. Où $D(x)$ est la distance euclidienne entre l'élément x et l'un des centroïdes déjà sélectionné le plus proche, considérer le centroïde c_j comme l'élément x pour lequel la probabilité :
$$\frac{D(x)^2}{\sum_{x \in X} D(x)^2} \tag{3.1}$$
est la plus élevé.
3. Répéter l'étape 2 jusqu'à l'initialisation de k centroïde.

La deuxième étape permet de sélectionner l'élément x pour lequel la distance avec le centroïde le plus proche est la plus élevé. Cette mesure assure que les centroïdes choisis sont

adéquatement dispersé.

À la suite de l'initialisation des centroïdes, la prochaine étape sert à minimiser la distance euclidienne entre les points au sein de chaque regroupement à travers la fonction suivante :

$$\sum_{i=0}^{N} \min_{c_j \in C} (\|\mathbf{x_i} - \mathbf{c_j}\|) \tag{3.2}$$

Pour cela, trois étapes importantes sont considérées :

1. Relier chaque élément appartenant à X au centroïde c_j le plus proche compte tenu la distance euclidienne ;
2. Pour chaque centroïde c_j, mettre à jour ses attributs comme la moyenne des attributs de ses éléments ;
3. Répéter les étapes 1 et 2 jusqu'à ce que les valeurs des attributs des centroïdes c_j ne changent plus, ou que leur taux de changement est plus faible qu'un seuil défini.

La simplicité de l'algorithme fait cependant place à certains inconvénients. En effet, la présence de valeurs aberrantes peut avoir un impact important sur le résultat final, notamment les moyennes des attributs de certains centroïdes. Pour contrer cet inconvénient, l'une des solutions les plus appropriées est d'éliminer ces observations du modèle en les considérant comme regroupements séparés. Une autre limitation est l'incapacité de l'algorithme de créer des regroupements d'éléments valides quand celle-ci ne forme pas des regroupements clairs convexes dans un espace euclidien.

Algorithme Hiérarchique

Contrairement à l'aspect aléatoire du K-Means, l'algorithme hiérarchique utilise une méthode de regroupement séquentiel où les éléments sont soit regroupés, soit séparés progressivement. Deux approches existent pour implémenter cet algorithme :

— Approche Ascendante : Commence avec un regroupement pour chaque élément avant de combiner les regroupement progressivement selon leur distance.

— Approche Descendante : Commence par un regroupement général avant de le diviser progressivement jusqu'à produire un regroupement par élément.

Chaque approche commence donc par le résultat du dernier. L'avantage de cet algorithme est sa capacité de mieux traiter des données distribuées plus aléatoirement dans un espace. Dû à sa simplicité cependant, l'approche ascendante est privilégiée puisqu'elle est aussi implémentée au sein de la librairie *scikit-learn*. Les détails de son fonctionnement et le calcul de la distance entre les différents regroupements sont expliqués plus en détail dans la section 4.4.1.

3.2.2 Mesures de distance

Les mesures de distances représentent un paramètre important pour le fonctionnement des algorithmes de regroupement susmentionnés. Dans le cadre de la recherche, seules deux distances clés sont étudiées pour la résolution des différents problèmes de regroupement dont leurs propriétés leur permettent une utilisation dans deux cas différents.

Distance Euclidienne

D'une part, la distance euclidienne est la distance la plus utilisée pour la résolution de problème de regroupement. Le but de cette distance est de calculer la distance point-à-point entre deux observations en considérant tous les attributs qui les définissent. Cette approche la rend optimale pour des données statiques. La section 4.4.1 explique plus en détail le fonctionnement cette distance.

Distance DTW

Tandis que la distance euclidienne prend en considération la distance point-à-point entre deux observations, le DTW permet de minimiser la distance entre deux séquences en prenant en compte une modélisation non-linéaire. Cette propriété la rend idéal pour des données temporelles. Son application est autant remarqué dans la comparaison de formes qu'elle ne l'est dans l'analyse de séquences [42]. En effet, il est même conseillé de l'utiliser à la place de comparer les attributs généraux de ces mêmes observations tels que l'heure d'arrivée ou départ moyenne ou la durée moyenne de connexion ou de chargement, etc. Tandis que son fonctionnement est détaillé dans la section 4.4.1, les paragraphes qui suivent détaillent l'algorithme développé pour optimiser son exécution.

Malgré ses avantages, la complexité du calcul de la distance DTW freine son utilisation dans plusieurs projets de recherche puisqu'elle requiert d'importantes ressources algorithmiques. Pour être exacte, chaque calcul de distance demande une complexité quadratique compte tenu de l'espace et le temps. Puisqu'il existe 211 018 profils de connexion à comparer, et que le DTW ne possède pas la propriété d'une métrique, il faudrait plus de $2{,}2310^{10}$ calculs pour compléter une matrice de distance de plus de 40 GO. En utilisant l'algorithme original du DTW, cette opération exigerait plusieurs semaines d'exécution. Il existe cependant plusieurs algorithmes d'optimisation du calcul du DTW tels que le *fastdtw* et le *awarp* qui généralise le calcul du DTW [43, 44]. En dépit de leur vitesse de calcul, ces algorithmes manquaient d'efficacité dans l'approximation de la distance entre les séquences considérées. Plusieurs mesures ont donc été prises pour optimiser l'algorithme original utilisé. En premier lieu,

l'optimisation algorithmique du calcul consistait à utiliser plusieurs librairies Python tels que *joblib* ainsi que *numba* pour tirer profit du CPU et remédier à la lenteur de Python. D'une part, *joblib* permet d'optimiser l'algorithme en introduisant la notion de *multiprocessing* ce qui permet de distribuer l'exécution de l'algorithme sous les nombreux cores du CPU. Selon le nombre de cores disponibles, il est possible de réduire le temps de calcul de 8 à 100 le temps de calcul normal. D'autre part, *numba* permet de remédier à la lenteur de Python en traduisant le code en langage machine optimisé rendant le programme aussi performant qu'un autre écrit en C ou Fortran. Enfin, la notion de *warping window* w est introduite dans le calcul du DTW pour non seulement réduire considérablement les calculs et ainsi le temps d'exécution mais aussi améliorer les résultats.

3.2.3 Représentation des centroïdes

Plusieurs méthodes de représentation des centroïdes existent, permettant de mieux définir les propriétés des éléments de chaque regroupement. La méthode la plus simple se fie sur le calcul de la moyenne des attributs des éléments appartenant à chaque centroïde. Cette approche est plus populaire lors de l'utilisation d'algorithmes de regroupement comme le K-Means. Elle est cependant moins appropriée pour des données temporelles. Pour ces types de données, il existe plusieurs alternatifs tels que le Partition Around Medoids (PAM) et le Dynamic Time Warping Barycenter Averaging (DBA) expliqué plus amplement dans la section 4.4.1.

3.2.4 Mesures d'évaluation

L'évaluation des résultats de regroupement est une étape primordiale pour l'analyse des hyperparamètres des algorithmes considérés, que ce soit pour le regroupement temporel ou celui d'attributs des BRVEs. Cette étape est surtout essentielle pour le choix du nombre k de regroupements optimaux à considérer. Ceci dit, le clustering est une méthode d'apprentissage non-supervisé, c'est-à-dire que contrairement aux méthodes d'apprentissage supervisé, il n'existe pas de valeurs de sorties exploitables pour mesurer la performance du modèle. Pour y remédier, plusieurs coefficients proposent l'évaluation de regroupement en prenant en compte la proximité des éléments au sein de regroupement et la distance entre les différents regroupements. Ainsi, le coefficient avec la valeur optimale représenterait un modèle pour lequel les éléments de chaque regroupement sont rapprochés et où les regroupements sont éloignés. Il existe cependant plusieurs limitations. En premier lieu, ces mesures ne sont tout de même pas exactes et se basent sur des modèles heuristiques qui apportent une valeur numérique approximative de la qualité du résultat. Il est donc conseillé de se fier à son propre jugement en traçant les mesures ou à se fier sur une distance spécifique en dessinant

le dendrogramme des résultats d'un regroupement hiérarchique. Or, l'approche suivie ne permet l'incorporation d'une étape manuelle pareil. En deuxième lieu, certaines mesures sont généralement applicables qu'avec les modèles qui emploient la distance euclidienne, or, le regroupement temporel emploie le DTW. Il est donc non seulement important de développer une solution automatisant le choix d'une valeur optimale de k mais aussi de se limiter à des mesures d'évaluation applicables au DTW dans le cas approprié.

Le développement d'une solution automatisé est particulièrement essentiel dans l'exécution du regroupement temporel pour lequel plus de 450 regroupements doivent être accomplis tels qu'expliqué dans la section 3.4.1. Il est toutefois aussi important de développer une solution pour laquelle le résultat est optimal. Ceci dit, l'un limite la faisabilité de l'autre. Pour remédier à ce compromis, il est donc important de développer une solution se basant sur plus d'une mesure. Un nombre important d'articles de recherches existent déjà faisant état des mesures les plus efficaces pour l'évaluation de regroupement. Malgré des résultats approximatifs et peu concluants, les coefficients Silhouette, Davies-Bouldin, Davies-Bouldin modifié et COP faisaient partie des mesures avec le plus haut taux de réussite. Ces mesures sont aussi considérées comme étant les plus efficaces puisqu'elles évaluent les distances au sein et entre les regroupements [45]. Enfin, elles sont utilisées dans plusieurs articles de recherche traitant la distance DTW [46, 47] et dans la librairie *dtwclust* qui sert à résoudre des problèmes de regroupement à partir de la distance DTW sur R [48]. Il existe toutefois des défis dans l'utilisation de ces coefficients. Premièrement, comme toute autre mesure prise en compte dans la recherche, elles ne rendent pas une évaluation élevée dans le cas où les données contiennent du bruit ou où les regroupements coïncident comme est le cas de certaines données utilisées. De plus, il est important de définir une approche pour réunir ces coefficients en une seule valeur. Pour cela, deux approches sont suivies.

En premier lieu, un modèle de classement est utilisé pour définir, pour chaque regroupement, la configuration idéale, qui dans le cas des algorithmes utilisés est la valeur k idéale. Pour cela, pour chaque regroupement, les mesures d'évaluation sont calculées pour chaque configuration. Les valeurs de chaque métrique sont ensuite traduites sous forme de classement selon la valeur de la métrique. Les meilleures valeurs sont associées à un classement plus faible tandis que les pires valeurs sont associées à des classements plus haut. La moyenne du classement des différentes mesures d'évaluation est ensuite calculée et définie comme classement final de chaque configuration. La configuration possédant la moyenne la plus faible est finalement choisie. Il existe cependant une contrainte importante avec cette approche puisqu'elle ne permet pas de prendre en considération la variance des valeurs des mesures d'évaluation, un aspect important qui pourrait biaiser les résultats finaux. Une autre approche a donc été considérée.

En deuxième lieu, une approche prenant en compte la moyenne des mesures d'évaluation est utilisée. Contrairement à l'approche susmentionnée, celle-ci prend en considération la variance des différentes mesures et permet donc de mieux capturer la configuration optimale à considérer. Pour cela, chaque mesure d'évaluation est redimensionnée sur une même échelle considérant les valeurs maximale et minimale de chaque métrique comme suit :

$$x_{scaled} = (x - min)/(max - min) \tag{3.3}$$

Cette méthode permet de comparer les différentes mesures. Cette propriété est de plus utilisée pour représenter une valeur globale évaluant la performance de chaque regroupement en prenant la moyenne de toutes les métriques. La configuration avec la valeur la plus faible est au final considérée la plus optimale, l'objectif étant de minimiser cette moyenne.

La section 4.4.1 offre plus de détails sur les différentes mesures utilisées dans la recherche en détaillant leur fonctionnement.

3.3 Problème d'apprentissage supervisé

D'autre part, le problème d'apprentissage supervisé constitue toute méthode d'apprentissage statistique pour laquelle la sortie est connue. Les méthodes de régression et de classification constituent cette catégorie. Alors que les méthodes de régression servent à prédirent des valeurs réelles, les méthodes de classification servent à prédire des classes. Les sections qui suivent servent à introduire le K-Nearest Neighbors ainsi que l'analyse SHAP utilisée pour évaluer la pertinence des valeurs d'entrées de modèles de prédiction.

3.3.1 K-Nearest Neighbors

Le modèle K-Nearest Neighbors (KNN) est un modèle de classification pouvant aussi servir comme modèle de régression dans certains cas. Le modèle est non-paramétrique et ne possède donc pas de nombre fini de paramètres à définir comme est le a cas d'un Support Vector Machine (SVM) ou d'une simple régression linéaire. La performance du modèle se fie donc sur la taille de la base de données. Cependant, en appliquant un réglage des hyperparamètres approprié, le modèle peut être très performant.

Le modèle fonctionne en définissant une distance et un nombre k de points à considérer pour estimer la classe dans laquelle une observation x_0 devrait appartenir étant donné les k points les plus proches représenté par un espace N_0. Pour ce faire, la probabilité conditionnelle qui

suit est évaluée pour chaque classe :

$$Pr(Y = \widehat{j|X} = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j) \tag{3.4}$$

Où l'objectif est d'évaluer la probabilité que l'observation appartienne à la classe j étant donné les valeurs d'entrées x_0. La classe avec la probabilité la plus importante, c'est-à-dire celle appartenant au plus de points avoisinants l'observation est finalement choisie pour identifier l'observation.

Plusieurs paramètres clés doivent être réglés pour l'optimisation du modèle implémenté dans la bibliothèque *scikit-learn* sur les données fournies, à savoir le *n_neighbors*, *weights*, *metric* et *leaf_size*. En premier lieu, le paramètre *n_neighbors* sert à définir le nombre de points à considérer dans le voisinage N_0. Une valeur trop basse pourrait entraîner à un surapprentissage et manquerait de généraliser les données d'entraînement. Tandis qu'une valeur trop large généraliserait beaucoup trop les données et entraînerait à un sous-apprentissage. Il est donc important d'entraîner le modèle sous plusieurs configurations de k. En deuxième lieu, le paramètre *weights* sert à définir un mécanisme de pondération au calcul de la probabilité présenté dans l'équation 3.5. Deux méthodes sont proposées ; l'une considère une pondération uniforme pour tous les points k, tandis que la deuxième pondération nommée pondération par distance se sert de l'inverse de la distance entre l'observation et les points avoisinants pour définir les probabilités des différentes classes. Cette approche permet de valoriser les classes des points les plus proches et utilise donc la fonction qui suit pour définir la probabilité conditionnelle :

$$Pr(Y = j|X = x_0) = \frac{1}{\sum_{i \in N_0} d(x_i, x_0)^{-1}} \sum_{i \in N_0} I(y_i = j) \times d(x_i, x_0)^{-1} \tag{3.5}$$

Où $d(x_i, x_0)$ représente la distance entre l'observation et chaque point l'avoisinant appartenant à l'espace N_0. Cette distance peut représenter diverses métriques définies à partir du paramètre *metric*. La bibliothèque *scikit-learn* permet de choisir parmi les distances euclidiennes et manhattan.

3.3.2 Analyse SHAP

L'analyse des facteurs importants étant un aspect primordial de la recherche, il est important de définir une méthode pour l'évaluation de leur pertinence dans la prédiction des différentes classes. Bien que des modèles basés sur l'algorithme des arbres de décisions permettent d'évaluer l'importance des variables d'entrées dans le modèle, cette caractéristique n'est applicable

qu'à un nombre limité d'algorithme de prédiction, de plus il n'est pas possible de l'utiliser pour évaluer l'importance des différentes variables dans la prédiction des différentes classes. L'analyse SHAP (SHapley Additive exPlanations) proposée par Lundberg and Lee [49] est une méthode applicable sur n'importe quel modèle de prédiction permettant d'analyser la prédiction de chaque observation séparément de sorte à mieux comprendre l'effet des valeurs d'entrée sur la prédiction. Pour y arriver, l'algorithme utilise la théorie des jeux. Son fonctionnement est résumé dans la section 4.4.2.

3.4 Détail de la solution

La solution présentée pour la résolution de la problématique de la recherche se divise en deux étapes clés tels qu'introduit dans la section 1. Ces axes utilisent différents modèles et données présentées dans les sections ci-dessus. Les sections qui suivent servent à réunir ces méthodes et données pour présenter l'approche technique utilisée pour résoudre chaque axe.

3.4.1 Partitionnement et analyse des données temporel

Le partitionnement ou *clustering* des données est le premier axe étudié, qui a pour objectif d'expliquer et de comparer l'utilisation des différents BRVEs. Contrairement au partitionnement de données statiques, celui de données temporelles nécessite une approche plus compliquée. Comme évalué dans la revue de littérature, plusieurs approches sont prises en compte pour réaliser cette tâche. Tandis que certains transforment ces données en forme d'attributs avant de les évaluer [28] d'autres prennent en compte l'entièreté des séquences, une approche jugée plus performante [50]. Dans ce cas, il existe plusieurs algorithmes utilisés dans la littérature tels que les Gaussian Mixture Models (GMMs). Cependant, dans le cadre de cette recherche, un modèle de regroupement hiérarchique utilisant le Dynamic Time Warping comme mesure de distance entre les éléments est utilisé. Plus précisément, une approche à étapes multiples est suivie pour la complétion de cette étape en raison de la complexité de l'algorithme DTW utilisé.

En premier lieu, les profils de connexion de chaque station sont regroupés de manière à préserver la granularité de ces données en créant des centroïdes aussi représentatifs des éléments de chaque regroupement que possible. Pour ce faire, $k = \{2, \ldots, 200\}$ est considéré ainsi qu'un *warping window* w de 1. Cette étape permet de donner une idée générale sur l'utilisation des différentes stations tout en réduisant le nombre de profils de connexion à évaluer pour l'analyse générale de toutes les stations. Ensuite, les centroïdes produits sont regroupés en utilisant des valeurs de k plus limité, à savoir $k = \{2, \ldots, 100\}$ de manière à générer le moins

de regroupements possible. De plus, un *warping window wC* de 2 est utilisé dans cette étape pour ne pas restreindre le calcul de la distance entre les profils de connexion. Cette étape permet de générer un nombre limité de profils de connexion général représentant la structure des profils de connexion des stations considérées. Ce résultat permet notamment de comparer les profils de connexion de toutes les stations et simplifie leur analyse.

Enfin, en associant le profil de connexion de chaque station aux profils généraux de connexion produits, un tableau de fréquence et de proportion peut être conçu. Le tableau de fréquence commence par détailler, pour chaque station, le nombre de jours associés à chaque profil de connexion général. Puis, le tableau des proportions divise ces valeurs par le nombre de jours totaux considérés pour chaque station de sorte à mieux refléter la pertinence de chaque profil de connexion général des différentes stations et mieux comparer les stations entre elles. Tandis que ces valeurs servent à mieux comparer les stations, elles sont aussi essentielles pour le deuxième axe de la solution.

La section 4.5 présente les étapes expliquées ci-dessus plus en détail.

3.4.2 Analyse de facteurs d'utilisation des BRVEs

L'étude des facteurs d'utilisation des BRVEs constitue une étape clé pour la compréhension du comportement des utilisateurs et du profil de connexion des différentes bornes. Les données des PoIs et de recensement détaillé dans la section 3.1.3 sont utilisées à cette fin. Plusieurs approches sont considérées et comparées pour remplir cette tâche :

- Analyse simultanée des partitions spatiales et temporelles ;
- Prédiction de la proportion d'appartenance de chaque site à chaque regroupement temporel ;
- Prédiction d'appartenance de chaque site à un regroupement.

Les sections qui suivent détaillent les défis et avantages de chaque approche en concluant avec la méthode choisie.

Problème non-supervisé

Plusieurs approches peuvent être considéré pour le regroupement de données spatiales. Dans un premier lieu, il est possible d'utiliser des méthodes de densité, telles que DBSCAN, pour le regroupement de points dans une carte. Cette approche se limite sur la situation géographique relative aux autres stations seulement cependant. La recherche considère plutôt une approche considérant les différentes données spatiales et socio-économiques. Celle-ci vise à regrouper les BRVEs possédant les attributs les plus similaires. Pour cela, plusieurs méthodes peuvent être

considérées en prenant en compte la distance euclidienne telles que le K-Means, le DBSCAN et la méthode ascendante du regroupement hiérarchique.

La solution requiert tout d'abord l'initialisation d'une analyse du Principal Component Analysis (PCA) qui sert à réduire la dimension des attributs à deux composantes représentant une importante proportion de la variance des données générales. Contrairement au regroupement de données temporel, l'utilisation d'une métrique telle que la distance euclidienne dans cette étape permet de représenter les données dans un espace euclidien et donc d'utiliser le PCA. Cette représentation permet non seulement de mieux choisir la méthode de regroupement appropriée, mais aussi de visualiser les regroupements créés par chaque modèle. Les méthodes de regroupement sont ensuite exécutées et leurs performances comparées pour choisir la plus appropriée. Le résultat constitue plusieurs regroupements différents représentant des stations pour lesquelles les PoIs et données de recensement sont très similaires.

Les différents regroupements spatiaux créés sont finalement comparés avec les regroupements temporels produits dans la section 3.4.1 pour en déduire les similitudes. Pour cela, les BRVEs appartenant à chaque regroupement temporel sont comparés en prenant en compte le regroupement spatial auquel ils appartiennent. Cette analyse permettrait de mieux comprendre la corrélation entre les regroupements temporels et les regroupements spatiaux. Ceci dit, cette approche présente certains inconvénients, dont la dépendance considérable de l'approche sur l'hypothèse que les regroupements temporels sont corrélés aux regroupements spatiaux. Si la corrélation est faible, les résultats seraient difficiles à interpréter puisque les regroupements seraient très différents.

Problème de régression

L'utilisation de modèles supervisés est utile pour l'analyse de facteurs pertinents dans l'identification du profil de connexion des BRVEs. Plusieurs modèles, notamment ceux se basant sur les arbres de décisions tels que le Random Forest, permettent de construire un tableau d'importance des différents attributs d'entrées utilisés. Pour suivre cette approche, les valeurs de sorties doivent être représentatives des résultats obtenus dans la section 3.4.1. Pour cela, le tableau de proportion, détaillant pour chaque station la probabilité d'un profil de connexion général, est utilisé. Cependant, ces données sont en formes de fractions dénommées données de composition, pour lesquels l'univers simplexe est défini comme suit par John Aitchison [51] :

$$S^D = \left\{ (x_1, ..., x_D)^T \middle| x_i \geq 0, \sum_{i=1}^{D} x_i = 1 \right\} \tag{3.6}$$

La somme de ces fractions, pour chaque observation, est égale à une constante, qui, dans le cas de la recherche est 1. Cette propriété rend toutes les valeurs de sortie corrélée. Les modèles de régression ne peuvent cependant pas prendre cette propriété en compte dans leur exécution. Pour donc assurer que les prédictions de n'importe quel modèle de régression sont des proportions valides, il est important d'exécuter pour ce genre de problème une transformation des fractions. Il existe plusieurs méthodes pour transformer ce type de données tel que les transformations logratio additive (alr) [51], isométrique (ilr) [52], ou centrée (clr) [51]. Ceci dit, la présence de 0 dans les données utilisées ne permet pas l'application de ces méthodes. Bien qu'il existe des méthodes d'imputation de ces valeurs, ces méthodes ne sont généralement valides que pour des données pour lesquelles les zéros sont de type arrondi. Ces zéros sont soit le résultat d'erreurs ou de limitations dans les outils de mesure utilisées dans l'expérimentation. Or, dans le cas des données, les zéros sont de type structurel, c'est-à-dire que les valeurs équivalent réellement à zéro. Il est donc important de considérer une méthode de transformation applicable à des données contenant des zéros structurels. Tsagris présente plusieurs alternatives aux méthodes susmentionnées pour résoudre ce problème tel que la transformation α [53].

Ceci dit, la complexité de cette approche, la nature stochastique de l'utilisation des différents BRVEs ainsi que le manque de maturité des algorithmes évalués nous ont dissuadés d'aller plus loin avec cette approche. L'utilisation d'un modèle de classification a été privilégiée plutôt, simplifiant la solution.

Problème de classification

Enfin, en raison de la complexité des approches susmentionnées, une approche à 2 étapes est utilisée. Cette approche consiste à étudier les similitudes entre les BRVEs en fonction du tableau des proportions de profils de connexion général à travers un modèle de regroupement, avant d'analyser la pertinence de leurs facteurs externes à partir d'un modèle de classification.

En premier lieu, l'analyse et la comparaison des différents BRVEs est effectué en utilisant le K-Means sur la matrice de proportions des profils de connexion général. Les regroupements résultants, référés comme regroupement spatial des BRVEs, représentent des stations pour lesquelles les proportions des différents profils sont très similaires. Le modèle est ainsi itéré sur $k = \{2, \ldots, 25\}$ et les résultats évalués à travers différentes métriques de regroupement, à savoir les indices Silhouette et Davies-Bouldin détaillés dans la section 4.4.1.

Ensuite, un modèle de classification dont l'objectif est de prédire le regroupement spatial auquel une station appartient est établi. Le K-Nearest Neighbors est utilisé à cette fin en raison de sa robustesse et son efficacité dans plusieurs problèmes de classification. Le modèle

est exécuté sous différentes configurations à savoir les paramètres expliqués dans la section 3.3.1 et le modèle performant le mieux en termes du *log loss*, *f1 score* et *accuracy* est choisi. Une analyse SHAP est ensuite complétée pour analyser la pertinence des différents facteurs dans chaque regroupement considéré dans la première étape.

La section 4.4.2 présente les étapes de cette approche plus en détail.

CHAPITRE 4 ARTICLE 1: MULTI-STAGE CLUSTERING AND ANALYSIS OF PUBLIC ELECTRIC VEHICLES CHARGING STATIONS' USAGE: A CASE STUDY FROM MONTREAL, CANADA

Authors: Ismail Zejli, Hanane Dagdougui and Martin Trépanier

Manuscript submitted to: Transportation Research Part C: Emerging Technologies

Abstract: The growing popularity of EVs has pushed for important investments for a more structured and available infrastructure of Electric Vehicle Charging Stations (EVCSs). An expansion of this infrastructure involves however a thorough study of various aspects, of which the required investment, the potential demand, and the implications of the use of the EVCSs on the grid. Leveraging the availability of more complete datasets, a more rigorous analysis can be completed considering the historical use of various EVCSs varying in their location and other aspects to formulate a clear understanding of the connection behavior of its users and tackle these challenges. The article constructs a 2-stage approach to tackle several of these concerns using data collected from Quebec's electricity provider Hydro-Quebec covering EVCSs in the island of Montreal. The first stage uses hierarchical clustering and the DTW distance measure to cluster daily connection profiles of EVCSs. Several clustering valuation indices, of which the Silhouette score, the Davis-Bouldin and the COP index are used to evaluate the clusters. The resulting clusters are then used to better understand and compare the way EVCSs are used in the city. The second stage uses a KNN model to better understand the impact of different external factors, particularly points of interests and socio-economic attributes, on EVCSs' connection profiles. A SHAP analysis is finally conducted on the classification model to understand the relevance of the different external factors on the connection profiles of EVCSs.

Keywords: Dynamic Time Warping (DTW), hierarchical clustering, electric vehicle charging stations

4.1 Introduction

Climate change, CO_2 emissions and fossil fuel dependency are growing concerns in the transportation sector, which have led many governments around the world to take appropriate actions to reduce the aggregated carbon footprint of their citizens. Amongst these efforts, was the promotion of electric vehicles, primarily Plug-in Hybrid Electric Vehicles (PHEVs) and Battery Electric Vehicles (BEVs). Many countries around the world initiated incentivization

plans for the quick penetration of these vehicles. In fact, with a land transportation sector accounting for more than 14% of global CO_2 emissions in 2019 [5], several initiatives and campaigns like the EV30@30 Campaign and the Electric Vehicles Initiative (EVI) have been endorsed by a number of countries, including Canada, to accelerate the deployment of EVs in the markets by 2030 [54]. Although Canada's EV sales represented only 3.52% of its total private vehicle sales in 2020, it has experienced an important growth in the past and expects to record an exponential growth in the coming years to reach its goal of 100% of the total light-duty vehicle sales in 2040 [55]. Moreover, of the 2020 EV sales, more than 48% were recorded in Quebec, with Montreal representing the most important EV market [7].

Although a quick adoption of EVs can be beneficial for the reduction of greenhouse gas emissions [56], one of the main challenges is to investigate the implications this will incur on the infrastructure, particularly the EV charging stations (EVCSs). Several aspects may affect its efficiency, including the deployment plan of EVCSs and their real-time operation amongst other issues. Today, these challenges contribute to increasing concerns for many stakeholders. For drivers, a lack of strategic planning of EVCSs augments range anxiety, but above all, accentuates the anxiety of finding a queue or unavailable EVCSs. Drivers' anxiety being a serious roadblock to EV sales, an indirect result of it is the delay of the government's prospect to achieve its 2030 EV goals and manufacturers' reduced incentive to invest on EV development. For operators, an inefficient expansion plan of EVCSs can lead to unfavorable impacts on the grid as peak load becomes a concern as well as significant losses as important investments on new EVCSs turn out unprofitable.

To remedy these issues, several approaches have been developed. Over the past years, long-term optimization models for the strategic planning of EVCSs have been developed taking into account the prospective demand for EVCSs and other conditions [14, 20]. Recent availability of private and public data on EVCS usage raised the interest of many researchers to develop models for the forecast of their popularity [23] and expected demand [57] as well as power consumption [24]. However, the charging behavior in EVCSs is extremely stochastic and differs greatly in time from one station to another. Therefore, models based on the processing of the entire historical data of the EV charging infrastructure are less accurate as they generalize the behavior of the whole EVCS network, ignoring the spatio-temporal correlations. On the other hand, station-dependent models are not only computationally expensive, they also limit the interpretability and generalization of the infrastructure's demand. Clustering serves as a middle ground for both approaches, offering a better compromise between interpretability, computational efficiency and accuracy. Indeed, clustering has shown promising results when used prior to regression or classification notably in medical [58] and energy [59] sectors, offering more accurate results. In the context of EVCSs, [60] used cluster-

ing to group geographically comparable stations to improve their location planning model for an electric bus transit system. Similarly, [31] uses clustered EV profiles to predict a day-ahead users' behavior. Additionally, clustering can also serve as an important approach to capture the charging behavior of stations according to their locations. Henceforth, understanding the spatial and temporal usages of EVCSs can help operators better plan future locations of EVCSs and manage the power demand through smart charging strategies including Vehicle-to-grid (V2G), Grid-to-Vehicle (G2V) and dynamic pricing and incentives [61,62]. Moreover, a better understanding of the user charging behavior can serve as an important basis for the roll-out of EVCSs [63].

This paper proposes to approach the analysis of EVCS behavior through a multi-step method. Stations are first clustered according to their demand, then a classification model is implemented to better understand external factors' influence on EVCSs' usage behavior, while considering diverse geographic, social and economic attributes. The proposed approach allows for not only the understanding of each EVCS's demand, but also for the comparison of disparate EVCSs' demand profiles. This multi-step approach will help planners select future EVCS locations considering the potential demand of stations within a predefined environment. It differs from common approaches which use predictive models to evaluate charging stations' popularity as done by [23].

The remaining sections are structured as follows. Section 4.2 reviews the literature on the analysis of EVCSs' behavior. Section 4.3 presents the used dataset, and section 4.4 describes the proposed methodology. The case study is then implemented in Section 4.5. Finally, a conclusion along with a discussion of the limitations and potential future research directions is presented in Sections 4.6 and 4.7.

4.2 Related Work

Several articles seek to understand EVCSs' charging behavior with the goal of better approaching more complex tasks. To achieve this goal, various techniques have been considered, notably, surveys, simulations and, as a result of the growingly available data and popularity of machine learning, clustering. In the following section, an analysis of the methods used as well as the attributes taken into account for clustering evaluations is conducted.

With a lack of labeled outputs, clustering is an unsupervised learning technique that can serve as an important way of understanding patterns and structures within the data set [34]. Multiple works implement this technique considering two different types of data: individual and time series data to cluster various aspects of EVCS' usage. On one hand, [28] use

K-Means with the Euclidean distance (ED) to cluster EV user behavior in the basis of the average charging time, connection times and standard deviation to solve the problem of centralized scheduling of the grid. The resulting clusters are used to classify the new users into the appropriate cluster using KNN while allowing the power system operators to optimally incorporate V2G and G2V strategies to the EVCS system. Likewise, authors in [29] cluster different aspects of user behavior such as charging time, arrival and departure times, and hours between charges using the Gaussian Mixture model (GMM). The paper concludes on an analysis of different clusters of users and their charging behavior. [30] use clustering to define hourly power demand profiles of EV charging in the UK. The clusters are found using K-Means through, assumingly, the ED for which the number of k clusters was validated through the Davies-Bouldin criterion. The resulting clusters are further used to define the charging demand of EVs' impact on the distribution network. [31] evaluate the state of charge (SOC) of EVs to cluster drivers with similar behaviors. The authors also use the Euclidean Distance and Hierarchical Clustering to define the different clusters, emphasizing on K-Means' weakness in treating outliers. The resulting clusters are then used to improve the accuracy of predictive models used for user behaviors' day ahead forecasting. Different distance measures, particularly the Euclidean distance (ED), the modified Euclidean Distance (MED) and the Dynamic Time Wrapping (DTW) are compared to cluster the tail of charging sessions' current in [32]'s research. The accuracy of MEDs is however shown to be superior than the other measures. [33] and [34] review different approaches used in the literature for the forecast and clustering of charging station data. [33] details research that uses clustering to define groups of elements, primarily users, by their charging profiles. The study emphasizes on the importance of clustering in the EVCS planning problem and grouping EVCSs with similar demand profiles. [34] also presents the clusteirng of EV users' charging behavior through methods like GMM, K-Means and Kernel Density Estimators (KDE) achieved by different articles.

Despite its computational complexity and its non-metric nature, DTW has been used extensively in the literature to solve various types of clustering problems involving time series data, particularly in the energy sector. For instance, [64] used a DTW K-Medoids approach to cluster similar load profiles before using Markov models to forecast next-day load profiles. Likewise, [65] uses the DTW to cluster household load profiles with the goal of reducing the number of clusters while improving the quality of the clusters. Consequently, the authors achieve a 50% reduction in the cluster size compared to K-Means and Gaussian based approaches. Moreover, values of the within-cluster sum, between cluster centers sum and the ratio of within cluster sum to between cluster variation are better than those achieved through the other approaches. DTW continues to also be a distance measure of choice in

various other domains such as finance [66].

Various limitations can be recorded from the above studies. The majority of works in the literature have been focusing on the clustering of users' charging behavior or specific charging sessions failing to cluster an EVCS' usage profile or power demand that play an important role for the operation of the power grid. Another unexploited feature is that authors tend to generalize user's charging behavior, limiting the efficacy of the proposed approaches such as in [28]. Finally, while the ED is widely used and appropriate for the clustering of individual data, it is significantly less reliable for the analysis of time series as it fails to grasp the temporal aspects of the data [35]. This aspect is however extremely essential, notably considering the randomness and noisiness of time series evaluated in this research.

While interest in the clustering analysis of EVCSs' charging behavior has grown, many aspects have not been studied further. Hence, to the best of our knowledge, the following is a list of the gaps in literature that are addressed in this paper:

1. Considering the attribute-based approach used by many authors, previous works have rarely considered the sequential aspect of the data. The present paper however considers the temporal aspect in the analysis of the connection behavior in EVCSs through exploiting the DTW distance as achieved in other domains by [64] and [65]. To the best of our knowledge, this is the first implementation of the approach in the evaluation of EVCSs.
2. Considering the approaches made by other authors, many seek to evaluate EV charging sessions separately or to consider driver-specific behavior in their clustering analysis. While very insightful, a station-specific approach can also provide ample information for the strategic planning of EVCSs by offering insight on the charging behavior of differently located stations. To the best of our knowledge, no other article has focused on evaluating stations' charging behavior considering the daily profiles.
3. Many attempts have also been made to better understand the popularity of EVCSs. Various variables are used in order to achieve this goal. However, quantifying the connection profiles of different EVCSs has not been considered yet. This article seeks to address this gap to better understand the effects of external factors on EVCSs' charging behavior.

4.3 Datasets

The research's success is greatly dependent on the availability and quality of the data. As a result, numerous data sets were considered and evaluated, while others were inaccessible or

incomplete. For ease and work efficiency, a Java application was developed to organize and extract useful data sets from .csv files and web scraped .json files. Once extracted, they were linked relationally into a MySQL database. The following subsections present the available data that was incorporated and processed into the database before introducing their role in the developed models. While the EVCS data set is used for the temporal clustering, the points of interest and census data are used for the analysis of the factors influencing the charging behavior of EVCSs. The choice of the geographic and census attributes is mainly based on their use in various articles treating EVCS' planning problem [19, 23].

4.3.1 Electric Vehicle Charging Stations

EVCS usage datasets are scarce as the world is yet to embrace a more global adoption of EVs. There exist however many complete and freely available datasets used by many authors like EVnetNL which covers the EVCS usage in the Netherlands [23] as well as an exponentially growing repository of open databases, 60 to be exact [36]. Many authors have also turned to private datasets unavailable to the public and surveys [19]. As this research focuses on studying Montreal's EVCSs, data provided by our industrial partner (Hydro-Québec) is investigated.

The dataset includes a comprehensive collection of (anonymized) customers and charging sessions for the province of Quebec. It covers over 2 million sessions from October 2011 to May 2020, detailing the session start and end times, session duration, and charge duration amongst other attributes. The sessions are representative of the use of unique ports, which aggregate to single EVCSs. Typically, a public station has a total of 2 ports or less (93.06% of the total EVCSs in Quebec), with certain exceptions, particularly those located in large parking lots or large venues for which the number of ports can be as high as 13.

The present study's goal is to evaluate the charging behavior of public EVCSs in Montreal. It represents 29.1% of a total of 1614 public stations in the Quebec Province. Stations are filtered by their availability to the general public, their power output, and their installation date as the primary objective is to look into activity-related charging habits which are generally represented mostly by Level 2 EVCSs [29]. Figure 4.1 presents a GIS-based representation of the considered public charging stations where the color is representative of the level while the size represents the number of ports associated to the station. The largest circle represents stations with 10 ports whereas the smallest circles represent stations with only 2 ports. Grey dots reflect the stations outside the scope of the case study.

Because various factors affect EVCSs' behavior, a series of data pre-processing steps are followed to ensure the homogeneity of charging sessions. First, considering the yearly expo-

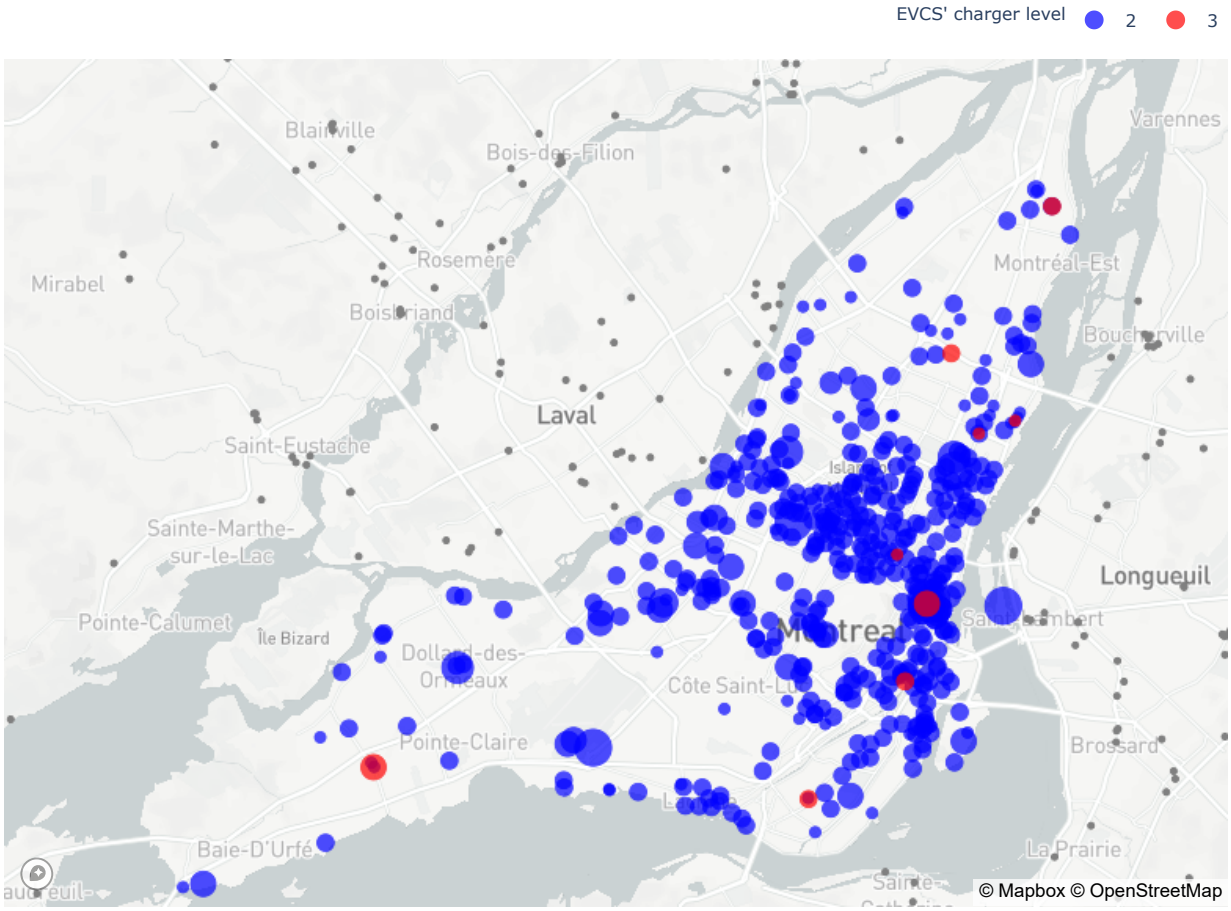


Figure 4.1 Geographic distribution of Level 2 and Level 3 EVCSs in the island of Montreal

nential growth of EVs' adoption in Quebec [7], only sessions beyond January 1st 2019 were considered. This growth is attested by the yearly sales of BEVs and PHEVs presented in figure 4.2 where growth slows down in 2020 but is highest in 2019. Concurrently, the number of unique customers taking advantage of EVCSs increased by 69% in 2019 to its highest point.

Second, Covid-19 restrictions had an important impact on transport behaviour and consequently, on the usage of EVCSs. To be exact, the overall demand, represented by the number of sessions, dropped by 392.8% between March 12th 2020 and April 12th, 2020 in Montreal island. Hence, only sessions prior to March 12th, 2020 are considered to avoid anomalies in travelling patterns. Additionally, as the deployment of EVCSs is done gradually, various EVCSs are installed during the considered time frame. Considering that new ports' charging behavior tends to stabilize after a week, any EVCSs installed after February 1st 2020 are

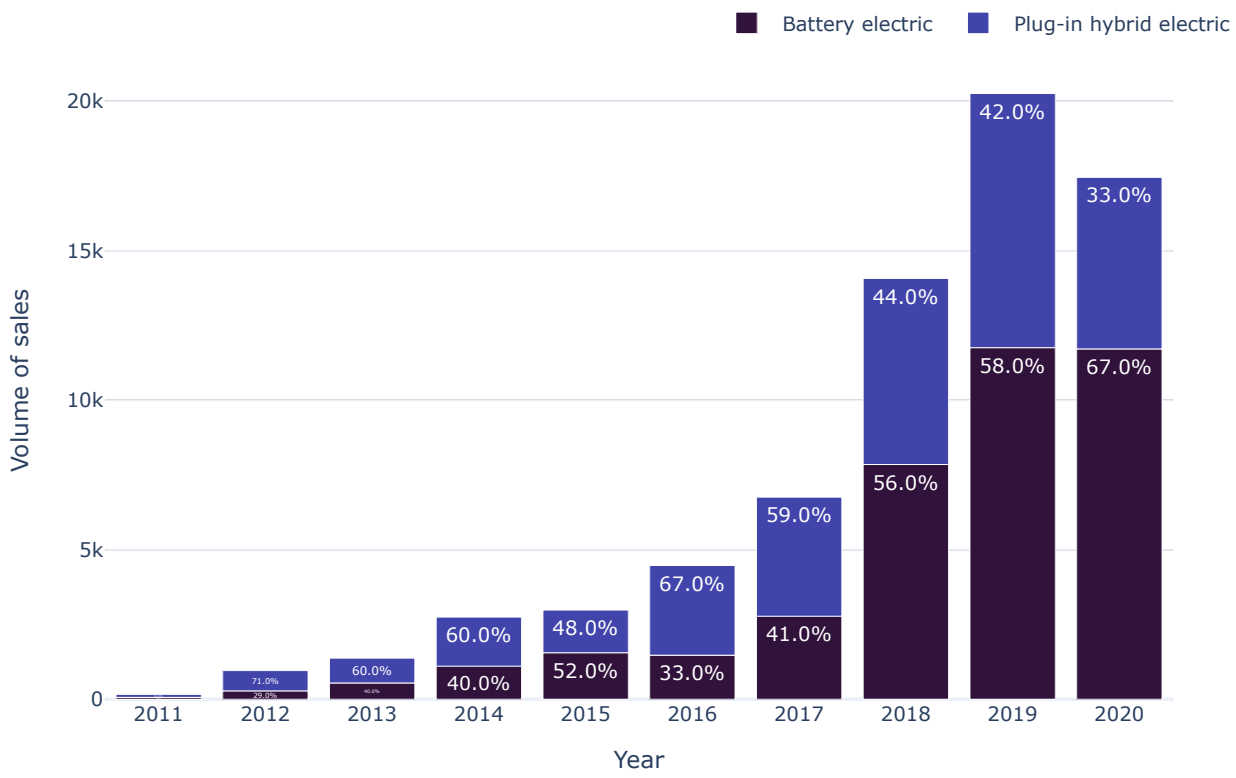


Figure 4.2 Volume of BEV and PHEV sales from 2016 to 2020 in Quebec where the bar labels are representative of the proportion of the total sales and the lines of the recorded percentage yearly growth.

discarded as their daily charging behavior cannot be generalized the same way older stations can. This is owing to their post-installation period which would have a greater influence on their overall profiles. Finally, any charging sessions lasting less than 60 seconds were removed to limit the amount of noise in the dataset. As a result of the pre-processing steps, a total of 339 386 sessions are considered.

The approach used in this study requires the representation of the sessions through a particular format. Contrary to other scholars seeking to analyse attributes separately, the present work builds a series of daily connection profiles for each port to be clustered as sequential data. While the data allows for a representation of the daily profiles precise to the second, in order to reduce the time and space complexities of the algorithms used in the analysis, a one-hour time step is considered. An example of random daily session profiles taken for the same station is shown in figure 4.3 where a binary value is set to each hour to represent the status of the port. A positive value represents a period in which the port is used, while

a value equal to zero is set to a period for which no vehicle is connected. This approach allows for the approximate analysis of periods in which the ports are in use during the day, considering the start time, end times as well as the duration of the connections.

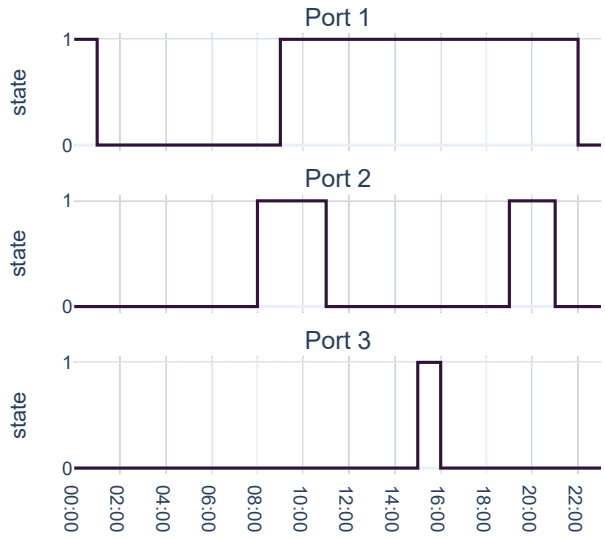


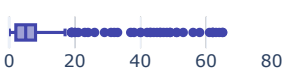
Figure 4.3 Graphic representation of ports' daily connection profiles sequence

4.3.2 Points of Interest

Open Street Maps (OSM) is used in order to gather the database with information on points of interests (PoIs). The data is first queried before it is organized into separate categories based on the category tree used by the Four Square service [40]. Additionally, the Four Square Places service is used to complement missing data from the OSM database, particularly data on offices. Consequently, more than 11000 PoIs are queried. Table 4.1 details the category hierarchy used, where the tertiary categories refer to OSM values, unless asterisked, in which case the category is retrieved from the Four Square Places API. In the scope of the project, to better evaluate the importance of separate PoIs while minimizing their volume, only secondary categories are used unless non-existent, in which case preliminary categories are considered.

Considering the uncertainty of users' willingness to walk to their final destinations, various radial distances are considered to link EVCSs to their neighboring PoIs. The present study however uses a 500 meter radius to associate PoIs to the EVCSs considering drivers' willingness to generally walk from their parking spot to their final destination (or vice-versa) [67].

Table 4.1 Description of categories used for the organisation of PoIs

Preliminary Category	Secondary Category	Tertiary Category	Distribution
Arts & Entertainment		cinemas, galleries, museums, theatres, bowling alleys and dances	
Education		colleges, libraries, schools and universities	
Services	Financial Services	banks, loan	
	Government Services	diplomatics, governments	
	Living Services	beauty shops, hairdressers, massages	
Groceries	Large Venues	marketplaces, supermarkets and wholesales	
	Small Venues	alcohols, bakeries, butchers and conveniences	
Shopping Services	Large Shopping Centers	department stores, malls	
	Household	beds, do-it-yourselfs, furnitures, hardwares, electronics	
	Miscelaneous Shops	clothes, boutiques, cosmetics, shoes, watches	
Car Services	Fuel Stations	fuel station	
	Car-specific	car repairs, car dealerships, car rentals	
Healthcare	Healtcare Service Centers	clinics, dentists, doctors, hospitals	
	Pharmacies	pharmacy	
Outdoors & Recreation		fitness centers, parks, sport centres, golfs	
Religious		place of worship	
Work		industrials, offices*	
Tourism		hotels, ...	
Food		bars, cafes, fast foods, food courts, ice cream, pubs, restaurants	
Nighlife		nightclubs, stripclubs	
Transportation Services		bicycle rentals, transport stations	

4.3.3 Census Data

To define the demographics of the region surrounding each EVCS, data from Canada's national statistical agency is integrated. The information within the dissemination areas of the Island of Montreal is retrieved from the 2016 census, which is also the latest census conducted as of the writing of the article. More specifically, data on the average revenue, age and household size of the population, population density and the type of households are considered. The data is first processed in QGIS before being linked to the EVCSs based on their proximity to the dissemination areas. In order to better capture the demographic profile of the EVCSs, the demographic data is aggregated amongst the dissemination areas for which the centers are at most 500 meters far from each EVCS.

4.4 Methodology

The completion of this research involves several key steps detailed further in figure 4.4. These steps comprise mainly the use of unsupervised learning models for the analysis of the connection profiles of the stations before using supervised learning models to study the effects of external factors on the stations' connection profiles. These models and their evaluations are detailed further in the following sub-sections.

4.4.1 Clustering

Clustering is representative of the first stages of the study, where the objective is to generalize granular data. These stages integrate different approaches for which the algorithms are explained further below, whereas section 4.4.1 details the aggregation of the algorithms into the set stages.

Distance Measures

Distance measures are an important parameter to consider for clustering algorithms as they define the similarity measures between different observations. One of the most widely used measure is the Euclidean distance. The main feature of this measure lies in its simplicity to capture point-to-point differences between two arrays of features. Considering J features, the distance between two observations A and B is computed as follows:

$$d(A, B) = \sqrt{\sum_{j=1}^{J} (a_j - b_j)^2} \tag{4.1}$$

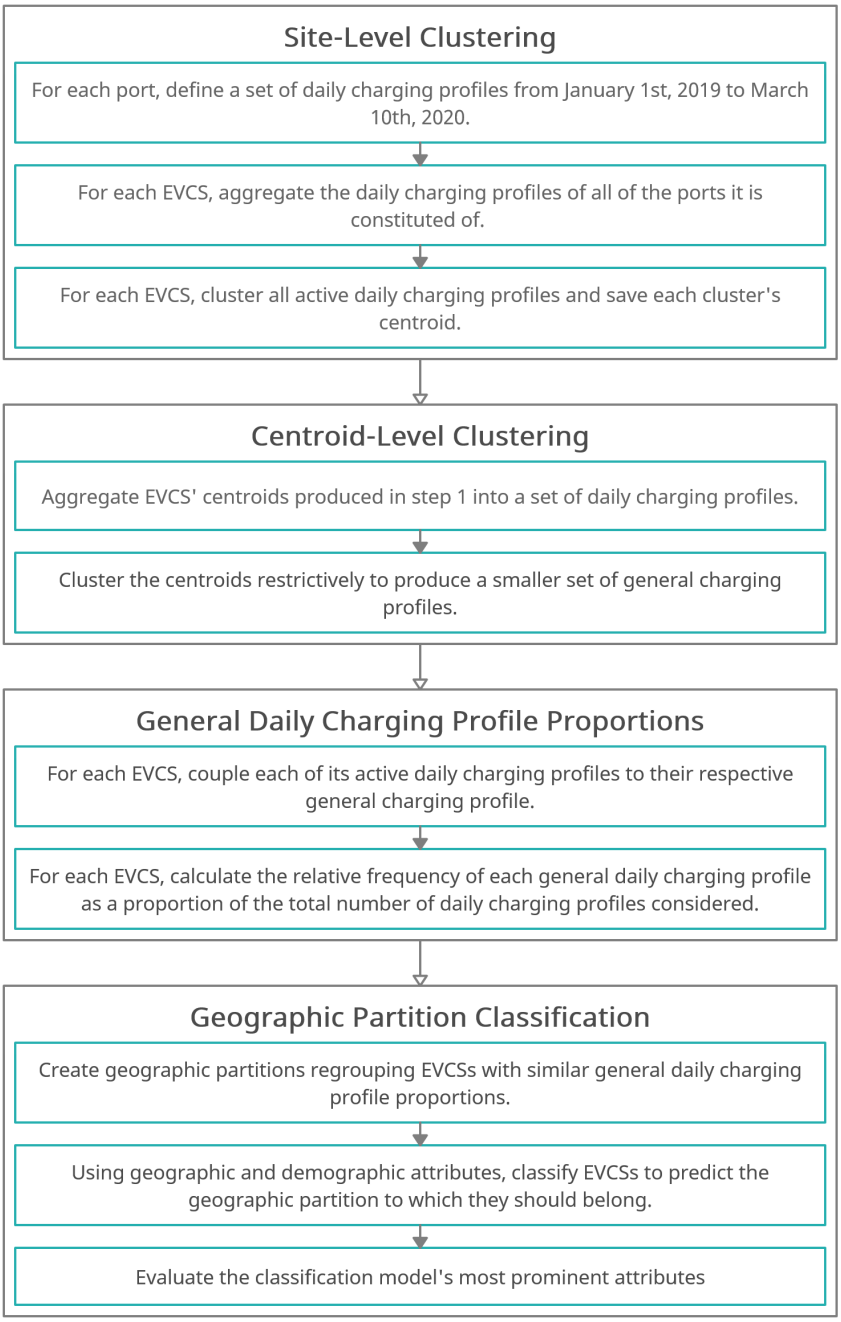


Figure 4.4 Flow chart representing the sequence of steps used to produce appropriate results.

As a result of its simplicity, this distance measure does not capture the temporal aspect of sequential data. DTW serves however as an appropriate algorithm to overcome this gap in spite of time and space complexity. As opposed to the Euclidean distance, DTW minimizes the distance between the two arrays by introducing a non-linear mapping between their points. While its use is very popular with series of different lengths, the approach is also

extremely important for evaluating series with similar patterns and temporal lags. The algorithm used to compute its distance bases its result on the least costly warping path W_p between two sequences A and B of lengths n and m, amongst P possible warping paths [68]:

$$DTW(A,B) = \min_{W_p \in P} \sum_{w_l \in W_p} d(w_l) \tag{4.2}$$

Where the warping path W_p is a sequence of grid points for which each w_l represents a point (i,j) depicting the association of points of sequences A to B. Moreover, $d(w_l)$ identifies the distance between the appropriate points of sequences A and B. As a result of the binary representation of the sequences, the following distance measure is used:

$$d(i,j) = |A_i - B_j| \tag{4.3}$$

To compute the optimal distance minimizing the cost between the two sequences, a dynamic programming approach can be used. This approach bases its result on the following recurrence relation, which defines the value associated to each cell (i,j) as $\gamma(i,j)$:

$$\gamma(i,j) = d(i,j) + \min \left\{ \begin{array}{c} \gamma(i-1,j) \\ \gamma(i,j-1) \\ \gamma(i-1,j-1) \end{array} \right\} \tag{4.4}$$

The DTW algorithm constructs a distance matrix D of size $n \times m$ using the aforementioned equation from which an optimal warping path w_p can be constructed by associating sequential cells with the smallest values considering the following constraints:

1. Boundary conditions: For each warping path, the starting point needs to be cell $(1,1)$ and the end point cell (n,m).
2. Monotonicity: Cells representing the warping path need to be arranged in a monotonous fashion satisfying the following index values conditions $i_{l-1} \leq i_l$ and $j_{l-1} \leq j_l$.
3. Continuity: Steps to construct the warping path need to be of a single unit satisfying the conditions $i_l - i_{l-1} \leq 1$ and $j_l - j_{l-1} \leq 1$.

Finally, the final DTW distance can be assumed to be represented by the cell $\gamma(n,m)$.

Despite offering great results, the complexity of the DTW algorithm also accounts for high computational complexity as a result of its $O(mn)$ complexity for each distance calculated. This can create a serious barrier considering the very high number of daily profiles to cluster. Moreover, given the absence of the metric property, a DTW distance has to be computed for each pair of daily profiles, rendering the approach time-consuming. While various algorithms

exist to optimise the computation of the distance, their approaches affect the final results for the type of discrete data used. A python algorithm, optimizing the original DTW algorithm using *numba* and *joblib* to leverage the computer's performance is thus developed. Moreover, in order to reduce space complexity, a warping window w constraint is introduced. This not only reduces the computational time, but also limits the comparison of data points of different profiles to a constrained range. A naive approach, is to use the largest value of w permitted by the available computational resources. However, figure 4.5 outlines the importance of considering this type of constraint strategically as it defines a better cost for dissimilar profiles, and thus efficiently helps the clustering algorithm to differ between sequences. Generally, authors suggest using a warping window equivalent to 4.5% of the length of the time series [69].

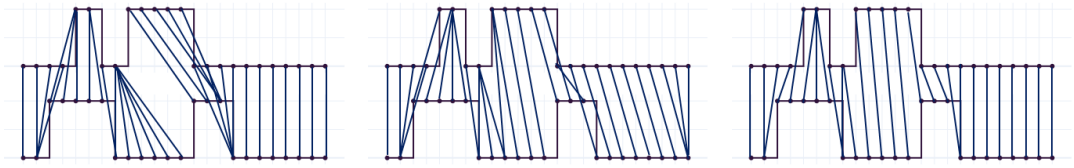
Clustering Algorithms

Depending on the distribution and source of the data, different clustering algorithms can be considered, of which the K-Means. Its simplicity and popularity as a clustering tool in various fields make it an optimal option for the clustering of non-sequential data. It constructs its result on the basis of minimising the inertia, otherwise known as the within cluster sum of squares, by iteratively looking into the euclidean distance between the centroids μ_i and the elements $\mathbf{x}$ constructing a set of K clusters referred to as C_k:

$$\min \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \mu_i\|^2 \tag{4.5}$$

A set of K centroids are first randomly generated through the *K-Means++* algorithm which optimises the initialisation of the centroids [41].

Although K-Means is a commonly used algorithm, it is not appropriate for temporal data [50].



(a) Unconstrained DTW computation. cost = 0.

(b) Constrained by window = 2 DTW computation. cost = 4.

(c) Constrained by window = 1 DTW computation. cost = 7.

Figure 4.5 Comparison of constrained and unconstrained DTW computations and their resulting costs.

Instead, other algorithms, particularly, hierarchical clustering can be used provided a pre-computed distance matrix. Two separate approaches make up this approach, of which the agglomerative clustering algorithm. It involves the assembling of single-element clusters to form multi-element clusters in a hierarchical manner to form a dendrogram of connections detailing all possible K clusters.

The algorithm uses a distance matrix to combine the closest elements together in an iterative fashion considering several linking approaches. Three are appropriate when using a non-euclidean affinity: average, complete and single. These criterion compute a distance between each pair of clusters, before the pair with the smallest value are joined together hierarchically. The complete criterion computes the maximum distance between clusters by analyzing their furthest elements. The single criterion on the other hand uses the minimum distance between elements of two clusters. Finally, the average criterion takes into account the average distance between all elements of separate clusters. Considering the use of a non-metric distance measure, the average criterion is most appropriate given its consideration of all elements.

Centroids

Centroids are not only essential for understanding the characteristics of the clusters produced but are also important components of Clustering Validation Indices (CVIs). When the Euclidean distance is appropriate, centroids can be represented as the average of the features of all elements making each cluster, an approach considered with the K-Means algorithm. Constrained by the temporal property of the data, using the point-to-point euclidean average or most prominent value with sequences can lead to the misrepresentation of the centers however. Instead, Partition Around Medoids (PAM) and DTW Baycenter Averaging (DBA) can be considered as they do not alter the temporal structure of the sequences. DBA on one hand is a time series averaging method which uses DTW to compute the mean of a set of temporal sequences [70]. Its iterative approach however engenders additional computation time as it necessitates a constant computation of the DTW distance matrix after each iteration to approximate the average of the sequences. PAM on the other hand sets the centroid of a cluster as the element closest to all other elements in the cluster. Using the distance matrix, the element with the smallest sum of distances to all other elements within the same cluster is set as the centroid. As opposed to DBA, it is simple, efficient and robust against outliers. PAM is thus used to define the centroids in CVIs while DBA is used to present the final centroids of the clusters.

Cluster Evaluation Metrics

Analysis of clusters is an important step to assess clustering algorithms' hyperparameters, particularly the number of K clusters to consider. Completing this step requires the use of different CVIs to evaluate and compare different results. CVIs are separated into internal and external cluster validation metrics, where external metrics consider labels of the observations whereas internal metrics only consider the features to validate the partitions. As a result of the approach used, only internal CVIs are considered for the evaluation of the clusters created. Moreover, while the use of the Euclidean distance permits the integration of a wide range of these measures, many are less appropriate for non-Euclidean distances as a result of their consideration of the clusters' centroids. However, various studies that used the DTW as a distance measure also used multiple popular evaluation metrics [46, 47]. Furthermore, the R library *dtwclust*, implemented to perform clustering using the DTW distance, also uses various CVIs such as the Silhouette, Davies-Bouldin, modified Davies-Bouldin and COP indices, amongst others, to evaluate the results [48]. The aforementioned metrics have also been generally proven to be some of the most efficient metrics [45], partly as a result of their evaluation of both the inter- and intra-cluster variance. Thus, they are considered for the evaluation of the created groups.

The Silhouette index [71] is one of the most popular metrics used for the evaluation of clustering results due to its ability to capture both, the nearest- and intra-cluster quality of the clusters through the following formula, where N represents the total number of elements being clustered:

$$S = \frac{1}{N} \sum_{i=0}^{N} \frac{b_{i,k} - a_{i,k}}{\max(a_{i,k}, b_{i,k})} \tag{4.6}$$

Two key components define the silhouette score, $a_{i,k}$ and $b_{i,k}$. The $a_{i,k}$ evaluates how close each element i within the same cluster C_k is whereas $b_{i,k}$ returns how far each element i within a cluster C_k is from the closest cluster C_l. In both formulas, $d(x_i, x_j)$ is the Euclidean or DTW distance computed between the elements x_i and another element x_j.

$$a_i = \frac{1}{|C_k| - 1} \sum_{x_i, x_j \in C_k, x_j \neq x_i} d(x_i, x_j) \tag{4.7}$$

$$b_i = \min_{l \neq k} \frac{1}{|C_l|} \sum_{x_j \in C_l} d(x_i, x_j) \tag{4.8}$$

The resulting metric lies within the range $[-1, 1]$, where the higher the range, the closer together the inter clusters are and the further apart the clusters are. Any negative value

represents a badly clustered set of elements.

Similar to the Calinski-Harabsaz index [72], the Davies-Boudin index evaluates the ratio of the within-cluster and between-cluster distances through the following formula:

$$DB = \frac{1}{K}\sum_{k=1}^{K}\max_{k\neq l} R_{kl} \tag{4.9}$$

where:

$$R_{kl} = \frac{1/|C_k|\sum_{x_i\in C_k} d(x_i,\bar{c_k}) + 1/|C_l|\sum_{x_j\in C_l} d(x_j,\bar{c_l})}{d(\bar{c_k},\bar{c_l})} \tag{4.10}$$

The objective is to compute the average ratio between each cluster and its nearest neighbor. The within-cluster variance is considered as the average distance between each element within the clusters k and l and their centroid, represented as $\bar{c_k}$ and $\bar{c_l}$ respectively. Whereas the separation is considered as the distance between the clusters' centroid. Hence, smaller values of the index, which lie within the range $[0,\infty]$, translate to better clusters.

Additionally, a modified Davies-Bouldin index proposed by [73] is also considered, which seeks to primarily resolve violations introduced by the averaging of the intra-cluster distances. Instead, the proposed formulation takes into account the maximum:

$$DB^* = \frac{1}{K}\sum_{k=1}^{K}\frac{\max_{k\neq l} 1/|C_k|\sum_{x_i\in C_k} d(x_i,\bar{c_k}) + 1/|C_l|\sum_{x_j\in C_l} d(x_j,\bar{c_l})}{\min_{k\neq l} d(\bar{c_k},\bar{c_l})} \tag{4.11}$$

The COP index is the last index considered, for which the within cluster variance is similarly computed to the aforementioned indices. Unlike other indices, the inter-cluster variance is defined as the smallest complete distance between elements of clusters:

$$COP = \frac{1}{N}\sum_{k=1}^{K}|C_k|\frac{1/|C_k|\sum_{x_i\in C_k} d(x_i,\bar{c_k})}{\min_{x_j\notin C_k}\max_{x_i\in C_k} d(x_i,x_j)} \tag{4.12}$$

Clustering Approach

Multiple stages of the proposed framework integrate different combinations of the aforementioned algorithms, distances and metrics. The first stage of the study involves the analysis of the daily connection profiles of EVCSs. As a result of the big data and the computational complexity of the DTW, an important consideration of the study is to separate the aforementioned task into multiple partitions before aggregating the results. Indeed, considering thousands of daily profiles when computing the distance matrix needs important computational resources as the generated matrix is not only very large but also requires a substantial

generation time. In particular, considering the 211 018 active daily sessions, more than 82.9 GB of space and several days of execution time would be required to complete the task. As can be seen from figure 4.4, the daily connection profiles of each EVCS are first generated by identifying the daily connection profiles of the ports that make up the EVCS. These daily connection profiles are then clustered as part of the site-level clustering step and the centroids of each EVCS are defined. These centroids are then clustered to create a broad set of centroids representative of the general connection profiles of all EVCSs as part of the centroid-level clustering step. The objective of this approach is not only to generalise similar daily connection profiles but also to subsequently ease the analysis of different EVCSs' daily connection profiles. Moreover, acknowledging that users' activities differ depending on the type of day, the clusters produced are further analysed on the basis of weekends and business days [74].

Once the aforementioned analysis are executed and the results generated, the algorithm produces a table linking each analysed daily connection profile to its respective final general connection profile. These results are manipulated further to create a table detailing the proportion of occurrence of each general connection profile in each EVCS as part of the general daily connection profile proportions step. This table provides ample data on the kinds of daily connection profiles recorded for each EVCS and more importantly, details the similarities amongst the different EVCSs. In order to better evaluate these similarities, a K-Means algorithm is executed to associate ports with similar proportions and most importantly realize the geographical partitioning of the stations.

4.4.2 Classification

Classification is the final stage of the study where the objective is to evaluate the relevance of different external factors of which the points of interests and census data collected. To complete this task, the labels produced in the aforementioned K-Means clustering of the proportion table are used as output, while the external factors are considered features. A number of algorithms are then hyper-parameter tuned and their performance is compared to select the optimal model. A Shapley Additive Explanations (SHAP) analysis [49] is finally conducted to better understand each label's most prominent attributes.

Classification Models

Due to limited number of observations, only random forest and K-Nearest Neighbors models (KNN) are evaluated. On one hand, random forest is referred to as an ensemble method as it uses multiple decision trees to deduce a final output. The model relies on bagging and the use

of a subset of the total features for each tree in order to create an ensemble of uncorrelated trees. This property further reduces the variance of the model, thus improving the results. K-Nearest Neighbors on the other hand deduces an observation's class by evaluating the most prominent class amongst the k nearest training observations' set.

Classification Evaluation Metrics

In order to choose the most appropriate model, several metrics are used, in particular, the accuracy, *f1* score and cross entropy loss. Only the cross entropy loss is however used to optimize the hyper-parameters of the models as it looks into how close the model's predicted probabilities are to the true label. Given the multi-class nature of the model, it can be calculated as follows:

$$Logloss(Y, P) = -\log Pr(Y|P) = -\frac{1}{N}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1} y_{n,k}\log p_{n,k} \tag{4.13}$$

where K is the number of labels and N the number of observations, Y is a binary matrix of size $K \times N$ representing the true values of the observations, whereas P is a probability matrix of size $K \times N$ representing the classification model's predicted probabilities. The closer the predicted probability of an observation's label is to the true value, the lower the loss is. The accuracy and f1 score, on the other hand, make use of the confusion matrix. The accuracy evaluates how accurate the model is at predicting the right label by dividing the number of correct predictions by the number of total observations. Whereas the f1 score computes the harmonic mean of the precision and recall which are more appropriate for unbalanced data than the accuracy measure. Where the precision is the number of true positives over the sum of true positives and false positives, and the recall is the number of true positives over the sum of true positives and false negatives.

SHAP Values

While a random forest can be very intuitive in the evaluation of feature importance, a SHAP analysis is more insightful, particularly in understanding the relevance of each attribute for the different labels. The analysis cannot only be conducted on a random forest but also on other models such as the K-Nearest Neighbors model. The algorithm relies on game theory's shapley values to evaluate the importance of each observation's features. To complete this, a power set representing multiple combinations of the model's features is created. Each observation is then evaluated on the basis of the features of the power set's nodes. In order to evaluate feature relevance, the model's results considering and ignoring the feature are

compared. Multiple graphs can then be constructed using the generated values to compare the features and their relevance to different classes of the classification model.

4.5 Results

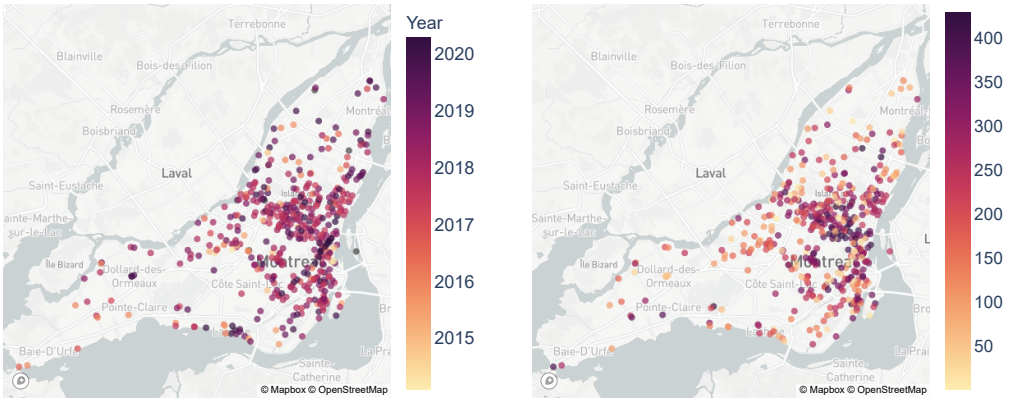
In order to evaluate the proposed methodology, several steps are considered to offer appropriate insight into the relationship between the daily connection profiles of an EVCS and external factors, mainly, PoIs and census data. The following section presents the simulations and an analysis of the results obtained by each step of the process.

4.5.1 Clustering of individual EVCSs

Initially, the daily connection profiles of each of the 450 sites are clustered separately. Depending on the number of ports they are associated with and their installation date, the number of daily connection profiles per site varies between 36 and 3480. While the number of profiles normalized to the number of ports ranges between 18 and 435 sessions, since the majority are installed between 2017 and 2020. More importantly, given the inconsideration of the daily connection profiles for which no sessions are recorded, the normalized number of active daily connection profiles for which at least one session is recorded ranged between 1 and 429. Figure 4.6 details the geographic representation of these numbers.

In order to ensure data granularity while avoiding information loss, a warping window of 1 is considered for the computation of the DTW distances between the sites' daily connection profiles. This is done to constrain the similarities between the temporal sequences. Concurrently, the agglomerative clustering algorithm is evaluated on values of K ranging between $[2, 200]$. This is not only justified by the important number of active daily connection profiles examined for certain EVCSs, but also by the highly stochastic charging behavior. Comparing the number of clusters generated for each station using the four CVIs to different distance thresholds, it can be depicted that elements within each cluster are generally less than 1 DTW unit away from one another. Figure 4.7 presents the clustering results of one station. Considering the centroids of each cluster in black and the elements that make up the clusters in blue, it is apparent that the use of a warping window of 1 reduces the variability in the elements that construct each cluster, and thus prevents the loss of substantial amount of information for the next step.

As a result of the proposed approach, multiple clusters are generated for separate EVCSs. Figure 4.8 details both the number of clusters produced for each EVCS and the median number of daily connection profile per cluster excluding the null cluster for which the daily



(a) Year EVCSs are in service since.

(b) Normalized number of active daily connection profiles to the size of the EVCS (number of ports).

Figure 4.6 Representation of the amount of data available for the different stations considering the installation date and the number of active sessions per EVCS.

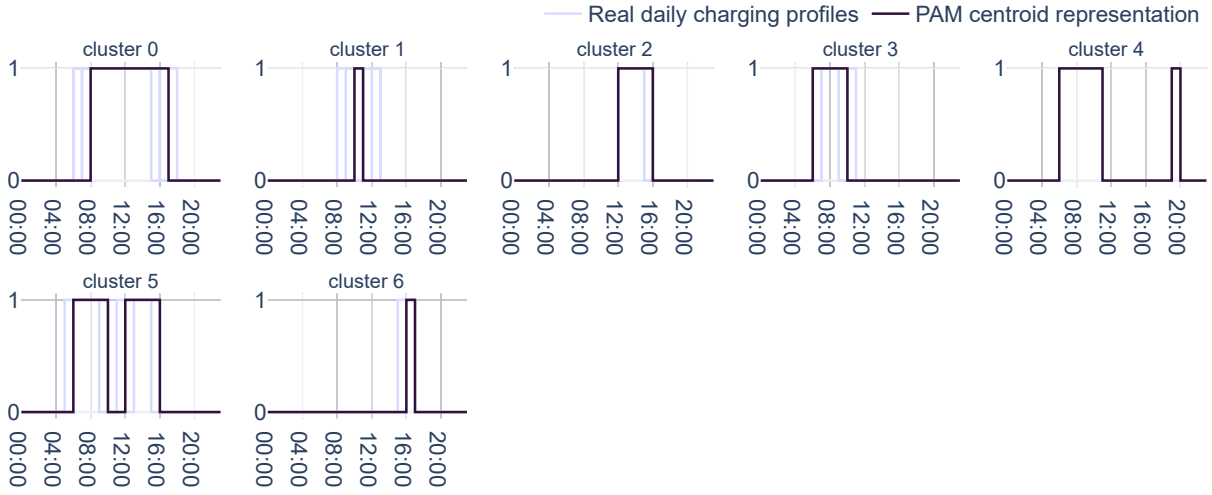
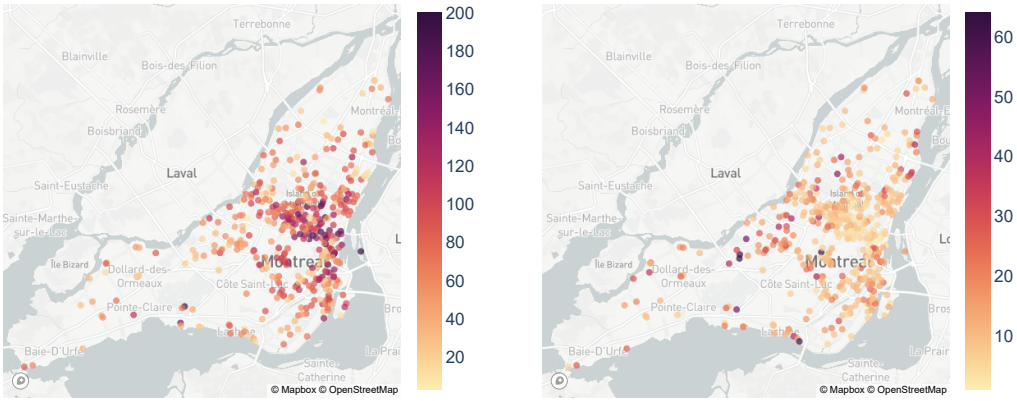


Figure 4.7 Cluster's centroids and the daily charging profiles that make them up of a considered station.

connection profiles have no sessions occuring during the day. Clearly, EVCSs closer to the center of the city (downtown) are more likely to have diverse daily connection profiles. These stations are more likely to have a higher number of clusters. Additionally, the median number of elements per any of their clusters is also generally very low. In fact, certain regions record a median of less than 10 daily connection profiles per cluster. Comparing figure 4.8(a) and figure 4.6(b), it may also be concluded that the higher the number of active daily connection profiles are associated with an EVCS, the higher the number of clusters will be generated



(a) Number of clusters per EVCS. (b) Median number of elements per cluster.

Figure 4.8 Results of the site-level clustering.

for the same station. However, further analysis shows that while a trend exists between EVCSs with less than 1000 active daily connection profiles and their number of clusters, the trend stagnates for EVCSs with higher active daily connection profiles. In fact, the pearson correlation between the two variables increases from 0.77 to 0.86 when removing data on EVCSs with more than 1000 active dqily connection profiles. Further analysis has also shown that the installation date has no effect on the diversity of the possible daily connection profiles.

4.5.2 Clustering of EVCSs' centroids

Considering the aforementioned results, a similar clustering process is followed to investigate the general connection profiles of all EVCSs. To complete this task, the centroids generated in the previous section for each EVCS are aggregated to make up the 32,945 centroids to be evaluated at once to finally create no more than 100 disparate general connection profiles. As opposed to the previous approach where the objective is to create clusters that generalized daily connection profiles the least, this step seeks to reduce the number of final clusters. Thus, alternatively, a warping window of 2 is used for the computation of the distance between the profiles. This complements the performance of the clustering algorithm by reducing and rendering the distance between certain temporal profiles to zero. For values of $k < 100$, the results of using a warping window of 2 are also better than 1. Consequently, considering a range of $k = [2, 100]$, an evaluation of the CVIs results in an optimal value of 14 clusters, regardless of the cluster representing daily connection profiles with no usage.

Figure 4.9 details the various clusters generated from this analysis, which, for clarity, are

referred to on the text as general connection profiles. Each cluster is represented by a numerical value and in parentheses, the proportion of daily connection profiles belonging to it. Many observations can be deduced from this figure. First, EVCSs are mostly used between 8am and 3pm represented by the general connection profile 3. Moreover, whereas day, afternoon and evening sessions are popular, EVCSs are less likely to be used during the entirety of the day or only at night. Finally, assuming DTW's capture of the shape of the daily profiles, it is also assumed that EVCSs are generally used in short periods of 3 to 6 hours.

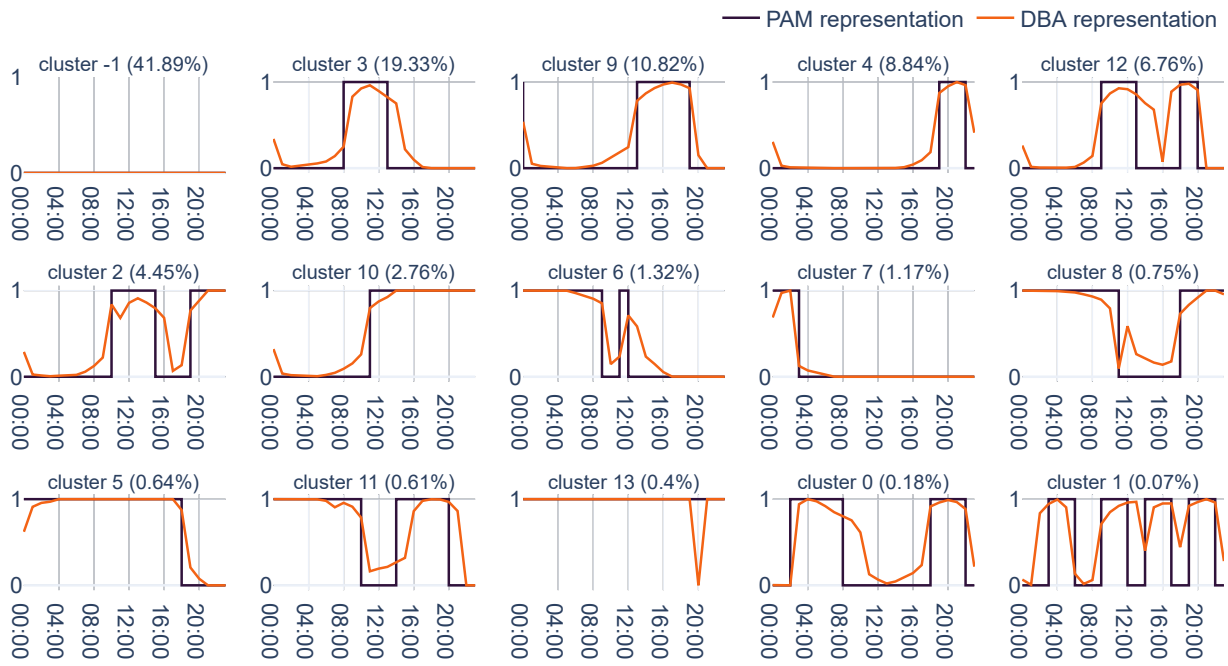
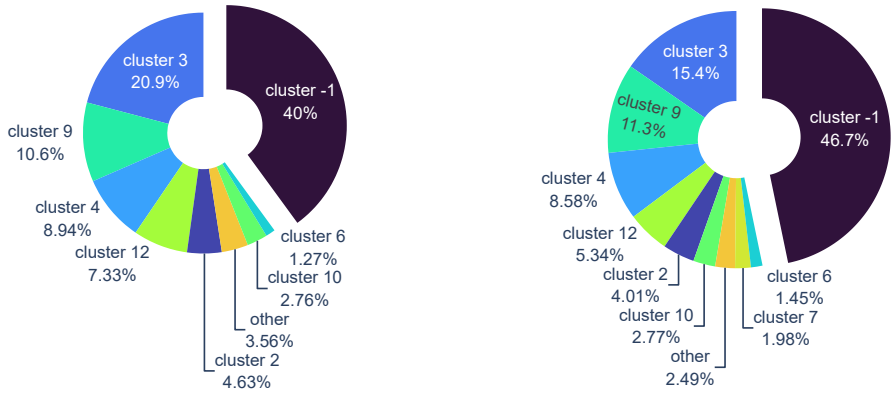


Figure 4.9 Centroids of the general connection profiles.

Further analysis of the temporal importance of different general connection profiles shows little distinction in their order of importance, but an important difference in their proportions as attested in figure 4.10. Indeed, on one hand, considering the type of day evaluated, while daily connection profiles generally belong to general connection profile 3 on business days, the proportion of this general profile is lower on weekends, whereas other profiles, such as general connection profile 7 are more important. Moreover, EVCSs are generally used much less on weekends than on business days.



(a) Proportion of the general connection profiles on business days.

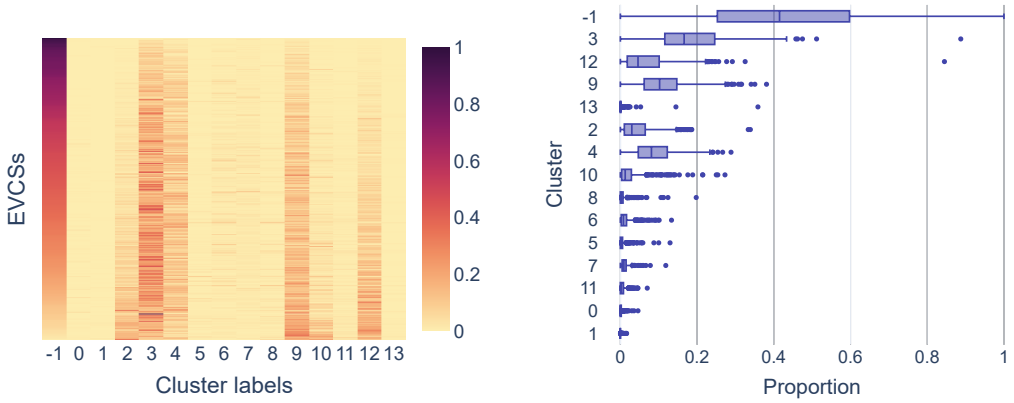
(b) Proportion of the general connection profiles on weekends and days off.

Figure 4.10 Temporal analysis of the proportion of the different general connection profiles.

4.5.3 Analysis of EVCSs' clusters proportions

Given the above general results, a more thorough analysis of the proportions of different general connection profiles is achieved for each EVCS. On one hand, figure 4.11(a) shows the general connection profile proportions through a heat map. This figure details the proportion of general connection profile that makes up each EVCS. On the other hand, figure 4.11(b) details the distribution of the general connection profiles' proportions through a box plot for the different EVCSs. These figures show that despite the low overall proportion of certain general connection profiles, such as profiles 13 or 8, they are very prominent in certain stations. This is not only attested by the isolated horizontal dark lines on the heat map but also by the numerous outliers in the box plots, emphasizing on the variability of different EVCSs' connection behavior trends.

Analyzing the geographic distribution of the general connection profiles' proportions per EVCS in figure 4.12 offers a better idea on their variability. Firstly, stations located around the city, more specifically in residence and industry dense areas, are less likely to be used than EVCSs located in the center of the city as represented by the darker shades of general connection profile -1, where restaurants, shops, offices and high and low residential buildings are less likely to be located. Secondly, while profiles 3 and 9 are more scattered around the map than others, the most prominent profiles in the center are representative of those for which drivers are most likely to charge their vehicles during the day and late in the evening such as profiles 2 and 12. Whereas profiles for which the connection behavior is more focused



(a) Heat map of the proportion of each general connection profile in the x-axis for each EVCS represented in the y-axis.

(b) Distribution of the general connection profile proportions for each EVCS.

Figure 4.11 General connection profile proportion distribution amongst the different EVCSs.

at night and early in the morning, such as profiles 6 and 8 are more popular around the center of the city. Finally, despite having an extremely low overall proportion, profiles 13 and 5, for which the usage behavior is concentrated during the whole day, are most prominent in EVCSs installed in the airport, for which taxi and ride-sharing drivers are most likely their primary customers. Similarly, profile 10 is an essential profile for which the proportion is highest at the airport and regions in the center of the city.

Considering the temporal analysis of the general connection profiles' proportions by taking into account the type of day offers fewer insight into the change in users' behavior. However, as can be seen in figure 4.10, despite minor overall variability of the proportions of profile -1 amongst the stations between business days and weekends, stations in the center and close to the center are generally very likely to be used during business days than on weekends. In fact, certain examined EVCSs are used up to 76% more often on business days than on weekends. Profile 3's proportions on the other hand are noticeably higher on business days in these same EVCSs. Other profiles however have a less noticeable change.

4.5.4 Classification

In order to offer a more thorough analysis of the impact of the geographic situation of the EVCSs on their daily connection profiles, the geographic attributes of EVCSs with similar general connection profile proportions are compared. Using K-Means and elbow methods, a total of 5 geographic partitions are chosen to represent groups of EVCSs with similar general connection profile proportions. Figure 4.13(a) offers a geographic representation

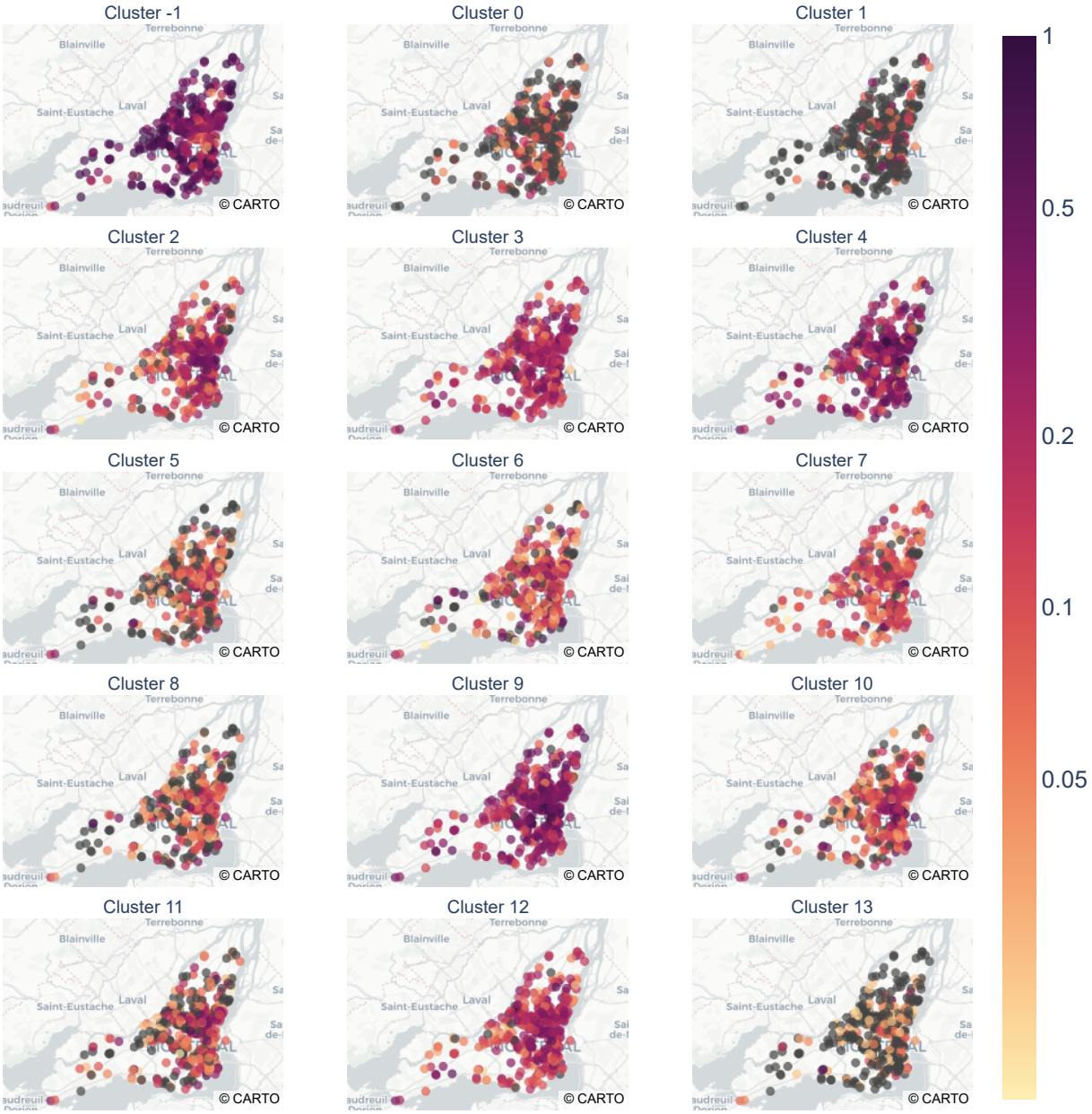
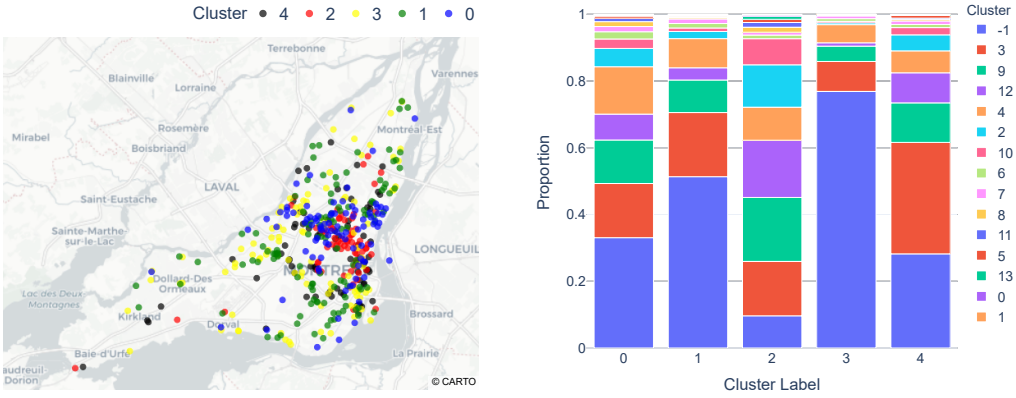


Figure 4.12 Geographic distribution of the proportion of the different clusters where the color is log scaled for ease of analysis.

of the partitions produced, whereas figure 4.13(b) presents the general connection profile proportions of the centers of the geographic partitions.

Despite the presence of a few geographically close EVCSs with similar profiles, a lack of consistency in the geographic distribution of different partitions can be observed. It could highlight the impact the connection behavior of certain EVCSs could have on nearby stations, especially in regions where the supply could possibly outweigh the demand. This observation



(a) Geographic representation of the EVCS clusters produced from K-Means.

(b) Representation of the cluster proportion of each geographic partition.

Figure 4.13 Representation of the geographic clustering of EVCSs.

is noticed particularly in the center which is composed of multiple partitions, especially 2, 4 and 0. These partitions are highlighted by their low proportion of the null general connection profile and higher proportion of the general connection profiles 3, 9, 12, 4 and 2, which generally consist of a more important use of the EVCSs during the day and the evening periods than on other times of the day. On the other hand, EVCSs further away from the center are more likely to belong to the geographic partition 3 and 1 for which the proportion of the null general connection profile is the highest.

A classification model helps to better understand the significance of the geographic partitions through the evaluation of the impact of 33 features representing PoI and demographic data on the connection profiles of the different EVCSs. Considering the labels presented in figure 4.13(a) as output, multiple classification models are generated. Among the three considered algorithms, the k nearest neighbor model considering a maximum distance of 500 meters to other PoIs shows the optimal solution. The test of the the model's weighted f1 score is 0.88, accuracy is 0.88 and log loss is 0.99. Individual labels' f1 scores are also consistent as they ranged between $[0.83, 0.91]$. In order to better understand the degree of impact of various features on the label, the SHAP analysis is conducted. Figure 4.14 shows the most prominent features, where financial services, transportation services and places of worship represent the model's most important attributes in classifying the different EVCSs.

Nonetheless, separate labels are characterized by different attributes as represented by figure E.1. This figure categorizes attributes by their level of importance for each geographic partition, where each attribute is represented by points belonging to the test set's observations. The darker the color of the point, the higher the attribute's value is for the specific

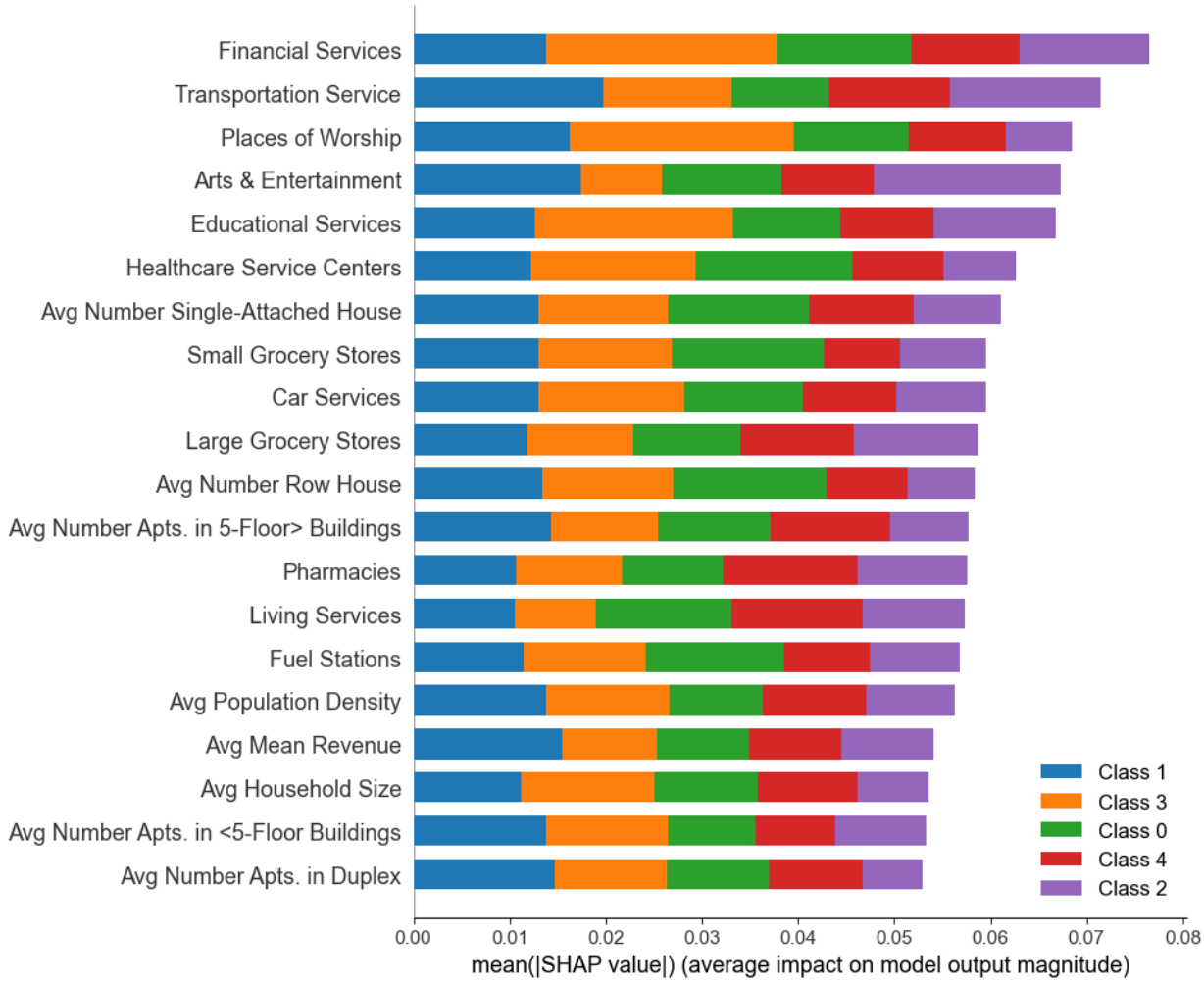


Figure 4.14 Feature importance of the classification model produced for the geographic partitions of the EVCSs.

observation. Moreover, the more positive the SHAP value of the observation, the greater is its contribution to the model's prediction of the geographic partition. Through the evaluating of the figures, various conclusions can be deduced. First, geographic partition 0 is distinguished by a higher number of healthcare services, small grocery stores and average number of single-attached houses. Instead, it is surrounded by less average number of row houses, arts and entertainment venues. Geographic partition 1 on the other hand is discerned by its higher number of worship places and average number of duplex apartments. Like partition 0 however, it is surrounded by less arts and entertainment venues and transportation services. Finally, the average revenue of the population within the radius of the partition's EVCSs is amongst the lowest evaluated. Contrary to other partitions, geographic partition 2's stations are more likely to be surrounded by more arts and entertainment venues, transportation ser-

vices, hotels, financial services, educational services and miscellaneous shops amongst other features. Moreover, the average mean revenue of the population within its stations' radius is amongst the highest. Fourth, geographic partition 3 is characterized by significant lower number of financial services, workship places, healthcare service centers and educational services. The average household size within its stations' radius is typically higher than the average. Finally, the geographic partition 4's stations are commonly surrounded by many offices and industries, transportation services, government offices and have a higher average number of apartments in high-rise buildings with more than 5 floors. On other hand, these stations have access to a fewer pharmacies and living services around.

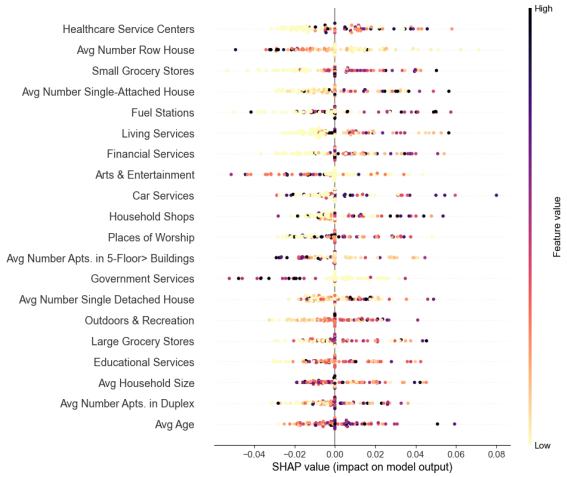
4.6 Conclusion

Considering the increasing number of daily recorded sessions, it is important to better understand the trends in the connection behavior within EVCSs. The proposed approach has permitted to reach various conclusions about these trends from evaluating each EVCS' daily connection profiles separately to delving into the impact of geographic and demographic attributes on EVCSs' connection profiles.

First, the unique analysis of EVCSs' connection profiles has shown the diversity in the connection behavior of multiple stations. While certain EVCSs, particularly those surrounding the center of the city, are rarely used, those located in the center are used more often with varying daily connection profiles. Indeed, not only is their number of clusters higher but the median number of daily connection profiles per cluster is also smaller.

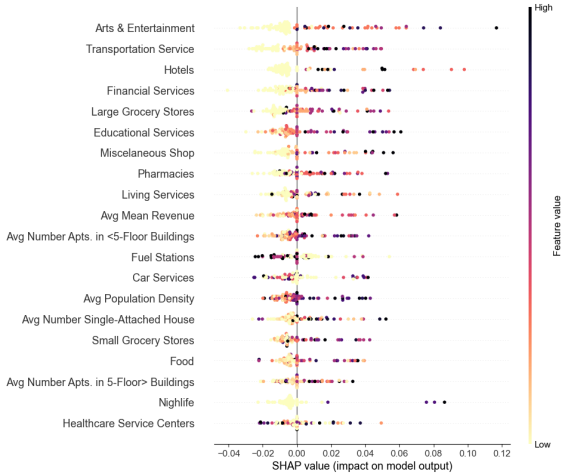
Second, the analysis of the general connection profiles offered a general perspective into the typical connection profiles of EVCSs. Their partition into 15 disparate groups presents daily profiles for which the connection typically occurs in the morning between 8am and 3pm as the most prominent before the null profile, which represents 41.89% of all daily sessions profiles. Daily profiles for which the usage is gathered in the afternoon and the evening, especially between 2pm and 7pm are also very popular. On the other hand, daily profiles for which EVCSs are used throughout the day are less important. Looking at the prominence of each general connection profile based on the type of day has shown a higher popularity of null connection profiles on weekends as opposed to business days. Moreover, a smaller proportion of general connection profile 3 on weekends leads to more distributed proportions amongst other profiles as opposed to business days.

Third, associating the general connection profiles to each EVCS lead to a more appropriate comparison of the profiles of different EVCSs. Indeed, the proportion table constructed from



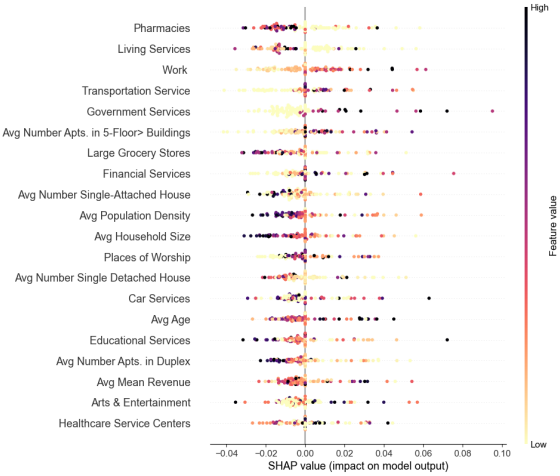
(a) SHAP values of geographic partition 0.

(b) SHAP values of geographic partition 1.



(c) SHAP values of geographic partition 2.

(d) SHAP values of geographic partition 3.



(e) SHAP values of geographic partition 4.

Figure 4.15 SHAP analysis of the classification model of each geographic partition.

this analysis depicted a better understanding of the relevance of different clusters beyond their general proportion amongst the general data. On one hand, general connection profiles which were least popular in the previous analysis were found to have a high proportions in EVCSs located in the airport. This assessment highlights the presence of many outliers amongst general connection profiles, for which the proportion is very high for particular stations, emphasizing on different connection behavior among the EVCSs. On the other hand, a comparison of all general connection profiles has shown that EVCSs located in the center are most likely to have higher connections during the day as opposed to those around the center. While EVCSs daily profiles with connections happening at night and in the morning are more popular around the city.

Finally, segmenting the EVCSs into 5 different partitions for which the proportion of daily connection profiles are most similar allowed for a more thorough understanding of the influence of demographic and geographic attributes on the usage trend of EVCSs. On one hand, the geographic distribution of the geographic partitions showed how dissimilar adjacent EVCSs can be. However, similar to the previous step's conclusion, EVCSs located in the center generally belong to geographic partitions for which the proportion of the null general connection profile is the smallest. On the other hand, the KNN model's SHAP analysis highlighted the importance particular attributes had on the different geographic partitions. As a result, EVCSs belonging to geographic partitions for which the usage normally happens during the day or in the evening are more likely to be surrounded by many offices and industries, transportation services, hotels and arts and entertainment venues. The demographic profile of this same group is a lower average number of row houses, but a higher average number of apartments in high-rise buildings and single attached houses. As opposed to other geographic partitions, which are mainly characterized by a higher average household size. The same analysis also allowed for a better understanding of relevant demographic and geographic attributes in the classification model. Whereas financial, transportation, educational and healthcare services are among the most important features, many demographic attributes like the average age or average revenue are less important.

4.7 Discussion and Future Work

Despite the highly accurate obtained results, valuable discarded information affecting EVCSs' demand may have weakened the final results. Information on the operational capacity of the EVCSs is for instance left behind, which is crucial to understand EVCSs' real potential demand. Moreover, constraints related to the parking space associated with the stations are also unknown. This information is very important since it varies from one station to another

and cannot only limit a user's connection time but also his potential arrival and departure times. Finally, various external factors may obstruct a port or an entire EVCS from being used. Indeed, in periods of high snowfall, multiple street parking spaces can be left unplowed for several consecutive hours or days. Moreover, evaluating randomly picked stations on Google Maps showed the presence of objects blocking the way to EV street parking spaces.

Apart for missing data, available data could have been used more optimally in order to better understand EVCSs' differing profiles. For instance, stations' proximity is an essential information needed to better understand the impact of their demand on one another. Additionally, the approach used can be generalised as a zone-based problem instead of a point-based problem, where the zone is representative of the aggregate demand for EVCSs in a small space in the map. Considering the supply and demand of each zone, this approach can provide important information on the planning of number of EVSCs.

The considered approach presents however numerous opportunities for further study. First machine learning models can be developed for short term and long term prediction of EVCSs' connection profiles using the site-level or centroid-level clusters. This approach is important as it can provide for simpler and more efficient models. This cannot only offer better judgement of day-ahead peak hours, but also helps plan for dynamic pricing of EVCSs. Moreover, a more thorough temporal analysis applied to the the proportions and classification models can offer valuable information on the difference of usage between business days and weekends, but also over different seasons, weeks and days of the week. Finally and most importantly, the presented results can be further used in order to produce a multi-period planning strategy of EVCSs considering the classification model's predicted geographic partition of potential locations. These results can be further used when studying the grid's ability to satisfy the electricity demand in the considered potential locations.

CHAPITRE 5 DISCUSSION GÉNÉRALE

Tandis que l'électrification de nos moyens de transport constitue un pas essentiel vers la réduction des émissions des gaz à effet de serre et donc le ralentissement du réchauffement climatique, elle introduit des défis variés. Notamment, la gestion de l'infrastructure des bornes de recharge de véhicules électriques. Ce défi présente plusieurs préoccupations se relatant à la planification et le fonctionnement optimal des BRVEs. La littérature présente plusieurs méthodes pour s'attaquer à la résolution de ces problèmes à travers diverses approches. Malgré la popularité des modèles d'optimisation se basant sur des méthodes de simulation pour la définition de la demande des BRVEs, un nombre croissant de recherches s'intéressent à l'utilisation des bases de données pour le développement des modèles d'apprentissage statistique. Ces recherches tentent de résoudre des problèmes de planification ou de gestion de la recharge de véhicules en considérant plusieurs attributs importants, et dans certains cas en étudiant des aspects importants sur l'utilisation des BRVEs. Le mémoire suit l'approche de certains articles en utilisant des données réelles sur l'utilisation des BRVEs. L'approche permet d'offrir des renseignements importants sur leur usage et impact. Les données sur les BRVEs installés dans l'île de Montréal sont fournies par Hydro-Québec, elles sont traitées et associées à diverses autres données sur les points d'intérêts et le profil de recensement de la population montréalaise dans une base de données MySQL. Puis, deux axes sont évalués à travers la complétion de différents points importants à savoir :

1. Analyse temporelle de l'utilisation des BRVEs à partir d'un algorithme de regroupement hiérarchique utilisant la distance DTW.
   (a) Analyse des profils de connexion de chaque BRVE et la production de centroïdes représentant les profils de connexion de chaque BRVE ;
   (b) Analyse des centroïdes produits et la production de profils généraux de connexion ;
   (c) Création d'un tableau de proportions des profils généraux de connexion.
2. Développement d'un modèle de classification pour l'évaluation de la pertinence des différents facteurs sur le profil des BRVEs grâce au K-Means et le modèle K-Nearest Neighbors.

A travers cette approche, les stations de la ville de Montréal ont pu être partitionné en plusieurs regroupement représentants des BRVEs pour lesquelles l'utilisation est similaire. D'une part, les profils généraux de connexion qui ont permis de compléter cette analyse ont aussi permis de différencier les stations par leurs périodes d'utilisation. On a remarqué que les stations installées dans le centre-ville possèdent des profils beaucoup plus variés que

celles installées autour. De plus, l'utilisation des BRVEs dans le centre est généralement plus probable en journée et en fin d'après-midi, tandis que les stations autour du centre-ville sont généralement utilisées moins souvent ou notamment le soir et tôt le matin. Enfin en évaluant l'importance des attributs externes sur l'utilisation des bornes, les PoIs se sont démarqués comme attributs pertinents dans la différenciation des bornes basé sur leur profil généraux de connexion. Plus précisément, les services financiers ainsi que les stations de métro et stations de location de bicyclette et les lieux religieux font partie des attributs les plus importants selon l'analyse SHAP. Enfin, selon la partition géographique d'une BRVE, certains attributs peuvent être plus pertinentes que d'autres.

CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS

Au final, plusieurs analyses pertinentes ont pu être produites. Celles-ci ont offerts des conseils importants aux opérateurs et donnent lieu à plusieurs amélioration et opportunités de travaux supplémentaires potentiels.

6.1 Résumé des résultats

En premier lieu, le regroupement séparé des différents sites nous a permis de mieux comprendre les profils de recharge de chaque station. L'approche utilisée a notamment permis, à travers l'utilisation d'un *warping window* restreint et des valeurs de k variées, de produire plusieurs regroupements de profils de connexion. Ces résultats ont particulièrement démontré que les stations installées dans le centre-ville sont non seulement nettement plus utilisées mais ont aussi des profils de connexion plus variés.

En deuxième lieu, le regroupement des centroïdes générés dans l'étape précédente a permis d'interpréter des profils généraux de connexion des stations. En effet, au total, 15 regroupements ont été déduites du modèle dont le profil général de connexion pour lequel l'état est nul tout au long de la journée. Ces résultats ont démontré que les stations ne sont généralement jamais utilisées puisque plus de 41% des profils de connexion sont nuls.

En outre, quand une station est utilisée, il est plus probable que celle-ci soit utilisée entre 8h et 15h ou 14h et 23h puisque ces regroupements représentent plus de 19% et 10% des profils de connexion évalués respectivement. D'autre part, les profils de connexion pour lesquelles l'utilisation est répandue tout au long de la journée sont nettement moins courant. Enfin, en prenant en compte l'aspect temporel, les BRVEs ont une proportion de profils de connexion général nulle nettement plus élevée durant les week-ends qu'en semaine. De plus, la distribution des proportions est plus uniformément distribuée les week-ends qu'en semaine.

La troisième étape étant l'association des profils généraux de connexion aux stations, elle a simplifié la comparaison des profils de connexion entre les BRVEs à travers la génération d'une matrice de proportion. Cette matrice a démontré que certains profils généraux de connexion pour lesquels la proportion est faible, sont en fait très populaires dans des BRVEs clés tels que celles installées à proximité de l'aéroport. Cette observation a aussi démontré la présence de plusieurs valeurs aberrantes compte tenu des proportions des profils des différentes stations, montrant qu'il existe non seulement une diversité de profil de connexion dans chaque station, mais aussi parmi toutes les stations. Enfin, une analyse a démontré que les stations installées

au centre de la ville sont plus susceptibles d'être utilisées durant la journée tandis que les BRVEs installées autour de la ville soient généralement utilisées le matin et le soir. Cette observation souligne la possibilité que les BRVEs installées en milieu résidentiel sont utilisées le soir alors que celles installées dans des régions à activités mixtes (bureaux, commerces, écoles, etc) sont plutôt utilisées la journée. Cette réflexion est étudiée dans la prochaine étape.

Enfin le développement d'une solution de prédiction du profil de connexion des BRVEs a non seulement mené à une meilleure compréhension de groupes de BRVEs possédant des proportions de profils généraux de connexion similaires, mais aussi à l'analyse de facteurs externes influençant le plus les différents regroupements de stations. En premier lieu, le regroupement à partir du K-Means de la matrice de proportions développée dans l'étape précédente a permis de représenter géographiquement les différents groupes de BRVEs ayant des profils de connexion similaires. À partir de cette analyse, il a pu être constaté que les stations installées dans le centre-ville sont généralement les plus utilisées compte tenu de la proportion très basse du profil de connexion général nul. De plus, d'une part, l'un des regroupements les plus populaires dans cette même région a été celui pour lequel le profil de connexion comporte une utilisation entre 8h et 15h. D'autre part un autre regroupement est représenté par des profils de connexion plus variés où la connexion entre 8h et 15h, 13h et 19h, et 9h à 20h. Les stations installées autour du centre-ville cependant font partie de regroupement pour lesquelles le profil de connexion général nul a une proportion très élevée dont l'une avoisine 80% des profils de connexion de chaque station. À l'aide de ces résultats, le modèle de classification KNN a pu être développé et calibré de sorte à évaluer les points d'intérêts et les données de recensement les plus pertinents dans la compréhension des différents regroupements de stations. En effet, à l'aide de l'analyse SHAP plusieurs conclusions ont pu être tirées. Globalement, les attributs affectant le plus l'usage des bornes sont la proximité de la station à un service financier, un métro ou une station de location de bicyclette, un endroit religieux, et un endroit de divertissement ou d'art parmi plusieurs autres attributs. En évaluant le problème par regroupement de stations cependant, l'importance des attributs diffère. En effet, pour les mêmes regroupements de stations pour lesquelles le profil de connexion général est le plus bas, les attributs les plus pertinents sont une proximité à des endroits de divertissement et d'art, des hôtels, des services de transports, de services gouvernementaux et de bureaux. Tandis que les groupes de stations pour lesquels le profil d'usage nul est le plus élevé sont isolés de n'importe quel service et celles-ci sont généralement installées dans des zones où la taille du ménage est élevée. Enfin, les stations pour lesquelles les profils généraux de connexion sont plus uniformément distribués, les attributs tels que le nombre moyen de maisons individuelles attenantes autour de la borne, la proximité à des services religieux et l'éloignement d'endroits de divertissement et d'art ainsi que les services de transport sont les plus pertinents. Grâce

à ces résultats, il est plus simple pour les opérateurs de prendre des décisions concernant la planification des BRVEs ainsi que l'impact de l'utilisation de certaines bornes dans le cas de la construction de restaurants, centre commercial ou centre religieux à proximité de la borne. Cette évaluation servirait aussi à conseiller les opérateurs sur le besoin en électricité ainsi que la profitabilité d'une BRVE potentielle.

6.2 Limitations et travaux futurs

Malgré des résultats complets, plusieurs limitations ont été rencontrées durant le développement d'une solution. Premièrement, plusieurs données pertinentes dans la réalisation d'un modèle de classification robuste manquaient. En effet, des données telles que les restrictions et l'indisponibilité des places de stationnement associées aux BRVEs sont cruciales pour la compréhension de l'usage de certaines BRVEs. Cette information indiquée dans les données extraites du circuit-électrique est pertinente pour mieux comprendre le comportement des usagers de BRVEs. En effet, des restrictions tels que les heures d'ouverture, la durée maximale de stationnement ainsi que le prix additionnel à payer pour la place pourraient être source de biais dans l'usage des BRVEs. De plus, bien que plusieurs BRVEs soient disponibles dans des places de stationnement sur la rue, certaines sont aussi placées dans des parcs de stationnement privés. Enfin, Montréal étant une ville où il neige beaucoup en hiver, le problème du déneigement peut restreindre l'utilisation de certaines bornes. Ainsi, des bornes rarement utilisées en période hivernale pourraient simplement être inaccessibles. Hormis les contraintes naturelles, un BRVEs peut aussi être entravé par un objet tel qu'une benne. C'est une observation qui a été faite à partir de la visualisation de BRVEs à travers le service Google Maps.

Deuxièmement, les données considérées auraient pu être utilisées de manière plus optimale pour mieux représenter les profils de connexion des BRVEs. Un aspect important considéré dans la planification des BRVEs est la proximité des bornes. Cette distance est très importante puisqu'elle définit la demande générale dans une zone. Alors qu'un nombre élevé de BRVEs dans un rayon peut diminuer l'utilisation de certaines bornes, des BRVEs isolées peuvent avoir l'effet opposé. La représentation de la distance entre les différentes BRVEs en tant qu'attribue aurait donc pu fournir une information pertinente dans le modèle de Classification.

Enfin, l'algorithme DTW a aussi présenté certaines restrictions dans la qualité des résultats. En effet, la complexité de l'algorithme nous a restreints à utiliser un pas de temps d'une heure pour la définition des profils de connexion. Cette approche généralise cependant les sessions de chaque borne en approximant les heures de début et de fin d'utilisation des bornes, ainsi

que la durée de connexion. L'utilisation d'un pas de 15 minutes aurait peut-être donné une meilleure représentation des profils de connexion des BRVEs.

Hormis les limitations, plusieurs travaux complémentaires pourraient être achevés. Premièrement, des modèles d'apprentissage statistique pourraient être utilisés pour réaliser les prédictions court-terme et long-terme des profils de connexion des BRVEs à partir des regroupements générés dans la première et deuxième étape du premier axe. Les prédictions de cette approche simple pourraient servir non seulement à la planification stratégique des BRVEs, mais aussi être utilisées pour l'approximation court-terme des périodes de pointes, et donc aideront à l'instauration des services V2G ou des tarifications dynamiques. Deuxièmement, une analyse temporelle plus approfondie pourrait être achevée pour la conception de la matrice proportions ainsi que dans le développement de modèles de classification en considérant les week-ends et jours ouvrables séparément. Enfin, les résultats du modèle de classification pourraient être utilisés pour le développement d'un plan d'expansion de l'infrastructure de BRVEs dans la ville de Montréal, en considérant les facteurs externes clés, le profil de connexion des différentes stations ainsi que l'infrastructure électrique et la capacité du réseau à satisfaire la demande en électricité des BRVEs potentielles.

RÉFÉRENCES

[1] "Climate change evidence : How do we know ?" Oct 2021. [En ligne]. Disponible : https://climate.nasa.gov/evidence/

[2] S. Kartha *et al.*, "The carbon inequality era : An assessment of the global distribution of consumption emissions among individuals from 1990 to 2015 and beyond," 2020. [En ligne]. Disponible : http://hdl.handle.net/10546/621049

[3] "Climate change and health," Feb 2018. [En ligne]. Disponible : https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

[4] "Paris agreement." [En ligne]. Disponible : https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en

[5] IEA, "Tracking transport 2020," 2020. [En ligne]. Disponible : https://www.iea.org/reports/tracking-transport-2020

[6] Iea, "Global electric car stock, 2010-2019 –nbsp ;charts – data amp ; statistics." [En ligne]. Disponible : https://www.iea.org/data-and-statistics/charts/global-electric-car-stock-2010-2019

[7] StatisticsCanada, "Table 20-10-0021-01 new motor vehicle registrations," 2021.

[8] IEA, "Global ev outlook 2021," 2021. [En ligne]. Disponible : https://www.iea.org/reports/global-ev-outlook-2021

[9] "New vehicle rebate." [En ligne]. Disponible : https://vehiculeselectriques.gouv.qc.ca/english/rabais/ve-neuf/programme-rabais-vehicule-neuf.asp

[10] F. Schulz et J. Rode, "Public charging infrastructure and electric vehicles in norway," *Energy Policy*, vol. 160, p. 112660, 2022. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0301421521005255

[11] T. Franke *et al.*, "Experiencing range in an electric vehicle : Understanding psychological barriers," *Applied Psychology*, vol. 61, n°. 3, p. 368–391, 2012. [En ligne]. Disponible : https://iaap-journals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1464-0597.2011.00474.x

[12] C. Csiszár *et al.*, "Location optimisation method for fast-charging stations along national roads," *Journal of Transport Geography*, vol. 88, p. 102833, 2020. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0966692319303801

[13] M. Kchaou-Boujelben et C. Gicquel, "Locating electric vehicle charging stations under uncertain battery energy status and power consumption," *Computers*

Industrial Engineering, vol. 149, p. 106752, 2020. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360835220304691

[14] J. He *et al.*, "An optimal charging station location model with the consideration of electric vehicle's driving range," *Transportation Research Part C : Emerging Technologies*, vol. 86, p. 641–654, 2018. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0968090X17303558

[15] L. Pan *et al.*, "A location model for electric vehicle (ev) public charging stations based on drivers' existing activities," *Sustainable Cities and Society*, vol. 59, p. 102192, 2020. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S2210670720301797

[16] A. A. Kadri *et al.*, "A multi-stage stochastic integer programming approach for locating electric vehicle charging stations," *Computers Operations Research*, vol. 117, p. 104888, 2020. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0305054820300058

[17] Y. Zhang *et al.*, "Review of the electric vehicle charging station location problem," dans *Dependability in Sensor, Cloud, and Big Data Systems and Applications*, G. Wang *et al.*, édit. Singapore : Springer Singapore, 2019, p. 435–445.

[18] X. Bai, K.-S. Chin et Z. Zhou, "A bi-objective model for location planning of electric vehicle charging stations with gps trajectory data," *Computers Industrial Engineering*, vol. 128, p. 591–604, 2019. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360835219300105

[19] C. Csiszár *et al.*, "Urban public charging station locating method for electric vehicles based on land use approach," *Journal of Transport Geography*, vol. 74, p. 173–180, 2019. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S096669231830471X

[20] J. Jordán *et al.*, "Localization of charging stations for electric vehicles using genetic algorithms," *Neurocomputing*, vol. 452, p. 416–423, 2021. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0925231220316556

[21] W. Kong *et al.*, "Optimal location planning method of fast charging station for electric vehicles considering operators, drivers, vehicles, traffic flow and power grid," *Energy*, vol. 186, p. 115826, 2019. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360544219314987

[22] C. Bian *et al.*, "Finding the optimal location for public charging stations – a gis-based milp approach," *Energy Procedia*, vol. 158, p. 6582–6588, 2019,

innovative Solutions for Energy Transitions. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1876610219300803

[23] M. Straka *et al.*, "Predicting popularity of electric vehicle charging infrastructure in urban context," *IEEE Access*, vol. 8, p. 11 315–11 327, 2020.

[24] ——, "Analysis of energy consumption at slow charging infrastructure for electric vehicles," *IEEE Access*, vol. 9, p. 53 885–53 901, 2021.

[25] S. Das, P. Acharjee et A. Bhattacharya, "Charging scheduling of electric vehicle incorporating grid-to-vehicle and vehicle-to-grid technology considering in smart grid," *IEEE Transactions on Industry Applications*, vol. 57, n°. 2, p. 1688–1702, 2021.

[26] F. A. Hashim *et al.*, "Henry gas solubility optimization : A novel physics-based algorithm," *Future Generation Computer Systems*, vol. 101, p. 646–667, 2019. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0167739X19306557

[27] M. S. Hashim *et al.*, "Priority-based vehicle-to-grid scheduling for minimization of power grid load variance," *Journal of Energy Storage*, vol. 39, p. 102607, 2021. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S2352152X21003492

[28] Y. Shen *et al.*, "Ev charging behavior analysis using hybrid intelligence for 5g smart grid," *Electronics*, vol. 9, n°. 1, 2020.

[29] J. R. Helmus, M. H. Lees et R. van den Hoed, "A data driven typology of electric vehicle user types and charging sessions," *Transportation Research Part C : Emerging Technologies*, vol. 115, p. 102637, 2020.

[30] E. Xydas *et al.*, "A data-driven approach for characterising the charging demand of electric vehicles : A uk case study," *Applied Energy*, vol. 162, p. 763–771, 2016.

[31] K. Miyazaki, T. Uchiba et K. Tanaka, "Clustering to predict electric vehicle behaviors using state of charge data," dans *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe)*, 2020, p. 1–6.

[32] C. Sun *et al.*, "Classification of electric vehicle charging time series with selective clustering," *Electric Power Systems Research*, vol. 189, p. 106695, 2020.

[33] A. S. Al-Ogaili *et al.*, "Review on scheduling, clustering, and forecasting strategies for controlling electric vehicle charging : Challenges and recommendations," *IEEE Access*, vol. 7, p. 128 353–128 371, 2019.

[34] S. Shahriar *et al.*, "Machine learning approaches for ev charging behavior : A review," *IEEE Access*, vol. 8, p. 168 980–168 993, 2020.

[35] C. Ratanamahatana et E. J. Keogh, “Making time-series classification more accurate using learned constraints,” dans *SDM*, 2004.

[36] Y. Amara-Ouali *et al.*, “A review of electric vehicle load open data and models,” *Energies*, vol. 14, p. 2233, 2021.

[37] [En ligne]. Disponible : https://lecircuitelectrique.com/en/

[38] “About us.” [En ligne]. Disponible : https://www.hydroquebec.com/data-center/about.html

[39] A. Lampert, “Hydro-québec, canada’s largest electricity producer, enters fast-growing energy storage business,” Dec 2020. [En ligne]. Disponible : https://financialpost.com/commodities/energy/canadas-biggest-electricity-producer-enters-energy-storage-market-2

[40] FourSquarePlaces, “Venue categories.” [En ligne]. Disponible : https://developer.foursquare.com/docs/build-with-foursquare/categories/

[41] D. Arthur et S. Vassilvitskii, “K-means++ : The advantages of careful seeding,” dans *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA ’07. USA : Society for Industrial and Applied Mathematics, 2007, p. 1027–1035.

[42] A. A. Mueen et E. J. Keogh, “Extracting optimal performance from dynamic time warping,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[43] S. Salvador et P. K.-F. Chan, “Fastdtw : Toward accurate dynamic time warping in linear time and space,” 2004.

[44] A. Mueen *et al.*, “Awarp : Fast warping distance for sparse time series,” dans *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, p. 350–359.

[45] O. Arbelaitz *et al.*, “An extensive comparative study of cluster validity indices,” *Pattern Recognition*, vol. 46, nº. 1, p. 243–256, 2013. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S003132031200338X

[46] S. Lee *et al.*, “Clustering of time series water quality data using dynamic time warping : A case study from the bukhan river water quality monitoring network,” *Water*, vol. 12, nº. 9, 2020. [En ligne]. Disponible : https://www.mdpi.com/2073-4441/12/9/2411

[47] G. Bottaz-Bosson *et al.*, “Continuous positive airway pressure adherence trajectories in sleep apnea : Clustering with summed discrete fréchet and dynamic time warping dissimilarities,” *Statistics in Medicine*, vol. 40, nº. 24, p. 5373–5396, 2021. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9130

[48] A. Sardá-Espinosa, “Time-series clustering in r using the dtwclust package,” *The R Journal*, 2019.

[49] S. M. Lundberg et S.-I. Lee, “A unified approach to interpreting model predictions,” dans *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA : Curran Associates Inc., 2017, p. 4768–4777.

[50] E. Keogh, J. Lin et W. Truppel, “Clustering of time series subsequences is meaningless : implications for previous and future research,” dans *Third IEEE International Conference on Data Mining*, 2003, p. 115–122.

[51] J. Aitchison, “The statistical analysis of compositional data,” *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 44, n°. 2, p. 139–160, 1982. [En ligne]. Disponible : https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1982.tb01195.x

[52] J. J. Egozcue, vol. 35, n°. 3, p. 279–300, 2003. [En ligne]. Disponible : https://doi.org/10.1023/a:1023818214614

[53] M. Tsagris, S. Preston et A. T. A. Wood, “Improved classification for compositional data using the -transformation,” vol. 33, n°. 2, p. 243–261, juill. 2016. [En ligne]. Disponible : https://doi.org/10.1007/s00357-016-9207-5

[54] IEA, “Global ev outlook 2020,” 2020. [En ligne]. Disponible : https://www.iea.org/reports/global-ev-outlook-2020

[55] TransportCanada, “Projected annual zev sales,” jan 2020. [En ligne]. Disponible : https://tc.canada.ca/en/road-transportation/innovative-technologies/zero-emission-vehicles#/find/nearest?country=CA

[56] W. J. Requia *et al.*, “How clean are electric vehicles ? evidence-based review of the effects of electric mobility on air pollutants, greenhouse gas emissions and human health,” *Atmospheric Environment*, vol. 185, p. 64–77, 2018.

[57] Y. Kim et S. Kim, “Forecasting charging demand of electric vehicles using time-series models,” *Energies*, vol. 14, n°. 5, 2021. [En ligne]. Disponible : https://www.mdpi.com/1996-1073/14/5/1487

[58] M. Rouzbahman, A. Jovicic et M. Chignell, “Can cluster-boosted regression improve prediction of death and length of stay in the icu ?” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, n°. 3, p. 851–858, 2017.

[59] M. Chaouch, “Clustering-based improvement of nonparametric functional time series forecasting : Application to intra-day household-level load curves,” *IEEE Transactions on Smart Grid*, vol. 5, n°. 1, p. 411–419, 2014.

[60] X. Wu *et al.*, “A novel fast-charging stations locational planning model for electric bus transit system,” *Energy*, vol. 224, p. 120106, 2021.

[61] K.-F. Chu, A. Y. S. Lam et V. O. K. Li, “Joint rebalancing and vehicle-to-grid coordination for autonomous vehicle public transportation system,” *IEEE Transactions on Intelligent Transportation Systems*, p. 1–14, 2021.

[62] W. Wu *et al.*, “Online ev charge scheduling based on time-of-use pricing and peak load minimization : Properties and efficient algorithms,” *IEEE Transactions on Intelligent Transportation Systems*, p. 1–15, 2020.

[63] P. Morrissey, P. Weldon et M. O’Mahony, “Future standard and fast charging infrastructure planning : An analysis of electric vehicle charging behaviour,” *Energy Policy*, vol. 89, p. 257–270, 2016.

[64] S. k. Shen, W. Liu et T. Zhang, “Load pattern recognition and prediction based on dtw k-mediods clustering and markov model,” dans *2019 IEEE International Conference on Energy Internet (ICEI)*, 2019, p. 403–408.

[65] T. Teeraratkul, D. O’Neill et S. Lall, “Shape-based approach to household electric load curve clustering and prediction,” *IEEE Transactions on Smart Grid*, vol. 9, n°. 5, p. 5196–5206, 2018.

[66] T. Han *et al.*, “A pattern representation of stock time series based on dtw,” *Physica A : Statistical Mechanics and its Applications*, vol. 550, p. 124161, 2020. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0378437120300157

[67] P. van der waerden, H. Timmermans et M. de Bruin-Verhoeven, “Car drivers’ characteristics and the maximum walking distance between parking facility and final destination,” *Journal of Transport and Land Use*, vol. 10, n°. 1, Sep. 2015. [En ligne]. Disponible : https://www.jtlu.org/index.php/jtlu/article/view/568

[68] D. J. Berndt et J. Clifford, “Using dynamic time warping to find patterns in time series,” dans *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, ser. AAAIWS’94. AAAI Press, 1994, p. 359–370.

[69] J. Paparrizos et L. Gravano, “K-shape : Efficient and accurate clustering of time series,” *SIGMOD Rec.*, vol. 45, n°. 1, p. 69–76, juin 2016. [En ligne]. Disponible : https://doi.org/10.1145/2949741.2949758

[70] F. Petitjean, A. Ketterlin et P. Gançarski, “A global averaging method for dynamic time warping, with applications to clustering,” *Pattern Recognition*, vol. 44, n°. 3, p. 678–693, 2011. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S003132031000453X

[71] P. J. Rousseeuw, “Silhouettes : A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, p. 53–65, 1987. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/0377042787901257

[72] D. L. Davies et D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, n°. 2, p. 224–227, 1979.

[73] M. Kim et R. Ramakrishna, “New indices for cluster validity assessment,” *Pattern Recognition Letters*, vol. 26, n°. 15, p. 2353–2363, 2005. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S016786550500125X

[74] C. Crozier, T. Morstyn et M. McCulloch, “Capturing diversity in electric vehicle charging behaviour for network capacity estimation,” *Transportation Research Part D : Transport and Environment*, vol. 93, p. 102762, 2021.

ANNEXE A MODÈLE ENTITÉ-ASSOCIATION

Modèle Entité-Association détaillant la base de données finale utilisée pour la représentation des données d'Hydro-Québec ainsi que toute donnée externe pertinente à l'étude tel que les PoIs et données de recensement.

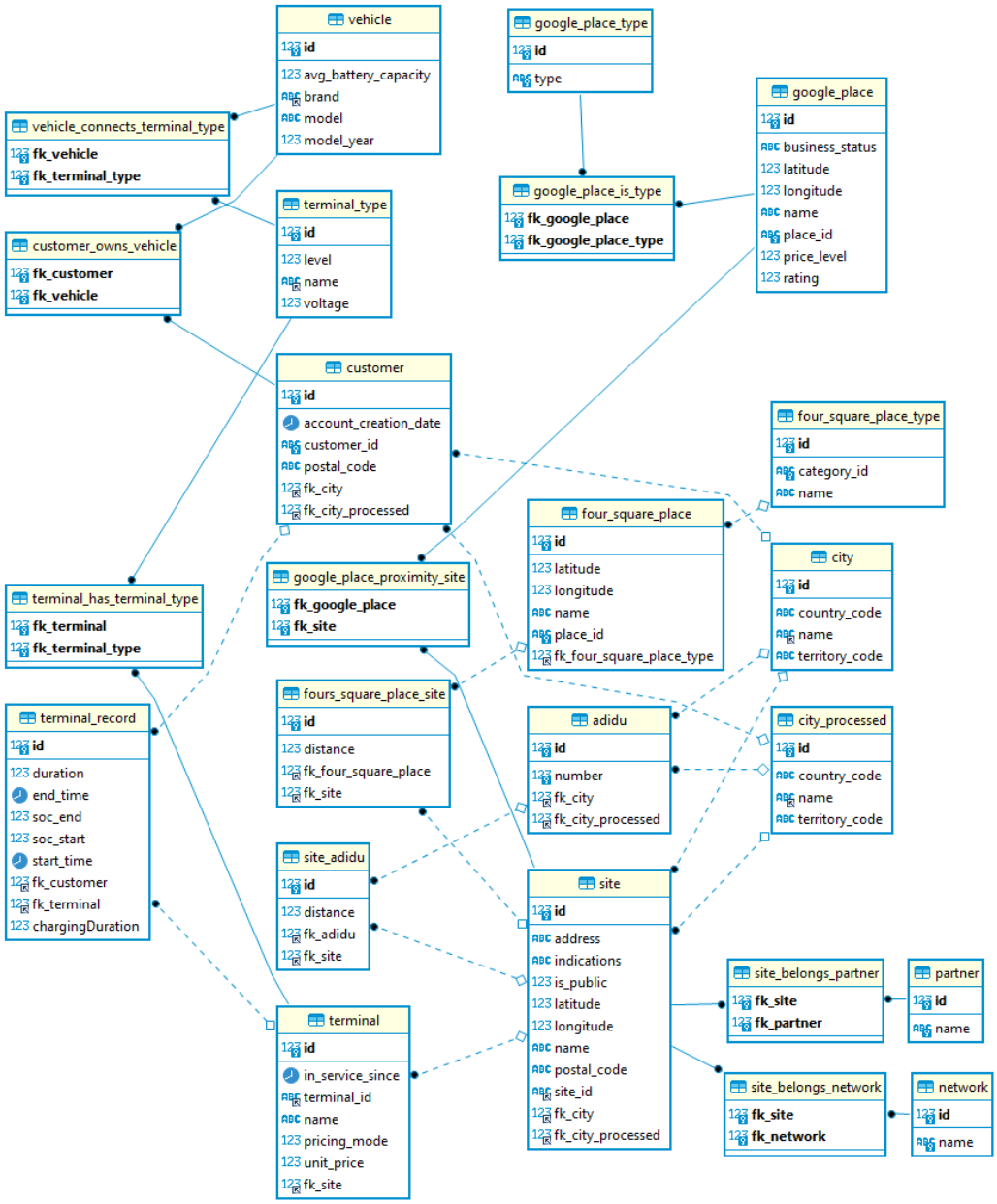


Figure A.1 Modèle Entité-Association de la base de données finale utilisée.

ANNEXE B SCHÉMA DES DONNÉES CAPTURÉES DU CIRCUIT-ÉLECTRIQUE SUR LES STATIONS

Le schéma qui suit représente la réponse de la requête POST pour la deuxième étape de collecte de données, soit la collecte de données sur les stations.

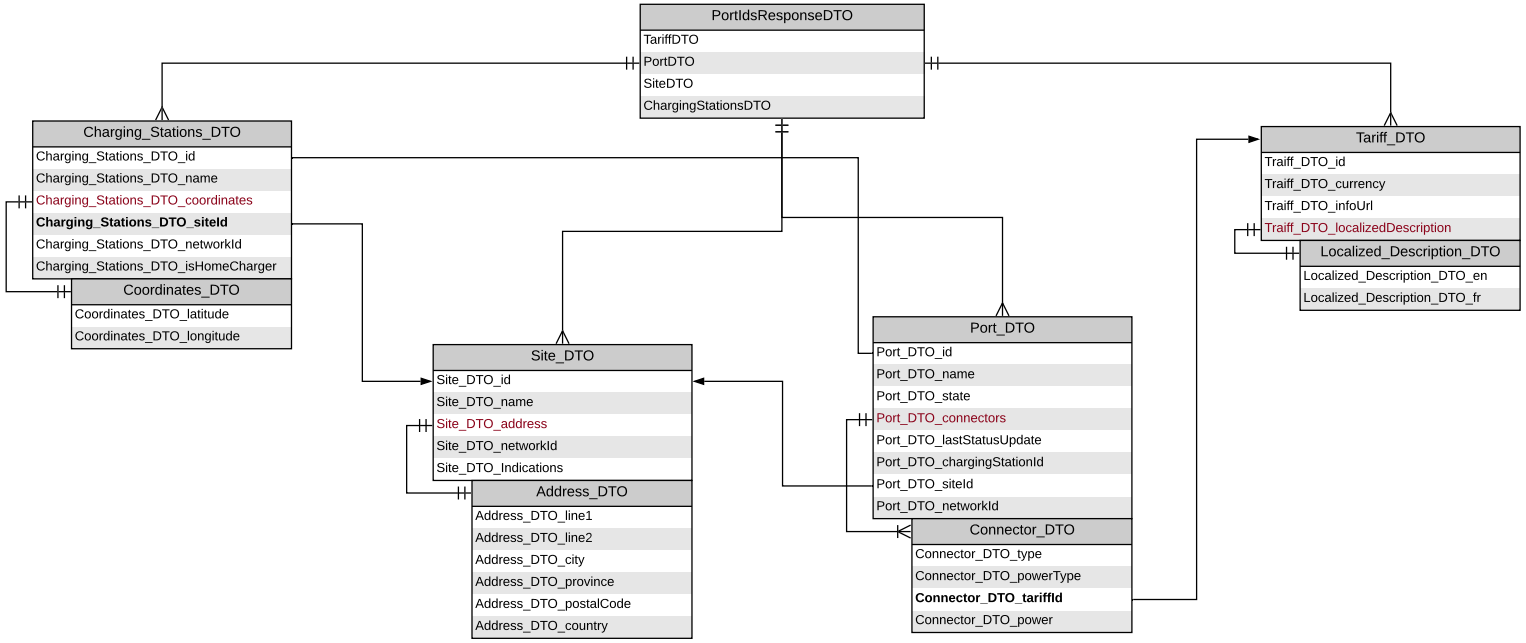


Figure B.1 Schéma d'Objet de transfert de données (DTO) des données collectées du site du circuit-électrique.

ANNEXE C STRUCTURE DES DONNÉES D'HYDRO-QUÉBEC

Structure des données extraites de la base de données d'Hydro-Québec

Tableau C.1 Description de l'entité *Parks*

Attribut	Description	Format	Traitement
Hourly Rate	Frais par heure	Numérique	
Id	Identifiant de la station	Chaîne de caractères	
Initial Fee	Frais initiaux	Numérique	
IsInDemandResponse		Booléen	
LastUpdated	Date la plus récente à laquelle l'information sur la station a été mis à jour	Date et Heure	
Name	Nom de la station	Chaîne de caractères	
OwnerId	Identifiant du propriétaire de la station	Chaîne de caractères	
TerritoryCode	Code de la province où est situé la station	Chaîne de 2 caractères	
TimeZone	Fuseau horaire utilisée	Numérique (0 ou -5)	Combinée à la valeur des dates d'arriver et de départ des sessions de différentes BRVEs pour définir leur fuseau horaire

Tableau C.2 Description de l'entité *Vehicles*

Attribut	Description	Format	Traitement
BatteryCapacity	Capacité de la batterie	Numérique	
Brand	Marque du véhicule	Chaîne de caractères	
ConnectorTypes	Compatibilité aux connecteurs	Chaîne de caractères	Lié à l'entité des connecteurs
Id	Identifiant unique de l'utilisateur	Numérique	
Model	Modèle du véhicule	Chaîne de caractères	
UserId	identifiant unique de l'utilisateur	Chaîne de caractères	
Year	Année du véhicule	Numérique	

Tableau C.3 Description de l'entité *Owners*

Attribut	Description	Format	Traitement
Id	Identifiant du propriétaire de la borne	Chaîne de caractères	
Name	Nom du propriétaire	Chaîne de caractères	
Alias		Chaîne de caractères	
LastUpdated	Date où l'entrée a été mise à jour pour la dernière fois	Date et Heure	

Tableau C.4 Description de l'entité *Members*

Attribut	Description	Format	Traitement
Id	Identifiant de l'utilisateur	Chaîne de caractères	
AccountCreationUtc	Date de création du compte	Date et Heure	
Language	Langage choisit par l'utilisateur	Chaîne de caractères	
City	Ville où le compte a été créé	Chaîne de caractères	
PostalCode	Code Postal de l'utilisateur	Chaîne de caractères	
ActiveCards	Nombre de carte associé à l'utilisateur	Numérique	
Currency	Devise de paiement	Chaîne de caractères	
NotifyOnLowbalance	Liée à l'utilisation de l'application	Booléen	
NotifyOnRechargeSessionCompleted	Liée à l'utilisation de l'application	Booléen	
NotifyOnRechargeSessionInterrupted	Liée à l'utilisation de l'application	Booléen	
Province	Province de l'utilisateur	Chaîne de caractères	

Tableau C.5 Description de l'entité *Stations*

Attribut	Description	Format	Traitement
Id	Identifiant de la borne	Chaîne de caractères	
ParkId	Identifiant de la station	Chaîne de caractères	
Name	Nom de la borne	Chaîne de caractères	
isPublic	Identifie si la borne est disponible au public	Booléen	
Longitude	Coordonnées géographique (Longitude)	Numérique	
Latitude	Coordonnées géographique (Latitude)	Numérique	
Level	Niveau de puissance	Chaîne de caractères	
Firmware	Nom du type de borne	Chaîne de caractères	
Voltage	Voltage de la borne	Numérique	
Amperage	Ampérage de la borne	Numérique	
City	Ville où la borne est installée	Chaîne de caractères	
TerritoryCode	Territoire où la borne est installée	Chaîne de caractères	
PostalCode	Code Postal où la borne est installée	Chaîne de caractères	
CreationDate	Date d'installation de la borne	Date et heure	
VisibleOnMap		Booléen	
Address	Adresse où la borne est installée	Chaîne de caractères	
Country	Pays où la borne est installée	Chaîne de caractères	
PricingMode	Si les frais sont par heure	Booléen	
Price	Frais de la recharge	Numérique	
TimeZone	Fuseau horaire	Numérique	
DayTimeSaving	Si l'heure d'été est considérée	Booléen	
InServiceSince	Date depuis que la borne est utilisée	Date & Heure	
LastUpdated	Dernière date à laquelle la borne a été mise à jour	Date & Heure	

Tableau C.6 Description de l'entité *Charging Sessions*

Attribut	Description	Format	Traitement
ID session d'utilisation	Identifiant de la session	Chaîne de caractères	
Borne	Nom unique de la borne	Chaîne de caractères	Les caractères sont rendues minuscule
Parc	Nom unique de la station	Chaîne de caractères	
Partenaire			
Durée Session (s)	Durée de la session	Numérique	
Energie kWh			
ID carte	Identifiant de la carte utilisée pour initier le chargement	Chaîne de caractères	
Réseau	Nom du réseau auquel appartient la station	Chaîne de caractères	Catégorisé
Connecteur	Type de bornes (Puissance)	Chaîne de caractères	Catégorisé
Coût	Coût de la recharge	Numérique	
Devise	Devise du coût	Chaîne de caractères	Catégorisé
Borne Type Name	Nom du type de borne	Chaîne de caractères	
Type	Type de borne	Chaîne de caractères	
StationId	Identifiant de la borne	Chaîne de caractères	
ParkId	Identifiant de la station	Chaîne de caractères	
Début session d'utilisation	Date et heure de début de session	Date & Heure	Divisé en date et heure de la session
Fin session d'utilisation	Date et heure de fin de session	Date & Heure	Divisé en date et heure de la session
StartSoc	Pourcentage de recharge du véhicule au début de la session	Numérique	
EndSoc	Pourcentage de recharge du véhicule à la fin de la session	Numérique	
Temps_recharge	Temps de recharge du véhicule (Cette valeur est égal ou inférieur au temps de connexion)	Numérique	

ANNEXE D DISTRIBUTION DES POINTS D'INTÉRÊTS DANS L'ÎLE DE MONTRÉAL

Distribution des points d'intérêts considérés dans la base de données dans une carte de la ville de Montréal.

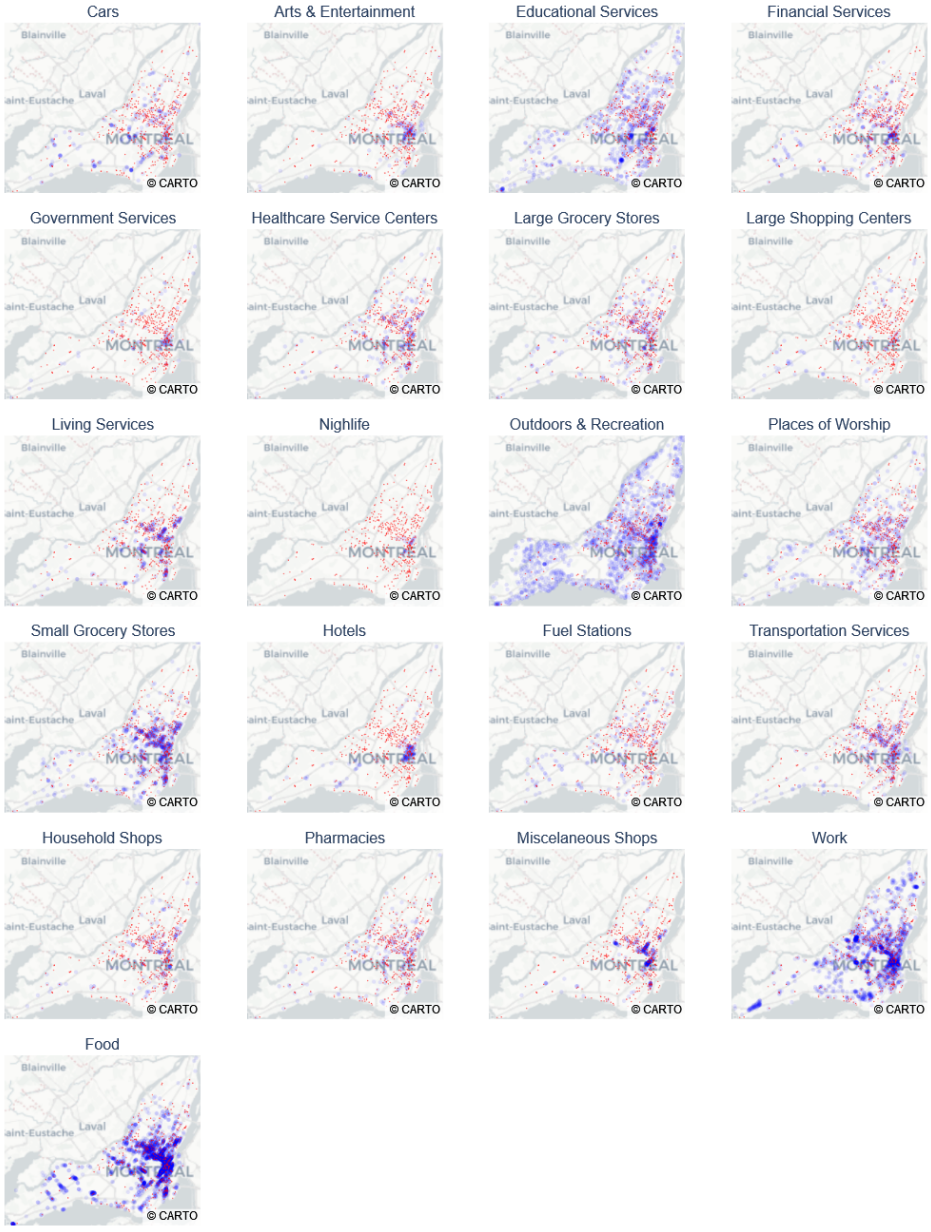
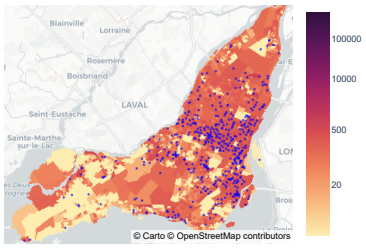


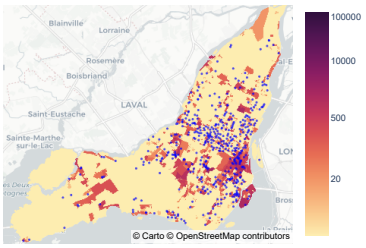
Figure D.1 Distribution des points d'intérêts dans la carte

ANNEXE E DISTRIBUTION DES DONNÉES DE RECENSEMENT DANS L'ÎLE DE MONTRÉAL

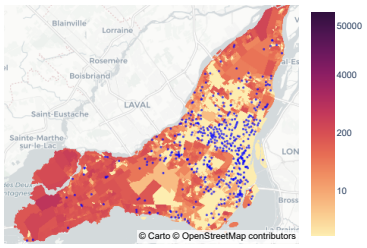
Représentation de données de recensement de chaque aire de diffusion avec en bleu ou en jaune les stations publics considérées dans la recherche.



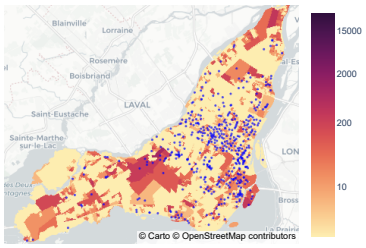
(a) Nombre moyen d'appartements dans un immeuble de moins de cinq étages



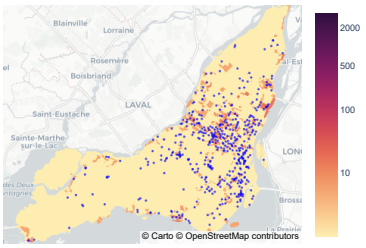
(b) Nombre moyen d'appartements dans un immeuble de cinq étages ou plus



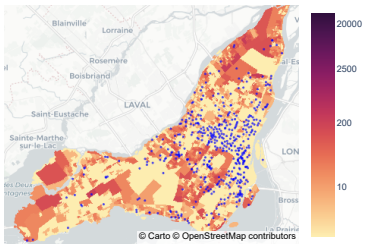
(c) Nombre moyen de maisons individuelles non attenantes



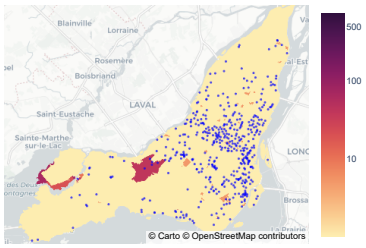
(d) Nombre moyen de maisons en rangées



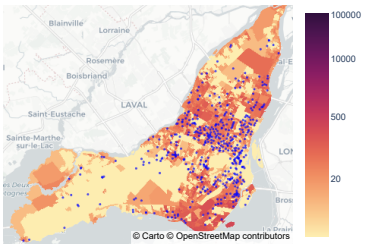
(e) Nombre moyen d'autre maisons individuelles attenantes



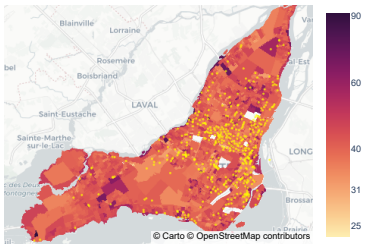
(f) Nombre moyen de maisons jumelées



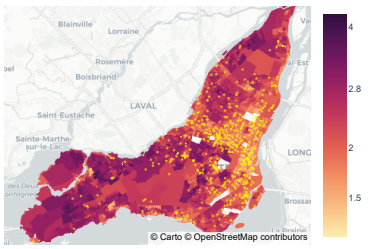
(g) Nombre moyen de logements mobiles



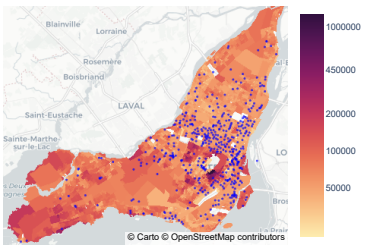
(h) Nombre moyen d'appartements ou plain-pieds dans un duplex



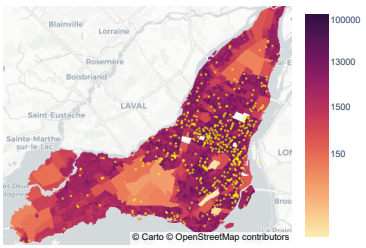
(i) Âge moyen de la population



(j) Taille moyenne des ménages privés



(k) Revenu total moyen avant impôts



(l) Densité de la population au kilomètre carré

Figure E.1 Distribution des données de recensement dans la carte