Titre: Title:	Spatio-temporal modeling of taxi trips
Auteur: Author:	Forouz Alahyari Fard
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Alahyari Fard, F. (2022). Spatio-temporal modeling of taxi trips [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10328/

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Spatio-temporal modeling of taxi trips

FOROUZ ALAHYARI FARD

Département des génies civil, géologique et des mines

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

Génie civil

Avril 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé:

Spatio-temporal modeling of taxi trips

présenté par **Forouz ALAHYARI FARD**

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

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

Francesco CIARI, président

Geneviève BOISJOLY, membre et directrice de recherche

Catherine MORENCY, membre et codirectrice de recherche

Martin TRÉPANIER, membre

DEDICATION

To my supervisors, Geneviève Boisjoly and Catherine Morency,

For their endless support,

And for the precious opportunity they gave to me,

And to all inspiring women in science…

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Geneviève Boisjoly, who was always supportive, from the first moment I started my Master at Polytechnique Montreal. I would like to express my deepest gratitude to her for her valuable time, guidance, patience, and for all the times she listened to me, and supported me, especially during the hard times of pandemic. I would also like to express my deepest appreciation to my co-supervisor, Catherine Morency, whose dedication, competence and advice were always inspiring to me. I would like to thank her for being such a supportive professor who always guided me and answered my questions. I am so grateful for the precious opportunity you both gave to me, to do my Master in such an amazing research group, and for trusting me, and helping me to believe in myself. I would be always thankful for your endless support, and for all you taught me during these two years.

I would also like to thank Bureau du taxi de Montréal and Revenu Québec for their financial support.

Furthermore, I would like to thank Elodie Deschaintres, for answering my questions, and for helping me with approaches of clustering in this study. I thank all my colleagues in office B-344. It was so unfortunate that pandemic did not let us work together more. I would also like to thank members of the transportation research group at Polytechnique Montreal who helped me during these two years: Anjeli Narrainen, Hamed Alizadeh, Jean-Simon Bourdeau, Vincent Obry-Legros, Yves Darly Mathieu, Yasser Amiour, Pénélope Renaud-blondeau and other colleagues.

I would also like to thank my roommate and cousin, Naz Alahyari, for having my back during difficult times. Times during pandemic would be much more difficult without you. Thanks for making those difficult times as tolerable as possible. I was so lucky to have you as my roommate. I would also like to thank my cousin, Elhaam, for her endless support. Thank you Elhaam for being always by my side, even when you were thousands of kilometers far from me. I am so lucky to have you in my life.

A great thank also to everyone who helped me before and during this journey. I would mention some as: Hanie Ehteshami, Mahtab Yaghouti, Chakavak Atrchian, Saba Hesam, Pardis Malekzadeh, Mojdeh Sharafi, Elmira Mirhallaj, Roghaye Fazeli, Kimberly Salt, Ehsan Yazdanpanah, Dr. Meysam Effati and everyone who helped me.

Finally, I would thank my mom and dad, for giving me the opportunity to follow my dreams. I know how hard it is to let your child live kilometres far away from you. I will always be grateful for your support. And thanks to my sister and brother, Forough and Farzan for being by my side from my childhood. This journey would not have been possible without any of you.

RÉSUMÉ

Les taxis sont une composante essentielle du système de transport urbain, car ils complètent les différents modes de transport public. Contrairement aux modes transport collectif tels que le métro ou les bus, qui fonctionnent selon des itinéraires et des horaires fixes, les taxis n'ont pas d'horaires, d'itinéraires ou d'arrêts à respecter. En outre, dans les zones métropolitaines, le service de taxi est devenu un mode de transport populaire, en raison de sa rapidité et de son service porte-à-porte, de sa flexibilité et de son caractère planifiable.

Une meilleure compréhension de la demande de taxis est nécessaire pour aider le secteur à améliorer l'offre de taxis afin de mieux répondre aux besoins des passagers de manière à ce que le taxi soit complémentaire aux autres modes de transport tels que le bus et le métro. En outre, la compréhension de la variation temporelle et spatiale de la demande de taxis peut aider les chauffeurs à fournir un service aux passagers au bon endroit et au bon moment. Cela peut empêcher les chauffeurs d'errer dans les rues pour trouver des passagers, ce qui entraîne une augmentation des embouteillages et de la consommation d’essence.

Selon la littérature, plusieurs facteurs peuvent expliquer la demande de taxis, tels que l'utilisation du sol, la conception des routes, l'environnement urbain, les arrêts de bus, la longueur des routes et, surtout, la météo. Deux raisons principales font de la prévision de la demande de taxis une tâche difficile. Premièrement, de nombreux facteurs qui expliquent la demande de taxi varient spatialement et temporellement, tout comme la demande de taxi qui fluctue dans l'espace et dans le temps. La deuxième raison est que le taxi est souvent utilisé pour des déplacements irréguliers et occasionnels, ce qui rend la prévision de son utilisation difficile.

En tant que telle, cette recherche vise à atteindre les objectifs suivants pour donner un aperçu de la méthodologie de la modélisation de la demande de taxi:

1. Exploration de l'association entre les variables qui varient spatialement et temporellement et les déplacements en taxi dans l'île de Montréal
2. Caractérisation de la variabilité temporelle hebdomadaire de la demande de taxi au niveau des secteurs de recensement (SR) et élaboration d’une typologie des semaines
3. Élaboration d’une typologie multi-semaines pour comparer l'utilisation du taxi entre les SRs au cours de plusieurs semaines

4. Identification des facteurs pouvant expliquer la typologie multi-semaines et de tester des modèles pour prédire à quels types multi-semaines les SRs appartiennent.

Une revue de la littérature présente tout d'abord les études existantes en matière de demande de taxi. Ainsi, la revue de la littérature présente les différentes variables qui ont été incluses dans différents modèles pour expliquer les fluctuations spatiales et temporelles de la demande de taxis. Ensuite, les différentes méthodologies existantes dans la littérature pour prévoir la demande de taxis sont présentées, y compris les algorithmes d'apprentissage automatique et les modèles statistiques.

Différentes analyses de données exploratoires sont ensuite menées pour mieux comprendre les fluctuations temporelles et spatiales de l'utilisation des taxis à l'aide des données du Bureau du taxi de Montréal (BTM), ainsi que les déterminants potentiels de la demande de taxis.

Après l'analyse exploratoire des données, des modèles typiques d'utilisation des taxis sont testés à l'aide des données de taxi de BTM sur 12 semaines d'avril, juillet et septembre 2019. Pour ce faire, cette recherche s'appuie sur l'analyse typologique des k-moyennes pour caractériser la variabilité temporelle hebdomadaire de la demande de taxis et développer des typologies de semaines. Ainsi, des vecteurs "SR-semaine" sont créés. Ces vecteurs comprennent des indicateurs de dispersion quotidienne (proportion de déplacements en taxi par jour de la semaine) et un indicateur d'intensité (moyenne quotidienne des déplacements en taxi). Les vecteurs sont normalisés pour être utilisés dans l'approche de l'analyse typologique des k-moyennes afin de développer des typologies hebdomadaires d'utilisation des taxis. Cinq types de modèles hebdomadaires sont identifiés. Ensuite, un indicateur d'entropie de Shannon est utilisé pour trouver les SRs avec des modèles temporels réguliers et irréguliers d'utilisation des taxis.

Dans l'étape suivante, une approche regroupement hiérarchique agglomératif est appliquée pour développer des typologies multi-semaines. Dans cette étape, des vecteurs "SR" sont créés sur la base de la séquence ordonnée des types observés sur 12 semaines. En d'autres termes, cette analyse conduit à comparer les SRs en fonction de leur séquence de patrons d'utilisation hebdomadaire des taxis (basée sur les types hebdomadaires précédemment développés). Ainsi, six types multi-semaines sont développés à partir des cinq types de semaines qui ont été développés à partir du premier type de vecteur.

Le dernier chapitre de cette étude applique une analyse par arbre de décision et un modèle logit multinomial (MNL) pour explorer les facteurs qui peuvent influencer l'appartenance aux différents types de multi-semaines et pour explorer si ces deux modèles peuvent prédire à quels types de multi-semaines les SRs appartiennent.

Cinq types de semaines sont introduits à la suite de l'application de l'analyse typologique des k-moyennes. Ces types de semaines sont les suivants : déplacements le jeudi et le vendredi (forte demande), déplacements le vendredi (faible demande), déplacements en semaine (faible demande), déplacements le jeudi et déplacements le week-end. Les résultats révèlent que le quatrième type de semaine (déplacements du jeudi) est le plus fréquent. L'analyse spatiale de la typologie de semaine développée indique que les SRs avec une forte proportion de modèles de semaine "déplacements le jeudi et le vendredi (W1: forte demande)" contiennent ou entourent l'aéroport et les zones industrielles. Les SRs avec une forte proportion de patrons hebdomadaires "déplacements du jeudi (W4)" sont concentrés dans les quartiers centraux. On constate également que les SRs ayant le pourcentage le plus élevé de patrons hebdomadaires "déplacements le week-end (W5)" sont concentrés dans le centre-ville. En ce qui concerne la régularité de l'utilisation du taxi, en utilisant l'indicateur d'entropie de Shannon, on constate que les SRs contenant et entourant l'aéroport, le centre-ville, le Plateau Mont-Royal, Westmount et le Mile-End présentent la variation temporelle la plus régulière de la demande de taxi sur les 12 semaines. D'autre part, on constate que les SRs présentant des variations irrégulières sont dispersés sur l'île de Montréal.

L'analyse temporelle de la typologie de semaine développée sur les 12 semaines d'analyse révèle que les patrons de demande de taxi varient au cours de l'année, car la distribution des SRs dans les différents clusters n'est pas stable temporellement. En ce qui concerne le patron de semaine "déplacements le jeudi (W4)", ce patron est prédominant dans la semaine de l'année comprenant le jour férié du Vendredi saint. La proportion de SRs avec le patron hebdomadaire "déplacements le jeudi et le vendredi (W1: forte demande)" augmente en été. On constate également que la proportion de SRs avec le patron "déplacements en semaine (W3: faible demande)" diminue remarquablement dans les semaines qui incluent les jours fériés de Pâques et de la fête du Canada.

La mise en grappes des séquences a donné lieu à six types multi-semaines. Parmi les six groupes multi-semaines proposés, deux présentent un patron complètement différent pour les semaines d'avril par rapport aux semaines de juillet et de septembre (types 3 et 5). Le type dominant

comprend principalement des déplacements le jeudi, et plus de 95 % des semaines dans le type le moins fréquent correspondent à W1 qui était des déplacements le jeudi et le vendredi (forte demande).

En ce qui concerne les résultats de la modélisation, certaines variables communes ont été identifiées avec le modèle de l'arbre de décision et le modèle MNL comme des déterminants clés du type multi-semaine. Il s'agit de la moyenne horaire des services de bus quotidien, de la densité des restaurants et du nombre de voitures par habitant. Les analyses révèlent également qu'il existe d'importantes difficultés à prédire le type de MW avec les variables indépendantes utilisées dans cette étude.

ABSTRACT

Taxis are an essential component of the urban transport system as they complement the various modes of public transport. Unlike public transport modes such as metro or buses, which operate based on fixed routes and scheduled times, taxis do not adhere to any schedules, routes, or stops. Furthermore, in metropolitan areas, taxi service has become a popular mode of transportation, due to its fast and door-to-door service, flexibility, and plannability.

Better understanding the taxi demand is necessary to help the industry improve the supply of taxis to better meet the needs of passengers in a way that taxi is complementary to other modes of transportation such as bus and metro. Furthermore, understanding the temporal and spatial variation of taxi demand can help drivers to provide service to passengers at the right location and time. This can prevent drivers from wandering the streets to find passengers which leads to more congestion and gasoline consumption.

According to the literature, several factors can explain taxi demand, such as land use, road design, urban environment, bus stops, road length and importantly weather. Two main reasons have made forecasting taxi demand a difficult task. Firstly, many of the factors which explain the taxi demand vary both spatially and temporally, similar to the taxi demand which fluctuates over space and time. The second reason is that taxi is often used for irregular and occasional trips, which makes forecasting the pattern of its usage a difficult task.

As such, this research aims to achieve the following objectives to provide insights into the methodology of the taxi demand modeling:

1. Exploring the association between factors which vary spatially and temporally and taxi trips on the Island of Montreal.
2. Characterizing the weekly temporal variability of taxi demand at the census tract (CT) level and developing a week typology.
3. Developing a multi-week typology to compare taxi usage over multiple weeks among CTs.
4. Identifying the factors which can explain the multi-week typologies and testing models to predict which multi-week clusters the CTs belong to.

A review of the literature first introduces the existing studies in terms of the taxi demand. Thus, the literature review presents different variables which were included in different models for

explaining spatial and temporal fluctuations of taxi demand. Then, different methodologies existing in the literature for forecasting taxi demand are presented, including both machine learning algorithms and statistical models.

Different exploratory data analyses are then conducted to better understand the temporal and spatial fluctuations of taxi usage using data from Bureau du taxi de Montréal (BTM), and the potential determinants of taxi demand.

After the exploratory data analysis, typical patterns of taxi usage are explored using the taxi data from BTM over 12 weeks of April, July, and September 2019. For this purpose, this research relies on k-means clustering to characterize the weekly temporal variability of taxi demand and develop week typologies. Thus, "CT-week" vectors are created. These vectors include daily dispersion indicators (proportion of taxi trips per day of the week) and an intensity indicator (daily mean of taxi trips). The vectors are normalized for being used in the approach of k-means clustering to develop week typologies of taxi usage. Five types of weekly patterns are identified. Then, an indicator of Shannon entropy is used to find CTs with regular and irregular temporal patterns of taxi usage.

In the next step, a hierarchical agglomerative clustering approach is applied to develop multi-week typologies. In this step, "CT" vectors are created based on the ordered sequence of the clusters observed over 12 weeks. In other words, this analysis leads to comparing CTs based on their sequence of weekly taxi usage patterns (based on previously developed weekly clusters). Hence, six multi-week clusters are developed based on the five types of weeks which were developed based on the first type of vector.

The last chapter of this study applies a decision tree analysis and multinomial logit model (MNL) for exploring the factors which can influence belonging to the different multi-week types and to explore if these two models can predict which multi-week clusters the CTs belong to.

Five types of weeks are introduced as a result of applying k-means clustering. These types of weeks are the following: end of week trips (high demand), Friday trips (low demand), weekday trips (low demand), Thursday trips and weekend trips. The results reveal that the 4th type of week (Thursday trips) is the most frequent one. The spatial analysis of the developed week typology indicates that the CTs with a high proportion of "end of the week trips (W1: high demand)" week patterns contain or surround the airport and industrial zones. The CTs with a high proportion of "Thursday trips

(W4)" week patterns are concentrated in central neighborhoods. It is also found that CTs with the highest percentage of "weekend trips (W5)" week patterns are concentrated downtown. In terms of the regularity of taxi usage, by using the indicator of Shannon entropy, it is found that CTs containing and surrounding the airport, downtown, Plateau Mont-Royal, Westmount and Mile End have the most regular temporal variation of taxi demand across the 12 weeks. On the other hand, it is found that CTs with irregular patterns are dispersed across the Island of Montreal.

The temporal analysis of the developed week typology across the 12 weeks of analysis reveals that the taxi demand patterns vary across the year, as the distribution of CTs in the different clusters is not stable temporally. With respect to the "Thursday trips (W4)" week patterns, this pattern is predominant in the week of the year includes the holiday of good Friday. The proportion of CTs with the "end of week trips (W1: high demand)" week pattern increases during summer. It is also found that the proportion of CTs with the "weekday trips (W3: low demand)" pattern decreases remarkably in weeks which include holidays of Easter and Canada day.

The clustering of the sequences resulted in six multi-week clusters. Among the six proposed multi-week clusters, two have a completely different pattern for weeks of April comparing to weeks of July and September (cluster 3 and 5). The dominant cluster mostly includes Thursday trips, and more than 95% of weeks in the least frequent cluster corresponds to W1 which was end of week trips (high demand).

With respect to the modeling results, some common variables were identified with the decision tree model and the MNL model as key determinants of the multi-week type. These include average of daily bus services, eating places and number of cars per population. The analyses also reveal that there are important difficulties in predicting the MW type with the independent variables used in this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANOVA	Analysis of variance
ARIMA	Autoregressive integrated moving average
BTM	Bureau du taxi de Montréal
CART	Classification and regression trees
CTree	Conditional inference tree
CT	Census tract
DA	Dissemination area
DBSCAN	Density-based spatial clustering of applications with noise
GAMM	Generalized additive mixed model
GTWR	Geographical and temporal weighted regression
GWR	Geographically weighted regression
LDA	Latent Dirichlet Allocation
NYC	New York city
OLS	Ordinary least squares
POI	Points of interest
SAAQ	Société de l'assurance automobile du Québec
STM	Société de transport de Montréal
TAZ	Traffic analysis zone

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Definition of taxi

Taxis provide different services, and they include several types of vehicles and institutional frameworks which make their definition variable depending on city they belong to (Lacombe, 2016). Since in this research the taxi demand of Montreal is explored, the definition of taxi is presented according to the province of Quebec.

The Assemblée Nationale du Québec (2019) defines a taxi as an automobile which is equipped with a lantern and taximeter and used to transport people on a regular basis, when the fare is calculated, under any circumstances or upon request from the customer, according to the rates established by the Commission des transports du Québec.

The term used for "taxi" is not the same in all cities in Canada and in other countries. For example, English speaking cities use different names for vehicles based on the service they provide, and the way that they are ordered. The provincial law on taxi transport provides the regulations for stakeholders such as license owners, drivers and intermediaries. It also provides information in terms of the required permit, obligations, pricing, inspections and the entities regulating the environment (Lacombe, 2016).

1.2 History of taxi

The appearance of the taxi industry goes back to the mid-seventeenth century in Paris and London, where horse-drawn carriages were operating as taxis to the public (Cooper, Mundy, & Nelson, 2010 as cited by Lacombe, 2016; Taxi Fare Finder, 2012). In Canada, the first traces of the existence of taxi services dates to 1837, in the city of Toronto (Lacombe, 2016). However, Harding, Kandlikar, and Gulati (2016) as cited by Laviolette (2017) mentioned that in the 1930s, when the U.S. was severely affected by the Great Depression, the government began to regulate taxis and fares as we know them today.

During the Great Depression, the taxi industry faced several market failures. For example, lack of information made it impossible for riders to compare the price and service quality of vehicles, which created a poor service quality. Over-competition also appeared due to the insufficient entry barriers in the taxi market which triggered drivers to behave aggressively and unsafely, and it also

led to congestion, poor vehicle maintenance and the appearance of illegal taxis (Harding et al., 2016; Rayle, Shaheen, Chan, Dai, & Cervero, 2014; Schaller, 2007). This situation led some American and Canadian cities such as New York, Chicago, Boston, Toronto, and Montreal to make some regulations in terms of fares, quality of vehicles and the system of permits (Haider, 2015; Pautler & Frankena, 1984).

The regulation of the taxi industry continued in North American cities until 1978, when a few American cities such as Atlanta, San Diego and Portland decided to retest the deregulation (Saponaro, 2013). This was imposed to reduce market entry restrictions and increase competition and innovation in services (Schaller, 2007). The idea to try again deregulation was inspired by some deregulation policies enacted in the transportation industry in 1978 and 1980 (Saponaro, 2013). In North American cities, deregulation of the taxi industry was followed by a significant rise in taxi supply, which impacted competition and entrepreneurship positively in some cities and negatively in others, resulting in low quality services and higher fares (Haider, 2015).

Nevertheless, deregulation was not limited to North American cities. In 1989 and 1990, New Zealand and Sweden imposed deregulations to the taxi industry. Thereafter, deregulation spread across Europe, and reached several cities such as Brussels, London, Oslo and several Irish and Dutch cities (Organisation for Economic Co-operation and Development, 2007).

1.3 Problem statement

Taxis are an essential component of the urban transport system as they complement the various modes of public transport. Unlike public transport modes such as metro or buses, which typically operate on the basis of fixed routes and schedules, taxis do not adhere to any schedules, routes, or stops. Furthermore, in metropolitan areas, taxi service has become a popular mode of transportation, due to its fast and door-to-door service, flexibility, and plannability (Q. Liu, Ding, & Chen, 2020; Rodrigues, Martins, Kalakou, & Moura, 2020).

According to the literature, several factors can explain the taxi demand, such as land use, road design (Q. Liu et al., 2020), urban environment, bus stops, road length (Zhang, Huang, & Zhu, 2019) and importantly weather (Kamga, Yazici, & Singhal, 2015; Q. Liu et al., 2020). Five main reasons have made forecasting taxi demand a difficult task. Firstly, many of the factors mentioned above vary both spatially and temporally, similarly to taxi demand which fluctuates over space and

time. The second reason is that taxi is often used for irregular and occasional trips which makes forecasting the taxi usage a difficult task. Additionally, there is a lack of literature on predicting taxi demand by simultaneously considering spatial and temporal fluctuations of taxi usage. There is a lack of research regarding the taxi demand in countries such as the United States, Canada, and Australia due to the affordability of private cars among most households (Conway, Salon, & King, 2018). Another reason is the difficulty of collecting taxi GPS trace data in large scale (Q. Liu et al., 2020). Finally, to better understand the taxi industry, it is critical to identify which groups of people use taxis the most; however, no demographic data is available about these users. Overall, given the complexity of forecasting the demand, more efforts are first required to understand and characterize the demand and the potential explanatory factors.

As mentioned earlier, taxi is an important mode of transportation since it can complement and supplement other modes of transportation. Better understanding the demand for taxi trips is necessary to help the industry improve the supply of taxis to better meet the needs of passengers in a way that is complementary to other modes of transportation such as bus and metro. Furthermore, understanding the temporal and spatial variations of taxi demand help drivers and taxi companies to provide service to passengers at the right location and time. This can prevent drivers from wandering the streets to find passengers which leads to more congestion and gasoline consumption (Lacombe, 2016). However, planning a taxi service able to respond to the needs of the passengers in an optimal manner is a complex task.

1.4 Objectives

The principal objective of this research is to characterize the patterns of taxi usage on the Island of Montreal, at the census tract level (CT) and to identify factors which can explain these patterns. Finally, this research aims to test models for predicting patterns at the CT level.

This study aims to achieve the following objectives:

1. Exploring the association between factors which vary spatially and temporally and taxi trips on the Island of Montreal.
2. Characterizing the weekly temporal variability of taxi demand at the CT level and developing a week typology.
3. Developing a multi-week typology to compare taxi usage over multiple weeks among CTs.

4. Identifying the factors which can explain the multi-week typologies and testing models to predict which multi-week clusters the CTs belong to based on their characteristics.

Overall, these four objectives contribute to developing a methodology to better understand how taxi demand varies over time and space and to examine the relationships between taxi demand and explanatory factors. Assessing the patterns and existing relationships is a first step to unravel the complexity of taxi demand and thereby provides insights for future taxi demand forecasting work.

1.5 Thesis structure

The first chapter of this document is an introduction to the taxi industry. It provides information about the definition of taxi and the history of the taxi industry. Then the importance of the research topic is explained in addition to the challenges of this research. Thereafter, the objectives of the research are presented.

Chapter 2 presents a review of the literature in terms of taxi demand. It provides information regarding explanatory factors which explain taxi demand, and different existing methodologies in the literature for modeling taxi demand.

Chapter 3 presents the data and methodology. In this chapter, the information regarding the data including their sources, and steps of data preparation are presented. Then, the general methodology applied in this project is presented. K-means clustering is used as an exploratory analysis for better understanding the temporal pattern of every single week in terms of taxi demand. Then hierarchical agglomerative clustering is applied to compare multi-week patterns among CTs since CTs may not always have the same weekly pattern. Two more methodologies are described in this chapter including decision tree analysis and multinomial logit model. These two models are applied to achieve objective 4 mentioned in the previous section.

Chapter 4 aims to achieve objective 1 by providing the descriptive analyses of taxi trips and the exploratory factors which are identified as potential key determinants to have significant impact on taxi demand according to the literature.

Chapter 5 achieves objective 2 and 3. It provides exploratory analysis of temporal pattern of taxi usage in two main sections. First, it explores the weekly temporal variability of taxi usage which leads to the development of a week typology. The second section focuses on developing a multi-

week typology using the week typology produced in the first section by considering the ordered sequences of clusters (week types) representing the twelve weeks.

Chapter 6 presents a machine learning algorithm and a statistical approach to identify different characteristics of CTs which can explain belonging to a specific multi-week cluster (objective 4). In other words, chapter 6 describes methods for determining which pattern of taxi usage a CT with certain characteristics belongs to.

Chapter 7 concludes this research project by summarizing its key findings. It also highlights the contributions of this study in terms of both literature review and methodology. The limitations of this research are also mentioned. Finally, the perspectives are noted to build on the current project.

CHAPTER 2 LITERATURE REVIEW

This chapter reviews the most recent research on the taxi industry, and more specifically the most important work on taxi demand. First, the existing literature is reviewed regarding different spatial and temporal factors which explain taxi demand. Then, different methodologies used in the literature for forecasting taxi demand, including both machine learning algorithms and statistical approaches are explored.

2.1 Determinant factors of taxi demand

According to the previous literature, several factors can explain taxi demand in both space and time. This section provides information regarding explanatory variables from different categories which explain spatial and temporal fluctuations of taxi demand. For this purpose, the variables can be grouped in four categories: socio-demographic, land use, transportation, and temporal variables.

2.1.1 Socio-demographic factors

Socio-demographic factors play an important role in variation of taxi demand. Previous studies have namely examined the impact of income on taxi demand. The results indicated that the correlation between income and taxi demand may be positive or negative. According to one study, there is a positive correlation between income and taxi trips meaning that more taxi trips are made by higher-income individuals (Lacombe, 2016; C. Yang & Gonzales, 2014). In another study, Qian and Ukkusuri (2015b) applied two models for assessing the spatial variation of taxi ridership including global model and geographically weighted regression (GWR). In the global model, they found that median income had a negative correlation with taxi ridership, which could be due to the fact that people with higher income may have their private car which reduces the probability of using taxi. This finding is inconsistent with previous research (C. Yang & Gonzales, 2014), which concluded that taxis are more affordable for higher-income individuals, so income and taxi are positively associated. On the other hand, GWR model resulted in the remarkable impact of geographical location on the sign of the coefficient for median income, so income may have positive or negative impact on taxi ridership (Qian & Ukkusuri, 2015b). Housing price is also found to have a positive impact on taxi usage. X. Liu, Sun, Sun, and Gao (2020) concluded in their study that taxi demand is higher in areas with higher housing prices.

Education is another socio-demographic characteristic which has been investigated in relation with taxi demand. C. Yang and Gonzales (2014) found that the education level has a positive impact on taxi trips. More specifically, they found that in areas with higher education levels, taxi demand is higher. Another study also suggested that a higher density of people with a bachelor's degree or higher is associated with more taxi trips (Qian & Ukkusuri, 2015b). which is consistent with the previous finding.

The number of employees commuting by subway, the number of households without a private car and the number of airport trips were also found as variables which have a significant association with taxi trips according to a study in the US (Schaller, 2005). Another research study found that commuting time is negatively associated with taxi ridership (Qian & Ukkusuri, 2015b).

Population density is another factor which plays an important role in taxi demand. In areas with higher population density, higher taxi demand is observed (Q. Liu et al., 2020; C. Yang & Gonzales, 2014).

2.1.2 Land use factors

A series of studies were conducted to find the impact of land use characteristics on taxi demand. For example, in areas with more concentrated human activities (Q. Liu et al., 2020), with higher residential (Q. Liu et al., 2020; X. Liu et al., 2020), commercial (Lacombe, 2016; Q. Liu et al., 2020; X. Liu et al., 2020; Qian & Ukkusuri, 2015b) and employment densities (C. Yang & Gonzales, 2014), more public spaces and mixed land use (Q. Liu et al., 2020), higher taxi demand is observed .

Land use variables can have a positive or negative impact on taxi trips according to different times of the day, week and models used. For example, Zhang et al. (2019) estimated three different models including ordinary least squares (OLS) , GWR and geographical and temporal weighted regression (GTWR) to identify the impact of urban environment on taxi ridership. They concluded that residential density was positively associated with taxi ridership in OLS model. However, GWR and GTWR models indicated that the temporal period and the geographical location play an important role in the sign of the coefficient of residential density. This means that the impact of residential density on taxi ridership can be positive or negative according to the spatial and temporal features.

A similar pattern was also observed for the density of tourist attractions. Zhang et al. (2019) found that the density of attractions was negatively correlated with taxi ridership in OLS model. However, the spatial visualization of the coefficients of density of attractions during weekdays in the GWR and GTWR models indicated that in most of the regions, coefficients were positive. This is inconsistent with the negative linear correlation found in OLS model. This indicates the important impact of the location and time on the coefficients. They also calculated the mean coefficient value for urban environmental variables during morning, afternoon and evening peak hours to identify if the independent variables had a temporal impact on taxi ridership. The results showed that the average coefficients of residential, employment and hotel density were positive during all three peak hours, while it was negative for density of attractions during all three peak hours.

2.1.3 Transportation factors

Since taxis supplement other modes of transportation as mentioned earlier, there is a relationship between taxi trips and other modes of transportation.

Road design is found as one of the factors which can explain taxi demand. Taxi trips are generally more frequent in areas with higher road density due to their higher populations (X. Liu et al., 2020; Qian & Ukkusuri, 2015b). This is consistent with findings of Zhang et al. (2019) which indicated that the average coefficient of road density was positive during morning, afternoon and evening peak hours. Furthermore, a recent study found that taxi demand is higher in areas with dense secondary roads and dense road junctions. This is due to the characteristics of the secondary roads such as their narrow road width, lower speed limit, and proximity to neighborhoods which make them more accessible for residents. Better accessibility of secondary roads leads to a lower probability of finding passengers on tertiary roads, which reduces tendency of drivers to drive on tertiary roads, and this leads to a lower demand on these roads (Q. Liu et al., 2020).

The impact of bus stops on taxi demand has been a disputed issue since some studies showed a negative association between density of bus stops and taxi trips, and in others, a positive correlation between density of bus stops and taxi trips was observed. In one study, it was mentioned that density of taxi trips was negatively associated with the density of bus stops (Q. Liu et al., 2020). However, another study concluded that in areas with a higher density of bus stops, more taxi trips occur (X. Liu et al., 2020). Two main reasons were put forward. First, areas with a higher density of bus stops include more human activities. Thus, they attract many residents which result in more

departing and arriving taxi trips. The next reason is the connectivity function of taxis. Transit users can take a taxi to arrive at the bus stop or depart from the bus stop (X. Liu et al., 2020). Zhang et al. (2019) found similar results in their study. They found that the coefficient of bus stop density had a positive sign during three peak hours including morning, afternoon, and evening peak hour. Transit access time is another factor which was found to have an impact on taxi demand (Lacombe, 2016). For example, one study showed that higher transit access time leads to higher taxi demand (C. Yang & Gonzales, 2014).

Regarding metro, different studies resulted in contradictory findings in terms of the impact of metro on taxi demand. Q. Liu et al. (2020) found in their study that the metro had an insignificant impact of taxi demand; however, Qian and Ukkusuri (2015b) found that there is a positive correlation between metro accessibility and taxi demand.

Taxi demand can also be influenced by parking lot. A recent research study presented that parking lot density is positively associated with taxi trips, and it can depend on their location in hotels, shopping malls, and transportation hubs (X. Liu et al., 2020).

2.1.4 Temporal factors

As mentioned at the beginning of this chapter, the determinants used in modeling taxi demand are not only limited to spatial factors. Although temporal variables have been included in the topic of taxi demand modeling, there is a lack of literature in terms of the impact of weekdays, weekends, event days and different hours of the day on the taxi demand (Q. Liu et al., 2020).

Several studies found that weather conditions have an impact on public transportation ridership (Changnon, 1996; Cravo, Cohen, & Williams, 2009; Guo, Wilson, & Rahbee, 2007; Kalkstein, Kuby, Gerrity, & Clancy, 2009; Stover & McCormack, 2012 as cited by Kamga et al., 2015), and since taxi can be a complementary mode to other modes of public transport, weather can also affect taxi demand. In terms of the impact of weather on taxi demand, previous research has shown how different hours of the day affect taxi demand on rainy days, while snow does not explain taxi demand significantly (Kamga, Yazici, & Singhal, 2013). Another study suggested that temperature has an impact on taxi demand (Q. Liu et al., 2020).

A study found that in New York City (NYC), people believe adverse weather makes it unlikely to find a vacant taxi, due to increased demand. Another reason for the difficulty of finding vacant

taxis is severe weather conditions which make the traffic move slowly (Kamga et al., 2015). Although based on the belief of the NYC residents, the unavailability of vacant taxis is associated with the inclement weather, there is not enough methodological study on the impacts of the inclement weather conditions on taxi demand (Kamga et al., 2015). Kamga et al. (2015) investigated the variation of taxi trips under different weather conditions using 147 million records of a taxi trip dataset. They showed that weather conditions can strongly affect the taxi revenues. In their study, during rainy weather, higher hourly revenues were observed because of more frequent taxi pick-ups with shorter distances.

Previous studies also considered how taxi demand varies significantly on weekdays and weekends (Lacombe, 2016; Zhao, Khryashchev, Freire, Silva, & Vo, 2016). The taxi demand was found to be higher during weekends than weekdays from 0:00 to 4:00 and from 14:00 to 22:00 (X. Liu et al., 2020), which indicates the significant impact of time of the day on taxi demand in addition to weekdays and weekends. Another study showed that taxi demand was higher during peak hours and weekdays(Q. Liu et al., 2020).

By reviewing the existing literature in terms of the impact of socio-demographic, land use, transportation and temporal factors on taxi demand, it was found that a broad range of factors have been studied to identify their influence on taxi demand. Among the factors, some varied over space, and some changed over time. Furthermore, it was found that the relationship between some factors and taxi demand was not similar in different studies. Some studies found a positive association between specific factors (e.g.: income, density of bus stops) and taxi usage, while some studies suggested that there was a negative relationship between those same factors and taxi demand. Thus, reviewing the literature reveals that the impact of different factors on taxi demand can vary depending on the case study and the applied methodology, and there is no consensus among researchers on the determinants of taxi demand.

2.2 Methodologies used for modeling taxi demand

Two main approaches have been implemented in the existing literature for modeling the taxi demand: machine learning algorithms (Shao, Wu, Xiang, & Lu, 2015; C. Yang & Gonzales, 2014) and statistical models (Q. Liu et al., 2020; Qian & Ukkusuri, 2015b; Schaller, 2005; Zhang et al., 2019). In this section, the most common machine learning algorithms used in transportation studies and the most common statistical approaches applied to taxi demand studies are presented.

2.2.1 Machine learning approaches

In terms of machine learning approaches, different algorithms have been applied in the literature to forecast and explain taxi demand. For example, Zhao et al. (2016) compared the prediction accuracy of two machine learning algorithms including Markov predictor and the Neural Network predictor, and the results revealed that the Markov predictor is more accurate for predicting taxi demand. The autoregressive integrated moving average (ARIMA) model is another machine learning approach which has been applied to forecast the spatial distribution of taxi travelers (Moreira-Matias, Gama, Ferreira, Mendes-Moreira, & Damas, 2013). Using historical data, an ARIMA is used to forecast future scenarios of taxi demand; the ARIMA model is the most widely used time-series model in traffic volume prediction (Q. Liu et al., 2020). Although these analyses offer forecast outcomes that provide instant road traffic conditions and enable matching supply and demand, they are unable to interpret the influence of significant factors on taxi demand and cannot help policymaking (Q. Liu et al., 2020).

2.2.1.1 The clustering in transportation

Clustering approaches are broadly applied in transportation studies, including bikesharing studies and analyses of smart card data. In one study, the mobility patterns of bikesharing users are characterized by k-means clustering method, using annual and weekly time patterns as well as intensity based on the number of trips each cyclist made, to develop a typology for users of a bikesharing system in Lyon, France. For this purpose, by calculating the intensity and regularity of the user's weekly and annual use, a vector of 21 attributes was created for each bike user. By applying k-means clustering on the vectors, different types of users were developed (Vogel et al., 2014).

A previous research used three different clustering methods including k-means clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and spatiotemporal clustering to identify the virtual stations of bike aggregation, and k-means clustering revealed the best clustering results by having largest silhouette coefficient and CH index (Hua, Chen, Zheng, Cheng, & Chen, 2020).

In another study, applying k-means clustering, the bikesharing stations were classified into five types based on their nearby points of interest (POI). Then, based on these 5 station types and smart card data, the Latent Dirichlet Allocation (LDA) was performed to determine bikesharing travel

patterns and trip purposes (Bao, Xu, Liu, & Wang, 2017). Xu, Duan, and Pu (2019) applied the same approach as Bao et al. (2017), k-means clustering, to classify the public bicycles' stations, and then for forecasting the number of check-outs of bikes in each cluster, random forest was conducted.

Morency, Trepanier, Frappier, and Bourdeau (2017) adopted k-means clustering to develop typologies of travel behavior of the bikesharing members in Montreal, Canada. Another study in Montreal, Canada (Tarpin-Pitre & Morency, 2020) applied k-medoid clustering to analyze the travel behavior of bikeshare-metro-bikeshare users. K-means clustering approach was also adopted in a study in Gatineau, Québec, Canada to classify transit users' behaviors using smart card data (Viallard, Trépanier, & Morency, 2019).

Clustering approaches were also used for understanding and measuring the individual variability in transit use based on the smart card data. Deschaintres (2018) and Deschaintres, Morency, and Trépanier (2019) applied a k-means clustering approach to develop typologies for transit card users to measure interpersonal variability. Then intrapersonal variability among users was measured based on weighted Hamming distance by developing typologies of sequences.

Spatial clustering approaches have been also adopted in the literature of taxi demand to find pick up locations' hot-spots, and to address several research questions such as dividing an urban area into pick-up zones and finding the best locations for picking up taxi passengers (Chang, Tai, & Hsu, 2010).

In another study, a hierarchical traffic prediction model was applied to forecast the number of bikes checked out and checked in for each station cluster (Feng, Chen, Du, Li, & Jing, 2018). For this purpose, an iterative spectral clustering algorithm was applied to cluster stations, and a gradient boosting regression tree was conducted to forecast the number of bike usage in the bike sharing system.

By reviewing different research studies which have used clustering algorithms, it is found that k-means clustering is a common method for addressing transportation issues and contributed to uncovering distinct usage patterns.

2.2.2 Statistical models

Several studies have used different statistical models for taxi demand analysis. A few probabilistic models which have been used in previous studies for taxi demand analysis are explained below.

2.2.2.1 Geographically weighted regression (GWR)

Ordinary least square (OLS) multiple regression models which are mostly applied as a traditional approach in ridership analysis, are cost-effective and appropriate for multi-scale analysis (Davis, 2008; Z. Yang et al., 2018). Assumptions in OLS model neglect the temporal and spatial variation of taxi trips. In OLS model, it is assumed that all variables are stationary and independent over the study area, however, due to spatial and temporal variation of urban functionality, it is difficult to find the factors which can explain taxi ridership (Zhang et al., 2019). Ignoring the spatial variation leads to the model's unreliability, and makes it difficult to understand how taxi trips vary over space (Qian & Ukkusuri, 2015a).

Geographically weighted regression (GWR) addresses this issue (O'Sullivan, 2003) by allowing independent variables to vary over space and explains unsteady trends. The GWR model is an extension of multiple regression models which accounts for spatially non-stationary variables and leads to visualization of space-varying coefficients. This has made it an appropriate model to describe the geographical data with spatial heterogeneity (Cardozo, García-Palomares, & Gutiérrez, 2012; Qian & Ukkusuri, 2015b).

With respect to taxi demand, Qian and Ukkusuri (2015b) applied geographically weighted regression (GWR) to find how space-varying variables such as socio-demographic and built-environment explain the special heterogeneity of the taxi ridership in NYC metropolitan area (Manhattan, Bronx, Brooklyn and Queens). In other words, in this study, the distributions of parameter estimations were spatially visualized. For example, it was found that the sign of coefficient of median income can be positive or negative over space. The heterogeneity of taxi ridership was also modeled spatially in the study. For this purpose, authors used New York City's taxi GPS data to calculate average taxi trip at the traffic analysis zone (TAZ) level. The limitation of this model is that it only considers spatial variables.

2.2.2.2 Geographical and temporal weighted regression (GTWR)

As mentioned in the previous section, GWR addresses the issue of the assumption of traditional approaches (all variables are stationary and independent over the study area) by including spatial non-stationary variables in the model. In order to model spatiotemporal data such as ridership using GWR, it is necessary to aggregate or average dependent variables based on a special timestamp (Chiou, Jou, & Yang, 2015; Chow, Zhao, Liu, Li, & Ubaka, 2006). Therefore, in GWR, temporal non-stationarity is ignored while aggregating the dependent variable (C. Chen, Varley, & Chen, 2011).

Taxi trips are time-sensitive which is indicated by a temporal non-stationarity, just as the space-sensitive nature of taxi trips that is represented by spatial non-stationarity. Due to the space and time sensitivity of taxi ridership, it is important to include spatial and temporal variations in the analysis of taxi trips (Zhang et al., 2019).

Time is not considered in the previous model, GWR, which is important when analysing hourly taxi ridership on a spatiotemporal basis. To address this issue in modeling the spatiotemporal hourly taxi demand, Zhang et al. (2019) implemented the geographically and temporally weighted regression (GTWR) to model the "spatiotemporal heterogeneity of hourly taxi ridership" in Xiamen city in China using weekday taxi pickup points data. This model also visualizes temporal and spatial variations of coefficients. Adding the temporal dimension to the GTWR model led to capturing the temporal fluctuation of coefficients. The authors analyzed the influence of spatial urban environment variables (POI) and transport factors on hourly taxi ridership using GWTR. They found that this model outperforms GWR and OLS due to its model fit and explanatory accuracy. Further, it was found that the coefficient of some variables such as residential density can have a positive or negative impact on taxi ridership depending over space and time (Zhang et al., 2019).

This approach only considers the temporal aspect of the dependent variable, and no time-varying independent variable was taken into account in this approach. All independent variables varied over space but not time.

2.2.2.3 Generalized additive mixed model (GAMM)

Q. Liu et al. (2020) applied another approach, a generalized additive mixed model (GAMM), to find the impact of urban environment characteristics on the spatiotemporal variation of taxi demand

in the central area of Beijing, China. They used daily density of taxi trips which were aggregated by traffic analysis zones (TAZ) as the dependent variable, and population density, land use, road design and temporal variables as independent variables. Taxi counts were derived from taxi trace GPS data.

Since the number of taxi trips change by temporal factors such as weather, hour of the day and day of the week, the authors argued that simple generalized linear models (Poisson) are not able to model these periodic characteristics of taxi trips (Q. Liu et al., 2020).

In the mentioned study, the authors applied GAMM which is a semi-parametric statistical model. This approach can address the concerns of heteroscedasticity and autocorrelation by assuming that the number of taxi trips follow a Poisson distribution. This approach is a panel model, which captures the interaction impacts between temporal variables and a time metric variable.

2.2.2.4 Multiple linear regression

(Lacombe, 2016) used two linear regression models to better understand the taxi demand by describing the origin and destination trip generation of taxi trips spatially and temporally. Different variables such as demographics, land use, transit accessibility, and weather were used to find their influence on taxi demand. The R2 (the coefficient of determination) which is used for evaluating the model was found to be 0.28 for origins and 0.27 for destinations. This suggests that, although these models provided some information regarding factors which can influence taxi demand, the models need to be improved due the low values of R2. Hence, since in a previous study the linear regression model was found an explanatory model, it was not tested in this study anymore.

CHAPTER 3 METHODOLOGY

The objective of this chapter is to explain the data used in this study, clarify the steps taken to process the data and explain the methodology applied to analyze and forecast the spatial and temporal demand of taxi.

3.1 Data

Two main groups of datasets are used in this study to better understand the taxi demand on the Island of Montreal including taxi data and explanatory data. Since the main objective of this study is to better understand the spatial and temporal demand of taxi in the Island of Montreal, all information regarding areas out of the Island of Montreal are excluded from the data sets.

3.1.1 Taxi data

The taxi data comes from the Bureau du taxi de Montréal (BTM). We used data of three months including April, July, and September 2019 to identify how taxi demand varies over different times of the year. This dataset shows the number of taxi trips per hour per census tract per date in Montreal. For each month two tables are available. One shows the trips that depart from the census tracts, and one includes trips that arrived in the census tracts in Montreal. A census tract (CT) is a small area with a population between 2,500 and 8,000 persons which is located in census metropolitan areas and in census agglomerations. (Statistics Canada, 2018).

In this study, we only used the data of departing trips from CTs in the Island of Montreal. We selected departing trips since they are more useful to adjust supply. Supply needs to be adjusted based on where people desire to start a trip. The dataset consists of four variables including “date_orig” which indicates the date that taxi trip was made, “heure” which is the departure time of trips, “sridu_orig” which is the census tracts that taxi trip was departed from, and “nb_courses” which shows the number of trips. The dataset that we received included the observations with at least one taxi trip. Since this research aims to forecast and analyze taxi demand, it is important to consider CTs, hours and days with no taxi trip. Hence, we added missing observations with zero “nb_courses” to the dataset. There are 533 CTs on the Island of Montreal, but we removed Dorval Island from our analysis since there were no taxi trips there.

Table 3.1 shows an extract from the original dataset of taxi trips including the four above-mentioned variables. We also added the day of the week to this data set for the purposes of descriptive analysis.

Table 3.1 Departing taxi trips data set in April 2019

date_orig <chr>	heure <int>	sridu_orig <dbl>	nb_courses <int>
2019-04-01	0	4620009	1
2019-04-01	0	4620014	2
2019-04-01	0	4620015	1
2019-04-01	0	4620017	1
2019-04-01	0	4620018	1

Table 3.2 shows the number of observations and the percentage of observations with zero departing taxi trips in each dataset. The number of observations is calculated by the number of days in the month (30 for April and September, and 31 for July) multiplied by the number of hours (24) multiplied by the number of census tracts on the Island of Montreal. According to Table 3.2, more than 40% of observations for all three datasets included no trips.

Table 3.2 Number of observations and percentage of observations with zero departing taxi trips for each dataset

	Number of observations	**Percentage of observations with zero trips for departing trips from census tracts (CT)**
April	30×24×532= 383,040	50%
July	31×24×532= 395,808	40%
September	30×24×532= 383,040	42%

3.1.2 Explanatory data

Four categories of explanatory data are used in this study including socio-demographic characteristics, land use, transportation and weather to explain the spatial and temporal fluctuation of taxi demand. For each category, several variables have been developed based on the literature review which are shown in Table 3.3. More details regarding developing these variables are provided in the next section.

Table 3.3 Explanatory data

	Variable	Description	Source
Socio-demographic	Median income	Median income among population aged 15 and over in private households	Census 2016
	Recent immigrants	Percentage of immigrants between years 2011-2016	Census 2016
	Elderly population	Percentage of population aged 65 and over	Census 2016
	Car ownership	Number of passenger cars and light trucks by the CTin which their owners live	SAAQ 2018
Land use	Density of hotel rooms	Number of hotel rooms in each census tract divided by the area of the corresponding CT (count/ km^2)	Données Québec
	Density of workers of health care services	Number of workers of health care centers in each census tract divided by the area of the corresponding CT (count/ km^2)	Census 2016
	Density of business workers	Number of people who work for business, finance and administration divided by the area of the corresponding CT (count/ km^2)	Census 2016
	Density of eating places	Number of eating places in each census tract divided by the area of the corresponding CT (count/ km^2)	DMTI 2019
	Density of drinking places	Number of drinking places centers in each census tract divided by the area of the corresponding CT (count/ km^2)	DMTI 2019
	Events	If there was an event on a specific day or not	STM 2019

Table 3.4 Explanatory data (continuous)

	Variable	Description	Source
Land use	Art, entertainment and recreational centers (dummy)	If there was an art, entertainment and recreational center or not	DMTI 2019
Transportation	Density of bus stops	Number of bus stops in each census tract divided by the area of the corresponding CT (count/ km^2)	STM
	Metro stations (dummy)	If there is a metro station in the census tract or not	STM
	Bus services	Number of bus passages per hour in each census tract	STM GTFS 2019
Weather	Rain	If it was rainy or not	Environment Canada 2019
	Snow	If it was snowy or not	Environment Canada 2019
	Clear	If it was clear or not	Environment Canada 2019

Table 3.5 shows weather conditions related to clear, rainy and snowy days in Table 3.3.The weather conditions of each category vary by month. The gray color in the box shows if the weather condition belongs to the corresponding category for each month or not.

Table 3.5 Weather conditions for each category

Category	Weather condition	April	July	September
Clear	Fog			
	Mostly cloudy			
	Mostly clear			
	Clear			
	Cloudy			
Rainy	Freezing drizzle			
	Drizzle			
	Rain showers			
	Moderate rain showers			
	Freezing rain			
	Moderate rain			
	Rain			
	Storms			
Snow	Snow showers			
	Snow			
	Moderate Snow			

3.1.2.1 Data preparation

The databases of median income, recent immigrants, elderly population, workers of health care services and business workers were derived from Statistics Canada. Business workers are those who work in business, finance et administration sectors. Regarding income, the median income among population aged 15 and over in private households was considered. The percentage of recent immigrants was calculated by dividing the number of immigrants between years 2011-2016 to the total number of immigrants and non-immigrants. The percentage of elderly population was

calculated by diving the number of people aged 65 and over by the total population of each CT. Car ownership data was obtained from the Société de l'assurance automobile du Québec (SAAQ) as the number of passenger cars and light trucks (e.g.: Sports Utility Vehicle (SUV)) as of December 31, 2018, by the dissemination area in which their owners live. Since the data was aggregated by dissemination area by SAAQ, we aggregated it by CT based on the CT IDs.

Regarding land use, the dataset of hotels including the number of rooms was derived from Données Québec. We calculated the number of hotel rooms in each census tract based on the latitude and longitude provided in the dataset.

Eating places, drinking places and art, entertainment and recreational centers were obtained from the DMTI database. eating and drinking places were derived from the dataset based on the standard industrial classification primary (SIC_1). Number of eating and drinking places in each CT was calculated based on their latitude and longitude provided by the dataset. Art, entertainment and recreational centers were derived based on the north American industry classification system (NAICS). In this study, theater companies and dinner theaters, museums, zoos and botanical gardens, amusement and theme parks, casinos (except casino hotels) and fitness and recreational sports centers were considered as art, entertainment, and recreational centers. Since number of these centers in CTs were not high, this variable was included in this study as a dummy variable.

Furthermore, the available events data which was obtained from Société de transport de Montréal (STM) included information of "name of event", "date of event", "start hour", "finish hour", "place" and "name of metro station". From this dataset, the date and hour were used for descriptive statistics (chapter 4) to represent if there was an event on a specific day or not. The events included different activities such as concerts, hockey and soccer matches, festivals and other large activities.

Regarding bus stops and metro stations, we used the shapefile provided by STM to count their number in each census tracts. Since the number of metro stations in each CT was very low, and there were also many CTs without metro station, metro station was used as a dummy variable. In terms of bus services, number of bus passages per hour per CT was calculated from GTFS (General Transit Feed Specification) data from STM. The GTFS was developed by Google in cooperation with the Portland, Oregon, public transit agency (TriMet). Google developed a standard publishing format for transit agency operational data such as stops, stop times and routes to allow users to access data more easily. This has also facilitated transit agencies to publish their data at a low cost

(McHugh, 2013). GTFS data are provided as 13 comma-separated values (CSV) files including 6 mandatory and 7 optional files. The files create a relational database. The GTFS dataset provides information about a transit system by describing its stops, routes, and schedules (Fortin, Morency, & Trépanier, 2016). Figure 3.1 indicates how different files of GTFS data are related to each other. Green color is an indication of mandatory files and orange color corresponds to optional files.

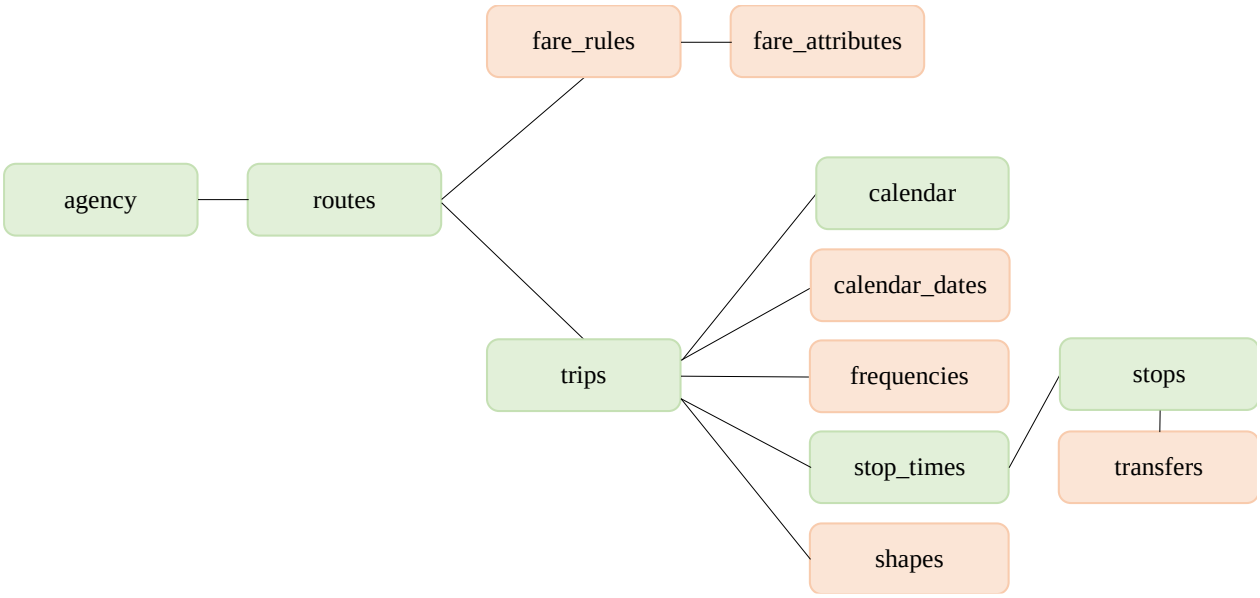


Figure 3.1 Diagram of mandatory and optional GTFS files and how they are related to each other, inspired by (Fortin et al., 2016)

For calculating the number of bus passages per hour per CT a few steps were taken for data preparation. Firstly, files of "stop_times" and "trips" were merged using a common variable which was "trip_id". Then, the produced file was merged with "stops" using "stop_id". This new merged file was again merged with "routes" file by "route_id". Finally, it was merged with "calendar" by the common "service_id". Since the purpose of this data preparation was to calculate the number of bus passages, all information regarding metro was excluded from the dataset. These steps were taken for GTFS files of April 5th, July 12th and October 16th (the planned service is similar for September and October) of 2019. The final file included "route_id", "stop_id", "trip_id", "departure_time", "service_id", "stop_lat", "stop_long" and 7 variables including days of the week from Monday to Sunday. Then, all three files of the three months were imported to QGIS based on the latitude and longitude of bus stops. Then the bus stops which included the information of GTFS datasets were intersected with the shape file of CTs of Montreal. After this step, using the attribute table of the intersected shapefile, the number of unique "trip_id"s per departure time and per CT

was calculated for each day of the week for each month using the "n_distinct" function in R. Then, the average of number of trips per CT was calculated for each day of the week. For the further purposes of this study, the daily mean of number of trips (including weekdays and weekends), the average of trips per weekday and the average of trips per weekend day was calculated for each month. Finally, in order to have one variable for each daily mean of number of trips, the average of trips per weekday and the average of trips per weekend day, the average of these three variables among April, July and September was calculated.

Weather data, which was retrieved from Environment Canada was included in the descriptive statistics based on different categories of weather conditions which was presented in Table 3.5

3.2 General methodology

Figure 3.2 illustrates that the methodology of this study contains four main steps including the descriptive statistics, the k-means clustering, the hierarchical agglomerative clustering, and modeling (decision tree and multinomial logit model). The figure also shows how different steps are related to each other. These steps are briefly explained in this chapter to present the overall methodology. The methodological details associated with each step are presenting in the following corresponding chapters.

As the first step, a descriptive analysis of the taxi trips and the potential explanatory factors which can explain the spatial and temporal fluctuations of taxi trips is provided.

In the next step, an automatic classification method, the k-means method, is used to characterize the weekly temporal variability of taxi demand at the CT level on the Island of Montreal, and to develop a week typology. To this end, a dataset including the 12 weeks of April, July and September of 2019, is used to create "CT-week" vectors. Then these vectors are normalized for being used in the process of k-means clustering, which results in a typology of weekly patterns. Finally, Shannon entropy is used as an indicator of temporal variability within each CT, by measuring the repetition of the same weekly patterns for each CT over the 3 months. In other words, it is used to find CTs with regular and irregular weekly temporal variation of taxi usage.

Then, "CT" vectors are created in the format of ordered sequences of 12 weeks using the week typology developed above. In other words, for each CT and each week (1 to 12), there is one type of week which was obtained from the previous step (the week typology). Then, the weighted

Hamming distance is used to calculate the dissimilarity matrix between each pair of sequences. The weighted Hamming distance is calculated by summing the Euclidean distances between the centers of the types of weeks (obtained from the previous step) over the 12 weeks. Finally, an agglomerative hierarchical clustering is applied to develop a multi-week typology to compare taxi usage among census tracts.

The last step of the methodology (presented in chapter 6) aims to test two models including decision tree and multinomial logit model to identify the factors which can explain the multi-week typology.

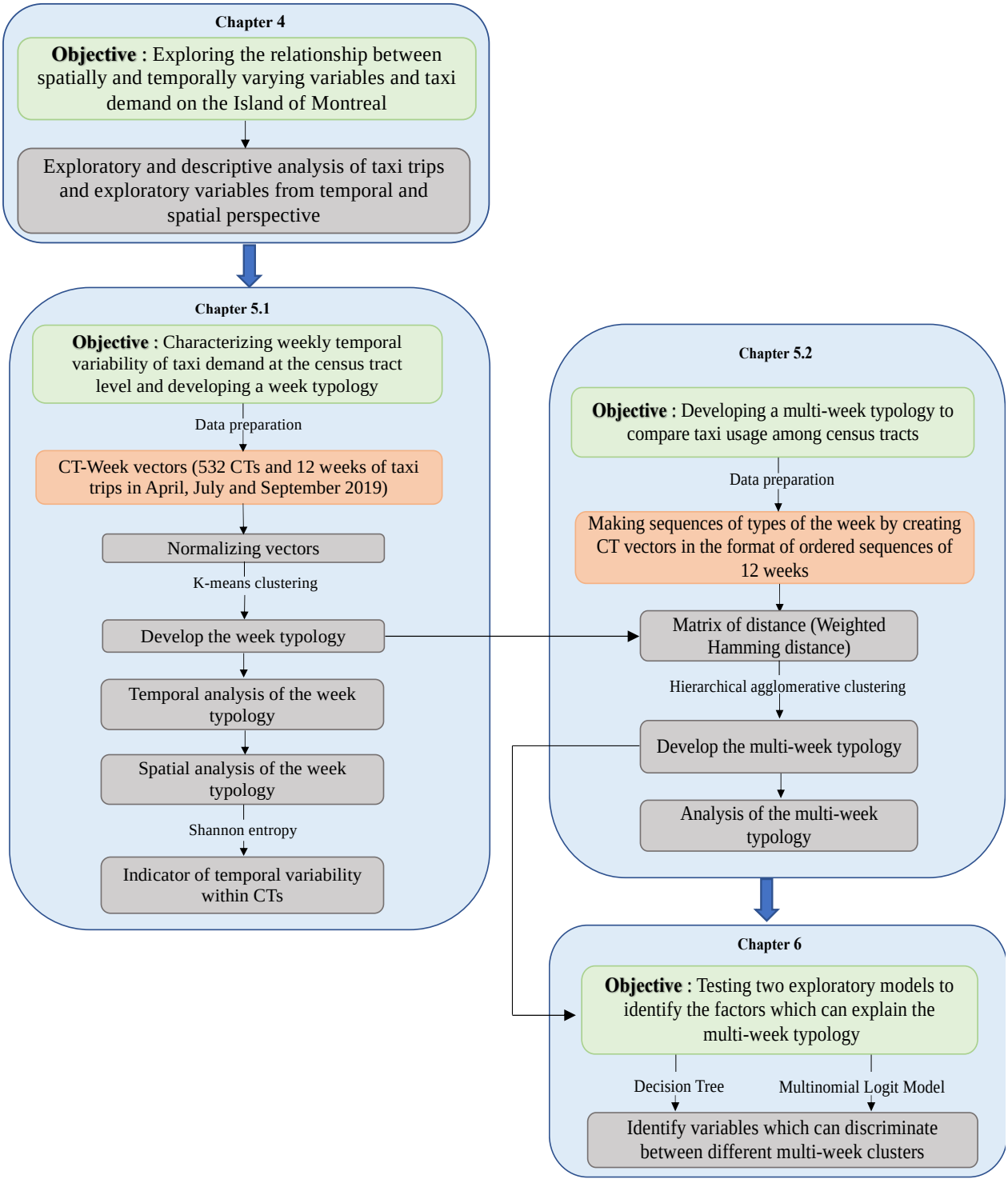


Figure 3.2 General methodology of this study

3.2.1 K-means clustering

According to James, Witten, Hastie, and Tibshirani (2013), clustering consists of a wide range of methods which is used to find subgroups or clusters in a dataset. Clustering aims to partition the observations of the data set into different groups so that the similarity between the observations within each group, and the difference between the observations between groups are quite high. In other words, by using clustering approaches, we seek to find homogeneous subgroups among the observations of a dataset.

K-means clustering is an approach which partitions observations of a dataset into K clusters, so that they are distinct and non-overlapping, and the within-cluster variation is as low as possible. In this approach, each observation is a member of at least one cluster among K clusters. Furthermore, since the clusters are non-overlapping, each observation only belongs to one cluster. To perform k-means clustering, it is necessary to pre-determine the number of clusters (K).

If C_1, …, C_K are defined as sets which include the attributes of the observations in each cluster, within cluster variation for cluster C_k is the difference between observations within a cluster as measured by W (C_k). For getting the lowest within-cluster variation, the aim is to partition the observations of a data set into K clusters, in a way that the sum of the total within-cluster variation for all clusters is as low as possible which is indicated in the following formula.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Equation 1

To better understand how k-means clustering works, it is necessary to provide a clear definition of within-cluster variation. There are several definitions available for this term in the literature, but the most common one is based on squared Euclidean distance (used in this study) which is defined by:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Equation 2

In the above equation, $|C_K|$ is the number of observations in the k_{th} cluster.

For each cluster, Division of the sum of the squared Euclidean distances between pairwise observations by the total number of observations in that cluster gives the within-cluster variation (James et al., 2013).

The following formula is a combination of equations 1 and 2 which defines the optimization problem of the k-means clustering.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Equation 3

The k-means clustering algorithm finds the solution for equation 3, which results in partitioning observations into K clusters. This algorithm includes the following steps:

1. Each observation is assigned a number between 1 and K randomly. This means that the observations are partitioned in K clusters randomly. The numbers between 1 and K are assigned to the observations as clustering initializations.
2. Iterate the following steps until the number of clusters which is assigned to observations do not change anymore:
    a) Calculate the center of each cluster which is the average of all observations in a group (this is why this approach is called k-means clustering).
    b) Place each observation in the cluster with the nearest center (where closest is defined in terms of the metric used, here Euclidean distance).

As mentioned earlier, for running the algorithm of k-means clustering, the number of clusters (K) must be determined beforehand. In this study, two methods were used for deciding the number of clusters: the elbow method and the dendrogram.

In the elbow method, the total within cluster squares is plotted on the y axis of a diagram, and the number of groups or cluster (K) is plotted on the x axis. According to this method, an optimal number of clusters (K) can be observed at the point where as K increases, the value on y axis (within cluster squares) does not change significantly (Deschaintres, 2018).

The second method, the dendrogram, is an agglomerative hierarchical algorithm. The K will be obtained in this method by cutting the dendrogram horizontally at any desired level of dissimilarity (Deschaintres, 2018). Since this method is not only used for determining the number of clusters, but also as a separate methodology in this study, it is explained more precisely in the next section.

In this study, the k-means clustering algorithm was applied in the programing tool of R, using the function of “kmeans” with the “Lloyd” algorithm from the “stats” package.

3.2.2 Hierarchical agglomerative clustering

Hierarchical clustering is another methodology applied in this study. It has two advantages comparing to k-means clustering. Firstly, it is not necessary to decide on the number of clusters before applying the approach. Secondly, it presents the results by a dendrogram which represents the observations by a tree-based diagram. There are different types of hierarchical clustering, and the most common one is the agglomerative clustering (also known as bottom-up) (James et al., 2013).

In a dendrogram, each leaf is one observation. The observations (leaves on the dendrogram) which are similar to each other, start to fuse from bottom of the tree to the top. Since the vertical axis of the tree indicates the dissimilarity, the dissimilarity between observations increases as observations fuse toward the top of the tree. Several different algorithms of hierarchical clustering exist, with different approaches for measuring dissimilarity, such as Euclidean distance. (James et al., 2013). In this study, Euclidean distance is used in chapter 5.1 using "hclust" function in R based on the "Lloyd" algorithm and weighted Hamming distance is used in chapter 5.2 for measuring the dissimilarity by using the "ward.D2" method and "hclust" function in R from the "stats" package.

The algorithm of hierarchical clustering runs iteratively. At first there are n observations (leaves) in n clusters (n corresponds to the total number of observations in the dataset). Then, the most two similar clusters are combined as one new cluster, leading to decreasing the number of clusters to n-1. The algorithm iterates until all observations belong to a single cluster at the very top of the dendrogram (tree) (James et al., 2013).

3.2.3 Decision tree

Decision tree analysis is a method for discovering characteristics and patterns of large databases for discrimination as well as predictive modeling. Decision trees can be interpreted intuitively. The mentioned advantages of the decision tree have made it a popular approach for "exploratory data analysis and predictive modeling applications" (Myles, Feudale, Liu, Woody, & Brown, 2004). Classification and Regression Trees (CART) introduced by Breiman, Friedman, Olshen, and Stone (2017) and Conditional Inference Tree (CTree) developed by Hothorn, Hornik, and Zeileis (2006) are the two most common methods for creating a decision tree. In this study, we used the CTree.

Unlike the k-means clustering algorithm, there is no need to normalize data for creating a decision tree.

Ctree, which is an abbreviation of the Conditional Inference Tree (CIT), is an algorithm that recursively divides data into binary fragments based on a decision tree. This creates a measurement based on change tests which try to differentiate between significant and insignificant improvements (VE & Cho, 2020). Unlike the CART algorithm which selects variables based on maximizing the Gini index, CTree applies a significance test to choose the variables. This method prevents variable selection bias which exists in the CART algorithm. In other words, CTree conducts a statistical testing between response and covariate for choosing the predictor variable for split. CTree applies the Chi-squared test (χ^2) if response variables and potential split variables are categorical. In the situation of having one categorical and one numerical variable, CTree uses one-way ANOVA (analysis of variance) for selecting variables. If both target (response) and potential split variables are numerical, CTree performs a Pearson correlation test (Schlosser, Hothorn, & Zeileis, 2019). In this study, the decision tree analysis was conducted in R using the "ctree" function from the "party" package.

To understand the effectiveness and performance of a decision tree, a confusion matrix is used. A confusion matrix measures the performance of classification in machine learning methods (Narkhede, 2018). In other words, the confusion matrix measures if the predicted values correspond to the actual values or not. It includes four components made up of four possible values resulting from the combination of actual and predicted values for binary classification as shown in Table 3.6 (Narkhede, 2018). Since in this study, multi-class classification will be adopted by decision tree, the explanations of the confusion matrix first will be presented for binary classification in this section, and then for a multi-class classification in chapter 6.

Table 3.6 Confusion matrix for binary classification

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

The components of Table 3.6 are defined as below (Mohajon, 2020):

- TP (true positive): The number of positive predictions correctly identified as “Positive” by the classifier.
- TN (true negative): The number of negative predictions correctly identified as “negative” by the classifier.
- FP (false positive): The number of negative predictions incorrectly identified as “positive” by the classifier.
- FN (false negative): The number of positive predictions incorrectly identified as “negative” by the classifier.

Based on the components of the confusion matrix, performance measures of the machine learning models are proposed. Among the existing performance measures, the most three common measures are accuracy, specificity and sensitivity (Mohajon, 2020). They are defined below.

Accuracy is used to understand the overall accuracy of the model. The accuracy of a model is the ratio of the number of samples which were correctly predicted by the classifier to the total number of samples. The following equation presents how accuracy is calculated mathematically.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Equation 4}$$

Specificity presents the ratio of samples which were correctly predicted as negative by the classifier to the total number of negative samples. Specificity is also known as true negative rate (TNR). It is calculated mathematically as below:

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad \text{Equation 5}$$

Sensitivity is the ratio of samples which were correctly predicted as positive by the classifier to the total number of positive samples. Sensitivity is also known as recall, true positive rate (TPR) and probability of detection. It is calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \text{Equation 6}$$

3.2.4 Multinomial logit model

In logistic regression, categorical data is used generally as the dependent variable. The dependent variable (Y) can be binary variables (variables which have only two possible values), or it can take

more than two values corresponding to r categories. Since in this study, the dependent variable is a categorical variable, a generalized logit model will be used. In this approach, each category of the dependent variable is compared with a category which has been selected as the reference category by using r-1 logits (Bham, Javvadi, & Manepalli, 2012).

The MNL model used in this study is as below (Y. Chen et al., 2016), and it was done in R using the "mlogit" function from "mlogit" package:

$$\log(\frac{\pi_i}{\pi_j}) = \alpha_i + X^T\beta_i \quad \text{Equation 7}$$

where π_i is the probability of the non-reference category i of the dependent variable,

$i = 1, \ldots p\ (i \neq j)$,

p is the number of categories of the dependent variable,

π_j is the probability of the reference category j of the dependent variable,

α_i is the intercept of the i^{th} equation,

X^T is the transpose of the vector X of the independent variable

and β_i is the coefficient vector for i^{th} equation.

CHAPTER 4 EXPLORATORY ANALYSIS

This chapter aims to better understand the spatial and temporal patterns of departing taxi trips from CTs on the Island of Montreal by performing several exploratory analyses on the taxi data of three months (April, July and September 2019), and potential explanatory variables. Firstly, an overall descriptive analysis is performed to compare taxi trips across different hourly and daily periods. Secondly, the temporal distribution of taxi trips is explored to assess the hourly variation throughout the day. Then, taxi trips during peak hour and non-peak hours, and on weekdays and weekends are analyzed spatially to find specific spatial patterns. Finally, the descriptive analysis of factors which are expected to explain taxi demand is provided.

4.1 Descriptive analysis of taxi data

The descriptive analysis of departing taxi trips from CTs is presented in Table 4.1. The table shows that six statistics are calculated for the taxi trips dataset which were aggregated by CT, date and hour. In other words, each observation in the dataset corresponds to a CT-date-hour. The column "Obs." provides the number of observations for each month which is made up of the number of CTs multiplied by the number of hours in a day multiplied by the number of days in the month.

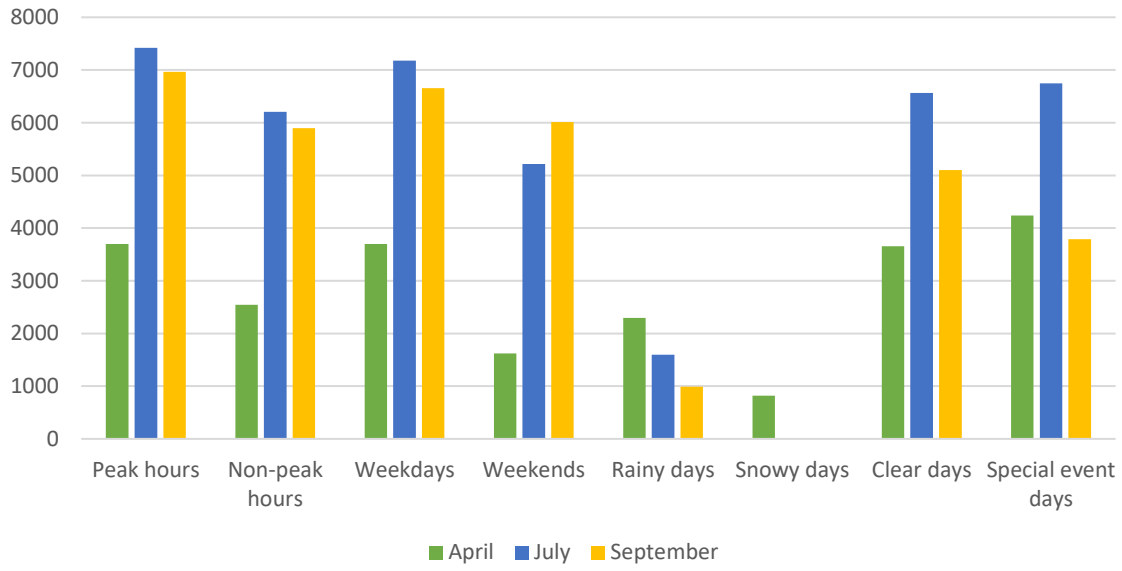
Table 4.1 Descriptive statistics of taxi trips in April, July and September 2019

Month	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Obs. (CT*h*d)
April	0.00	0.00	0.00	1.15	1.00	69.00	532*24*30
July	0.00	0.00	1.00	1.74	2.00	112.00	532*24*31
September	0.00	0.00	1.00	1.70	2.00	102.00	532*24*30

Table 4.1 shows, importantly, according to the low value of the 3rd quartile, it is found that 75% of the observations had only one or two trips. It is also important to note that 50% of the observations is 0 in April, and 1 in July and September, while the maximum values range from 69 to 102. This highlights the skewed distribution of taxi trips across CTs and hours of the day.

Figure 4.1 presents the hourly average of taxi trips during different hours and days in April, July and September. In the mentioned figure, peak hours refer to hours 6:00-8:59 and 15:00-17:59, and non-peak hours include hours 0:00-5:59, 9:00-14:59 and 18:00-23:59. Special event days are days which had at least one event such as a concert or hockey game and several other types of events. The information about rainy day, clear day and snowy day is presented in chapter 3.1.2.

Figure 4.1 Hourly average of taxi trips during different hours and days



According to Figure 4.1, the highest hourly average of taxi trips in April occurred on days with special events, while in July and September the highest hourly average of taxi trips were observed during peak hours. Snowy days in April had the lowest hourly average of taxi trips, and in July and September, rainy days had the lowest hourly average.

4.2 Temporal distribution of taxi trips

This section first compares the temporal distribution of average hourly taxi trips per day between weekdays and weekends with and without events in April, and then provides the equivalent information for July and September respectively. This analysis leads to find the hours of the day with the highest and lowest demand of taxi for each group of weekday and weekend.

Figure 4.2 shows the weekend and weekday taxi trips pattern considering the events in April 2019. Both weekday groups exhibit a peak at around 8h, which is morning peak hour, and then it starts falling from after 8h. Weekends with events accounts for the highest average taxi trips at midnight

from 0h to 3h due the “night-out” activities at weekends. The number of trips then falls from 3h to 5h. The pattern at weekends with events is almost stable from 10h to 18h. The lowest number of trips during midnight (0h-4h) belong to weekdays without events.

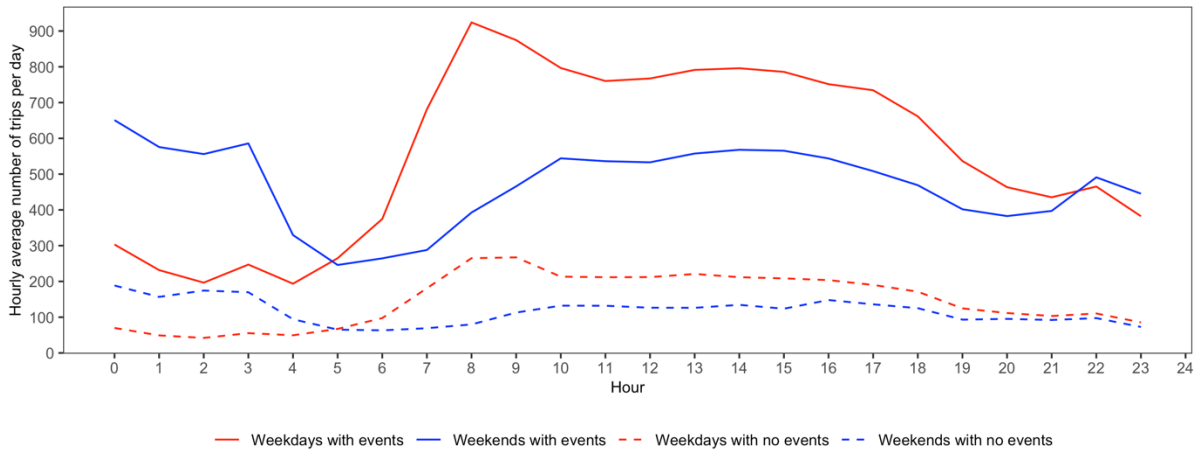


Figure 4.2 Average number of departing taxi trips per hour per day of the week in April 2019

Figure 4.3 provides the temporal distribution of taxi trips in July 2019. Since in July, all Saturdays and Sundays were also event days, there is only one line for weekend. Since in Montreal, there are many “night-out” activities during the weekends in summer, hourly average trips per day at weekends with events is higher than weekdays during the day, with the exception of 7h to 9h which is known as morning peak hour. The highest weekend hourly average taxi trips are observed at midnight between 0h and 3h. Then it falls and reaches its lowest point at 6h. Trips on weekdays with events and without events follow a similar trend. Weekday trips are lowest between 0h and 5h, and highest between 8h and 15h.

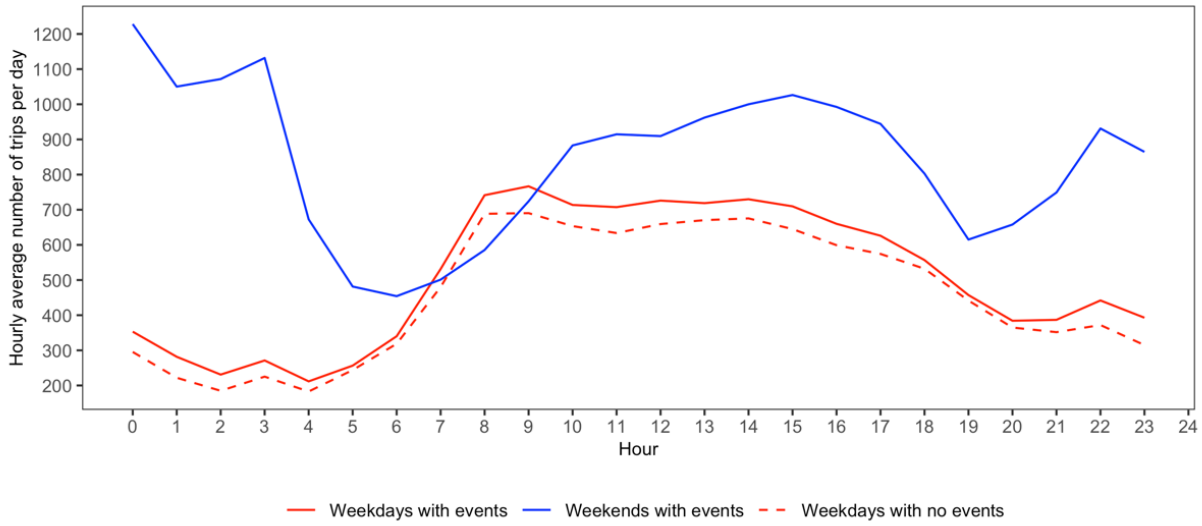


Figure 4.3 Average number of taxi trips per hour per day of the week in July 2019

Figure 4.4 presents the hourly distribution of taxi trips in September 2019. In September, same as July, there was no weekend without event. The highest hourly average of trips during weekends happened at midnight, and between 10h and 17h. Weekdays with events have more trips in the morning peak hour from 8h to 12h, and less taxi trips at midnight. Weekdays without events have the lowest number of trips compared with the two other categories.

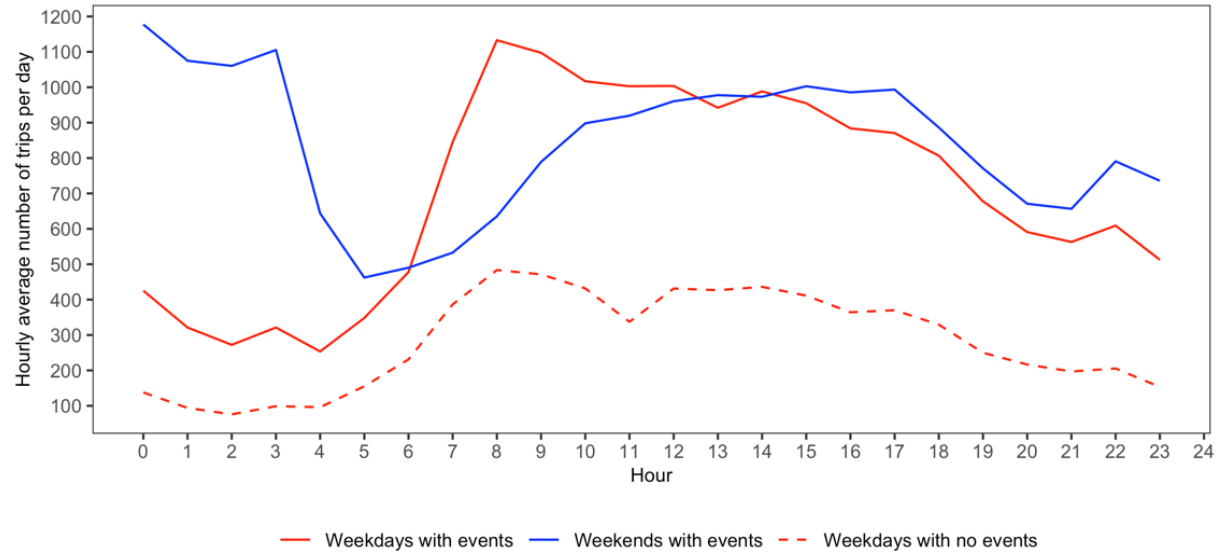


Figure 4.4 Average number of taxi trips per hour per day of the week in September 2019

By comparing these three figures, the following results were found.

- July and September have similar temporal patterns for weekends with events.

- Weekdays with events have the same pattern in April and September.
- In all three months, midnight trips happened predominantly during weekends with events.
- In April, the highest averages are found on weekdays with events between 5h and 21h.
- In July, there are more trips during weekends than weekdays with or without events.
- In September, there are more trips during weekends than weekdays with or without events at 0h to 6h and 15h to 23h.

The hourly patterns of taxi trips vary over different months and weekday groups. This shows how complex it is to develop a model which can find factors that explain both the temporal and spatial fluctuations of taxi demand.

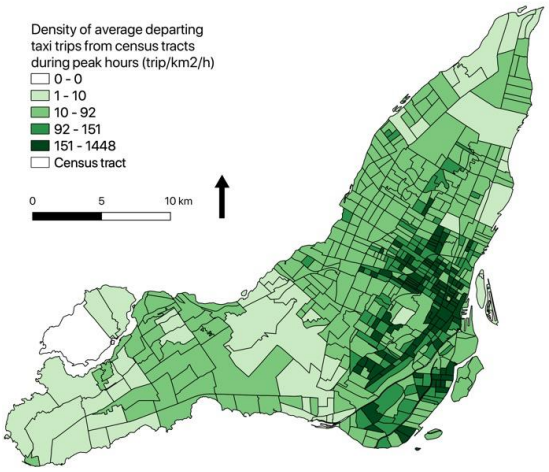
4.3 Spatial distribution of taxi trips

This section presents the spatial distribution of taxi trips during peak hours, non-peak hours, weekdays, and weekends. For this purpose, the density of average hourly taxi trips was calculated by dividing the average hourly taxi trips by the area of the CTs (trips/km^2/hour).

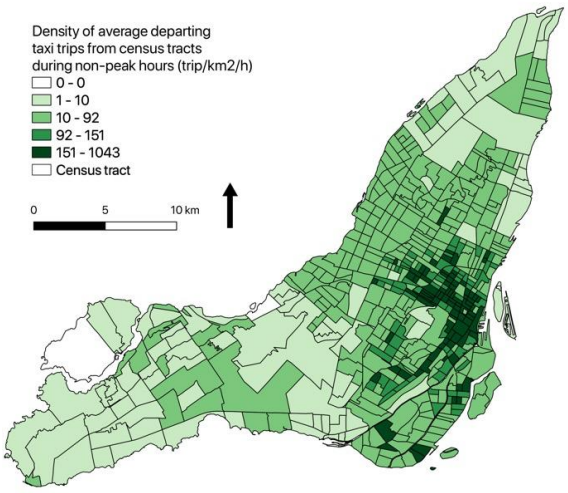
Figure 4.5 and Figure 4.6 and

Figure 4.7 present the spatial distribution of taxi trips during peak hours of the whole week (Monday to Sunday), non-peak hours of the whole week, weekdays and weekends including all hours in April, July and September 2019.

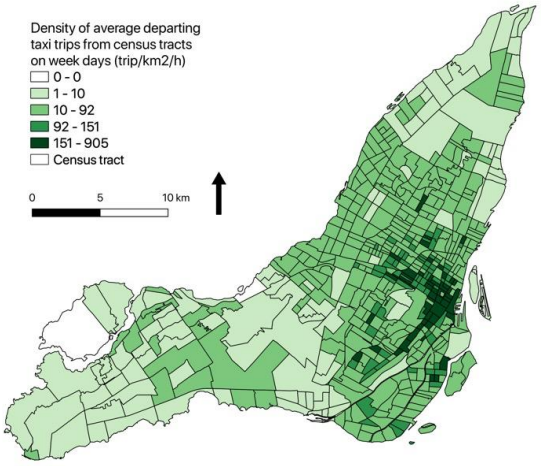
Commencing with April, Figure 4.5 shows that the spatial pattern of taxi trips is almost identical during peak hours, non-peak hours and on weekdays since most of the trips are concentrated in downtown and central neighborhoods. However, the pattern of trips during weekends is different as they follow a more decentralized pattern.



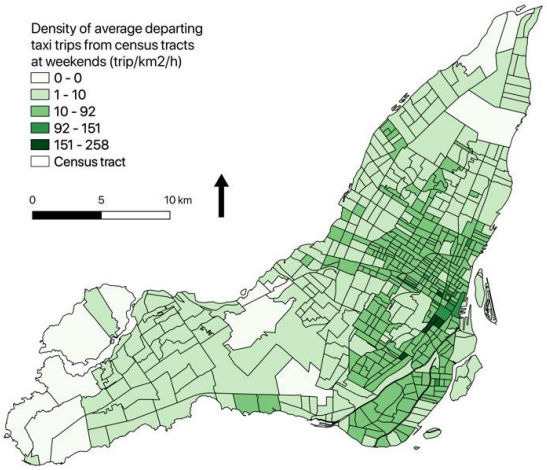
a) Peak hour (6:00-8:59 and 15:00-17:59)



b) Non-peak hour (0:00-5:59, 9:00-14:59 and 18:00-23:59)



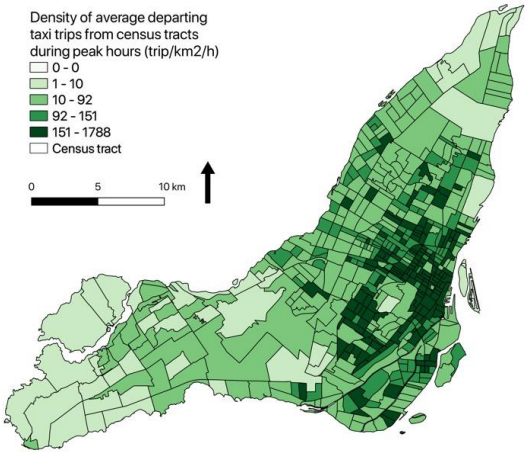
c) Weekdays (Monday to Friday)



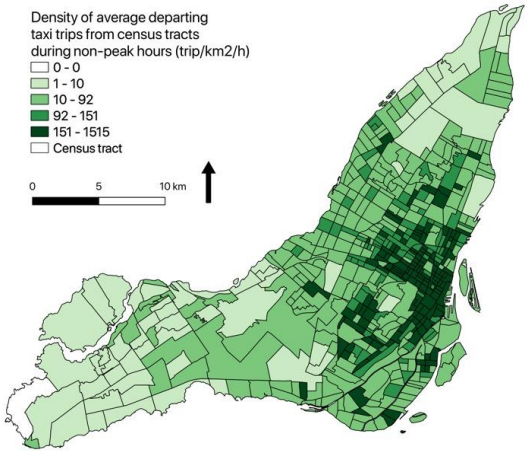
d) Weekends (Saturday and Sunday)

Figure 4.5 Spatial distribution of average taxi trips (origin) per hour (April 2019)

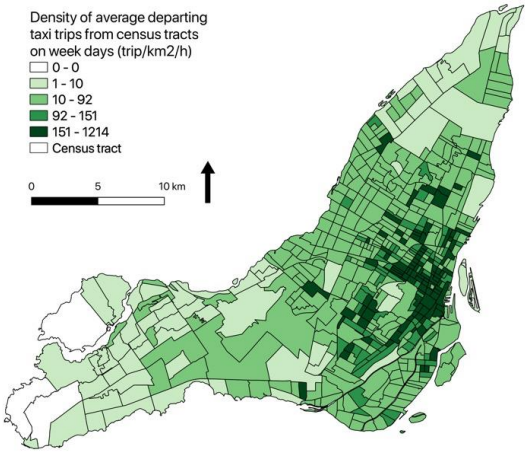
In July (Figure 4.6), same as April, most of the trips happened in downtown and central neighborhoods during peak hours, non-peak hours and on weekdays, but they were also concentrated in other areas such as Côte-des-Neiges. Weekends also follows a different trend from other categories with more decentralized trips, but more trips were concentrated in downtown at weekend comparing to April.



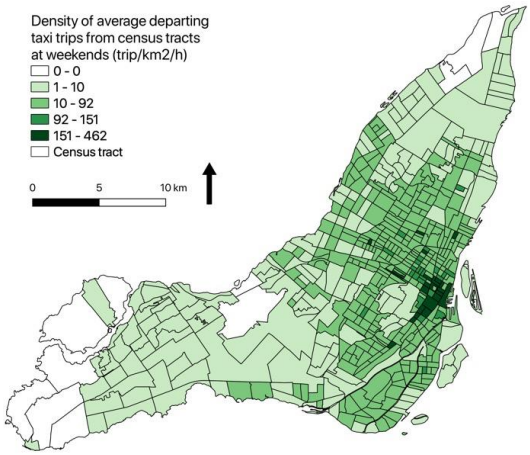
a) Peak hour (6:00-8:59 and 15:00-17:59)



b) Non-peak hour (0:00-5:59, 9:00-14:59 and 18:00-23:59)



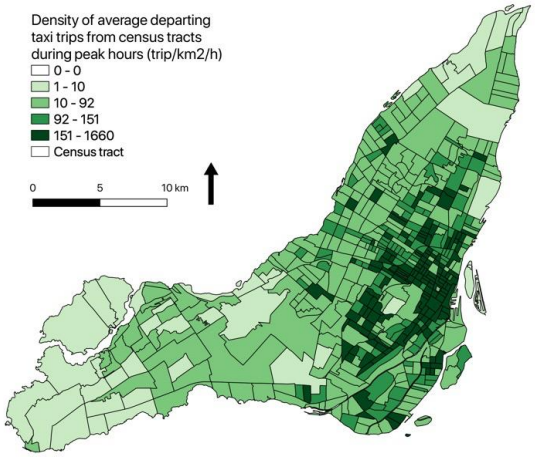
c) Weekdays (Monday to Friday)



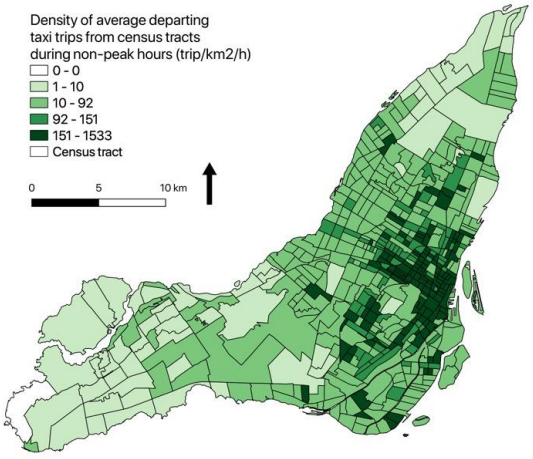
d) Weekends (Saturday and Sunday)

Figure 4.6 Spatial distribution of average taxi trips (origin) per hour (July 2019)

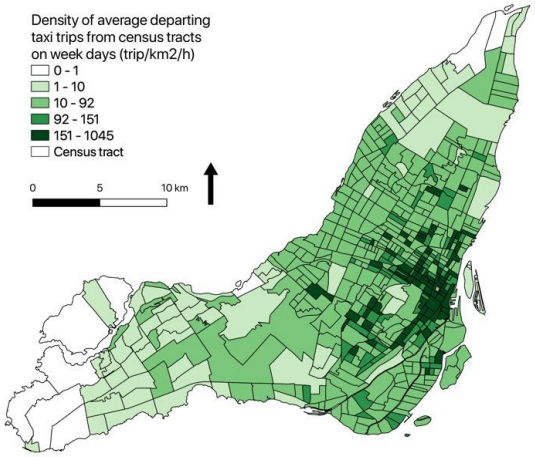
Figure 4.7 shows the spatial distribution of taxi trips in September 2019. Even though this month is known as opening of schools, no remarkable difference was observed in the spatial pattern of taxi trips between this month and July.



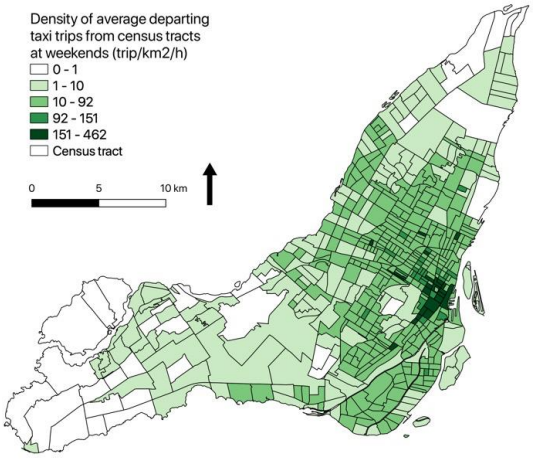
a) Peak hour (6:00-8:59 and 15:00-17:59)



b) Non-peak hour (0:00-5:59, 9:00-14:59 and 18:00-23:59)



c) Weekdays (Monday to Friday)



d) Weekends (Saturday and Sunday)

Figure 4.7 Spatial distribution of average taxi trips (origin) per hour (September 2019)

Comparing all three figures it was found that September and July had more similar patterns than April. Furthermore, in all three months, trips had a decentralized pattern at weekends, but in July and September more trips were concentrated in downtown.

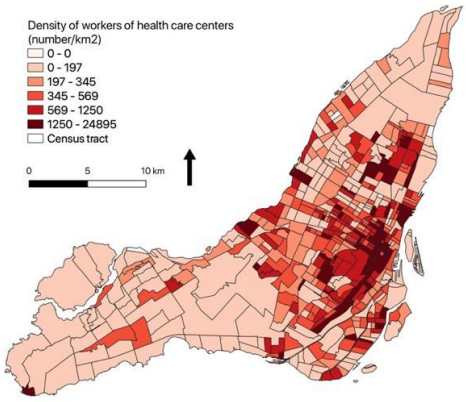
4.4 Descriptive analysis of independent variables

In this section, Figure 4.8 presents the spatial distribution of some exploratory variables used in this study which are expected to have a significant impact on taxi demand due to the literature (land use was found as one of the most significant factors in explaining taxi demand). The descriptive statistics of all exploratory variables which are used for forecasting taxi demand in this study are

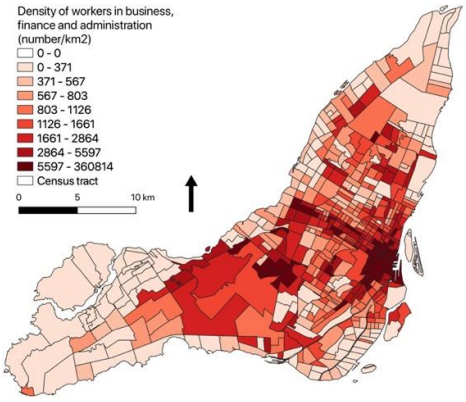
also provided in Table 4.2. Among 532 CTs, 11 CTs with 0 and very low population were excluded from the analysis.

According to Figure 4.8, density of workers of health care centers and density of workers in business, finance and administration are mostly located in central neighborhoods. The difference between these two groups is that as we move toward the western parts of the Island, density of health care workers decreases, while density of business workers is quite high in some western CTs. The figure also shows how drinking places and restaurant vary throughout CTs. Drinking places are denser in downtown, while the highest density of restaurants is found in the Vieux-Port. The figure also shows the same pattern for hotel rooms. This figure shows that hotel rooms are mostly concentrated in the Vieux-Port of Montreal, and that there are many CTs without any hotel rooms.

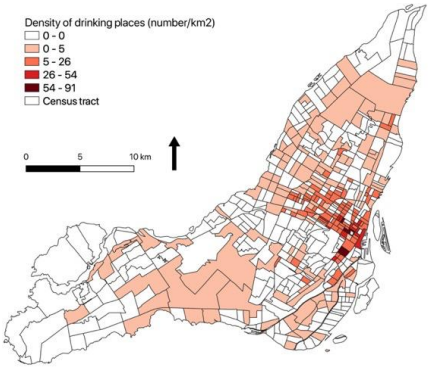
By comparing Figure 4.8 with Figure 4.5, Figure 4.6 and Figure 4.7 it can be concluded that the spatial distribution of density of workers of health care centers and density of workers in business, finance and administration is similar to distribution of taxi trips during peak hour and non-peak hour. Furthermore, the way that eating places are distributed is close to the spatial distribution of taxi trips at weekends. It is also notable that the CTs with denser drinking places and hotel rooms were also the CTs with higher taxi demand.



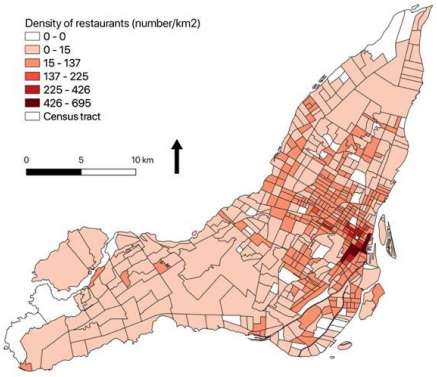
a) Workers of health care centers



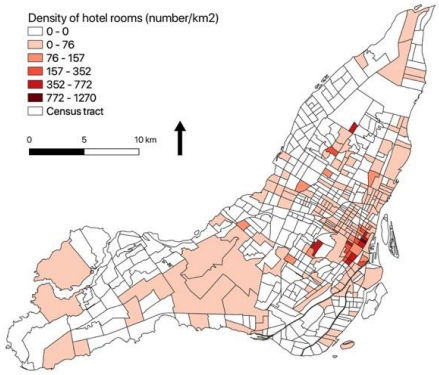
b) Business workers



c) Drinking places



d) Density of eating places



c) Hotel rooms

Figure 4.8 Spatial distribution of exploratory variables

According to Table 4.2, regarding the socio-demographic characteristics of CTs, less than 75% of CTs have a median income of 34,700$. The maximum percentage of recent immigrants is 30.45% in all CTs. In 75% of CTs, the percentage of elderly people is less than 20%. Regarding car ownership, the maximum number of passenger cars and light trucks per person for all CTs is less than 1. The remarkable difference between the 3rd quartile and maximum of hotel rooms' density, health workers' density, density of business workers, drinking places and restaurants shows that they are concentrated in a few CTs of Montreal. This shows that the land use variables are only concentrated in some specific CTs. The 0 median of drinking places and hotel rooms means that there is no drinking place and hotel room in 50% of CTs. In terms of bus services, there are some CTs with zero average of bus services per weekend, but there is no CT with 0 average of bus service per weekday and 0 average of daily bus service. Regarding subway, 50 CTs out of 521 have at least one metro station which is about 10% of the total.

Table 4.2 Descriptive statistics of exploratory variables

	Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Socio-demographic	Median income*10000	1.21	2.58	2.96	3.11	3.47	7.86
	Percentage of recent immigrants	0.00	4.01	6.22	7.29	9.59	30.45
	Percentage of people aged 65 years old and above	1.70	10.70	15.30	15.88	19.20	60.00
	Number of cars per population	0.09	0.29	0.37	0.38	0.46	0.69
Land use	Hotel rooms density	0.00	0.00	0.00	23.75	5.81	1270.11
	Health workers density	0.00	145.00	355.40	1139.60	797.50	24895.40
	Business workers density	12.3	407.30	959.00	5085.50	2438.20	360813.60
	Density of drinking places	0.00	0.00	0.00	3.82	3.55	90.94
	Density of eating places	0.00	4.20	13.93	35.50	33.96	695.43
		0	**1**				
	Arts, entertainment, and recreational centers (*D)	356	165				
Transportation	Density of bus stops	0.13	18.13	26.56	29.54	37.79	116.48
	Average of bus services per weekend	0.00	11.57	16.39	18.27	22.83	87.98
	Average of bus services per weekday	2.88	23.36	31.63	36.04	45.16	194.48
	Average of daily bus services	2.06	19.41	27.11	30.83	38.93	162.22
		0	**1**				
	Metro stations (*D)	471	50				

*D represents dummy in the table

CHAPTER 5 TYPICAL PATTERN OF TAXI USAGE

This chapter presents the two main components of this study. Firstly, it focuses on characterizing weekly temporal variability of taxi trips departing from CTs by developing a week typology which will be presented in the first section of this chapter. In the first section, at first, a typology of weeks is developed and analyzed from both the temporal and spatial perspectives. Then, the regularity or irregularity of weekly temporal variation of taxi usage of each CT is identified by measuring the repetition of the same weekly patterns of each CT over the 12 weeks during April, July and September 2019.

Secondly, this section focuses on developing a multi-week typology which will be presented in the second section. In this section, a sequence of types of weeks (obtained from the previous section) over 12 weeks is created for each CT, and then a multi-week typology is developed. In other words, this multi-week typology is developed based on the week typology developed in the previous section by re-clustering the clusters. This will lead to comparing the temporal distribution of taxi usage over the three months between CTs.

5.1 Typology of weeks for departing trips from census tracts

5.1.1 Methodology

5.1.1.1 Method of clustering

The k-means clustering method is applied to create a typology of weeks. The k-means approach is implemented with the Lloyd's algorithm, using the Euclidean distance as the metric. For performing the k-means clustering, 50 iterations are used, and for finding the number of clusters, two approaches are used: the elbow method and the dendrogram.

The k-means clustering method is applied to the observations of the dataset corresponding to departing taxi trips from CTs over 12 weeks in 2019. For this purpose, we summarized the data by week, and created normalized CT-week vectors. By using this method, we aim to classify these CT-week vectors to develop a typology of weekly taxi usage.

5.1.1.2 Creation of vectors

Since the main purpose of this section is to characterize the weekly distribution of taxi trips departing from CTs, we created "CT-week" vectors to find the weekly usage of taxi trips during 12 weeks in 2019 (April, July and September). This method of creating vectors was inspired by Morency et al. (2017). "CT-hour" vectors were tested as well, but since it made the interpretation of the results more complex, and this work was inspired by another research as mentioned earlier, the "CT-week" vectors were taken into account. The CT-week indicator includes seven dispersion indicators of taxi trips and one intensity indicator. The dispersion indicators are the number of taxi trips departing from census tracts in each day of every week over a total of 12 weeks, and the intensity indicator is the average number of taxi trips per day of each week. These indicators are then normalized for applying the k-means clustering.

Finally, the database created for applying the k-means clustering approach consists of 532×12 vectors, for a total of 6384 vectors, where 532 is the number of CTs on the Island of Montreal and 12 is the number of weeks. To develop the week typology, only complete weeks are included in this analysis (from Monday to Sunday).

To include complete weeks in our dataset, we assigned a number from 1 to 12 to the weeks of all the months examined. We started from April, and since we wanted to have complete weeks, we excluded April 29th and 30th which were Monday and Tuesday. As a result, the days of April included in our analysis started from April 1st, which was Monday, to April 28th, which was Sunday. We assigned week numbers 1 to 4 to April's weeks. As for July, we considered July 1st to July 28th to have four complete weeks. So, July 29th, 30th and 31st were excluded. We assigned week number 5 to 8 to July's weeks. We also excluded September 1st which was Sunday, and September 30th which was Monday. We added week numbers 9 to 12 to September's weeks covering September 2nd to 29th.

Finally, we calculated the number of trips per day of the week for each CT and week, and then we merged all data of three months. This led to having a dataset including 12 weeks for each CT.

5.1.1.3 Normalization of vectors

To apply k-means clustering, it is necessary to normalize the variables to give them comparable weights. Different normalization methods were tested, and finally the normalization method which led to the best clustering was selected. Such normalization methods as using Z-score for dispersion

indicators and rescaling intensity indicator between 0 and 100 did not provide distinguishable clusters. So finally, the number of trips per day of each week were converted into percentage of taxi trips per week. For normalizing the intensity indicator (daily mean), we first used the logarithmic function to reduce the impact of outliers (the log function flattens the lowest and highest values), and then we rescaled the values between 0 and 100. This normalization method was inspired by Deschaintres et al. (2019) and Deschaintres (2018). The following equations indicate how the vectors were normalized.

$$V_{d,n} = \frac{\text{Number of trips per day of the week}}{\text{Total number of trips per week}} \times 100 \qquad \text{Equation 8}$$

where $V_{d,n}$ is the normalized dispersion indicator.

$$V_{i,log} = log_{10} V_i \qquad \text{Equation 9}$$

where V_i is the intensity indicator and $V_{i,log}$ is the intensity indicator after applying the logarithmic function.

$$V_{i,\ n} = \left(\frac{V_{i,log} - \min(V_{i,log})}{\max(V_{i,log}) - \min(V_{i,log})} \right) \times 100 \qquad \text{Equation 10}$$

where $V_{i,\ n}$ is the normalized intensity indicator.

Table 5.1 presents an extract of the CT-week vectors before normalization, and Table 5.2 indicates the components of Table 5.1 after normalization. Table 5.3 and Table 5.4 provide descriptive statistics of the indicators before and after normalization, respectively.

According to Table 5.3, Thursday had the highest maximum number of daily taxi trips per day, while the lowest maximum of the equivalent value was seen on Sunday. Since the value of log 0 with base 10 is not defined, it is important to identify if there are any observations with the value of 0. By looking at the daily mean (intensity indicator), we see that there is no observation with the value of 0 which means that there is at least one trip per week in each CT. Thus, for using logarithmic function to normalize the intensity indicator, there is no need to make any changes to the value of the intensity indicator.

As Table 5.3 indicates, the third quartile of daily mean (intensity indicator) before normalization was 41.14. This means that 75% of the observations had a daily mean of less than 41.14, while the maximum of daily mean was 874. The great different between the third quartile and the maximum

shows the existence of outliers or of a skewed distribution. However, Table 5.4 shows that daily mean was almost normally distributed after normalization.

Figure 5.1 shows the histogram of the intensity indicator normalized by rescaling between 0 and 100 without applying the logarithmic function. This figure shows that although the indicator was normalized, it even did not tend to have a normal distribution.

Figure 5.2 shows the histogram of the intensity indicator which was normalized using equations 9 and 10 by applying the logarithmic function. As this figure shows, the normalized intensity indicator tends to have a normal distribution which shows that the logarithmic function has reduced the impact of outliers.

Table 5.1 Extract of the CT-week vectors (taxi usage vectors) before normalization

Dispersion indicators

Intensity indicator

sridu_orig	week_no	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	daily_mean
4620001	1	8	7	4	9	5	3	2	5.428571
4620001	2	4	3	6	4	4	3	2	3.714286
4620001	3	4	3	4	4	2	4	4	3.571429
4620001	4	1	0	4	4	3	2	4	2.571429
4620001	5	8	16	11	13	10	11	4	10.428571
4620001	6	6	12	7	11	13	9	6	9.142857
4620001	7	6	8	1	6	18	6	7	7.428571
4620001	8	4	1	10	9	5	4	5	5.428571
4620001	9	4	3	5	4	5	11	8	5.714286
4620001	10	1	7	6	5	8	7	4	5.428571
4620001	11	4	4	3	6	6	4	4	4.428571
4620001	12	7	4	3	10	3	4	8	5.571429

Table 5.2 Extract of the CT-week vectors (taxi usage vectors) after normalization

Normalized dispersion indicators

Normalized intensity indictor

sridu_orig	week_no	Monday_stan	Tuesday_stan	Wednesday_stan	Thursday_stan	Friday_stan	Saturday_stan	Sunday_stan	daily_mean_st
4620001	1	21.052632	18.421053	10.526316	23.684211	13.157895	7.894737	5.263158	41.7202
4620001	2	15.384615	11.538462	23.076923	15.384615	15.384615	11.538462	7.692308	37.3678
4620001	3	16.000000	12.000000	16.000000	16.000000	8.000000	16.000000	16.000000	36.9179
4620001	4	5.555556	0.000000	22.222222	22.222222	16.666667	11.111111	22.222222	33.1503
4620001	5	10.958904	21.917808	15.068493	17.808219	13.698630	15.068493	5.479452	49.2082
4620001	6	9.375000	18.750000	10.937500	17.187500	20.312500	14.062500	9.375000	47.6991
4620001	7	11.538462	15.384615	1.923077	11.538462	34.615385	11.538462	13.461538	45.3176
4620001	8	10.526316	2.631579	26.315789	23.684211	13.157895	10.526316	13.157895	41.7202
4620001	9	10.000000	7.500000	12.500000	10.000000	12.500000	27.500000	20.000000	42.3085
4620001	10	2.631579	18.421053	15.789474	13.157895	21.052632	18.421053	10.526316	41.7202
4620001	11	12.903226	12.903226	9.677419	19.354839	19.354839	12.903226	12.903226	39.3851
4620001	12	17.948718	10.256410	7.692308	25.641026	7.692308	10.256410	20.512821	42.0181

Table 5.3 Descriptive statistics of dispersion and intensity indicators before normalization

Indicators	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Monday	0.00	11.00	20.00	32.26	36.00	915.00
Tuesday	0.00	13.00	22.00	36.45	41.00	934.00
Wednesday	0.00	13.00	23.00	37.90	42.00	1025.00
Thursday	0.00	15.00	26.00	43.53	48.00	1175.00
Friday	0.00	14.00	26.00	42.86	48.00	1004.00
Saturday	0.00	12.00	23.00	37.10	44.00	805.00
Sunday	0.00	10.00	18.00	29.11	33.00	624.00
Daily mean	0.14	13.43	22.71	37.03	41.14	874.00

Table 5.4 Descriptive statistics of dispersion and intensity indicators after normalization

Indicators	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Monday	0.00	10.00	12.71	12.82	15.38	100.00
Tuesday	0.00	11.59	13.95	14.11	16.38	54.55
Wednesday	0.00	12.21	14.40	14.54	16.67	42.86
Thursday	0.00	14.18	16.48	16.64	18.92	66.67
Friday	0.00	13.71	16.43	16.27	18.89	43.75
Saturday	0.00	11.46	14.25	14.30	16.94	71.65
Sunday	0.00	8.70	11.01	11.32	13.59	55.71
Daily mean	0.00	52.11	58.14	58.70	64.95	100.00

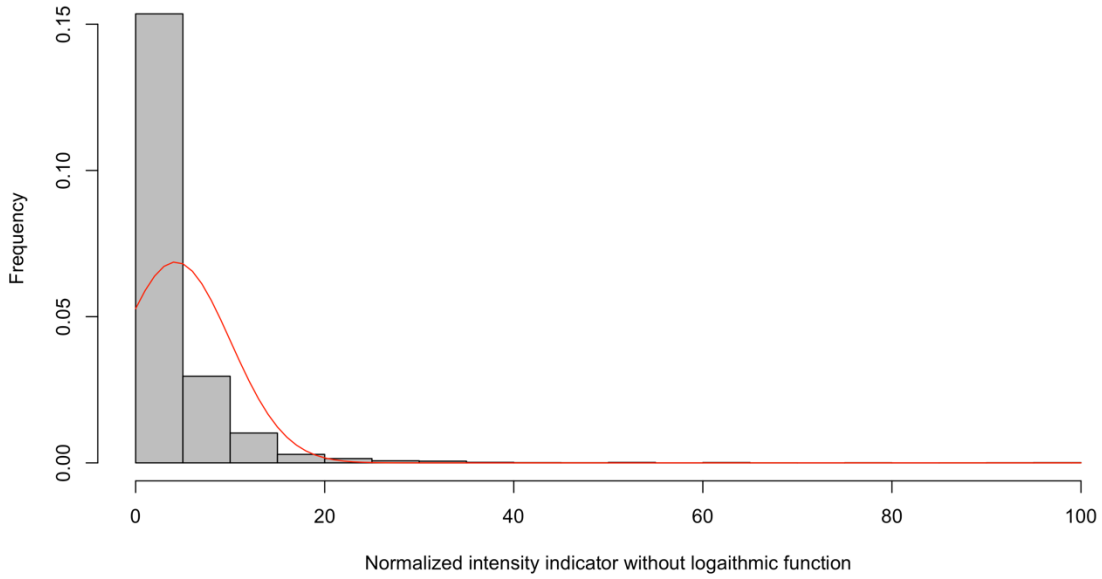


Figure 5.1 Histogram of the normalized intensity indicator without applying the logarithmic function

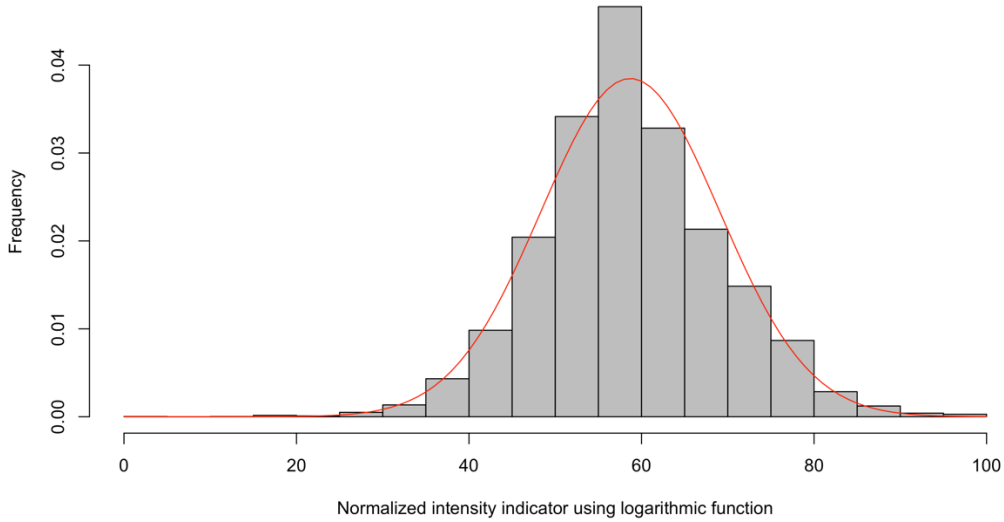


Figure 5.2 Histogram of the normalized intensity indicator using the logarithmic function

5.1.1.4 Develop an indicator to assess the temporal variability across weeks for CTs

This section aims to develop an indicator to assess the temporal variability across weeks for each CT, and compare the temporal variability within CTs which is inspired by Deschaintres et al. (2019) and Deschaintres (2018). For this purpose, we need to find the regularity of weekly patterns for each CT. By finding the repetition of the same weekly patterns of each CT over the 12 weeks, the regularity of a given CT can be analyzed. If all 12 weeks corresponding to each CT belong to

the same type of weeks (cluster), the CT is considered to have a regular temporal weekly pattern of taxi usage.

Equation 11 presents the Shannon entropy as the proposed indicator to measure the temporal variability across weeks for each CT, and equation 12 is used to normalize the Shannon entropy by rescaling the values between 0 and 1. In other words, this indicator is used to estimate the distribution of weeks of each CT over the five types of weeks (clusters). A lower entropy means that the weekly pattern of each CT is less diverse over clusters, so they have more regular temporal variation in terms of taxi trips.

$$H_i(X) = -\sum_{j=1}^{n} P_{ij}\, log\, P_{ij} \qquad \text{Equation 11}$$

$$H_i^*(X) = \frac{H_i(X)}{log(n)} \qquad \text{Equation 12}$$

where

H_i = Shannon entropy index,

H_i^* = Normalized Shannon entropy,

n = Number of different types of weeks (clusters) which is 5 in this study,

P_{ij} = Proportion of weeks in each CT_i (i=1:532) for the type of week j.

5.1.2 Results of developing a week typology

In this section, the results of the k-means clustering are presented. First, choosing the number of clusters is explained by the elbow method and the dendrogram from the hierarchical agglomerative clustering. Then the obtained clusters are presented which leads to developing a week typology, and then types are analyzed from both the temporal and spatial perspectives. The results of the normalized Shannon entropy which assesses the temporal variability across weeks in CTs are also presented.

As mentioned earlier, for performing the k-means clustering, it is necessary to choose the number of clusters a priori. In this study, the elbow method and hierarchical agglomerative clustering are used to determine the number of clusters.

Figure 5.3 shows the results of the elbow method for choosing the number of clusters K. In the elbow method, the optimal number of clusters is where the total within-clusters sum of squares

does not change remarkably after that K. In other words, the number of clusters is where the elbow is created on the graph. Figure 5.3 shows that the elbow is made where K=2, so based on this figure, we need to choose two as the number of clusters. But since two clusters did not capture the distinct weekly distributions that exist in dataset, we tried different number of clusters to find out which one can show different patterns with more granularity. Finally, we set K=5 which means five types of weeks are developed for the taxi usage. This is in line with the dendrogram as discussed below.

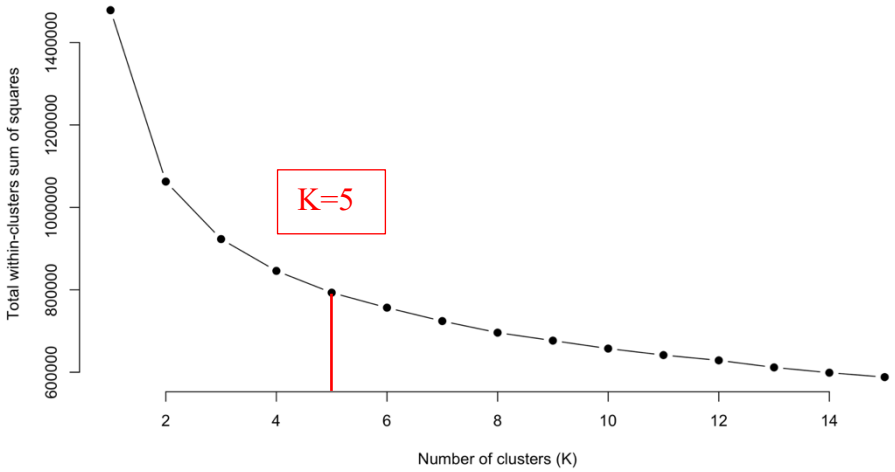


Figure 5.3 Choosing the number of clusters (K) with elbow method

Figure 5.4 shows the results of the hierarchical agglomerative clustering based on Euclidean distance for choosing the number of clusters K. According to the figure, K=5 is chosen since 5 distinguishable clusters are observed in the dendrogram.

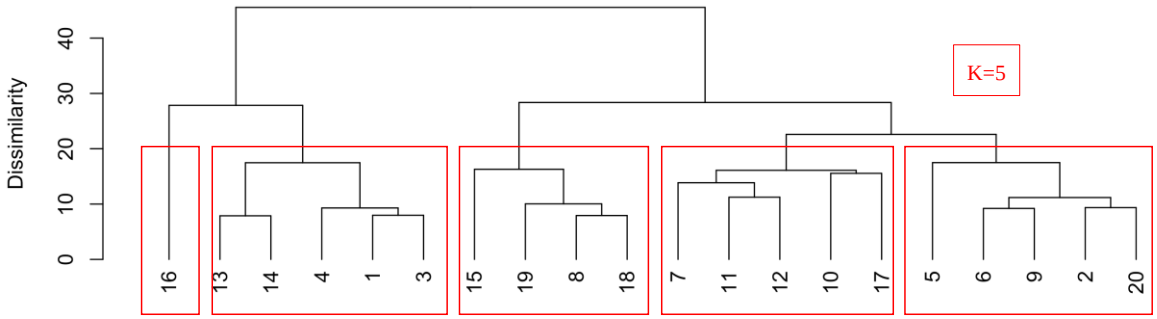


Figure 5.4 Choosing the number of clusters (K) with hierarchical agglomerative clustering

Figure 5.5 presents the results of k-means clustering by showing the temporal variation of departing taxi trips (average proportion of departing taxi trips per day) on the right and the average daily intensity on the left which is the normalized average number of departing taxi trips from each CT per day of each week.

The results of the k-means clustering including the percentage of CT-weeks in each cluster and the centers of the five clusters are also shown in Table 5.5. For each center, the average of normalized percentage of taxi trips per day of each week and the average of normalized daily mean (intensity indicator) are calculated.

According to Figure 5.5 and Table 5.5, the fourth group, W4, accounts for 27.4% of all the CT-week observations, which means that it is the most common type of week for using taxi trips. This group represents the week with the highest proportion of trips on Thursdays, which is the reason that this type of week is defined as “Thursday trips”. In this group, taxi trips are evenly distributed among Tuesday, Wednesday and Friday. The proportion of trips on Monday in this group is less than the other weekdays of this group. The lowest proportion of trips on Sundays occurs in this group.

In W1, the temporal variation from Monday to Tuesday and Wednesday to Thursday follows a similar pattern as the temporal variation in W4, with a lower proportion of trips before Thursday comparing to W4. From Friday to Sunday, the proportion of trips in W1 is higher than in W4. The proportion of trips on Saturday in W1 is higher than on Monday. It may be due to some events on Saturdays. Considering the intensity indicator of these two clusters helps us to compare them in terms of the number of trips: the intensity of trips is much higher for W1. According to the explanations provided above, this type of week is called end of week trips and due to its highest intensity indicator, this group has a high demand in terms of taxi usage. This is the reason that this group is called “end of week trips (high demand)”.

W2 has by far the highest proportion of trips on Fridays among all clusters, and it has the highest proportion of trips on Friday among all weekdays and weekend in this group. The intensity indicator of this group is very low. These are the two reasons for defining this group as “Friday trips (low demand)”. It is notable that in this group, the proportion of trips on Monday is less than on other days. The proportion of trips on Saturday is higher than the first weekdays in this group including Monday, Tuesday and Wednesday.

In W3, Monday to Thursday account for the majority of the trips with the lowest proportion of trips on Friday to Sunday. This cluster is the least common type of week for taxi usage, accounting for 11.5% of all CT-weeks, and it also has the lowest intensity indicator. This leads to defining this cluster as “weekday trips (low demand)”.

In W5, almost half of the trips are made on Friday, Saturday and Sunday. Since in this cluster, Saturday has the highest proportion, this cluster represents “weekend trips”.

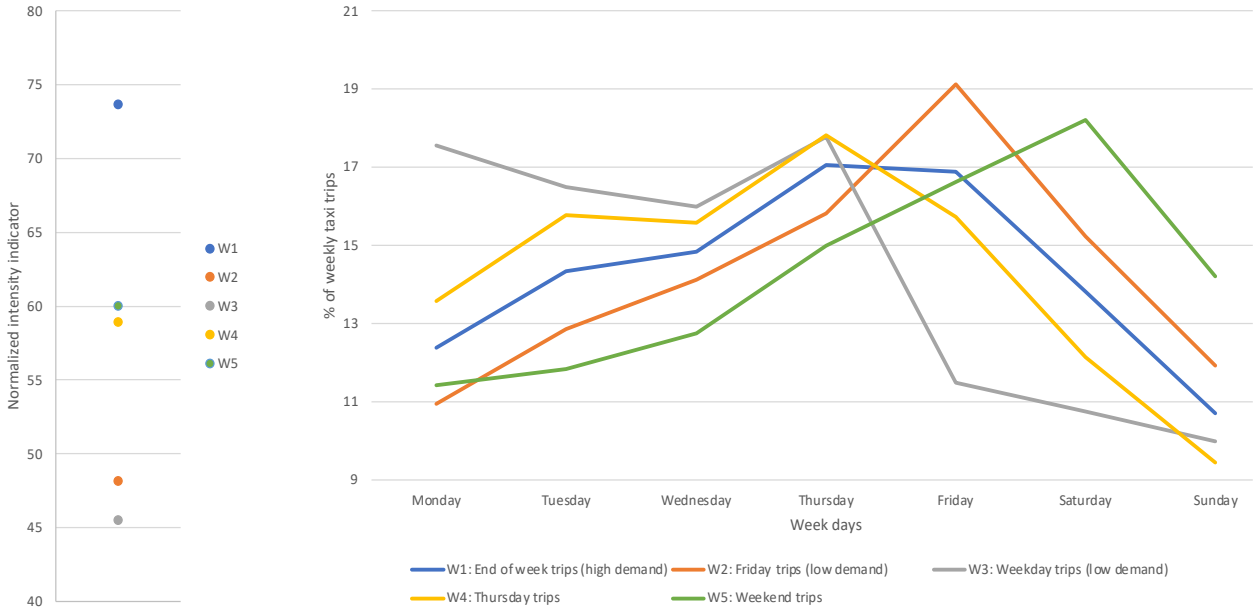


Figure 5.5 Centers of dispersion indicators (right) and the intensity indicator (left) of five clusters (types of the week)

Table 5.5 Percentage of CT-week vectors for each cluster (type of the week) and the average value of indicators (cluster centers)

Cluster	Nb CT-Week	Percentage	M (%)	T (%)	W (%)	Th (%)	F (%)	S (%)	S (%)	Intensity
W1: End of week trips (high demand)	1256	19.70	12.40	14.30	14.80	17.10	16.90	13.80	10.70	73.7
W2: Friday trips (low demand)	1103	17.30	10.90	12.90	14.10	15.80	19.10	15.20	11.90	48.2
W3: Weekday trips (low demand)	734	11.50	17.60	16.50	16.00	17.80	11.50	10.70	10.00	45.5
W4: Thursday trips	1751	27.40	13.60	15.80	15.60	17.80	15.70	12.10	9.40	59
W5: Weekend trips	1540	24.10	11.40	11.80	12.70	15.00	16.60	18.20	14.20	60

5.1.2.1 Temporal analysis of clusters

In this section, the previously obtained clusters are analyzed from a temporal perspective.

To better understand the temporal variation of taxi usage, the proportion of CTs in each cluster for the 12 weeks are shown in Table 5.6, and then the results are plotted in Figure 5.6.

Each column of the table shows for each week, the percentage of CTs that belong to each cluster. The darkest green color shows the highest percentage of CTs, and the darkest red corresponds to the lowest proportion of CTs per week (each column of the table).

Table 5.6 Proportion of census tracts in each cluster (type of the week) per week

	April				July				September			
Cluster	1	2	3	4	5	6	7	8	9	10	11	12
W1: End of week trips (high demand)	10	12	11	10	28	26	25	20	24	23	23	23
W2: Friday trips (low demand)	11	16	19	37	16	12	18	9	19	15	19	16
W3: Weekday trips (low demand)	26	18	19	5	3	8	6	25	6	7	6	9
W4: Thursday trips	32	38	35	18	27	32	21	31	28	30	21	18
W5: Weekend trips	21	16	16	29	27	22	30	15	23	24	31	35

Table 5.6 illustrates that the taxi demand patterns vary across the year, as the distribution of CTs in the different clusters is not stable temporally. Since in this research we have considered three different months of the year, including one in the spring, one at the beginning of summer, and one at the end of the summer when schools open, it is important to understand what is happening at different times of the year in terms of taxi usage.

In Figure 5.6, the x axis shows the four weeks of each month, and the y axis indicates the percentage of CTs belonging to each cluster. The weeks with (H) are the ones including a holiday in Quebec. A table including holidays in April, July and September 2019 is presented in Appendix A.

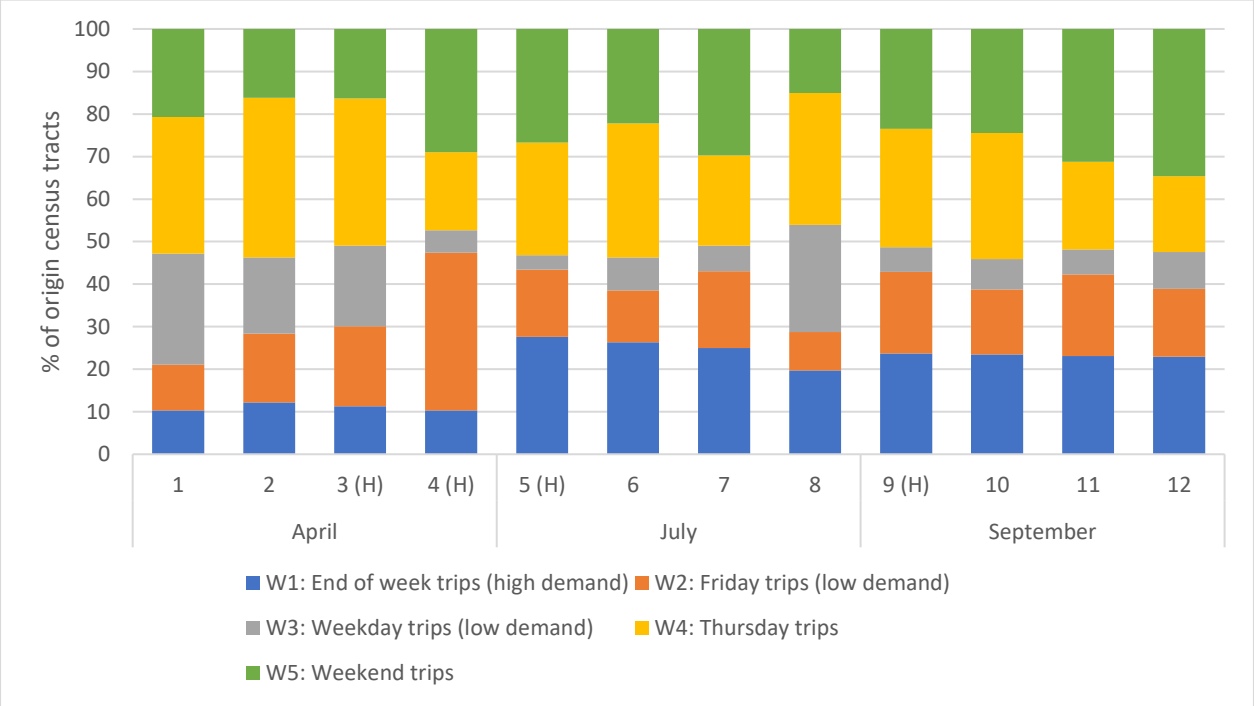


Figure 5.6 Proportion of CTs per week in 5 clusters

According to Figure 5.6, the proportion of CTs in W1 (end of week trips with high demand), is almost evenly distributed among April weeks, which is spring, and it increases during summer (July and September). In W1, the highest percentage of trips belonged to Thursday and Friday.

W2 (Friday trips) which included a higher proportion of taxi trips on Saturday than Monday, Tuesday and Wednesday, consists of most of the CTs during week 4 (last week of April). It is not clear why W2 is predominant in week 4 since the good Friday holiday is in the week 3.

The proportion of CTs belonging to W3 (Weekday trips with low demand) decreases remarkably during weeks 4 and 5. Weeks 4 and 5 included holidays of Easter and Canada day respectively, both on Monday. W3 has also a low proportion of CTs during weeks of September. W3 was the type of week with the highest proportion of trips from Monday to Thursday and the lowest proportion of trips on Friday.

W4 (Thursday trips) represents the highest proportion of CTs for 7 weeks out of 12. One of the highest proportions of CTs in this group belongs to week 3 including the holiday of good Friday.

Since this group had the highest proportion of CT-week trips on Thursday, it may correspond to the holiday of good Friday, where many taxi trips could have been made before the long weekend.

W5 (weekend trips), contains a high proportion of CTs in weeks 4, 5, 7, 11 and 12. Among these five weeks, two include a holiday. This can show that weekend trips do not depend on the holiday which exists in the week.

5.1.2.2 Spatial analysis of cluster

In this section, for each CT, we calculated the proportion of weeks that belong to each cluster which is presented in Figure 5.7. The spatial analysis of the clusters in addition to the temporal analysis help observe which areas have which temporal patterns of taxi usage.

According to Figure 5.7, the map for W1 (end of week trips with high demand) indicates that there is high percentage of weeks in the CTs containing or surrounding the airport. Percentage of weeks in W2 (Friday trips with low demand) and W3 (Weekday trips with low demand) follow a decentralized pattern, and high percentages are mostly concentrated on the eastern and western tips of the Island. In W4 (Thursday trips) and W5 (weekend trips), higher proportions of weeks are observed in central and pericentral neighborhoods.

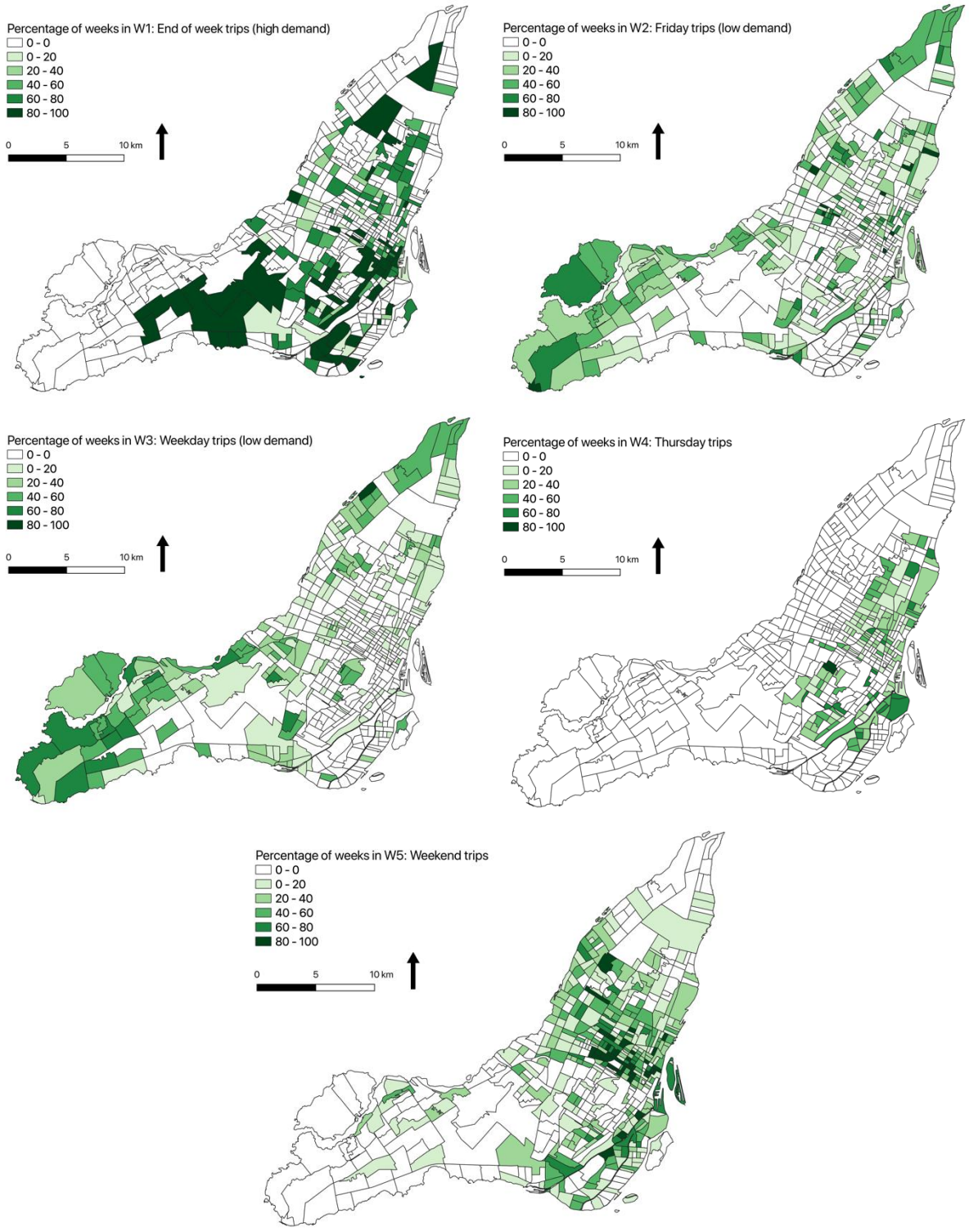


Figure 5.7 Percentage of weeks in each type of week (cluster) per census tract

5.1.2.3 Analysis of spatial variables in each cluster

In this section, for understanding the relationship between the determinants of taxi trips (independent variables discussed in chapter 3), and the developed week typology, the weighted mean of each variable is calculated for each week type (cluster). Table 5.7 shows the weighted mean of independent variables for each type of week. The darkest green shows the highest weighted mean of the variable, and the darkest red corresponds to the lowest value of weighted mean over the five clusters. This is an exploratory analysis for finding a relationship between these variables and taxi usage. The more precise models for finding the most significant variables are presented in chapter 6.

According to Table 5.7, density of health centers' workers, density of business workers, density of hotel rooms, density of eating places, metro stations, average of daily, weekday and weekend bus services have the highest weighted mean in W1 (end of week trips with high demand). This means that most of the variables have the highest weighted mean in the type of week that had the highest intensity indicator or the highest demand. This suggests that high density of activities and transport service are conducive to a high demand, and a higher proportion of trips at the end of the week. Density of drinking places, arts, entertainment, and recreational centers and density of bus stops have the greatest value of their weighted mean in W5 (weekend trips). This shows that high density of drinking places and bus stops and presence of recreational centers lead to a higher proportion of trips at weekends. Density of eating places have also a very high mean in this group. It was also found that high median income, percentage of recent immigrants, percentage of elderly people and number of cars per population are associated with higher trips on Thursdays (W4).

On the other hand, W3 (Weekday trips with low demand), has the lowest mean for all variables. This suggests that low density of activities and transport services, and low income, immigrants, elderly people and number of cars per population leads to a low demand, and a lower proportion of trips during weekdays.

Table 5.7 Weighted mean of independent variables

	Variables	W1: End of week trips (high demand)	W2: Friday trips (low demand)	W3: Weekday trips (low demand)	W4: Thursday trips	W5: Weekend trips
Socio-demographic	Median income*1000	53.89	44.17	26.64	74.48	66.10
	Percentage of recent immigrants	148.92	81.99	44.65	170.73	164.61
	Percentage of people aged 65 years old and over	310.92	208.15	133.32	367.15	289.08
	Number of cars per population	6.12	5.59	3.64	8.74	7.06
Land use	Health workers density	37643.11	8389.00	3958.43	23090.29	32443.96
	Business workers density	192002.60	37714.13	12910.43	107446.70	124160.90
	Density of hotel rooms	1027.37	153.39	76.27	293.45	507.13
	Density of drinking places	126.11	28.09	8.69	38.50	164.81
	Density of eating places	1288.86	254.84	88.39	428.14	1228.59
	Arts, entertainment, and recreational centers (dummy)	7.16	2.81	1.34	7.45	8.55
Transportation	Density of bus stops	653.37	375.52	166.43	627.52	789.32
	Metro stations (dummy)	4.55	0.56	0.14	1.65	2.38
	Average of daily bus services	778.89	310.23	187.35	695.94	635.53
	Average of weekday bus services	905.04	365.03	222.54	813.19	738.77
	Average of weekend bus services	463.53	184.59	109.16	415.30	385.24

5.1.2.4 Analysis of Shannon entropy

Figure 5.8 shows the temporal variability of taxi usage across weeks for CTs, based on the normalized Shannon entropy index. As Figure 5.8 shows, CTs with white color are those with regular weekly pattern of taxi usage. It means that all 12 weeks of these CTs belong to the same type of week. According to the figure, the CTs containing and surrounding the airport, downtown, Plateau Mont-Royal, Mount Royal Park, Westmount and Mile End have regular temporal variations of taxi demand. On the other hand, CTs with dark green color have irregular weekly pattern of taxi usage, meaning that their corresponding 12 weeks are assigned to different types of weeks. As figure shows, CTs with irregular pattern are dispersed across the Island of Montreal.

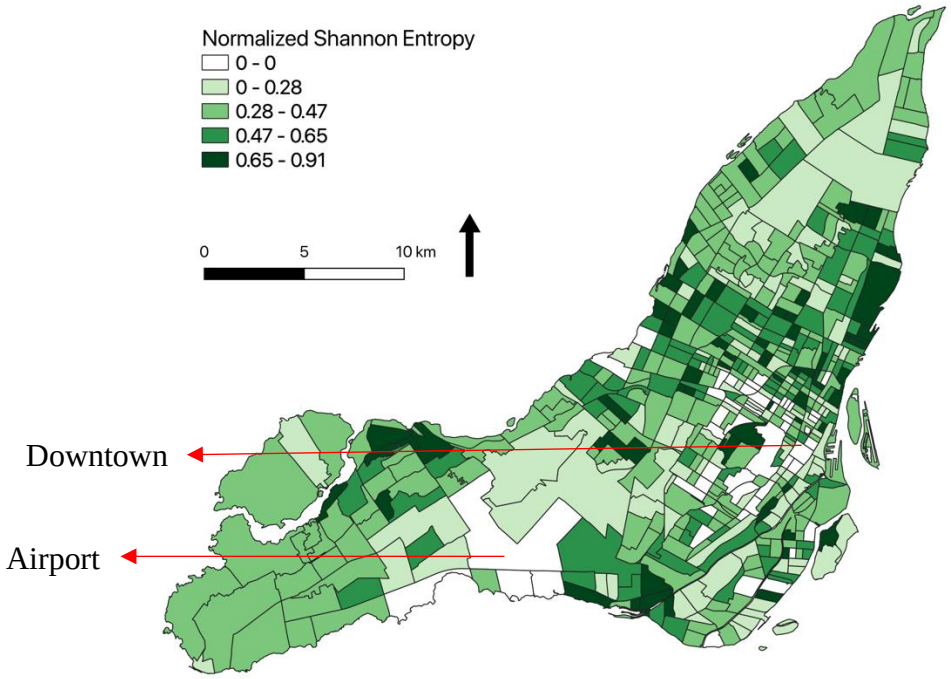


Figure 5.8 Temporal variability across weeks for CTs according to the normalized Shannon entropy

5.2 Analysis of sequences of week type

In the first section of this chapter, a week typology including five types of weeks was developed to measure the weekly temporal variability of taxi demand for each CT, over 12 weeks. In other words, five weekly patterns of taxi usage were developed at the CT level. In this section, the sequences of these five types of week are analyzed, and a six-multi-week (MW) typology is developed based on the five-week typology from the previous section.

5.2.1 Methodology

The following diagram indicates the steps of the methodology of this section which is applied to analyze the sequences of week types. Each step will be explained more precisely in the following sections of this chapter.

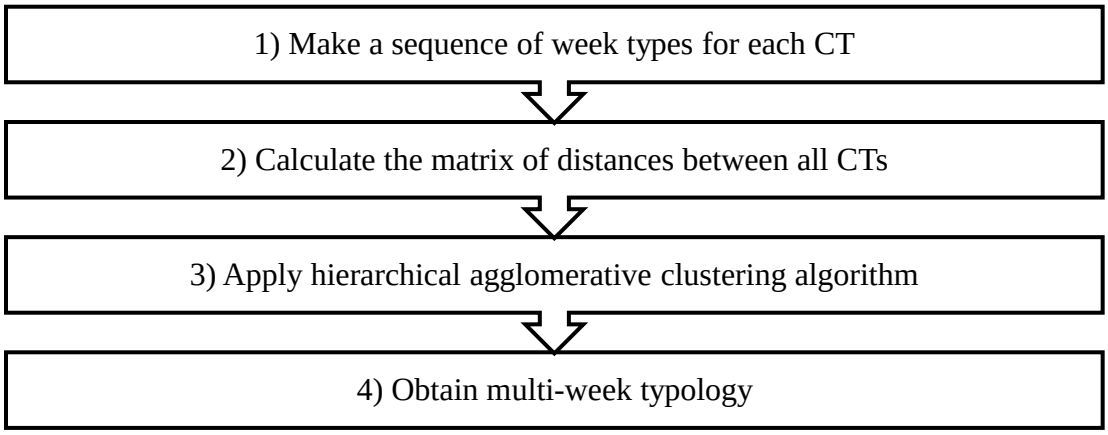


Figure 5.9 Steps of methodology applied for developing a multi-week typology

5.2.1.1 Making sequences

In section 5.1.1.2, we created “CT-week” vectors (Table 5.1) to measure the weekly temporal variability of taxi usage for each CT. For the purposes of this section, “CT” vectors are created in the format of ordered sequences of 12 weeks. In the other words, a table of 532 sequences is created which includes one sequence per CT. For each CT and each week (1 to 12), there is one type of week, cluster 1 (W1) to cluster 5 (W5), developed in the previous section using k-means clustering. Table 5.8 shows an extract of the table of sequences of CT vectors. Each row represents a CT, and each column represents a week. The value then represents that week type to which the CT belong for each of the 12 weeks.

For example, for the first CT (4620001.00), weeks 1, 2, 3, 5, 8, 12 corresponds to W3, and weeks 4, 6, 7, 9, 10, 11 belong to W2 developed in the previous section. Hence, the sequence for the first CT is 333232232223.

Table 5.8 Extract of the CT vectors in the format of ordered sequences of 12 weeks

sridu_orig	1	2	3	4	5	6	7	8	9	10	11	12
4620001.00	3	3	3	2	3	2	2	3	2	2	2	3
4620002.00	3	3	3	3	2	2	2	3	2	2	2	3
4620003.00	3	2	3	2	4	4	5	4	4	4	4	2
4620004.00	2	3	3	3	1	1	1	1	1	1	1	1
4620005.00	3	2	3	2	4	4	4	4	4	4	4	4

As Deschaintres (2018) explained, each vector performs as a string. In this study, each vector is a string with the length of 12, which is the number of weeks for 3 months, and each string is made of 12 characters that represent one of the 5 week types developed in the previous section. According to the length of 12, and 5 week types, 5^{12} unique sequences are possible.

5.2.1.2 Calculating the matrix of distances between all CTs

In the previous section, 532 sequences were created based on the new CT vectors. The purpose of this section is to calculate the distance matrix between all pairs of sequences. This matrix will later be used to re-cluster the sequences of the types of weeks developed in section 5.1. Deschaintres (2018) tried both Levenshtein distance and modified Hamming distance, known as weighted Hamming distance, to calculate the distance between strings of characters using smart card data and found the modified Hamming distance provided more interesting results. Due to this reason, in this section, we use modified Hamming distance to calculate the distance between pairs of sequences.

The traditional Hamming distance calculates the distance between two sequences of equal length by counting the number of different characters. If the characters are similar, "0" is added, and if the characters are different, "1" is added. The sum of all 0 and 1 then provides the Hamming distance (Deschaintres, 2018). For example, the traditional Hamming distance between the 1st and 2nd sequence of Table 5.8 is calculated by summing the distance between each character of the sequences which is presented in Table 5.9.

The traditional Hamming distance between two CTs of Table 5.9 is as below:

$$d(CT4620001.00, CT4620002.00) = d(333232232223, 333322232223) = 0+0+0+1+1+0+0+0+0+0+0+0 = 2 \quad \text{Equation 13}$$

Table 5.9 Example of calculating the traditional Hamming distance between 1st and 2nd sequence of Table 5.8

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
4620001.00	3	3	3	2	3	2	2	3	2	2	2	3
4620002.00	3	3	3	3	2	2	2	3	2	2	2	3
Distance	0	0	0	1	1	0	0	0	0	0	0	0

As mentioned earlier, in this study, a weighted Hamming distance is applied rather than the traditional Hamming distance. In the weighted Hamming distance, the Euclidean distance between each character of the string in calculated. In this study, the Euclidean distance between the center of each week type between the two sequences from the same week (1:12) is calculated, and then the sum of all Euclidean distances between centers of the types of the week is summed over 12 weeks. The following equation represents the mathematical formula for calculating the distance between two sequences (Deschaintres, 2018).

$$d(i, j) = \sum_{k=1}^{N} d_E\left(W_{(i,k)}, W_{(j,k)}\right) \quad \text{Equation 14}$$

where N is the number of weeks which is 12 here,

d_E is the Euclidean distance,

$W_{(i,k)}$ is the type of week belonging to the kth week and ith CT,

$W_{(j,k)}$ is the type of week belonging to the kth week and jth CT.

Table 5.10 and equation 15 show how equation 14 is applied to compute the weighted Hamming distance between two sequences.

$$d(CT4620001.00, CT4620002.00) = d_E(3,3) + d_E(3,3) + d_E(3,3) + d_E(2,3) + d_E(3,2) + d_E(2,2) + d_E(2,2) + d_E(3,3) + d_E(2,2) + d_E(2,2) + d_E(2,2) + d_E(3,3) \quad \text{Equation 15}$$

Finally, a matrix of weighted Hamming distances was made between all pairs of CTs with the dimension of 532×532. The hierarchical agglomerative clustering is then applied on this matrix.

Table 5.10 Example of calculating the weighted Hamming distance between 1st and 2nd sequence of Table 5.8

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
4620001.00	3 $W_{(1,1)}$	3 $W_{(1,2)}$	3 $W_{(1,3)}$	2 $W_{(1,4)}$	3 $W_{(1,5)}$	2 $W_{(1,6)}$	2 $W_{(1,7)}$	3 $W_{(1,8)}$	2 $W_{(1,9)}$	2 $W_{(1,10)}$	2 $W_{(1,11)}$	3 $W_{(1,12)}$
4620002.00	3 $W_{(2,1)}$	3 $W_{(2,2)}$	3 $W_{(2,3)}$	3 $W_{(2,4)}$	2 $W_{(2,5)}$	2 $W_{(2,6)}$	2 $W_{(2,7)}$	3 $W_{(2,8)}$	2 $W_{(2,9)}$	2 $W_{(2,10)}$	2 $W_{(2,11)}$	3 $W_{(2,12)}$
Distance	0	0	0	$d_E(2,3)$	$d_E(3,2)$	0	0	0	0	0	0	0

5.2.1.3 Applying Hierarchical agglomerative clustering

Deschaintres (2018) tested different algorithms such as single linkage, complete linkage, centroid linkage on a matrix of weighted Hamming distances created using smart card data. Comparing the results obtained from different methods indicated that the Ward's method (Ward Jr, 1963) was the most appropriate one in terms of interpreting the results. For this reason, we also used the Ward's algorithm which functions by minimizing the total within-cluster variance for developing a multi-week typology using the *hclust* function and the *ward.D2* method in R. Ward's method is more precisely explained by James et al. (2013), Murtagh and Legendre (2014) and (Ward Jr, 1963).

5.2.2 Results of developing sequences of week type

5.2.2.1 Analysis of sequences

This section aims to analyze the sequences which were constructed in 5.2.1.1 from different perspectives. Among the 532 sequences created in this study, 419 unique sequences were observed. This suggests that the distribution of week types in sequences varies greatly.

Table 5.11 shows the most frequent sequences, and the proportion of CTs belonging to each of them.

Table 5.11 Frequency of the most frequent sequences

Sequence	**Frequency**	**Proportion (%)**
111111111111	40	7.5
555555555555	16	3.0
444444444444	11	2.1
444411111111	9	1.7
555511111111	5	0.9
555555545555	4	0.8
		Total=16%

The table demonstrates that the most frequent sequence is observed among 40 census tracts which accounts for 7.5% of all census tracts. This most frequent sequence includes only type W1 (end of week trips with high demand). According to the table, the most frequent sequences only accounts for 16% of all CTs which means that sequences of weekly patterns of taxi demand is dispersed among CTs.

Table 5.12 shows the percentage of CTs having different numbers of unique clusters and Figure 5.10 illustrates the number of unique clusters per CT. For example, CT 4620001.00 in the first row of Table 5.8 includes two unique clusters: W2 and W3. As Table 5.12 presents, about 70% of CTs have 2 or 3 unique clusters (week types), and less than 1% of CTs have all 5 week types.

Table 5.12 Percentage of CTs with different number of unique clusters

Count of unique clusters	Number of CTs	Percentage of CTs
1	68	12.8
2	208	39.1
3	167	31.4
4	87	16.4
5	2	0.4

Figure 5.10 demonstrates the spatial distribution of number of unique clusters (number of week types developed in the previous section). Figure 5.10 shows that the CTs with only 1 unique cluster (CTs with only 1 weekly pattern over 12 weeks) are mostly concentrated around the CTs containing and surrounding the airport, and industrial zones. Moreover, CTs with 2 clusters (2 weekly patterns over 12 weeks) are mostly located in the East, West and central neighborhoods.

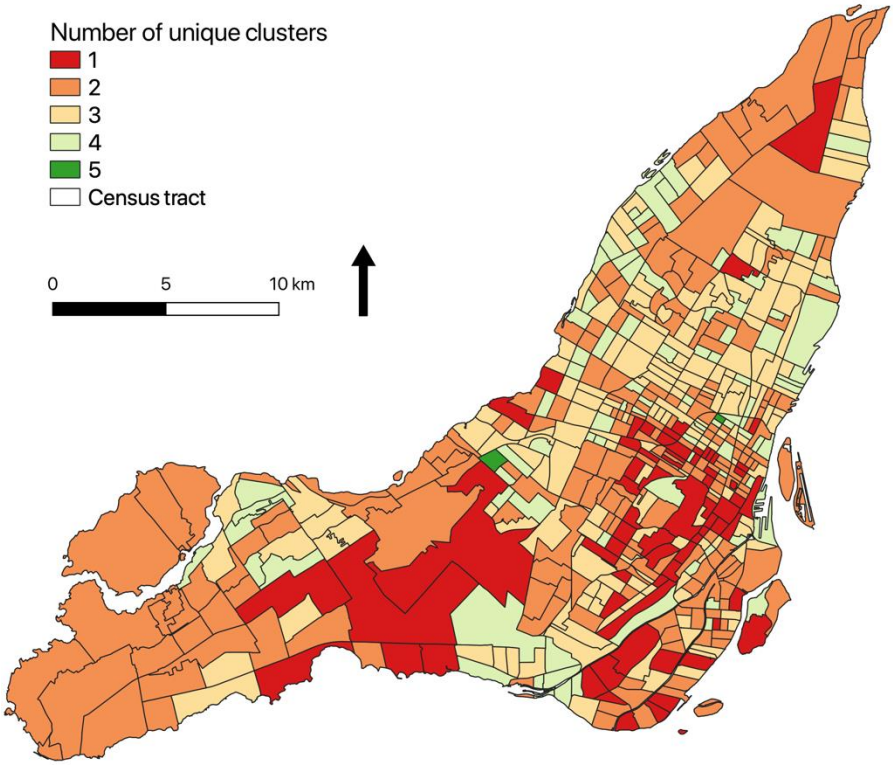


Figure 5.10 Number of unique clusters per CT

As another step of exploratory analysis of the sequences, the dominant cluster (the week type with the highest frequency) in each sequence was identified which is presented in Table 5.13. Since in some sequences more than one week type had the same frequency, the dominant cluster was identified randomly among the week types with the higher frequency. Based on Table 5.13, in more than 50% of CTs, the dominant clusters were W1 and W4 which were end of week trips with high demand and Thursday trips respectively.

Table 5.13 Percentage of CTs for each dominant cluster (type of the week)

Dominant cluster	Number of CTs	Percentage of CTs
1	125	23.5
2	94	17.7
3	38	7.1
4	163	30.6
5	112	21.1

5.2.2.2 Multi-week typology

The Ward's algorithm produced the dendrogram of Figure 5.11. We tried different number of clusters, and compared their results, and decided that K=6 better presents different multi-week types which will be presented in Figure 5.11.

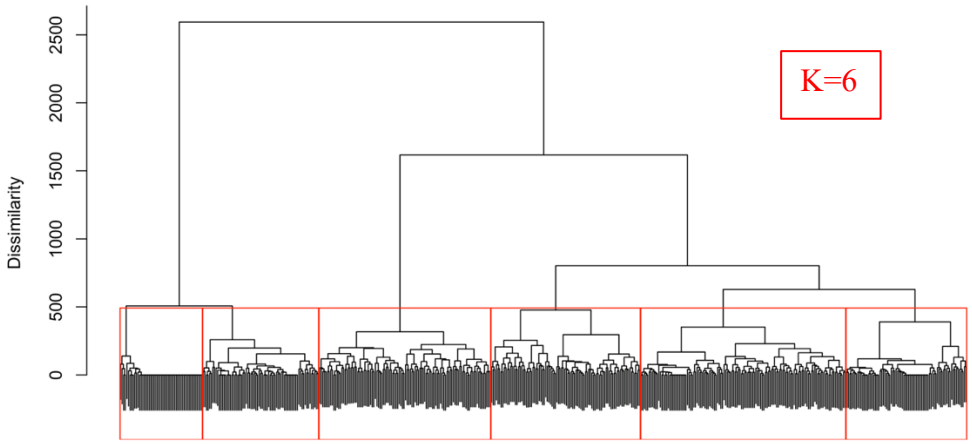


Figure 5.11 Dendrogram as a result of hierarchical agglomerative clustering based on ward's method

Figure 5.12 shows the results of re-clustering the clusters obtained in the first section of this chapter. This re-clustering resulted in 6 clusters which are ordered based on the percentage of CT in each cluster. The x axis on the map shows the 12 weeks, and the y axis presents the census tracts. The percentage of each week type developed in section 5.1.2 are presented at the right side of the figure of each cluster.

In cluster 1 (MW1), more than half of the weeks correspond to W4. This means that taxi trips departed from CTs of this group were mostly made on Thursdays. W5 (Weekend trips) is the second most common week in this group. W1 (End of week trips with high demand), W2 (Friday trips with low demand) and W3 (Weekday trips with low demand) account for less than 7% of the weeks in this cluster.

Cluster 2 (MW2) is mostly made up of W2 (Friday trips) and W3 (weekdays trips), both with low demand, which account for more than 92% of CT-weeks.

Cluster 3 (MW3) can be interpreted separately for weeks in the spring (1 to 4), and weeks in the summer (5 to 12). In the first 4 weeks (April), most of the weeks are observed as W2 which is Friday trips (low demand). For the rest of the weeks (July and September), the dominant types of week are W4 (Thursday trips) and W5 (Weekend trips).

In cluster 4 (MW4), W5 is the most common type of week observed over the 12 weeks, accounting for about 74% of the weeks. This means that CTs which exist in this cluster, are CTs with a high proportion of trips made during the weekend.

Like cluster 3 (MW3), cluster 5 (MW5) has two completely different patterns for spring and summer. In April (weeks 1 to 4), W4 and W5, which are Thursday and Weekend trips respectively, are the most significant types of the week. However, in July and September End of week trips with high demand (W1) is remarkably the most common type of week.

Finally, the last cluster (MW6) which accounts for less than 10% of the sequences, has W1 as its very dominant cluster. It means that in this cluster, more than 95% of CT-weeks belong to W1 (End of week trips with high demand).

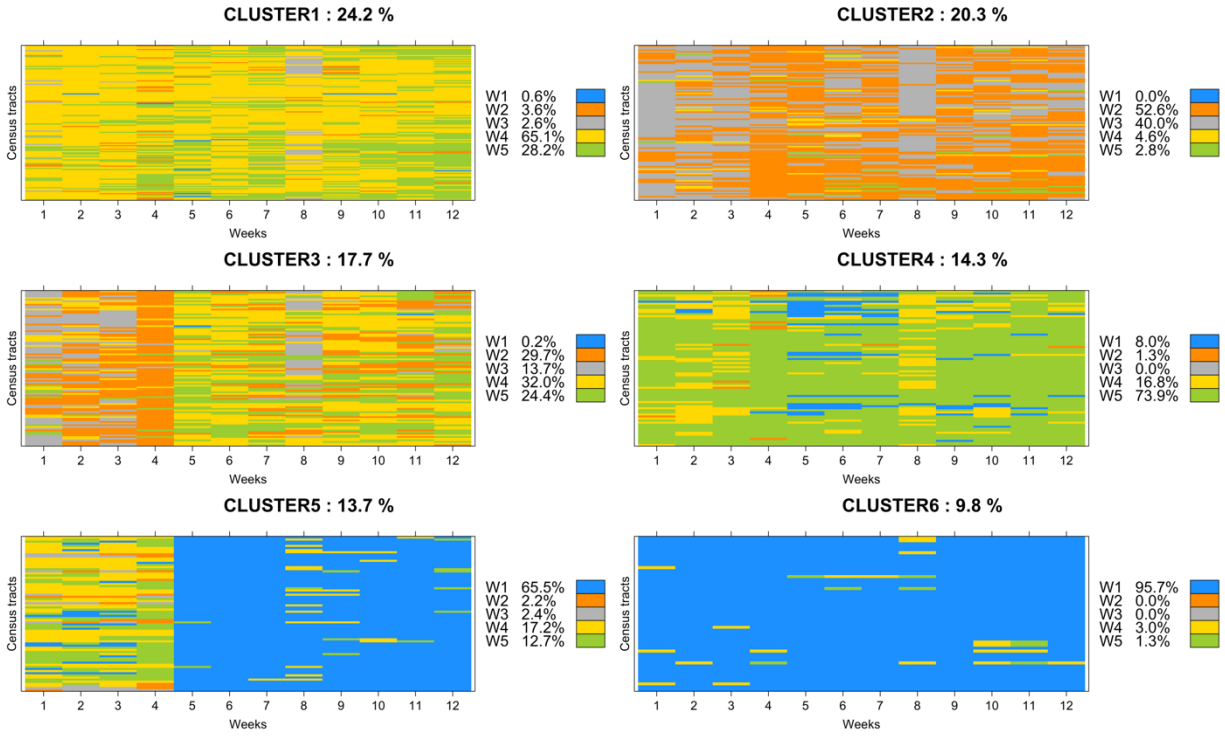


Figure 5.12 Multi-week types

Figure 5.13 shows the spatial distribution of the clusters interpreted above. According to Figure 5.13, Cluster 1 (MW1) (Thursday trips), which was the most frequent multi-week cluster follows a decentralized pattern scattered across the Island. Cluster 2 (MW2) (Friday trips and weekday trips with low demand) is mostly concentrated to the East and West of Island of Montreal. For Cluster 3 (MW3) (Friday trips for winter and Thursday and weekend trips in Summer), no special pattern has been observed. Cluster 4 (MW4), which mostly includes weekend trips, is concentrated at downtown, Plateau Mont-Royal and Rosemont. Cluster 5 (MW5) (Thursday and weekend trips during Winter and end of week trips during summer) was found close to metro stations. Finally, cluster 6, which was composed of high demand trips at the end of the weeks, was observed at the CTs containing and surrounding the airport, and industrial zones.

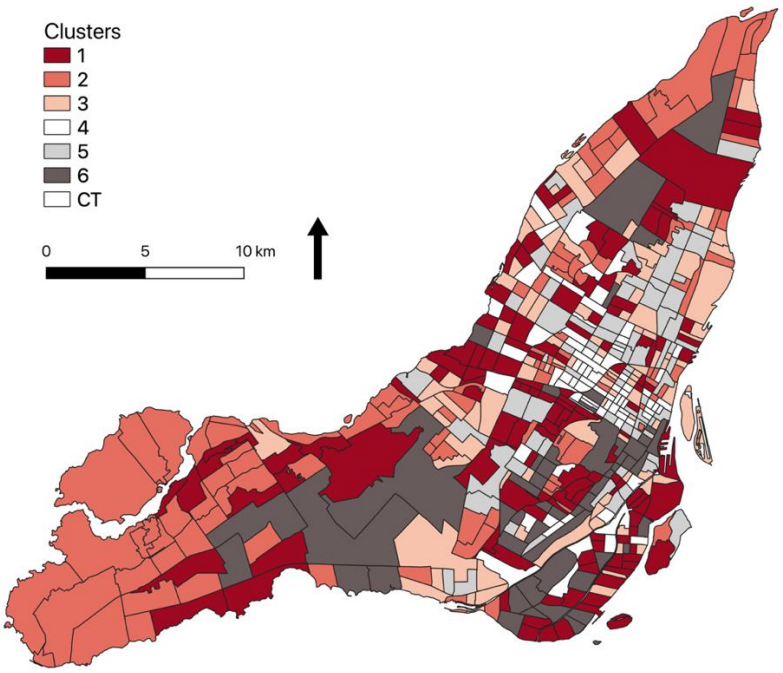


Figure 5.13 Spatial distribution of multi-week clusters

Developing a week typology, and a multi-week typology for taxi usage in this chapter, in addition to analyzing them indicate that patterns do vary importantly both spatially and temporally which makes understanding the variation of taxi demand a complex task.

CHAPTER 6 MODELING TAXI DEMAND

This chapter focuses on identifying the factors which can explain the multi-week typology obtained previously, and testing models to predict which multi-week clusters the CTs belong to. To this end, this chapter explores how different models such as a decision tree and a multinomial logit model (MNL) can contribute to identifying key determinants and predict patterns.

6.1 Data

The independent variables used for the analyses of this chapter include median income, percentage of recent immigrants, percentage of elderly people, ratio of cars to people, density of hotel rooms, density of health care workers, density of business, finance and administration workers, density of drinking places, density of eating places, presence of any art and recreational centers, presence of metro stations, average of daily bus services, average of bus services on weekdays and average of bus services during weekends. All these variables were aggregated at the CT level. The observations with 0 and few populations were excluded from the analysis. These data are presented in Table 3.3 with their description and sources. The dependent variable or the target variable of the decision tree and MNL is the multi-week type obtained in chapter 5.2.2.2, and it is used as a categorical variable with six categories (clusters).

6.2 Methods

As the first methodology in this chapter, a decision tree analysis is applied, and then a MNL model is conducted to compare the results and the performances of the models in terms of predicting which multi-week clusters the CTs belong to.

6.2.1 Decision tree

Decision tree is explained in detail in chapter 3.2.3. In this study, the decision tree is generated using the Ctree method. For making the decision tree using the Ctree method, different percentages were tested for dividing the dataset into a test and a train data set, and the percentage which leads to the highest accuracy of the model was selected. Finally, 85% of the dataset was selected as the training dataset, and the remaining 15% was used as the test dataset for validation and determining the accuracy of the model. The maximum depth for making the decision tree was selected as four after trying different numbers for the maximum depth which did not change the accuracy of the

model remarkably. A tree's depth indicates how many splits it can make before arriving at a prediction. By selecting a specific maximum depth, the tree stops growing after it reaches the certain depth. This would avoid the tree from having many repetitions which can lead to overfitting on the training dataset (Galarnyk, 2019).The results of the decision tree analysis are presented in the next section.

6.2.2 Multinomial logit model

As a first step of applying MNL model, a Pearson correlation analysis is applied between the exploratory variables. In order to prevent multicollinearity issues from happening, it is important to assess existing correlations between independent variables. Figure 6.1 illustrates the correlation between exploratory variables. As shown in the figure, density of drinking places (dn_dri_pl) and eating places (dn_eat_pl) are highly correlated. Both variables were tested in the model, and only density of eating places had significant coefficients. Thus, only density of eating places was included in the model. Furthermore, average daily bus services (avg_GTFS) is highly correlated with average weekend bus services (Weekend_GTFS_Mean) and average weekday bus services (Weekday_GTFS_Mean). Due to this high correlation, and for focusing on the overall service for the whole week, average weekend and weekday bus services were not included in the model. Other variables were included in the model, and finally, only those which were significant in any cluster were kept in the model.

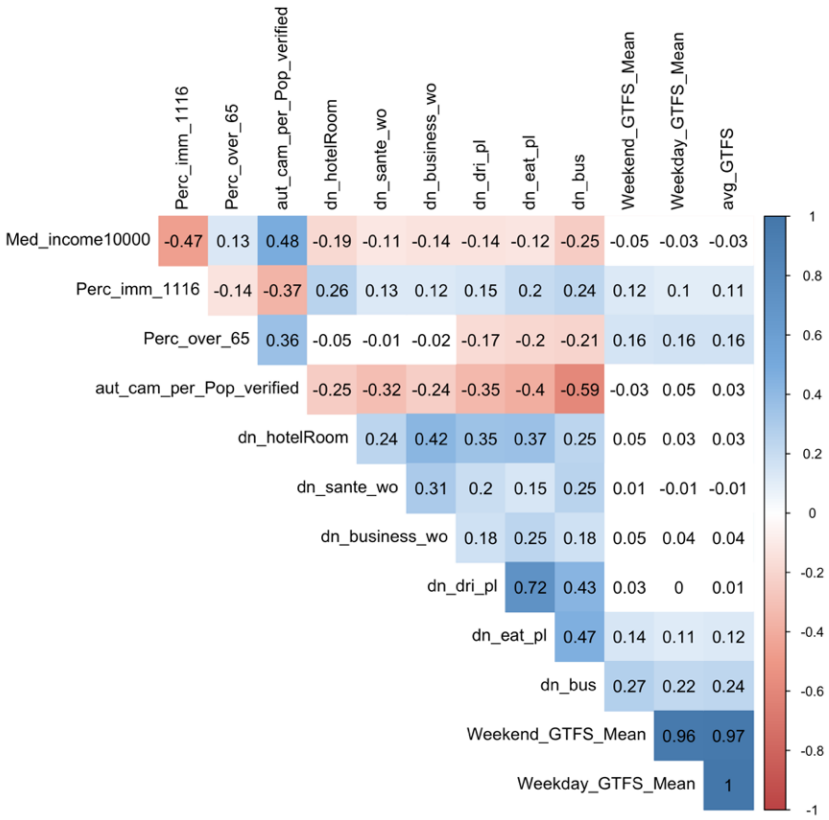


Figure 6.1 Correlation between independent variables of MNL model

6.3 Results

6.3.1 Results of the decision tree

In this section, firstly, the decision tree is interpreted, and then the results of the validation of the decision tree analysis are provided. As mentioned in section 6.2.2, since the focus was on the overall bus service for the whole week, average of bus services per weekday and weekend was not included in the analysis.

Figure 6.2 represents the decision tree generated by the CTree method with a controlled depth of four levels. The figure shows that number of cars per population, average of daily bus services, density of drinking places and density of eating places are the significant variables in the model. According to the figure, a CT with a lower number of cars per population, lower average of daily bus services, and higher density of eating places is more likely to belong to cluster 4 (MW4) than any other clusters. On the other hand, CTs with a higher number of cars per population, low density

of drinking places and low average of daily bus services are more likely to belong to cluster 2 (MW2). Furthermore, in CTs with low number of cars per population, and average of daily bus services less than 32.77, density of eating places plays an important role for belonging to a specific cluster. CTs with the above-mentioned characteristics, and the density of eating places between 11.99 and 40.72 are mostly likely to belong to MW1. On the other hand, if the density of eating places is less than 11.99, MW1 is no longer the dominant cluster. However, the patterns are not clear. In most of the circles there are all 6 clusters or 5 clusters, and there is not one cluster that is clearly dominant.

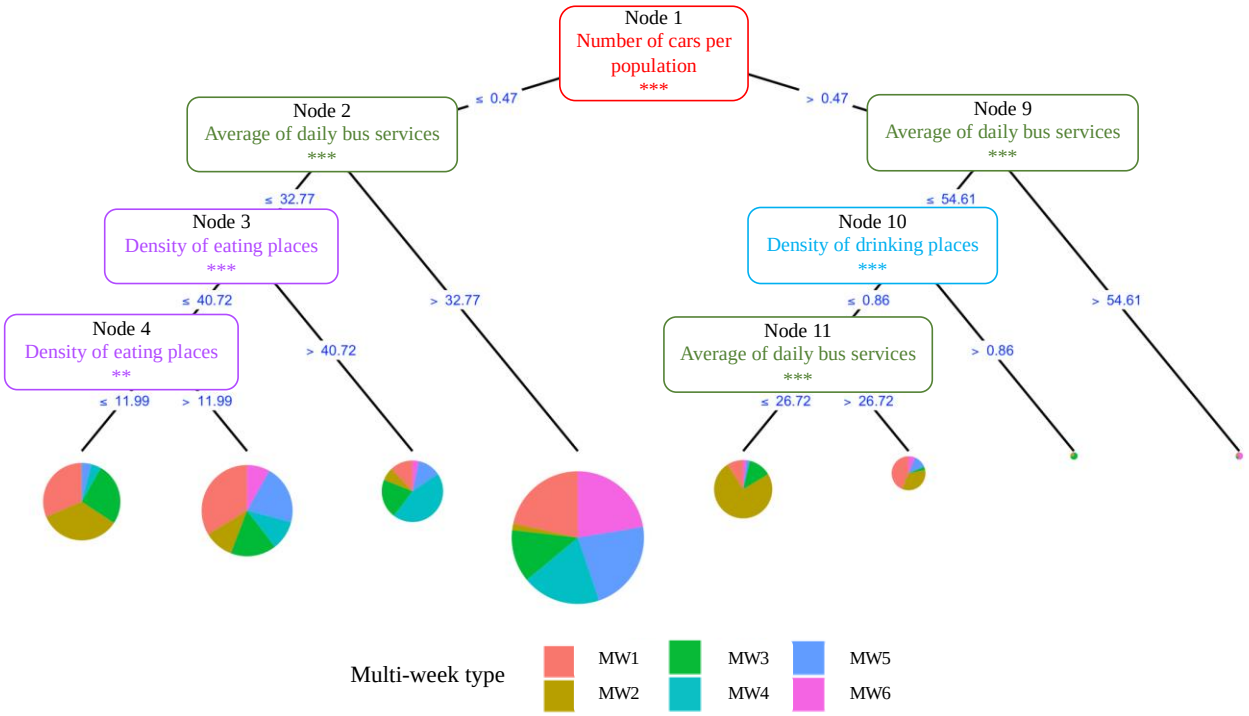


Figure 6.2 CTree decision tree with a controlled depth of four levels
*** = |p| < 0.01, ** = |p| < 0.05, * = |p| < 0.10.

The accuracy of this model is 0.36 which is low. This suggests that the model is not able to predict which multi-week clusters the CTs belong to. To further understand the performance of the model, the confusion matrix is presented.

Table 6.1 illustrates the confusion matrix of the decision tree with the multi-class target variable. In a multi-class classification model, the performance measures are calculated separately for each

class. In this study, the classes correspond to each MW type, which was developed in the previous chapter.

Since the confusion matrix explained in chapter 3.2.3 was a confusion matrix for a binary classification, finding TP (true positive: the number of positive predictions correctly identified as "Positive" by the classifier), TN (true negative: the number of negative predictions correctly identified as "negative" by the classifier), FP (false positive: the number of negative predictions incorrectly identified as "positive" by the classifier) and FN (false negative: the number of positive predictions incorrectly identified as "negative" by the classifier) in Table 6.1 is not as clear as it was for the binary class confusion matrix. In a multi-class machine learning model, TP occurs when the actual value and the predicted value are the same. For a given class, the TN will be the sum of all values except the values of that class. A class's FP value is derived from the sum of its columns excluding the TP value, and the FN value of a class is the sum of its rows except the TP value (Bharathi, 2021).

As an example of how to derive the different components of the confusion matrix for each individual class in Table 6.1, the values of the confusion matrix used to derive the four components are highlighted for cluster 1 with different colors. The colors corresponding to TP, TN, FP and FN are green, red, blue and orange respectively in the Table 6.1.

Table 6.1 Confusion matrix of the multi-class classification of this study

		True class					
	Clusters (MW)	**1**	**2**	**3**	**4**	**5**	**6**
Predicted class	**1**	2	3	5	2	2	0
	2	6	9	7	2	0	0
	3	2	0	0	0	1	0
	4	2	0	0	7	2	0
	5	8	1	2	2	6	3
	6	0	0	0	0	0	4

The sensitivity and specificity are presented in Table 6.2 for each cluster (each target variable). Table 6.2 illustrates that the highest performance of the model in terms of the true predicted positive samples (sensitivity) is observed in MW2. The 0 sensitivity of MW3 means that it was the cluster which observations were not able to be forecasted. Thus, MW3 has the lowest performance in terms of true predicted positive observations. However, MW6 has the best performance regarding specificity which focuses on true predicted negative samples.

Table 6.2 Model performance

	MW1	**MW2**	**MW3**	**MW4**	**MW5**	**MW6**
Sensitivity	0.10	0.69	0.00	0.54	0.56	0.57
Specificity	0.79	0.77	0.95	0.94	0.76	1.00

6.3.2 Results of MNL

Table 6.3 presents the results of the MNL model. Since MW1 (having Thursday trips as the most common week type and weekend trips as the most second common week type) was the most frequent multi-week cluster, it was selected as the reference category in the model. The numbers in each column of the table are the coefficient, and the number in parentheses are the P-values. MacFadden R^2 and Log-likelihood show the fit of the model which are presented below the table. The closer MacFadden R^2 to 1 makes Log-likelihood closer to 0, which means the better fit of the model.

In terms of socio-demographic characteristics of the CTs, median income has the most significant impact on MW2 (mostly made up of Friday trips and weekdays trips both with low demand) and MW3 (in MW3, most of the weeks are observed as Friday trips (low demand) in April, and Thursday trips and weekend trips in July and September) comparing to MW1. The negative coefficients show that a higher median income in a CT decreases the probability of a CT to belong to MW2 and MW3 as compared to MW1. The percentage of elderly people has a significant impact on belonging to MW2, MW4 (mostly including weekend trips) and MW6 (the least frequent cluster which mostly includes end of week trips with high demand): a lower percentage of elderly population in a CT increases its probability of belonging to MW2 and MW4, and a higher

percentage of elderly people in a CT makes it significantly more likely to belong to MW6. This finding suggests that the higher percentage of elderly people is associated with higher taxi demand. It is also found that CTs with a higher number of cars per person are significantly more likely to belong to MW2 comparing to MW1. This finding highlights that in CTs with higher car ownership there is lower taxi demand.

Regarding land use variables, only density of restaurants and absence of art, entertainment and recreational centers had a significant impact on some clusters. CTs with a higher density of restaurants are significantly likely to belong to clusters MW4, MW5 (This MW has Thursday and Weekend trips as the most significant types of the week in April and end of week trips with high demand in July and September), and MW6. This result shows that more eating places in a CT is associated with weekend trips and high taxi demand. Furthermore, CTs where there is art, entertainment and recreational center are less likely to follow the pattern of MW2. This is in line with the fact that MW2 corresponds to a high proportion of trips on Fridays and during the weekdays with low demand.

Among transportation variables, the positive and significant coefficient of the presence of at least one metro station (compared to the reference category which is no metro station) in MW5 indicates that CTs with at least one metro station are more likely to belong to MW5. This means that in CTs with at least one metro station, there is a high taxi demand at the end of week during summer. Finally, the negative significant coefficient of average of daily bus services in MW2 means that CTs with less bus services are more likely to belong to MW2. In addition, due to the positive significant coefficients of average of daily bus services in MW5 and MW6, it can be concluded that CTs with higher daily bus services are more likely to belong to MW5 and MW6 comparing to MW1.

By looking at the accuracy of two models applied in this chapter, it is clear that neither of these models are able to predict the belonging of CTs to different multi-week clusters, and they can only be used as exploratory models to identify which factors can explain the multi-week clusters. Furthermore, it is difficult to interpret the results of the models. Key determinants with the two models were identified including average of daily bus services, eating places and number of cars per population, but there are some discrepancies and further work needs to be done.

Table 6.3 Results of Multinomial logit model

Independent variables	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Socio-demographic					
Median income	-0.602 (0.003**)	-0.516 (0.018*)	-0.391 (0.127)	-0.044 (0.840)	-0.203 (0.464)
Percentage of elderly people	-0.053 (0.048*)	-0.005 (0.845)	-0.066 (0.037*)	0.041 (0.072.)	0.077 (0.003**)
Percentage of recent immigrants	-0.046 (0.217)	-0.110 (0.004**)	-0.049 (0.203)	-0.025 (0.528)	0.049 (0.276)
Number of cars per population	11.306 (0.000***)	0.730 (0.715)	-1.200 (0.609)	-1.554 (0.483)	3.336 (0.184)
Land use					
Density of restaurants	-0.013 (0.249)	0.005 (0.463)	0.026 (0.000***)	0.024 (0.000***)	0.030 (0.000***)
Presence of art, entertainment, and recreational center	-1.116 (0.004**)	-0.388 (0.230)	0.542 (0.103)	0.003 (0.993)	0.180 (0.647)
Transportation					
Presence of metro station	0.887 (0.262)	-1.674 (0.139)	0.826 (0.158)	1.302 (0.017*)	0.991 (0.134)
Average of daily bus services	-0.079 (0.000***)	-0.022 (0.067.)	0.012 (0.322)	0.035 (0.001***)	0.053 (0.000***)

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Cluster 1= Reference category, MacFadden R^2 : 0.21, Log-likelihood: -716.2

CHAPTER 7 CONCLUSION

This chapter presents a summary of this research project, which focused on better understanding taxi demand, characterizing weekly temporal variability of taxi demand, compare taxi usage among census tracts by developing a week typology and identify factors which can explain multi-week patterns. In this chapter, a summary of the methodology applied in this study and the associated results are presented. The contribution of this thesis is then highlighted. The limitations of this research project are also mentioned. Finally, perspectives and recommendations for future research are proposed to improve the current research.

7.1 Summary of the research

A review of the literature first introduced the existing studies in terms of the taxi demand. Thus, the literature review presented different variables which were included in different models for explaining spatial and temporal fluctuations of taxi demand. Then, different methodologies existing in the literature for forecasting taxi demand were presented from both the machine learning algorithms and statistical models.

Different exploratory data analyses were conducted on the taxi data (departing taxi trips from census tracts in the Island of Montreal for the months of April, July and September 2019). The temporal analysis of taxi trips revealed that July and September had similar temporal patterns for weekends with events. However, weekdays with events had the same pattern in April and September. Furthermore, in all three months, midnight trips happened predominantly during weekends with events. According to the spatial analysis of taxi trips, September and July had more similar patterns than April. Furthermore, in all three months, trips had a decentralized pattern during weekends, but in July and September more trips were concentrated downtown.

After the exploratory data analysis, the typical weekly patterns of taxi usage were explored. For this purpose, this research relied on k-means clustering to characterize the weekly temporal variability of taxi demand and developing a week typology. Thus, “CT-week” vectors were created which included daily dispersion indicators (taxi trips per day for all 12 weeks) and an intensity indicator (daily mean of taxi trips for each 12 weeks). These indicators were then normalized for being used in the approach of k-means clustering. Five types of weeks were obtained as a result of applying k-means: end of week trips (W1: high demand), Friday trips (W2: low demand), weekday

trips (W3: low demand), Thursday trips (W4) and weekend trips (W5). The results were then analyzed from both the temporal and spatial perspectives. The results revealed that the 4th type of week (W4: Thursday trips) was the most frequent one.

The temporal analysis of the developed week typology revealed that with respect to the "Thursday trips (W4)" week patterns, one of the highest proportions of CTs in this week type belongs to the week which includes the holiday of good Friday. The proportion of CTs in "end of week trips (W1: high demand)" week pattern increased during summer. It was also found that the proportion of CTs in "weekday trips (W3: low demand)" week pattern decreased remarkably in weeks which included holidays of Easter and Canada day.

The spatial analysis of the developed week typology indicated that the CTs with a high proportion of "end of the week trips (W1: high demand)" week patterns contained or surrounded the airport and industrial zones. The CTs with a high proportion of "Thursday trips (W4)" week patterns were concentrated in central neighborhoods. Interestingly, it was found that CTs with the highest percentage of "weekend trips (W5)" week patterns were concentrated in downtown. In terms of the regularity of taxi usage, by using the indicator of Shannon entropy, it was found that CTs containing and surrounding the airport, downtown, Plateau Mont-Royal, Mount Royal Park, Westmount and Mile End had the regular temporal variation of taxi demand. On the other hand, it was found that CTs with irregular pattern are dispersed across the Island of Montreal

In the next step, a hierarchical agglomerative clustering approach was applied on the second type of vectors to develop multi-week typologies. In this step, "CT" vectors were created based on the ordered sequence of the clusters observed over 12 weeks. In other words, this analysis led to comparing CTs based on their sequence of weekly taxi usage patterns (based on previously developed weekly clusters). Hence, six multi-week clusters were obtained based on the five types of weeks obtained in the first clustering approach. Among these six multi-week clusters, two multi-week clusters had a completely different pattern for spring and summer (MW3 and MW5). The most frequent MW mostly included "Thursday trips" weekly patterns, and more than 95% of weeks in the least frequent MW corresponded to type W1 of the weeks which was "End of week trips (high demand)". Then, a spatial analysis of these multi-week clusters indicated that MW1 (Thursday trips), which was the most frequent multi-week cluster follows a decentralized pattern scattered across the Island. Cluster 2 (MW2) (Friday trips and weekday trips with low demand) is

mostly concentrated to the East and West of Island of Montreal. For Cluster 3 (MW3) (Friday trips for winter and Thursday and weekend trips in Summer), no special pattern has been observed. Cluster 4 (MW4), which mostly included weekend trips, is concentrated at Plateau Mont-Royal and Rosemont. Cluster 5 (MW5) (Thursday and weekend trips during Winter and end of week trips during summer) was found close to metro stations. Finally, cluster 6, which was composed of high demand trips at the end of the weeks, was observed at the CTs containing and surrounding the airport and industrial zones.

The last chapter of this study applied a decision tree analysis and MNL for exploring the factors which can influence belonging to the different multi-week types and to explore if these two models can predict which multi-week clusters the CTs belong to. The decision tree model showed that average of daily bus services, density of eating places, density of drinking places and number of cars per population were the most significant variables in explaining the belonging of each CT to a specific MW cluster. In the MNL, average of daily bus services, density of eating places and number of cars per population were also found to be significant. In addition, socio-economic characteristics (median income, percentage of elderly people and percentage of recent immigrants) also appeared to be significant determinants, as well as the presence of art, entertainment and recreational centers and presence of a metro station. Comparing the significant variables in these two models indicate that average of daily bus services, density of eating places and number of cars per population were common in both models. Chapter 6 also concluded that there were important difficulties in predicting the MW type with the independent variables used in this study.

7.2 Contributions

This research project provides some contributions in terms of reviewing literature and the proposed methodology in the topic of taxi demand modeling. In terms of the literature, this study presents a review of the recent literature regarding the variables which can explain the taxi demand, and methodologies which can be applied to forecast and understand spatial and temporal variations of the taxi demand. For modeling taxi demand, it is necessary to choose variables which can be included in the model to identify if they are significant in the taxi usage or not. Reviewing the literature has revealed the fact that the significancy of variables can vary depending on the methodology which is applied in the study. As mentioned at the beginning of this dissertation, one of the main challenges in forecasting taxi demand is its variation over space and time. In this study,

several machine learning algorithms and statistical modelling techniques which have been used in the existing literature have been reviewed. Including temporal aspect of taxi usage had been found a challenging part in modeling taxi demand, however, some methodologies have proposed approaches to address these issues.

In terms of methodology, clustering methods are rarely applied in taxi demand modeling studies, however, there are many studies at the subject of bike-sharing which have used clustering algorithms. Inspired by the method used by Deschaintres (2018) and Deschaintres et al. (2019), a k-means clustering algorithm was applied to identify different weekly patterns of taxi usage at the CT level, and a hierarchical agglomerative clustering algorithm was then used to compare the week sequences across CTs. These approaches made it possible to assess the existing weekly patterns for each CT, and to compare the temporal variation of taxi trips among CTs. Based on our knowledge of the existing literature, this methodology has been applied on taxi data for the first time.

Finally, a decision tree analysis and multinomial logit model was used to explore which factors can explain the multi-week clusters. In the existing literature of taxi demand modeling, the number of taxi trips were used as the dependent variable to find the significant determinants explaining the taxi demand. In this study, the multi-week clusters which were produced as a result of re-clustering previous clusters were used as dependent variables, which had not been used before in the subject of taxi demand modeling.

7.3 Limitations

This research project has some limitations regarding the data and methodology which are discussed below.

Weather and events are two important variables for modeling taxi demand. In this study, weather was only used in an exploratory data analysis in chapter 4. The reason for not including weather in the subsequent analyses is that for developing the week and multiweek typologies, data were aggregated at the CT level, which only change over space. Due to this reason, it was not possible to include weather as the factors explaining the multi-week clusters. Another issue was that there was not any appropriate dataset for events, so the events could only be used in the temporal exploratory analysis of taxi trips (section 4.2). Furthermore, due to the lack of events data, the

impact of more than one event at the same day and time on taxi demand was ignored. There was also no access to data of Uber trips. Since Ubers provide a similar service to taxis, they could affect the results.

Furthermore, the temporal aspect of taxi trips was considered at the week level for the clustering approaches. This led to ignoring the hourly patterns of taxi trips. In the k-means clustering approach, the indicators were based on days of the weeks and weeks of the year, and in the decision tree and MNL model, temporal variables were calculated based on the week temporal unit. Thus, GTFS data which vary both spatially and temporally was altered to the average of bus services per week, per weekdays, per weekend and aggregated by CT. This leads to ignoring the hourly variation of bus services.

Another limitation of the data was that there was no access to the exact origins and destinations of trips. This information could result in more precise modeling of taxi demand, both in terms of demand and explanatory variables

The final models which were used in chapter 6, were unable to predict the results meaning that the models were exploratory rather than conclusive models.

7.4 Perspectives

It is possible to improve the current research project by considering following recommendations.

All the analyses and approaches of this study were applied on taxi trips which departed from CTs. By applying the same methodology on the taxi trips which arrived in the CTs, a comparison between the results could lead to a better understanding of the temporal patterns of taxi trips.

Another future work can be developing an hourly temporal typology, and then use it to develop a multi-hour typology. Since hour plays an important role in taxi trips, including hours in patterns is liable to increase the precision of the results.

It is also important to experiment other models which can include time-varying variables in addition to spatial variables only.

BIBLIOGRAPHY

Assemblée Nationale du Québec. (2019). *Loi concernant le transport rémunéré de personnes par automobile*. Retrieved from http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=5&file=2019C18F.PDF

Bao, J., Xu, C., Liu, P., & Wang, W. (2017). Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. *Networks and Spatial Economics, 17*(4), 1231-1253.

Bham, G. H., Javvadi, B. S., & Manepalli, U. R. (2012). Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban US highways in Arkansas. *Journal of Transportation Engineering, 138*(6), 786-797.

Bharathi. (2021, 24 June). Confusion Matrix for Multi-Class Classification. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*: Routledge.

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied geography, 34*, 548-558.

Chang, H.-w., Tai, Y.-c., & Hsu, J. Y.-j. (2010). Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining, 5*(1), 3-18.

Changnon, S. A. (1996). Effects of summer precipitation on urban transportation. *Climatic Change, 32*(4), 481-494. doi:10.1007/BF00140357

Chen, C., Varley, D., & Chen, J. (2011). What affects transit ridership? A dynamic analysis involving multiple factors, lags and asymmetric behaviour. *Urban Studies, 48*(9), 1893-1908.

Chen, Y., Wang, K., King, M., He, J., Ding, J., Shi, Q., . . . Li, P. (2016). Differences in factors affecting various crash types with high numbers of fatalities and injuries in China. *PLoS one, 11*(7), e0158559. doi:10.1371/journal.pone.0158559

Chiou, Y.-C., Jou, R.-C., & Yang, C.-H. (2015). Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice, 78*, 161-177.

Chow, L.-F., Zhao, F., Liu, X., Li, M.-T., & Ubaka, I. (2006). Transit ridership model based on geographically weighted regression. *Transportation Research Record, 1972*(1), 105-114.

Conway, M. W., Salon, D., & King, D. A. (2018). Trends in taxi use and the advent of ridehailing, 1995–2017: Evidence from the US National Household Travel Survey. *Urban Science, 2*(3), 79.

Cooper, J., Mundy, R., & Nelson, J. (2010). *Taxi! Urban economies and the social and transport impacts of the taxicab*: Ashgate Publishing.

Cravo, V. S., Cohen, J., & Williams, A. (2009). *Impact of weather on transit revenue in New York city.* Paper presented at the Transportation Research Board 88th Annual Meeting.

Davis, L. W. (2008). The effect of driving restrictions on air quality in Mexico City. *Journal of Political Economy, 116*(1), 38-81.

Deschaintres, E. (2018). *Analyse de la variabilité individuelle d'utilisation du transport en commun à l'aide de données de cartes à puce.* (École Polytechnique de Montréal).

Deschaintres, E., Morency, C., & Trépanier, M. (2019). Analyzing transit user behavior with 51 weeks of smart card data. *Transportation Research Record, 2673*(6), 33-45.

DMTI spatial Inc. (2019). DMTI CanMap Content Suite.

Données Québec. Établissements d'hébergement touristique au Québec. Retrieved from https://www.donneesquebec.ca/recherche/fr/dataset/etablissements-d-hebergement-touristique-au-quebec#

Feng, S., Chen, H., Du, C., Li, J., & Jing, N. (2018). *A hierarchical demand prediction method with station clustering for bike sharing system.* Paper presented at the 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC).

Fortin, P., Morency, C., & Trépanier, M. (2016). Innovative GTFS data application for transit network analysis using a graph-oriented method. *Journal of Public Transportation, 19*(4), 2. doi:10.5038/2375-0901.19.4.2

Galarnyk, M. (2019, 31 July). Understanding Decision Trees for Classification (Python). Retrieved from https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952

Guo, Z., Wilson, N. H., & Rahbee, A. (2007). Impact of weather on transit ridership in Chicago, Illinois. *Transportation Research Record, 2034*(1), 3-10. doi:10.3141/2034-01

Haider, M. (2015). To Uber or Not to Uber: That is the Question. Retrieved from https://stream1.newswire.ca/media/2015/09/29/20150929_C6395_PDF_EN_508957.pdf

Harding, S., Kandlikar, M., & Gulati, S. (2016). Taxi apps, regulation, and the market for taxi journeys. *Transportation Research Part A: Policy and Practice, 88*, 15-25.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics, 15*(3), 651-674.

Hua, M., Chen, X., Zheng, S., Cheng, L., & Chen, J. (2020). Estimating the parking demand of free-floating bike sharing: A journey-data-based study of Nanjing, China. *Journal of Cleaner Production, 244*, 118764.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Unsupervised Learning. In *An introduction to statistical learning: with applications in R* (Vol. 103). New York, NY: Springer.

Kalkstein, A. J., Kuby, M., Gerrity, D., & Clancy, J. J. (2009). An analysis of air mass effects on rail ridership in three US cities. *Journal of transport geography, 17*(3), 198-207. doi:10.1016/j.jtrangeo.2008.07.003

Kamga, C., Yazici, M. A., & Singhal, A. (2013). *Hailing in the rain: Temporal and weather-related variations in taxi ridership and taxi demand-supply equilibrium*. Paper presented at the Transportation Research Board 92nd Annual Meeting.

Kamga, C., Yazici, M. A., & Singhal, A. (2015). Analysis of taxi demand and supply in New York City: implications of recent taxi regulations. *Transportation Planning and Technology, 38*(6), 601-625. doi:10.1080/03081060.2015.1048944

Lacombe, A. (2016). *Méthodologie d'analyse et de suivi d'un système de transport par taxi.* (École Polytechnique de Montréal).

Laviolette, J. (2017). *Planification stratégique d'un système de transport par taxi.* (École Polytechnique de Montréal).

Liu, Q., Ding, C., & Chen, P. (2020). A panel analysis of the effect of the urban environment on the spatiotemporal pattern of taxi demand. *Travel Behaviour and Society, 18*, 29-36. doi:10.1016/j.tbs.2019.09.003

Liu, X., Sun, L., Sun, Q., & Gao, G. (2020). Spatial variation of taxi demand using GPS trajectories and POI data. *Journal of Advanced Transportation, 2020*. doi:10.1155/2020/7621576

McHugh, B. (2013). Pioneering open data standards: The GTFS Story. *Beyond transparency: open data and the future of civic innovation*, 125-135.

Mohajon, J. (2020, 29 May). Confusion Matrix for Your Multi-Class Machine Learning Model. Retrieved from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems, 14*(3), 1393-1402.

Morency, C., Trepanier, M., Frappier, A., & Bourdeau, J.-S. (2017). *Longitudinal analysis of bikesharing usage in Montreal, Canada.* Paper presented at the Transportation Research Board 96th Annual Meeting.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification, 31*(3), 274-295.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society, 18*(6), 275-285.

Narkhede, S. (2018, 9 May). Understanding Confusion Matrix. Retrieved from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

O'Sullivan, D. (2003). Geographically weighted regression: the analysis of spatially varying relationships. *Geographical analysis, 35*(3), 272-275.

Organisation for Economic Co-operation and Development. (2007). *ECMT Round Tables (De)Regulation of the Taxi Industry. Paris, France*. European Conference of Ministers of Transport.

Pautler, P. A., & Frankena, M. W. (1984). An Economic Analysis of Taxicab Regulation. *Bureau of Economics Staff Report*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.6001&rep=rep1&type=pdf

Qian, X., & Ukkusuri, S. V. (2015a). *Exploring spatial variation of urban taxi ridership using geographically weighted regression.* Paper presented at the Transportation Research Board 94th Annual Meeting.

Qian, X., & Ukkusuri, S. V. (2015b). Spatial variation of the urban taxi ridership using GPS data. *Applied geography, 59*, 31-42. doi:10.1016/j.apgeog.2015.02.011

Rayle, L., Shaheen, S., Chan, N., Dai, D., & Cervero, R. (2014). App-based, on-demand ride services: Comparing taxi and ridesourcing trips and user characteristics in san francisco university of california transportation center (uctc). *University of California: Berkeley, CA, USA*.

Rodrigues, P., Martins, A., Kalakou, S., & Moura, F. (2020). Spatiotemporal variation of taxi demand. *Transportation Research Procedia, 47*, 664-671.

Saponaro, M. A. (2013). *Economic regulation in the taxicab industry: a case study of Iowa City, Iowa.* (The University of Iowa).

Schaller, B. (2005). A regression model of the number of taxicabs in US cities. *Journal of Public Transportation, 8*(5), 4.

Schaller, B. (2007). Entry controls in taxi regulation: Implications of US and Canadian experience for taxi regulation and deregulation. *Transport policy, 14*(6), 490-506.

Schlosser, L., Hothorn, T., & Zeileis, A. (2019). The power of unbiased recursive partitioning: a unifying view of CTree, MOB, and GUIDE. *arXiv preprint arXiv:1906.10179*.

Shao, D., Wu, W., Xiang, S., & Lu, Y. (2015). *Estimating taxi demand-supply level using taxi trajectory data stream.* Paper presented at the 2015 IEEE International Conference on Data Mining Workshop (ICDMW).

Statistics Canada. (2016). 2016 Census of Population. Using CHASS. Retrieved from https://datacentre.chass.utoronto.ca/census/ct.html

Statistics Canada. (2018). Census tract (CT). Retrieved from https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/ct-sr/ct-sr-eng.htm

STM GTFS. 2019. Retrieved from https://transitfeeds.com/p/societe-de-transport-de-montreal/39?p=3

Stover, V. W., & McCormack, E. D. (2012). The impact of weather on bus ridership in Pierce County, Washington. *Journal of Public Transportation, 15*(1), 95-110. doi:10.5038/2375-0901.15.1.6

Tarpin-Pitre, L., & Morency, C. (2020). Typology of bikeshare users combining bikeshare and transit. *Transportation Research Record, 2674*(10), 475-483.

Taxi Fare Finder. (2012, November 1). The History of the Taxi Industry. Retrieved from https://www.taxifarefinder.com/newsroom/2012/11/01/the-history-of-the-taxi-industry/

VE, S., & Cho, Y. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing, 53*(sup1), 166-183. doi:10.1080/22797254.2020.1725789

Viallard, A., Trépanier, M., & Morency, C. (2019). Assessing the evolution of transit user behavior from smart card data. *Transportation Research Record, 2673*(4), 184-194.

Vogel, M., Hamon, R., Lozenguez, G., Merchez, L., Abry, P., Barnier, J., . . . Robardet, C. (2014). From bicycle sharing system movements to users: a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset. *Journal of transport geography, 41*, 280-291.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236-244. doi:10.1080/01621459.1963.10500845

Xu, H., Duan, F., & Pu, P. (2019). Dynamic bicycle scheduling problem based on short-term demand prediction. *Applied Intelligence, 49*(5), 1968-1981.

Yang, C., & Gonzales, E. J. (2014). Modeling taxi trip demand by time of day in New York City. *Transportation Research Record, 2429*(1), 110-120.

Yang, Z., Franz, M. L., Zhu, S., Mahmoudi, J., Nasri, A., & Zhang, L. (2018). Analysis of Washington, DC taxi demand using GPS and land-use data. *Journal of transport geography, 66*, 35-44.

Zhang, X., Huang, B., & Zhu, S. (2019). Spatiotemporal influence of urban environment on taxi ridership using geographically and temporally weighted regression. *ISPRS international journal of geo-information, 8*(1), 23. doi:10.3390/ijgi8010023

Zhao, K., Khryashchev, D., Freire, J., Silva, C., & Vo, H. (2016). *Predicting taxi demand at high spatial resolution: Approaching the limit of predictability.* Paper presented at the 2016 ieee international conference on Big data (big data).

APPENDIX A HOLIDAYS OF QUEBEC IN APRIL, JULY AND SEPTEMBER 2019

Date	Day	Week	Holiday
19 April	Friday	3	Good Friday
22 April	Monday	4	Easter Monday
1 July	Monday	5	Canada day
2 September	Monday	9	Labour day

APPENDIX B DESCRIPTIVE STATISTICS OF THE INDEPENDENT VARIABLES BASED ON THE MULTI-WEEK TYPOLOGY

Descriptive statistics of the independent variables based on the multi-week typology (observations with 0 and very few population are excluded)

	Cluster	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Median income *10000	**C1**	1.67	2.55	2.90	3.18	3.54	7.86
	C2	1.73	2.77	3.22	3.20	3.71	6.07
	C3	1.30	2.57	2.85	3.022	3.40	7.36
	C4	1.99	2.62	2.86	2.92	3.23	4.61
	C5	1.46	2.48	2.98	3.07	3.37	5.88
	C6	1.21	2.41	2.82	3.01	3.47	5.71
Percentage of recent immigrants	**C1**	1.12	4.32	6.67	7.83	10.89	21.67
	C2	0.00	2.40	5.09	5.84	8.00	30.45
	C3	0.97	3.74	6.11	6.47	8.50	25.89
	C4	1.52	4.96	6.90	7.79	10.09	21.03
	C5	1.26	4.47	6.69	7.68	9.97	26.91
	C6	1.13	5.10	7.42	9.07	11.48	28.70
Percentage of elderly people	**C1**	5.10	11.45	15.85	16.10	19.15	40.60
	C2	4.10	12.05	15.95	16.27	19.80	41.90
	C3	3.40	10.60	15.60	15.34	19.10	40.50
	C4	4.40	9.25	11.30	12.26	14.40	32.00
	C5	1.70	10.97	14.80	16.99	21.55	46.80
	C6	6.30	14.15	16.80	19.20	22.27	60.00

	Cluster	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Number of cars per population	**C1**	0.11	0.30	0.38	0.38	0.45	0.66
	C2	0.12	0.37	0.49	0.47	0.58	0.69
	C3	0.09	0.30	0.37	0.38	0.44	0.60
	C4	0.14	0.24	0.29	0.30	0.35	0.48
	C5	0.13	0.28	0.34	0.34	0.41	0.56
	C6	0.12	0.24	0.37	0.37	0.46	0.66
Density of hotel rooms	**C1**	0.00	0.00	0.00	3.92	0.00	208.96
	C2	0.00	0.00	0.00	8.34	0.00	536.26
	C3	0.00	0.00	0.00	22.56	3.22	771.94
	C4	0.00	0.00	6.80	24.03	26.46	251.82
	C5	0.00	0.00	0.39	47.34	10.74	1068.62
	C6	0.00	0.00	0.00	72.84	19.00	1270.11
Density of health care workers	**C1**	0.00	152.70	331.30	877.40	631.40	20479.00
	C2	0.00	60.09	135.13	373.51	517.68	3287.23
	C3	0.00	145.00	386.30	896.70	696.40	14995.80
	C4	93.08	228.05	407.88	1463.53	851.11	24895.41
	C5	28.29	239.65	561.96	1973.13	1556.33	21833.94
	C6	35.65	196.00	434.07	2160.73	22.35.89	22948.72

	Cluster	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Density of business workers	**C1**	50.70	361.60	838.50	4409.30	1612.70	360813.60
	C2	12.34	172.19	424.43	1784.04	1606.70	39980.25
	C3	67.42	407.26	1017.62	4769.58	2043.12	249627.71
	C4	171.30	767.50	1432.10	2902.30	4011.90	18168.20
	C5	141.40	681.80	1471.80	9461.10	4065.80	208803.00
	C6	119.40	625.8	1383.00	11228.70	5086.60	176568.10
Density of drinking places	**C1**	0.00	0.00	0.00	1.66	1.72	43.23
	C2	0.00	0.00	0.00	0.74	0.00	19.33
	C3	0.00	0.00	0.00	2.34	3.62	21.06
	C4	0.00	0.00	4.23	9.55	12.36	90.94
	C5	0.000	0.000	1.51	5.26	5.67	54.24
	C6	0.00	0.00	1.18	7.67	4.97	83.28
Density of eating places	**C1**	0.00	4.43	12.16	17.80	23.03	120.93
	C2	0.00	0.61	3.31	9.24	9.39	105.88
	C3	0.00	5.15	10.42	20.80	25.10	151.09
	C4	2.19	20.36	45.42	71.15	99.27	426.06
	C5	0.00	12.97	22.11	47.99	44.79	266.27
	C6	0.00	9.35	23.23	90.15	59.37	695.43

	Cluster	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Density of bus stops	**C1**	4.57	17.47	24.49	27.44	32.58	77.55
	C2	0.13	10.40	18.95	21.21	27.52	80.82
	C3	5.21	17.83	25.37	28.12	35.12	73.67
	C4	12.24	24.64	35.83	36.13	42.05	116.48
	C5	9.53	23.69	33.12	36.78	42.36	112.17
	C6	8.22	23.92	28.53	34.66	41.31	83.26
Average of weekday bus services	**C1**	4.57	25.99	34.45	35.55	43.80	74.51
	C2	2.88	17.87	25.06	25.59	30.56	60.95
	C3	5.52	20.99	29.38	31.13	38.51	72.73
	C4	9.60	21.50	32.57	34.04	40.68	92.00
	C5	8.88	29.39	41.19	44.23	53.73	194.48
	C6	9.49	41.63	49.46	58.93	69.03	144.29
Average of weekend bus services	**C1**	3.21	13.66	17.32	18.25	22.62	42.81
	C2	0.00	7.73	11.90	12.28	15.86	27.79
	C3	4.52	10.83	13.95	15.75	19.35	35.54
	C4	5.09	11.71	16.82	17.77	21.96	48.03
	C5	4.59	12.45	20.79	22.28	26.88	81.56
	C6	6.76	22.17	26.51	30.24	35.90	87.98

			(0)	**(1)**		**(0)**	**(1)**
	C1	Art, entertainment, and recreational centers (dummy)	86	42	Metro station (dummy)	122	6
	C2		89	17		103	3
	C3		68	21		88	1
	C4		39	36		64	11
	C5		46	26		55	17
	C6		28	23		39	12