POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Intercity travel in Québec: corridor analysis and demand modelling

HAMED ALI ZADEH

Département des génies civil, géologique et des mines

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

Génie Civil

Août 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé:

Intercity travel in Québec: corridor analysis and demand modelling

présenté **Hamed ALI ZADEH**

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

Catherine MORENCY, membre et directrice de recherche

Martin TRÉPANIER, membre et codirecteur de recherche

Yan CIMON, membre

DEDICATION

To my beloved wife for all her support during my study

To my supervisors Catherine Morency and Martin Trépanier

To all my family and friends

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis supervisors, Professor Catherine Morency and Professor Martin Trépanier for their invaluable support, accessibility, patience, and encouragement through pursuing my master's at Polytechnique. However, these two years were accompanied by a COVID-19 pandemic period your endless effort to keep me on track and being accessible and hardworking have been an incredible source of inspiration and an exceptional example of dedication. I would always be thankful for all your guidance and support.

I am also very grateful for the funding sources for my research project. I was funded by the ministère des Transports du Québec (MTQ) and the Mobility Chair of Polytechnique Montreal, headed by Professor Catherine Morency.

I would also want to thank all my colleagues for being supportive and being accessible throughout my master's, specifically I would like to say thanks to Elodie Deschaintres for coordinating the biweekly online seminars to have some social interaction with other students during the pandemic.

Furthermore, I would like to thank my friend for helping me to get adapted to a new country, thanks for making the difficult moments as tolerable as possible for me, specifically I would like to express my gratitude to my friend Masoud who I always can count on him, I am so thankful to have you.

This journey would not have been possible without the support of my parents who always were supportive of me. Mom and dad, your endless support in every possible way. I have always been indebted to you for all the selfless love, care, pain and sacrifice you did to shape my personality and my life, and I will always be! I might understand how hard it is to let your child lives kilometres far away from you.

Also, I would like to say thank you to all my brothers and sisters for all their support, my heaviest regret for these years is being too far from you and seeing you through the phone. Thank you for always being there beside the last-born of family.

Last but not least, words cannot express my gratitude and heartfelt to my wife, my dearest Aaliye! I believe that this journey would not be possible without you. Your support, kindness, and patience inspired me to finish my thesis. I deeply owe you for passing your wishes away for these years to

support both of us in any way you could. You were always there beside me, during happy and hard moments, and motivated me with your delightful smile and your genuine belief in me.

RÉSUMÉ

Les voyages de longue distance (LD) ou interurbains attirent moins l'attention des chercheurs que les déplacements quotidiens habituels. Ce type de voyage s’applique aux niveaux nationaux, provincial et interrégional. Parallèlement, ces types de déplacements constituent une composante importante de l'étude des transports puisque leur contribution est très élevée en termes de kilométrage et d’émission de gaz à effets de serre, même si leur fréquence est inférieure à celle des déplacements urbains quotidiens.

Actuellement, il n'y a pas de définition spécifique pour les voyages de LD : chaque étude et enquête, en fonction de son objectif, en propose une nouvelle. Par exemple, dans l'Enquête sur les voyages pour les résidents du Canada (EVRC), le voyage de LD est défini comme un voyage quotidien non récurrent, avec une nuit passée à destination. Pour différencier les déplacements urbains et les déplacements de LD, deux critères sont généralement utilisés : la distance parcourue depuis le point d'origine et la durée du séjour (Gerike, 2018). Les enquêtes sur les déplacements de LD sont développées pour deux raisons principales : les études sur le tourisme et les études sur la planification des transports. La définition du voyage de LD varie donc selon ces deux usages. Dans le cas des études sur le tourisme, le critère de définition d'un voyage de LD est généralement la durée du séjour, tandis que la distance parcourue est principalement utilisée pour les études de planification des transports.

L'objectif principal de ce mémoire est de développer une méthode d'analyse de la demande et de l'offre de déplacements de LD au Québec, pouvant être utilisée par les planificateurs et les décideurs en transport afin d’améliorer les services de transport en commun de LD et de rendre ces déplacements plus durables. En particulier, ce projet de recherche aborde trois sujets :

1. Identification préliminaire des principaux corridors interrégionaux du Québec et utilisation des caractéristiques socio-économiques et sociodémographiques des villes pour hiérarchiser ces corridors.

2. Utilisation d’une méthode d'apprentissage automatique pour développer un modèle de génération de voyages à l'aide des données de l'Enquête sur les voyages des résidents du Canada (EVRC).

3. Analyse de la compétitivité modale des principaux corridors interrégionaux du Québec.

En premier lieu, une méthodologie en 6 étapes est proposée pour l'identification des principaux corridors interrégionaux entre 28 villes du Québec. Ces corridors comprennent 50 paires origine-destination (OD) reliant ces villes. En outre, une méthode de classement est utilisée pour classer ces paires OD en fonction des caractéristiques socio-économiques et socio-démographiques des villes d'origine et de destination. Cette méthode de classement mesure l'importance de chaque paire OD en termes de demande potentielle. Les données du recensement de 2016 ont été utilisées pour évaluer et classer les corridors. Il a été constaté que, dans les deux méthodes de classement, les villes d'origine et de destination sont très importantes, l'itinéraire le plus important ayant une ou deux extrémités dans les villes centrales. Cependant, lorsque l'évaluation est basée sur les municipalités le long du corridor, toutes les paires importantes traversent des zones à haute densité.

En second lieu, ce mémoire présente un modèle de génération de déplacements et une méthode de comparaison de modèles pour la génération des déplacements de LD effectués par les résidents du Canada. Le terme de voyage « longue distance » basé sur l'Enquête sur les voyages des résidents du Canada (EVRC) comprend les voyages non fréquents d'une nuit ou d'une journée. Cette étude a comparé quatre méthodes d'apprentissage automatique pour le modèle de génération de déplacements. Étant donné que le voyage de LD est plutôt rare, l'ensemble de données de l'EVRC correspond à un ensemble de données débalancé. Ainsi, trois techniques différentes pour préparer des données dans le cadre de la modélisation d'événements rares (sur et sous-échantillonnage ainsi que suréchantillonnage synthétique) sont utilisées pour pallier ce problème. Les données de l'EVRC de 2012 à 2017 ont été utilisées pour l'estimation du modèle. Parmi les modèles testés (forêt aléatoire, CART, CTree et logit), il a été constaté que la forêt aléatoire a les meilleures performances en matière de prévision et que les modèles d'arbre de décision obtiennent la meilleure précision globale. De plus, le niveau de revenu et le niveau d'éducation jouent un rôle majeur dans l'occurrence d'un déplacement interurbain. Le mémoire souligne l'importance d'améliorer les méthodes d'enquête sur les déplacements interurbains et de développer d'autres méthodes de collecte de données.

Enfin, une analyse de la compétitivité des modes est proposée pour évaluer dans quelle mesure les modes de transport en commun (TC) sont compétitifs avec le véhicule privé pour les corridors importants du Québec. La variation des temps d'accès à l'origine et à la destination pour les déplacements interurbains en autobus et en train est un élément très important. Le temps et la distance de déplacement sont les variables utilisées dans cette étude pour évaluer la compétitivité. Chaque déplacement interurbain en TC comprend trois parties : l'accès à la gare (ou la station), le tronçon interurbain (avec ou sans correspondance) et l'accès à la destination. Un déplacement de LD est défini comme un déplacement du centroïde d'un secteur de recensement (SR) d'origine au centroïde du SR dans la ville de destination, pour les villes incluses dans une agglomération de recensement (AR) ou une région métropolitaine de recensement (RMR) du Québec. Les données sur le temps de trajet pour l'analyse ont été extraites de trois sources différentes mentionnées dans la section méthodologie. Le principal résultat de l'analyse montre que le temps de déplacement en transport en commun (TC) (bus ou train) est supérieur à celui du véhicule privé, alors que le ratio de ces temps de déplacement est plus élevé dans les trajets plus courts pour les deux modes TC. Cependant, pour les distances plus longues, ce ratio va diminuer pour le mode bus, et rester élevé pour le train. Un autre élément important dans les déplacements de LD est le temps d'accès à l’origine et à la destination qui, pour les trajets de moins de 100 km, peut être égal au temps de trajet.

Ce projet de recherche présente certaines limites dans chaque section de l'étude. Dans l'analyse des corridors, l'accès aux données de comptage de véhicules pourrait aider à mieux comprendre la demande potentielle dans les corridors; dans l'estimation du modèle de génération de trajets, le fait d'avoir un ensemble de données avec un accent principal sur l'étude du tourisme et étant un ensemble de données déséquilibré était la partie la plus difficile; pour l'étude sur la compétitivité des modes, l'utilisation de différentes plateformes pour l'estimation du temps de trajet peut biaiser les résultats.

Ce projet de recherche a mis en évidence l'importance des questions liées aux données concernant les études de déplacements de LD. La conception d'une nouvelle enquête ciblant ces déplacements peut aider à l'amélioration des études et de la planification des déplacements de LD. En outre, dans la partie compétitivité des modes, l'ajout d'autres modes de transport tels que le covoiturage et les avions peut être utile. Sur la partie analyse du corridor, avoir la même analyse sur le réseau routier et TC actuel avec l'ajout de données de comptage peut améliorer les résultats et l'interprétation de la situation actuelle de la demande potentielle.

ABSTRACT

Long-distance (LD) or intercity travel is getting less attention from researchers than usual daily trips. There is no specific definition for this kind of trip at the national, provincial, and inter-regional levels. At the same time, these kinds of trips are an important component of transportation study since their contribution is very high in terms of mileage, even if their frequency is lower than the one of daily urban trips.

Currently, each study and survey, depending on its aim, proposes a new definition. For example, in the Travel Survey for Residents in Canada (TSRC), a LD trip is defined as an overnight or daily non-recurrent trip. To differentiate between urban commuting and LD travel, two criteria are generally used, distance travelled from the point of origin and length of stay or both (Gerike, 2018). LD travel surveys are developed for two main reasons: tourism studies and transportation planning studies. The definition of LD travel varies according to these two reasons. In the case of tourism studies, the criterion for the definition of a LD trip is usually the length of stay while the distance travel is mostly used for transportation planning studies.

The main purpose of this study is to develop a method to analyze the LD travel demand and supply in Québec, which can be used by transportation planners and decision-makers to improve the share of public transportation and make LD travel more sustainable. Particularly, this research project investigates a more in-depth study on three topics:

1. Preliminary development of the main interregional corridors of Québec and using socio-economic and socio-demographic characteristics of the cities for ranking these corridors.
2. Using machine learning method to develop a trip generation model using Travel Survey for Residents in Canada (TSRC) survey data.
3. Mode competitiveness analysis of main interregional corridor in Québec.

A methodology consisting of 6 steps is used for the development of the main interregional corridors through 28 cities in Québec. These corridors include 50 OD pairs connecting these cities. Also, a ranking method is used to rank the OD pairs based on socio-economic and socio-demographic characteristics of the origin and destination cities and the cities among each OD line. This ranking method can represent how each OD pair is important in terms of potential demand. Data from the

2016 census was used for the evaluation and ranking of the corridors. It was found that, in both cases, the city of origin and destination is highly important, the most important route has one or two ends in the core cities. However, all the CSDs along each OD pair become more important when the ranking is based on the characteristics of CSDs along the corridor.

In addition, this research developed a trip generation model and presented a model comparison method for (LD) trip generation model for LD trips performed by residents of Canada. The terms of "long-distance" trip based on the Travel Survey for Residents in Canada (TSRC) survey is considered non-frequent overnight and day trips. This study compared several machine learning methods for the trip generation model. Since LD trip is relatively rare, the data set of TSRC is considered as imbalanced data. Three different techniques for data preparation as part of rare event modelling (over, under, and synthetically oversampling) are employed to handle the issue of imbalanced data. TSRC data from 2012 to 2017 was used for model estimation.

Among the random forest, CART, CTree, and logit models, it was found that the random forest has the best performance in prediction, and decision tree models have the best overall accuracy. Also, income level and educational level play an essential role in the occurrence of an intercity trip. The paper highlights the importance of improving intercity travel survey methods and proposes other data collection methods.

A mode competitiveness analysis is used to assess how competitive public transportation (PT) modes are with the private vehicle for important corridors of Québec. The variation of access and egress travel time for travel by intercity bus and train are very important elements. Travel time and distance are the variables used in this study to assess competitiveness. Each intercity public transport trip includes three parts, access to the station, intercity leg (with or without a transfer) and access to the destination.

LD travel is defined as a trip from the centroid of an origin census tract (CT) to the centroid of destination CT of cities consisting of Census agglomeration (CA) and census metropolitan agglomeration (CMA) in Québec. The travel time data for analysis was extracted from three different sources mentioned in the methodology section.

The key finding from the analysis shows that the public transit (PT) mode travel time (bus or train) is higher than the private vehicle, while this increase in travel time is more significant in shorter trips for both PT modes. However, for longer distances, this ratio will decrease for bus mode, and it remains high for the train. Another important component in LD trips is the access and egress time which for trips of less than 100 km, can be equal to travel time.

This research project has some limitations in each section of the study, in corridor analysis, having access to the traffic count data could help to have a better understanding of potential demand for corridors, in the estimation of trip generation model, the having a data set with main focus on tourism study and being an imbalanced data set was the most challenging part. As for the mode competitiveness study, using different platforms for estimation of travel time can bias the results.

This research project highlighted the importance of data-related issues regarding the LD travel studies, designing a new survey based on transportation planning studies can help the improvement of LD studies. Also, in the mode competitiveness part adding other modes of transportation such as carpooling, and airplanes can be helpful. On the corridor analysis part, having the same analysis on the current highway and road network with adding traffic count data can improve the results and interpretation of the current situation of potential demand.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CA	Census Agglomeration
CART	Classification and Regression Trees
CD	Census Division
CMA	Census Metropolitan Area
CSD	Census Subdivision
CT	Census Tract
CTree	Conditional Inference Trees
DT	Decision Tree
FN	False-negative
FP	False-positive
GLM	Generalized Linear Model
LD	Long-Distance
NTS	National Travel Survey
OD	Origin-Destination
OOB	Out-of-Bag
OSRM	Open Street Routing Machine
PT	Public Transportation
RF	Random Forest
TN	True negative
TP	True positive
TSRC	Travel Survey for Residents in Canada

CHAPTER 1 INTRODUCTION

Long-distance (LD) or intercity trips are getting more attention in recent years since their contribution is very high in milage even if their frequency is lower than the one of daily urban trips. A study in Great Britain shows that less than 2% of travel performed by British residents are considered long-distance trips within a country; however, this 2% is responsible for about 30% of distance travelled by British residents (Dargay, 2012). Canada covers a broad territory which can probably lead to having a high proportion of kilometres travelled by residents related to long-distance trips. Also, with a growing population, wealth and education, the travel demand for LD trips will probably rise as well. With a potential increase in LD trips, it is crucial to improve the knowledge of LD trips and travel demand modelling to correctly assess their environmental impacts and identify opportunities for service improvement. Researchers have paid much less attention to LD trips during the past decades than to within metropolitan areas trips since the latter are much more numerous. Additionally, data collection methods and surveys have continued to improve during recent years, namely for regional trips, while LD still suffers from a lack of datasets to support relevant research and models. It is more typical for people to make daily urban trips than to do a LD trip, making it harder to capture over a short period of observation. As is the case in other countries, survey methods for LD trips are less developed (and lack interest from decision-makers) in Canada than regional surveys covering trips during a typical weekday.

Québec has developed an origin-destination (OD) survey in several populated metropolitan areas of the province. This survey collects OD data every five to ten years (depending on the region), and these surveys provide important information for researchers and decision-makers to better understand the mobility situation within metropolitan areas. However, this type of survey provide very scares information on intercity trips (only those conducted on the day of the week surveyed and with one trip-end within the survey area), which probably cover a high proportion of milage travelled within the province. The lack of data on LD trips is leading to less intention by researchers to study LD trips. In Canada, some periodic surveys have been conducted at the federal and provincial levels, but most of the time, they only cover a particular period or a particular mode of transportation (Guillemette Y. , 2015).

The Travel Survey for Residents in Canada (TSRC) is one of the surveys that is collecting data on LD trips among Canadian residents. This survey considers LD trips within and outside of the border of Canada. According to this survey, a “long-distance” trip is considered as a non-frequent overnight and day trip. The TSRC is collected with the international guidelines recommended by the World Tourism Organization (WTO) and mainly to conduct research in tourism study. It is not specifically designed to model LD trips and would need improvement for more in-depth analysis but still provides some relevant insights into the intensity of LD travel.

Currently, there is no consensual definition of LD travel. The definition of a LD trip varies depending on the geographical location or the focus of the study. To differentiate between urban commuting and LD travel, two criteria are generally used: distance travelled from the point of origin and length of stay at the destination (Gerike, 2018). The first criterion is mainly used for the study with a focus on transportation planning, where the travel demand modelling is important, and the second criterion focuses on tourism study where marketing becomes important. These two main criteria do not have any exact definition to differentiate LD travel from the urban daily commute and the number of nights for stay or distance travelled from the point of origin varies from one study to another.

This study covers intercity or long-distance (LD) trips and goes into more detail about the province of Quebec. In this study, the intercity trip considers all non-frequent day trips and overnight trips as reported in the Travel Survey for Residents in Canada (TSRC).

1.1 Objective

The goal of this project is to develop methods and tools that will be used to analyze the travel demand in Québec. Also, it develops a procedure to have a better understanding of the current situation of travel with private vehicles and public transportation modes. The main sub-objectives of this research are as follows:

- Propose a typology of corridors for intercity trips in Québec.
- Develop a method to estimate travel demand in the main intercity Québec corridors.

Investigate travel time for different modes of transport for LD travel to have a better understanding of how public transportation

1.2 Research problem

To our knowledge, there are only a few studies regarding LD travel in Canada. This lack of study in this field can be related to the lack of data for this kind of travel. There is no survey for LD travel with aim of transportation study in Canada. The only available systematic data is the TSRC (which has been replaced by National Travel Survey (NTS)) survey which is conducted with a focus on tourism study. This issue made us investigate more deeply into this type of trip.

The main questions that arise in LD trips that are discussed in this study are:

- What is the current situation of intercity travel patterns in the main corridors of Québec?
  - Which cities should be included in the corridor development studies?
  - How do identify intercity travel corridors in Québec?
  - Which corridor has the highest potential for travel demand?
  - What criteria should be used to find the importance of a defined corridor?
- How to model trip generation in LD tip with TSRC data?
  - Determine the available dataset for the estimation of a LD trip generation model.
  - How to deal with the LD trip generation model; should it be modelled like urban daily travel?
  - What kind of modelling approach should be used for LD travel demand?
  - What is the most appropriate model considering the available dataset?
- How competitive are the public modes of transport in contrast to private mode?
  - Is it possible to use the current TSRC data set for mode choice modelling?
  - Are the public transportation modes competitive with the private vehicle in the Québec corridors?
  - What is the ratio of travel time for public vs private modes of transportation for LD trips?

These questions are discussed in three main parts of this study which are further explained in the thesis structure section.

- modes are competitive with cars.

1.3 Thesis structure

This research is mainly focused on these sub-objectives. The first section relates to the development of corridors with socio-economic and socio-demographic characteristics of the areas in Québec, the second part focuses on the development of a trip generation model for LD travel in Quebec and the last component proposes an investigation of mode competitiveness analysis in the main interregional corridors of Québec. These three main parts are presented in 5 chapters as follows.

In the first chapter, there is a brief introduction to LD travel and why it is important to do a more in-depth study on LD travel, also it represents what problem made us do this research followed by the main objectives of the research.

State of the art in LD travel is mainly divided into four different sections in chapter two. Firstly, the definition of the LD travel is discussed, the data collection method used by earlier studies is presented in the next section and followed by LD travel demand modelling. And the last section is dedicated to mode competitiveness studies.

The third chapter presents the methodology used in each part of the study. This chapter also includes the explanation of data and the study area in each part. Chapter four discusses the results from the employed data and methodology in chapter three. This chapter discusses thoroughly the findings and the outcomes. Basically, this chapter is divided into four sections, three chapters are about each research problem starting with a descriptive analysis of the TSRC dataset to have a better understanding of the LD travel in Canada.

The conclusion of the research is exposed in the last chapter, which contains a summary of the results in each part of the study, followed by the contribution of the research, the limitations, and the perspectives of the study. The diagram of the research methodology is presented in Figure 1.1.

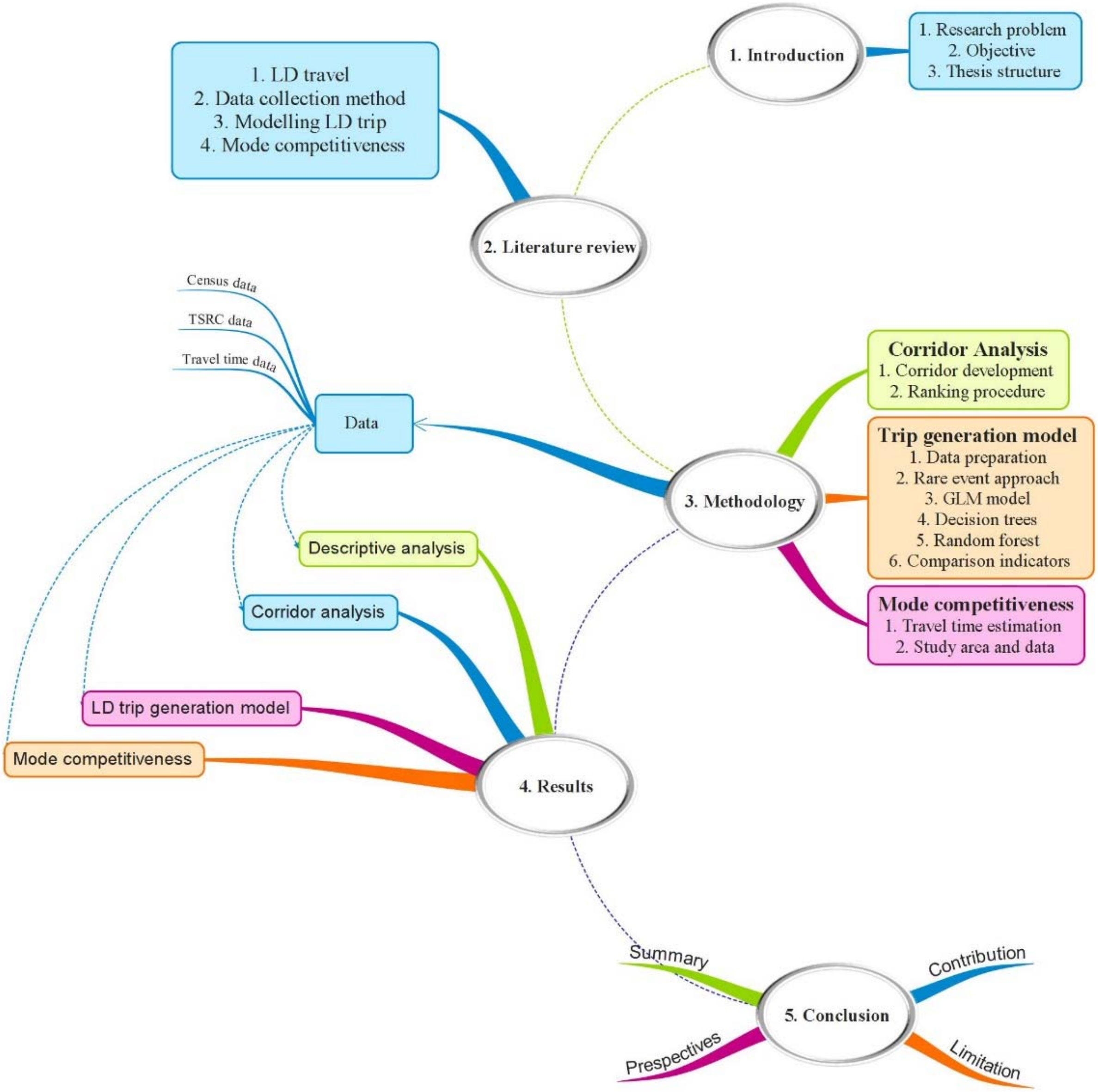


Figure 1.1 Research methodology

CHAPTER 2 LITERATURE REVIEW

2.1 Long-distance travel

Long-distance (LD) or intercity trips have received more attention in recent years since their contribution is very high in milage, even if their frequency is lower than the one of daily urban trips. A study in Great Britain shows that less than 2% of travel performed by British residents is considered long-distance trips; however, this 2% is responsible for about 30% of distance travelled by British residents (Joyce M. Dargay, 2012). Canada covers a broad territory which can probably lead to having a high proportion of kilometers travelled by residents related to long-distance trips. Also, with a growing population, the travel demand for LD trips will probably rise as well. With a potential increase in LD trips, it is crucial to improve knowledge of LD trips and travel demand modelling to correctly assess their environmental impacts and identify opportunities for service improvement. To our knowledge, researchers have paid much less attention to LD trips during the past decades than to within metropolitan areas trips since the latter are much more numerous. Additionally, data collection methods and surveys have continued to improve, namely for regional trips, while LD still suffers from a lack of datasets to support relevant research and modelling.

LD travel surveys are conducted primarily for two reasons, the first being tourism studies and the second being transportation planning studies. To differentiate between urban commuting and LD travel, two criteria are generally used, distance travelled from the point of origin and length of stay or both (Gerike, 2018). Surveys for tourism studies mainly use the length of stay because marketing is an important factor in this type of study. In transportation studies, the focus is on trip generation, mode and route choice, so the most used boundary is the distance travelled. The decision as to which boundary to use is based on the objective of the study. In the case of distance travelled, the distance boundary varies from survey to survey depending on the geographical characteristics of the study area.

Long-distance (LD) or intercity travel is an important subject in transportation study which has been declined by researchers, mostly because of a lack of data and literature. However, it is a highly important component in terms of mileage travelled. Consequently, it is crucial to do more research on LD trips in all aspects, sustainability, LD travel modelling, mode competitiveness analysis, etc.

LD travel is quite different from urban daily travel in terms of spatial-temporal features, regularity, and frequency. Also, it is not easy to capture LD trips in typical surveys. Since they are not typical trips, they are not necessarily conducted during the usual survey periods (fall, weekday) and are less easy for respondents to remember due to their less usual occurrence. These characteristics make this type of trip more complex to observe and analyze.

LD travel is defined differently from paper to paper, depending on the available data, geographic characteristics of the study area and network features. Most of the earlier studies consider distance or overnight travel as a boundary to differentiate a LD trip from other typical trips (Aguiléra, 2015), (J.J. LaMondia, 2015). This definition of a LD trip is important since it can lead to different results (Aultman, 2018).

2.2 Data collection method

Lack of data or poor-quality data is the most challenging part of LD travel studies. There are many efforts to capture LD trips but because of the nature of LD travel, it is not easy to capture it both in terms of spatial and temporal levels. Surveys are the most used method for collecting data in LD studies. However, the uncertainty in remembering LD travel by individuals and the recall period of the survey are some of the issues regarding data collection.

In recent years, data has become more imperative in transportation industries, ranging from real-time reports to collecting data for different kinds of studies. As mentioned before the demand for LD travel is growing over time while our information and data collection for these kinds of trips are not improved at the same level. In this section, the different data collection methods and their applications were discussed.

However, travel surveys for LD trips are quite rare compared to urban travel, but the primary data source used by researchers is the survey. These surveys vary based on location and the study purpose. Frei (2010) has investigated different LD travel surveys conducted in Europe, these surveys vary based on the definition of LD travel, reporting period, and everyday travel diary.

Data sources for LD travel basically can be divided into three categories, survey data, location-based data (cellular phone), and operation data from public agencies such as buses, trains, and aviation companies. Gerike (2018) has summarized a workshop on data collection methods for LD travel and they found three different kinds of surveys for LD travel are used in studies. Diary surveys, in which respondents are enquired regarding their trips on the interview day, single protocol surveys targeting respondents to provide information regarding trips, retrospectively, over a specific recall period and multi-protocol surveys where respondents are categorized using their LD travel patterns and are questioned regarding their trips similar to what is done in single protocol surveys.

In the first survey method, it is quite hard to capture LD travel since this kind of event is rather rare. Hence, the probability of respondents having travelled can be very low on the same day. To address this issue Anderson (2017) represents a mobile phone-based survey that allows a respondent to add LD travel over a long-time period. Kuhnimhof (2009) used a national travel survey (NTS) for LD travel in Europe and found that the recall period varies between 2 weeks and 12 months. The TSRC (Travel Survey of Residents of Canada (TSRC), 2022) can be classified as a single protocol survey since respondents are questioned about the trips they made in the last two months. However, this survey is repeatedly conducted every month with different respondents.

2.3 Modelling Long-Distance trips

The methodological approach for long-distance travel demand modelling usually follows the ones used to model urban trips. In contrast, LD travel differs from daily trips both in terms of frequency and regularity. It is more typical for people to make daily urban trips than to do a LD trip, making it harder to capture over a short observation period. In general, survey methods for LD trips are less developed (and lack interest from decision-makers) rather than regional surveys covering trips during a typical weekday. Currently, few sources of data are available to analyze LD trips. In Canada, some periodic surveys have been conducted at the federal and provincial levels. Still, they only cover a particular period or a particular mode of transportation (Guillemette Y. , 2015). The Travel Survey of Residents of Canada (TSRC) is the only systematic survey conducted in Canada every year for the purpose of tourism study. It is not explicitly designed to model LD trips and needs more in-depth analysis but still provides some relevant insights into travel intensity.

In long-distance (LD) travel, there are no significant features to distinguish a LD trip from other trips. Previous studies are considering distance as a variable to identify LD and non-LD trips. Some studies are considering overnight trips as LD travel, such as (Aguiléra, 2015; J.J. LaMondia, 2015). One study found that various definitions for LD trip can lead to different results (Aultman, 2018).

Miller's study (Miller, 2004) presents several challenges regarding LD travel demand model estimation. It states that higher resolution is needed at both spatial and temporal levels and that data on accessibility should also be collected. The author also mentions that access to private transportation company data should be facilitated (Miller, 2004). Data collection is one of the critical points of each travel demand model estimation. Several studies mention the need to improve data collection regarding the features of attractions (destinations) or the smartphone-based data collection methodologies such as presented in Van Nostrand (2013) and Outwater (2015).

Many studies stated that variables having a significant effect on LD travel patterns include age, gender, having kids, income, population density, and proximity to train stations and airports (Rickard, 1988). Also, authors say that income has a significant positive relation with performing LD trips (Berliner, 2018; Czepkiewicz, 2020; Aultman-Hall, 2018). Two studies mentioned that accessibility to airports plays a crucial role in LD travel behaviours (Enzler, 2017; Aultman-Hall, 2018), but fewer studies have dealt with this issue. It can be due to limited data availability and not having access to individuals' distance and travel time to airports, bus, and train stations. LIorca (Llorca, 2018) has conducted a study on LD travel demand modelling with the same data source that we use in this study, and to deal with the challenge of data collection, despite TSRC, they used Foursquare and Rome2rio data. He found that Foursquare check-in data can improve the goodness of fit of models, particularly for leisure trips. Also, Rome2rio data can improve the mode choice model. However, results might be biased since data from both Rome2rio and Foursquare are searches performed by people (not confirmed travels), especially in mode selection related to Rome2rio data since each search result contains all potential options from origin to destination.

Since LD trips are getting less attention by researchers for many reasons like lack of a good dataset for these kinds of trips, the modelling of LD trips is more or less treated the same way as usual daily trips. (Yao, 2005) developed an integrated intercity travel demand model which considers that all components of LD travel demand are interrelated together. He states that the LD travel choice is related to the destination, trip frequency, mode, route choice, etc. Among the four-step of trip modelling for LD trips, mode choice is getting the highest attention by researchers while trip generation is important as well. (Hess, 2018) developed a hybrid choice model to understand the mode choice of drivers for intercity trips. Results show how the traveller's attitudes like privacy and anti-car attitude can affect mode choice which can also be influenced by longer travel time.

Trip generation is the first step of the four-step travel demand model. Several studies considered different approaches to the trip generation models to find which and how variables impact LD travel. (LaMondia, 2014) used a non-distance-based LD trip threshold to define LD trips by purpose, duration, mode, and destination and used an ordered probit methodology to model trip generation. Some studies used negative binomial regression models to a model annual trip generation, number of trips by purpose, and number of domestic and international ground trips (Berliner, 2018; Czepkiewicz, 2020; Aultman-Hall, 2018).

A classification data set where one class has significantly more observations than the others is defined as an imbalance or rare event data set (Cieslak, 2008). To our knowledge, few studies are considering LD trips as a rare event in their models. Accident occurrence is widely modelled in the transportation field using a rare event approach (Theofilatos, 2016). This study (Theofilatos, 2016) uses a rare event approach to predict accidents on the road. It mentions that the condition of non-event data collection might vary by event condition; hence, they used a rare event logit model package in R software to estimate the model. (Vilaça, 2019) used the rare event method for modelling injury severity risk of vulnerable road users; he used three methods of data preparation (under, over, and synthetically oversampling method) to deal with imbalanced data and decision trees and logistic regression model were used for model estimation.

A rare event dataset can be considered a dataset when one class's occurrences are significantly lower than those of another class. In this case, data can be defined as a rare event or an imbalanced dataset (Chawla, 2008).

The problem with rare event datasets arises when we are interested in the rare class. The majority class biases the decision tree. It leads to a reasonable model accuracy for the majority class. In contrast, it results in poor performance for the minority class (Chang, 2005; Zheng, 2016). To overrepresent the rare event class, it is possible to set up the prior probability for both categories (Enterprise, 2018). "Increasing the prior probability of the rare event class increases the posterior probability of the class, which moves the classification boundary for that class so that more observations are classified into the class." (Zheng, 2016).

Since the LD trip is a relatively rare event, traditional trip generation modelling approaches may be biased by the majority group in terms of model performance. So, in this study, machine learning methods with a rare event approach are employed for model estimation.

2.4 Mode competitiveness

The modal share studies can cover a variety of areas, including mode choice modelling or mode competitiveness analysis. Earlier research has mentioned that socio-demographic and socioeconomic characteristics have a significant influence on mode choice in LD travel (Georggi, 2000) and (Mallett, 2001). A study in China has investigated the passenger's mode preferences for LD travel based on travel cost, safety, efficiency, punctuality, and comfort (Y. Wang, 2015).

Besides the qualitative feature like safety, efficiency, etc., other factors are also essential in passengers' behaviour for mode choice, distance and cost of travel are other important factors influencing the mode choice. A study in China (Wang, 2017) found that higher-income individuals are more likely to use a high-speed railway (HSR) while those with less income prefer regular trains. The plane is also mostly used by the higher-income group. They also found that distance can be a significant feature for individuals in mode choice. The most competitive distance for public transportation mode (bus, regular train, HSR, plane) is 500 km, 500–1000 km, 500–1500 km and over 1500 km. This finding can highlight the importance of travel time and possibility which shows the difference between travel time by private vehicle and other public transportation can have a significant impact on mode choice.

Earlier studies' findings show that in some distances there is not always an alternative for other modes. (Rothengatter, 2010) demonstrates that HSR is competitive with the airplane in distances between 400 and 2000. Another feature influencing the mode choice is the departure time. (Abdel-Aty, 1998) conducted a survey on elderly travel mode choice and found that distance and departure time have a significant impact on elderly travel behaviour. However, the distance and the departure time seem to be significantly important. A study by Savignat (2004) regarding the competitiveness of HSR and airplane found that the total travel time is the most important determinant of mode share.

CHAPTER 3 METHODOLOGY

This study consists of three main parts. In the first step, the analysis of the main corridors over Québec is carried out through a socio-demographic analysis. For the analysis of this part, the data of the 2016 Canadian Census for 2016 is employed at the census tract level. In the next part, trip generation modelling was developed with four different machine learning methods through a rare-event approach. In the modelling part, a generalized linear model (GLM), decision trees and random forest are used for model estimation. This part is followed by the mode competitiveness analysis, which analyzes the travel mode competitiveness for long-distance travel using two different metrics, travel time and distance.

3.1 Data

In this research project, several datasets were used for the descriptive analysis, corridor analysis, modelling and mode competitiveness study. Each part of this research based on the objective and its requirements is using one or a combination of these datasets. These datasets are census data for 2016, TSRC data from 2012 to 2017, access travel time bus and train stations from the Transition tool (www.transition.city), developed by Chaire Mobilité, (Montreal, n.d.), and the LD travel time data between centroids of each CT of CMA and CA.

3.1.1 Travel survey for residents of Canada (TSRC)

The travel survey for residents of Canada (TSRC) is a quarterly survey that supplements the Canadian Labour Force Survey (LFS) which is the main source for tourism study in Canada. This survey replaced the Canadian Travel Survey (CTS) in 2005. This survey provides information about the volume, trip characteristics, the expenditure of travel, and the socio-demographic and socio-economic characteristics of travellers (Statistic Canada, Travel Survey of Residents of Canada (TSRC), 2022). Since 2011, this survey has been totally redesigned which made us use the data from 2012 for this research project. Also, after 2017, the National Travel Survey (NTS) was replaced with TSRC, so the latest accessible data of TSRC is 2017 which is included in this study.

This survey is basically developed focusing on tourism study with the international guidelines recommended by the World Tourism Organization (WTO). However, it is still possible to use this data for the LD travel study. The information in this section is extracted from the user guide for this survey (Statistic Canada, Microdata User Guide Travel Survey of Residents of Canada 2017, 2017).

This survey evaluates the volume of domestic and international travel in Canada made by Canadian residents; it includes data on the trip's origin and destination, trip characteristics, duration in case of overnight trips, activities conducted during the trip, expenditures, and socio-demographic characteristics of the respondent. The highest spatial resolution of the origin and destination points is the Census Division (CD), which corresponds to small regions and metropolitan areas. In this study, the TSRC survey is the main data source used for the LD study. Surveys from 2012 to 2017 were employed for the model proposed in this thesis.

3.1.1.1 Definition of Travel

In this survey, travel is defined as same-day trips and overnight trips. On the same-day trip the distance travelled must be more than at least 40 km (one way), and for an overnight trip there is no boundary for distance; the only requirement is that the traveller spends one night away from home (Statistic Canada, Travel Survey of Residents of Canada (TSRC), 2022). In summary, a trip can be defined as non-frequent travel.

Three types of trips in terms of origin and destination are covered by TSRC:

- "Type A: domestic trips (origin and destination in Canada) with no international component
- Type B: domestic trips (origin and destination in Canada) with an international component
- Type C: international trips (origin in Canada and destination outside Canada) with a domestic component (at least one night is spent in Canada)" (Statistic Canada, Microdata User Guide Travel Survey of Residents of Canada 2017, 2017)

The TSRC is not covering the international trips with no overnight stay in Canada, these trips are covered by the International Travel Survey (ITS).

3.1.1.2 Structure of data

The TSRC consists of three datasets, the person data, trip data and the visit data. The person data provides information on travellers and non-travellers including socio-demographic information of the respondents and the household such as gender, employment, age, income, the highest level of education, number of people in the household, and number of children. So, this file cannot be used for the analysis of the volume of trips, however, it can be used to find the volume of travellers and non-travellers based on their characteristics.

The trip file contains all the information about the trips such as origin, destination, mode of travel, number of nights staying, expenditures during the trip, etc. Each traveller has a record for each trip in the trip file. If a respondent has no trip reported, there is no record for that individual in the trip file.

Information about the places visited by the travellers whether it is the main destination or an overnight stay during the trip is covered in the visit file.

This research project is using the person file and trip file. The trip file is used for descriptive analysis of the volume of trips, and the combination of these two files is used to assess the volume of travellers and non-travellers based on their socio-demographic characteristics. Also, for the development of the trip generation model, the combination of these two files was employed.

3.1.1.3 Sampling and data collection

This survey is covering the population of 18 years old and over as a supplement to the LFS survey. This survey covers 10 provinces, and the territories are excluded from the survey. The sampling consists of six rotations from the LFS survey which means that the respondents are eligible for this survey in the second month of their LFS.

The survey was conducted by computer-assisted telephone interview (CATI) until April 2015, after this date, the electronic questionnaire (EQ) was introduced, so, the non-respondent to EQ was eligible for CATI. The respondents are asked for the overnight trips that ended in the last two months and the same-day trips that ended in the previous month.

“The target population for the TSRC EQ are all LFS respondents aged 18 years and older living in Canada. In order to be selected as a viable household for the survey, the following criteria must be met:

- the entire birth survey must have been completed by the same respondent.
- The household must only have one family.
- a valid phone number must be available.
- a valid postal code must be available; none of the family members can have fictitious names” (Statistic Canada, Microdata User Guide Travel Survey of Residents of Canada 2017, 2017)

3.1.2 Census data

The national census in Canada is conducted by Statistics Canada every five years in mid-May of that year. This data will be used for various planning such as schools, daycare, public service, transportation planning, etc. In this research project, census, population, and census geography are employed in different parts. The census population is mainly used for socio-economic and socio-demographic analysis and the census geography is used for the definition of LD travel between census areas and visualizing the maps.

Census population data for the first time took place in New France in 1666 by Intendant Jean Talon and it only includes sex, age, occupation, and marital status covering a population of 3,215 but the first national census in Canada was administered in 1871. The first online questionnaire was developed in 2006 and approximately 18.5% of respondents welcomed the online version.

For collecting data in the census there is a geographic component, this concept is defined by Statistics Canada at different levels. These geographic areas can be defined by statistical areas and administrative areas. Table 3.1 shows the different components of administrative areas and statistical areas.

In this research project, based on the requirements and available data, the highest resolution of areas possible was used for analysis. In this section, a trip is defined as travel from the centroid of each Census metropolitan area (CMA) or Census agglomeration (CA) in origin to the centroid of each CMA or CA of destination. Also, to analyze the ranking of the corridors, two different methods were proposed. In the first method, the importance of origin and destination cities is considered, and in the second method, the importance of the cities along each corridor is considered. For the first method, the CMA and CA were used and in the second method, the Census subdivision (CSD) was used.

The development of the trip generation model was limited to the highest resolution available in the TSRC which is the CSD. Hence, individual levels were employed for the development of the trip generation model.

Table 3.1 Geographic division of the Census by Statistics Canada

Administrative areas	**Statistical areas**
Canada	Region
Province or territory	Census agricultural region (CAR)
Federal electoral district (FED)	Economic region (ER)
Census division (CD)	Census consolidated subdivision (CCS)
Census subdivision (CSD)	Aggregated dissemination area (ADA)
Designated place (DPL)	Dissemination area (DA)
Forward sortation area (FSA)	Dissemination block (DB)
Postal code	Statistical Area Classification (SAC)
	Census metropolitan area (CMA)
	Census agglomeration (CA)
	Census tract (CT)

In the third section on mode competitiveness, the trip is defined as a travel from the centroid of each CT in the city of origin to the centroid of each CT in the destination city. In this section, since the access time component of LD travel to and from the train and bus station is crucial, the highest possible resolution was considered as the start point for each trip.

3.1.3 Travel time data

Based on the literature, a mode competitiveness study can be assessed by qualitative and quantitative features of the services such as fare, travel time, distance, comfort, efficiency, etc. The qualitative features must rely on a survey among travellers. In this research project, we analyze the mode competitiveness of the bus and train with regards to the private vehicle in Québec using travel time and distance.

The LD travel in this section was defined as a trip from the CT centroid of an origin city to the CT centroid of the destination city. Calculation of the travel time for private vehicles has two components while the travel time by PT mode has four components: access time from the origin to the station, travel time from station to station and egress time from the station to the destination and the waiting time for PT and break time for private vehicle mode, The waiting times for the bus and train are based on company`s website recommendations which is 30 minutes for train and 15 minutes for bus, and the rest time for drivers is a safety recommendation (Lisa, 2021). Three different sources were used to extract data for LD travel time and access/egress time. Travel time from station to station was extracted from the PT companies' timetables, two bus service companies including Orléans Express and Limocar for travel by bus and Via Rail for travel by train. The access and egress time was extracted from the Transition tool, developed by Chaire Mobilité, (Montreal, n.d.). For estimation of LD trip by car, the OSRM (Open Street Routing Machine) engine was used with corrected speeds from historical GPS points in the case of the Montreal region.

The Transition tool allows for estimating travel time by different modes from point A to point B. For estimation of travel time by PT mode with this tool, General Transit Feed Specification (GTFS) files from the agency responsible for the transit service in the area must be imported. "The GTFS is a data specification that allows public transit agencies to publish their transit data in a format that can be consumed by a wide variety of software applications" (Specification, 2022). For this research project, the scenario of 2018 from STM is used for the estimation of access time in Montréal by PT and for the area around Montréal the STL (donneesquebec, 2022), RTL (RTL, 2022) and EXO (EXO, 2022) GTFS file were used.

For estimation of access and egress time, since for cities other than Montréal, it is only possible to estimate the travel time by car, the estimation by other modes like bike, PT, and walk was excluded from the Montréal region for assessment of the mode competitiveness. However, access time to the Montréal central bus station and Montréal central train station was estimated with all the possible modes to assess the ratio of access time with the shortest mode regarding private vehicles. The estimation of travel time by the Transition tool requires the preparation of a CSV file including the coordinate of origin and destination point, unique identifier and the departure time.

The OSRM (Open Street Routing Machine) engine was used to estimate the total trip duration and travel times by car. The OSRM uses the shortest path, under free-flow conditions, for the calculation of travel time. Historical GPS data from *Registre du Taxi* from the Bureau du Taxi de Montréal (BTM) was used for the estimation of travel time and distance in the Greater Montreal Area.

3.2 Corridor analysis

The corridor analysis section attempts to assess the potential demand of the main travel corridors in Québec in terms of the importance of these corridors from a socio-demographic point of view. This part is mainly focused on identifying how potential demand can be assessed independently from the current road network. Analyzing demand under the constraint of the current transportation network could bias the full understanding of potential (or latent) and not allow to imagine new travel corridors. Therefore, this section proposes to develop a corridor network which connects all main cities in the province of Quebec without a priori. The main idea for defining new corridors is to find how demand can be evaluated without having any road, train, and air network. For developing these corridors, we consider there is no barrier for individuals if they want to travel from a city to another one. With these assumptions, a new methodology is developed to define corridors. Then characteristics of the city of origin and destination and the characteristics of the cities along each OD pair are used for ranking corridors to approximate the potential demand.

Since there is no data for demand on LD travel in Quebec and because of the objective of this part of the research, which is trying to find the demand independent of any constraint, this section relies on a logical methodology, and mostly uses try and error method to find relevant corridors.

In the first phase, identification of the cities and the network which connects those cities is done, then socio-demographic data is used for ranking each link of the network.

To analyze the main corridors, two different data sets were used, the 2016 census data for the ranking analysis and the census boundary file in different resolutions (census tract, census agglomeration and census metropolitan agglomeration) to visualize the analysis.

As a result, the analysis describes the potential demand based on several indicators defined in the study and found in the literature. Also, this study includes a new preliminary concept of intercity travel corridors in Québec.

3.2.1 Development of Corridors

In the first step, the identification of important cities is conducted to find which cities need to be included in the study. Three different statistical levels namely Census Metropolitan Area (CMA), Census Agglomeration (CA) and Census Division (CD) are considered the main target point of studies. Then those CMAs and CAs which cover a part of CDs are considered the same target point. For example, Laval is considered part of the Montréal region. In the next step, the study was limited to areas with more than 15,000 people. This filtering process reduces the number of cities considered in this study to 27. Figure 3.1 shows the methodology that has been used for the development of the main corridors in Québec step by step.

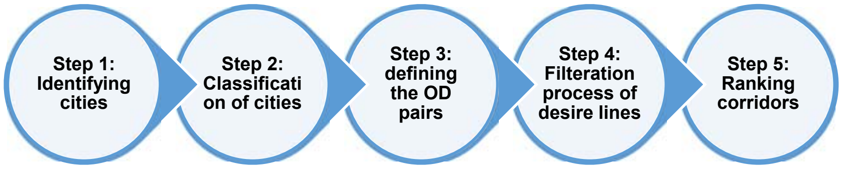

Step 1: Identifying cities
Step 2: Classificati on of cities
Step 3: defining the OD pairs
Step 4: Filteration process of desire lines
Step 5: Ranking corridors

Figure 3.1 Development of corridors methodology

In the second step, all the selected cities were categorized into three levels by their population: core cities, medium-size cities, and small cities. The cities with a population (census 2016) of 100,000 and more are considered main or core cities and cities with a population between 50,000 to less than 100,000 are considered medium-sized cities, and small cities have a population of fewer than 50,000 people. However, the lower boundary for a city to be included in the study is 15,000 people.

Figure 3.2 presents the filtration procedure developed to achieve the main corridors from the desired lines. This process was developed by a try and error procedure. The idea of having the main corridors is to have an optimized connected network between selected cities in Québec with the least amount of OD pairs. Also, it is considered that the corridor network covers the current road network.

The filtration process was employed in five main levels with classified cities as core, mid-size, and small cities to have a connection between all classes of cities. Also, another step has been added to the procedure to connect cities which are on the different sides of the Saint-Laurent River.

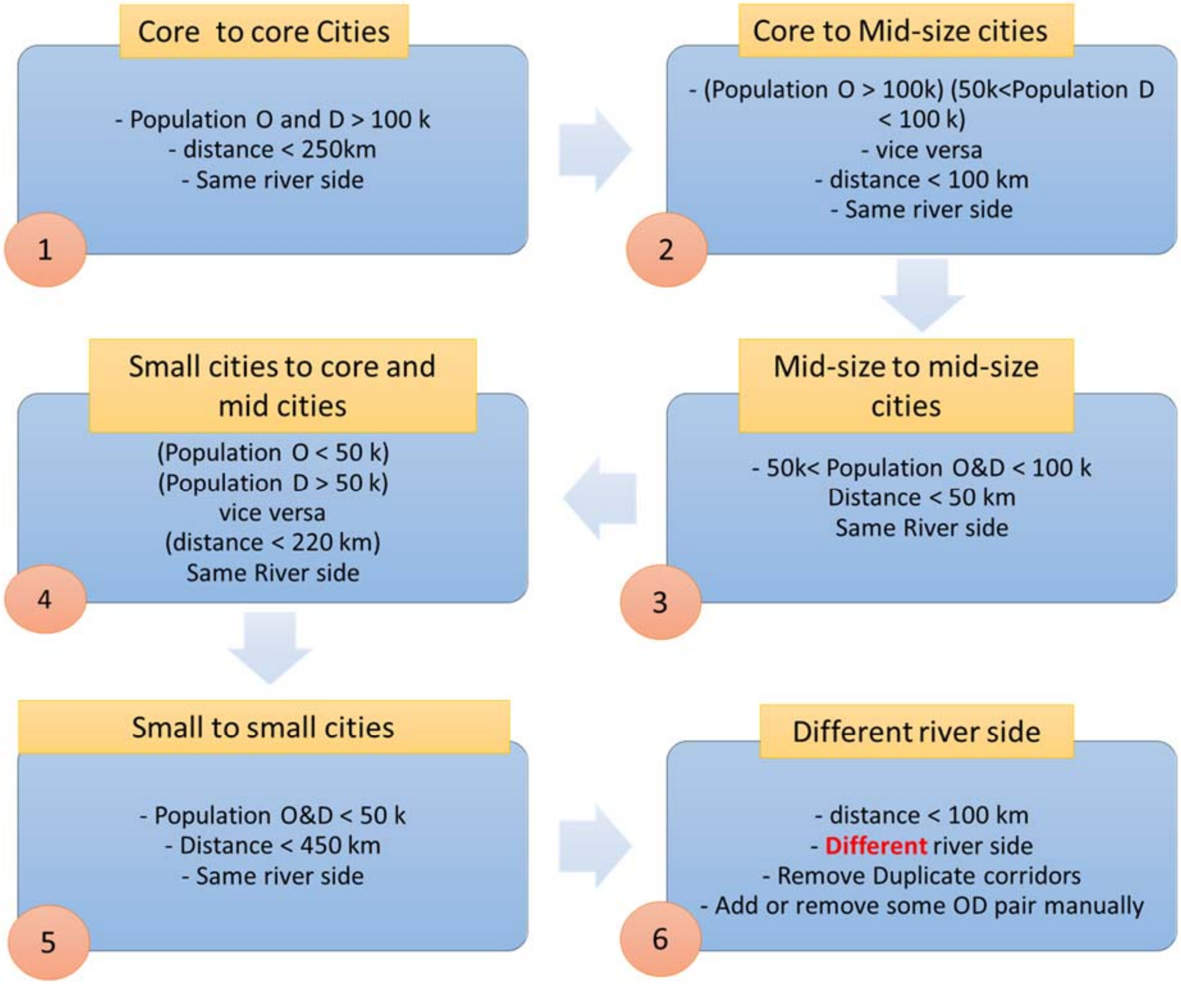


Figure 3.2 Filtration procedure from desired lines to the main corridors

In the first step, core cities are connected to each other with a distance of less than 250 km if they are on the same side of the Saint-Laurent River. Then, core cities are linked to mid-size cities. These cities are a maximum of 100 km away from each other and they are in the same river side. In the third phase, small cities are connected to the core and mid-size cities with less than 220 km distance and being on the same side of the river. Then, mid-size cities are connected with a distance smaller than 50 km. The fifth phase is the connection of small size cities with a distance smaller than 450 kilometres. The last phase is dedicated to linking cities which are on different sides of the river and that are not more than 100 km away from each other. Despite these main steps, the links which are passing more than once over the Saint-Laurent River were removed. Also, a link crossing from a CMA or CA to another one was removed due to connectivity through another

CMA or CA, for example, Montréal to Trois-Rivières is crossing from Sorel-Tracy so this link was removed due to duplicate line.

3.2.2 Ranking Procedure

For the analysis of corridors, several variables were used as intercity travel capacity indicators. These criteria were used in two different categories to rank the corridors. These criteria are based on population and demographics characteristics.

Table 3.2 presents a description of each criterion used in two different ranking systems. The total number of employees of Census Subdivision (CSD) criterion is the sum of employees in each CSD that an OD line crosses. This indicator was selected because more employees result in more business trips; also, it leads to having higher income which results in making leisure trips. The population and young population (15-64) are the other criteria used for ranking analysis. The total population of CSDs through which the OD line passes is considered in the study. The population can be determined as the market size of intercity travel for bus networks and other ridership. A larger total population of a corridor indicates a potentially higher demand throughout the corridor.

Table 3.2 Variable used for ranking corridors

Ranking using the average of Rank	**Ranking using Sum of Rank**
Total employees of CSD along the OD pairs	Number of CSD along Corridor
Total populations of CSD aged between 15-64 along the OD pairs	The population along the corridor by CSD
Average of the median income of CSDs along the OD pairs	Average Density of O&D
Total number of CSDs along the OD pairs	Average population growth of O&D
Total Population of CSDs along the OD pairs	The average number of young people in O&D
The average density of CSDs along the OD pairs	The average percent of employment of O&D
	The average median income of O&D

The average of median income is another indicator used for ranking; this criterion was selected as the average of all median income over all CSDs that the OD line crosses. According to the literature, income is the most important element in conducting an intercity trip.

To rank the corridors, two different methods have been used: in the first method, the average of ranks was considered as the base to rank the corridors and in the second method, the sum of ranks was used to rank them.

In the average ranking method, six different variables were employed. First, each corridor was evaluated based on each variable, for example, each OD pair was ranked based on the total employees of CSD along that OD pair and consecutively for all variables. Then, an average was calculated over all variables for each OD pair. The lowest average was considered the most important variable and the highest number was considered the least important OD pair.

$$R_{OD} = \frac{R_{V1}+R_{V2}+\cdots+R_{Vn}}{N} \qquad \text{Equation 1}$$

$$R_1 = Min\ (R_{OD1} + R_{OD2} + \cdots + R_{ODn}) \qquad \text{Equation 2}$$

$$R_N = Max\ (R_{OD1} + R_{OD2} + \cdots + R_{ODn})$$

Where R_{OD} is the rank of each OD pair and V is each variable and N is the number of variables.

In the second method, each OD pair is assigned a rank like Equation 1 then instead of using average ranking, a summation of the score is considered for the final ranking. For example, the OD with the highest score has the first and most important rank while the OD pair with the lowest score has the lowest rank and is the least important pair. The lowest number of summations was considered the most important OD pair and the highest number was the least important OD pair.

$$R_{OD} = Sum\ (R_{V1} + R_{V2} + \cdots + R_{Vn}) \qquad \text{Equation 3}$$

$$R_1 = Min\ (R_{OD1}, R_{OD2}, \dots, R_{ODn}) \qquad \text{Equation 4}$$

$$R_N = Max\ (R_{OD1}, R_{OD2}, \dots, R_{ODn})$$

Where R_{OD} is the rank of each OD pair and V is each variable.

3.3 Trip generation model estimation

The trip generation model is the first step in the traditional four-step travel forecasting model. Fundamentally, in the urban trip generation model, the study area is divided into transportation analysis zones and the model is developed to determine the relationship between the number of trips generated from each zone and the characteristics of the population and land use of that zone. LD trips differ from daily trips both in terms of frequency and regularity.

It is more typical for people to make daily urban trips than to do a LD trip. This makes it harder to capture over a short observation period. Therefore, a LD trip can be considered a rare event for individuals. In general, a dataset with a binary class when it is not 50%-50% for each class can be considered as an imbalanced dataset, but this consideration is a relative issue. When we have a 40% and 60% for classes, we may not need a balancing technic but when the distribution of minority class reaches less than 30% based on the nature of the data, we may need to balance the dataset. For example, when characteristics of the minority class is quite different from the majority class, even if the 5% of data is for minority class, we may not need to balance dataset, but when there is no difference between the feature of two classes a 30 % of minority class could be considered as an imbalanced data.

This research project aims to develop a LD trip generation model to study the relationship between an individual's completion of a LD trip and socio-demographic characteristics, also it investigates among different models to find the most appropriate trip generation model for LD trips by considering the LD trip as a rare event phenomenon using TSRC data.

On one hand developing a trip generation model for LD travel can highlight what characteristics of people have more impact on conducting a LD trip and support optimization of public transportation services. On the other hand, this part, by evaluating the performance of the models with TSRC data, can demonstrate how important it is to improve data collection methods such as new surveys with more focus on transportation planning studies. Following, data preparation and methodology used for trip generation model are discussed.

3.3.1 Data Preparation

As previously mentioned, the occurrence of LD trips is considered a rare event. Consequently, the dataset contains imbalanced data with the minority (or positive group) of individuals who make at least one LD trip and the majority (or negative group) of individuals who do not make any LD trip during the month of observation. Resampling of training data set is a frequently used method to tackle the issue of the imbalanced data set (He, 2009). Three resampling methods are commonly used: under-sampling, oversampling, and synthetically oversampling. These methods take into consideration having a more balanced dataset on the training level. The under-sampling technique aims to reconstruct a more balanced dataset by randomly removing cases from the majority class to reach the desired ratio of class distribution (Haixiang, 2017).

In contrast, the oversampling technique aims to reconstruct a more balanced dataset by randomly duplicating cases from the minority class to reach the anticipated ratio of class distribution (Haixiang, 2017). The oversampling technique's challenge could be an overfitting model by improving recognition of the minority class (He, 2009). The synthetic oversampling method is another technique to tackle the issue of overfitting caused by oversampling (Menardi, 2014). In this technique, synthetic cases with a feature of minority class according to the smoothed-bootstrapping method are generated.

The first step in the trip generation modelling is to evaluate the proportion of people who made at least one LD trip against those who did not during the study period. It was found that each month, around 25% of people are making at least one LD trip. Since those who do LD trips are the minority group in the dataset, traditional statistical models underestimate the probability of this minor class.

In this study, the three techniques mentioned earlier are employed. We found that the under-sampling technique led to better model performance. So, the cases were randomly removed from the majority class, individuals with no trip cases, to have a more normally distributed classification dataset. The difference between these techniques is represented in Figure 3.3.

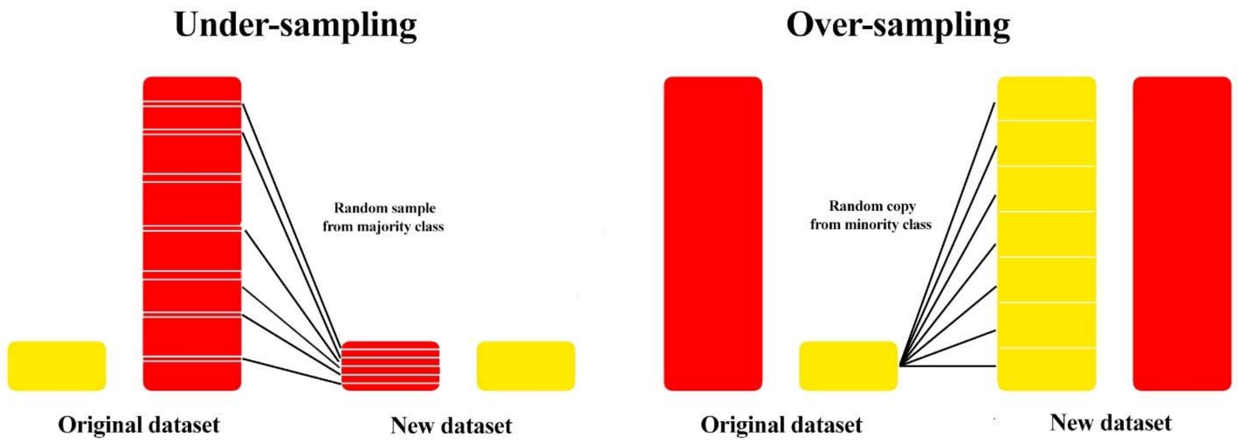


Figure 3.3 Under-sampling and over-sampling methodology

The TSRC data from 2012 to 2017 were used for modelling. This dataset contains three files consisting of person, trip and visit files. In this part of the research, the person file and the trip file are combined to find the people who are not travelling either. If a respondent does not report any trip, the trip file does not cover that respondent, so by a combination of these two files, it is possible to have all individuals notwithstanding if they have reported any LD travel or not.

Figure 3.4 presents the procedure of the combination of these two datasets, the common column in both files is the public use microdata file number (PUMFID) which uniquely represents each respondent.

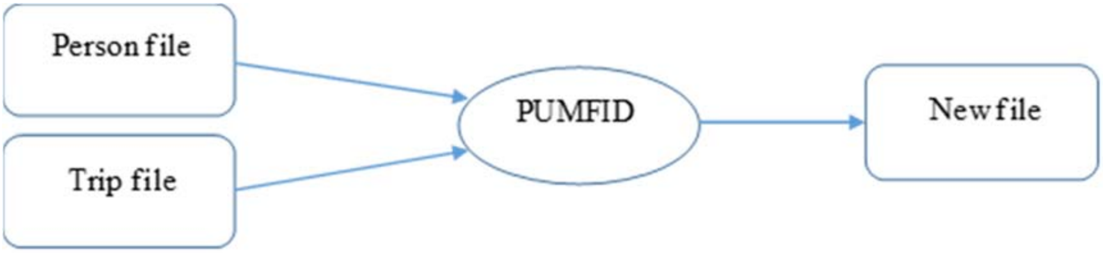


Figure 3.4 Combination of TSRC data procedure

Rare event modelling has received less attention in long-distance trip generation modelling. In this study, a priori mentioned methods that are common in rare event data preparation were assessed. It was found that under-sampling the majority class leads to better model results. Different machine learning metrics are used to find the best-fitted model. They are described hereafter.

3.3.2 Logistic Regression Model

John Nelder and Robert Wedderburn proposed an advanced statistical modelling technique called Generalized Linear Model (GLM) in 1972 and it includes Linear Regression, Logistic Regression, and Poisson Regression. It is suggested to use GLM instead of a regular linear model when the relationship between x and y is not linear or when the variance of errors in y varies with x. GLMs are an extension of the traditional linear models. An ordinary linear model requires that the error term be normally distributed, while in GLM, this assumption is relaxed. In the GLM model, the response variable can be binomial, Poisson, and other kinds of distributions from the exponential family. In this study, because we have two classes, the response variable is set up as binomial. In the linear model, the predictions estimate the response variable, while in GLM, it is the function of the response variable prediction.

In general, a generalized linear model (GLM) for binary data is the inverse of the standard logistic function:

$$logit(p) = \log\left(\frac{p}{1-p}\right) \qquad \text{Equation 5}$$

Where p is the probability, and logit(p) is the logarithm of the corresponding odds. The logistic models can lead to biased results on classification problems with imbalanced datasets. Ma and Lukas (2021) compared the performance of several machine learning methods and GLM on rare event datasets (Ma, Lukas, 2021). They found almost the same performance for classification problems in both ways. The GLM method has better sensitivity for the classification method of rare events. However, some studies are stating that the machine learning method has better results in general. Lu (2021) compared the different approaches to analyze the hesitancy in the choice of transfer airport and the results show that the random forest and deep reinforcement learning models have a more accurate and structured results. Further discussion on GLM can be found in the paper “A Tutorial on Generalized Linear Models” (Myers, 1997).

In this study, generating a LD trip is assumed as a rare event phenomenon, and "Have Trip" corresponds to a trip made by a respondent and is assumed to be a positive class in the model.

3.3.3 Decision Tree

Decision trees (DT) are a valuable method to identify homogeneous subgroups distinct by individual characteristics. This study utilized the Classification and Regression tree (CART) technique and the Conditional Inference tree (CTree) technique. These DT approaches have the advantage of being easy to explain and interpret. Results can be represented graphically, and, for qualitative variables, there is no need to create dummy variables. The performance of these techniques is compared with other models in the process of model selection.

One of the most used methods for building a decision tree, "CART", was developed by Breiman (Breiman, 2017). A split in CART aims to minimize the relative sum of squared errors in the two partitions of a split. The splitting process consists of two steps: 1) the best split will be found across all covariates; then, 2) the point will be split up for those covariates.

CART searches across all splits generated by predictor variables for split selection. Then the split with the most significant criterion is selected to transfer samples into corresponding sub-nodes. It has been mentioned by studies that when there are numerous split points for each variable, this method might be biased with variable selection. This issue of variable selection with many possible splits is widely discussed in previous studies (Breiman, 2017; Shih, 2004; Loh, 1997).

The CART approach basically consists of three main steps: 1) growing the three, 2) pruning, and 3) selecting an optimal tree. In the first step, based on the values of a set of covariates, it recursively executes univariate splits of the dependent variable. This splitting of a feature results in two splits by choosing one variable and its split value. Child nodes are then treated like parent nodes, and this process continues until some criterion is met (Mishra, 2003). In the pruning process, CART aims to reduce the tree's complexity by replacing nodes and subtrees with leaves. This process can reduce the size of the tree and, in some cases, improve the classification accuracy (Patil, 2010). CART algorithm is using GINI index to determine the how well a split in decision tree is made, this value is varying between 0 to .50, basically it helps to find which variable make purer split in decision tree where lower GINI index led to purer split node.

A conditional inference tree (CTree) algorithm was proposed by (Hothorn, 2006). CTree avoids variable selection bias in the CART algorithm; instead of selecting a variable that maximizes the Gini index, it uses a significance test method to choose that variable. CTree selects the predictor variable for split by statistical testing between response and covariate. CTree, in each step, uses traditional statistical procedures; in the case where both response variables and possible split variables are categorical, it uses the Chi-squared test (χ^2), in case where one variable is categorical and one is numeric, one-way ANOVA (analysis of variance) is performed, and for both numerical variables, the Pearson correlation test is employed (Schlosser, 2019).

To assess the association of each covariate and outcome, the mentioned test is employed in CART. If sufficient evidence is found to reject the global null hypothesis, the node will be selected to be split. The covariate which has the strongest association with the outcome of interest is chosen as a candidate for splitting. Song (2015) is providing insightful information regarding the introduction to decision trees and Dobilas (2021) provides a perfect example of application of CART.

In this study, the CART algorithm was performed from the Rpart R package, the party R package was employed for the CTree algorithm, and the caret R package was used to achieve the predictive performance of both algorithms.

3.3.4 Random Forest

Random forest (RF) is one of the classifications and regression models widely used for binary class datasets. The random forest model uses many decisions, tree-like models, bootstrapped data, and the model decision based on the average prediction of all decision trees. The random forest model, by reducing the correlation between decision trees, aims to improve variance reduction. The study by Kabir (2018) expresses how the random forest method improves the predictive performance by deciding based on the growth of numerous trees. Random forest allows each tree to individually sample from the data set randomly with replacement with the same sample size and different variables, which results in different trees. Hence, these decision trees are susceptible to "train" data sample that changes in the training set, resulting in various structures in decision trees. This process is called Bootstrap Aggregation (Bagging).

Generally, the RF development process consists of four steps. Firstly, it uses bootstrapping to choose a sample from a training set with the same sample size. It then uses a subspace method to select different dependent variables from the total set of variables. Having a new sample will build a decision tree, and finally, the RF model performs the tree steps repeatedly to make many trees. The number of trees is determined by an error called OOB (Out-of-Bag) error.

Having more trees in the model leads to a lower OOB error rate in an RF tree, which is desired because a lower error rate results in better accuracy in the RF model. On the other hand, having more trees will increase the possibility of similar trees. This issue is tackled in the random forest by restricting the number of variable selections with the subspace method. In this process, adding more trees does not result in overfitting, but there is not much benefit to growing more trees (Friedman, 2009).

RF model uses a Gini indicator which is a non-parametric value to evaluate the prediction power of variables in classification data based on impurity reduction (Strobl, 2007). The Gini index of a node n is calculated as follows:

$$\mathrm{Gini(n)} = 1 - \sum_{k=1}^{2} (P_k)^2 \qquad \text{Equation 6}$$

where pk is the relative frequency of class k in the node n

To measure the importance of variables, random forest utilizes the MeanDecreaseGini index. This index is calculated based on the Gini impurity index used to calculate splits (Atkinson, 1970). This study uses the MeanDecreaseGini to identify the importance of explanatory variables that contribute to the model. MeanDecreaseGini is considered as the average decrease of the Gini impurity index over all trees in the model. In this study, the random forest algorithm was used from the Random Forest R package, and the caret R package was used to achieve the predictive performance of the random forest. To go further in detail on random forest the paper “A random forest guided tour” provides more details (Biau, 2016), Also, the website of “ Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review ” provides a broad explanation of random forest (Dementias, 2017)

3.3.5 Comparison Indicators

To compare the results of models, classification accuracy is employed. Table 3.3 presents the agreement of observed and predicted conditions of the test dataset for "having trip" and "no trip."

Having a LD trip refers to a positive or event class, and not having a LD trip refers to a negative or non-event class.

Table 3.3 Confusion matrix components

	Predicted Condition	
Actual Condition	True positive (TP), Hit	False-negative (FN), Miss, underestimation
	False-positive (FP) Overestimation	True negative (TN) Correct rejection

As it can be deduced from Table 3.3, the components of the confusion matrix are represented below:

- True positive (TP) = the number of instances correctly identified as "have trip."
- False-positive (FP) = the number of instances incorrectly identified as "have trip."
- True negative (TN) = the number of instances correctly identified as "no trip."
- False-negative (FN) = the number of instances incorrectly identified as "no trip."

To compare the models, result accuracy is the most common value found in the literature (Cieslak, 2008; Rachman, 2019). It is defined as the ability to differentiate the event and non-event correctly; the mathematical calculation of the accuracy is presented in equation 2.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

The accuracy represents the overall performance of the model, but it does not measure the accuracy of prediction on each class of events and non-events. The model's overall accuracy can be biased by the majority class because most of the observations are in the "no trip" class. To tackle this issue and avoid misinterpretation of prediction accuracy with overall accuracy, sensitivity, specificity, F1 score, and precision are considered prediction performances of the negative and positive classes of the model.

The sensitivity of a test states its ability to determine the "have trip" or positive class cases correctly, and the specificity of a test is its ability to determine the "no trip" or negative cases correctly. Mathematically these two parameters can be stated as equations 3 and 4 consecutively.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

Precision is another factor that describes how well a model predicts the "have trip" or positive class. It evaluates the proportion of correct positive predictions to the overall wrong and correct positive class. This score is mathematically stated as equation 5.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

A harmonic mean of precision and sensitivity called the F1-score is used. It allows having a better measure of accuracy on negative classified cases. There are several advantages of using an F1-score instead of accuracy. In the F1-score, the importance of false positive and false negative is also taken into consideration, while in the accuracy indicator, the focus is just on true positive and true negative classes. Also, it is more relevant for an imbalanced dataset to use F1-score, while for a normally distributed dataset, accuracy can be employed. In this study, the "have trip" class is considered rare, so the F1-score is a better metric for evaluating the model performance. The F1-score is estimated using equation 6.

$$\text{F1} - \text{score} = 2 * \frac{sensitivity * precision}{sensitivity + precision} \tag{6}$$

The result of these indicators is stated in the result chapter.

3.4 Mode competitiveness

The main purpose in this section is to analyze the competitiveness of modes of transportation for LD trips in Québec to determine the extent to which alternative modes of transportation are competitive with the private vehicle mode. Many factors influence the willingness of individuals to use an alternative to the private car. Hence, to further investigate these factors, it is relevant to determine if there is an appropriate alternative to the car in terms of availability and travel time. The following methodology section proposes a brief description of the study area and data, as well as a description of the method used in this research.

3.4.1 Travel time estimation

Model share studies aim to evaluate the factors influencing people choosing one mode over another. This study is trying to assess whether alternatives to the car are competitive or not, using total travel times.

Figure 3.5 shows the different phases of a long-distance trip by public transportation service. The first part is the access from the origin which is considered the centroid of the resident census tract to the bus or train station which can be accessible by walk, bike, public transportation or car. The second phase is the main travel from the station in the origin city to the station in the destination city. The last phase is the egress from the station to the centroid of the destination census tract. As a first step to analyzing the competitiveness of the public transportation option concerning the private vehicle, we assume the access and egress segments of the trips are traveling by car. There are many other factors influencing LD travel mode choice and those are typically gathered using surveys. A study by Van Can (2013) presents the impact of personal characteristics on mode choice (Van Can, 2013).

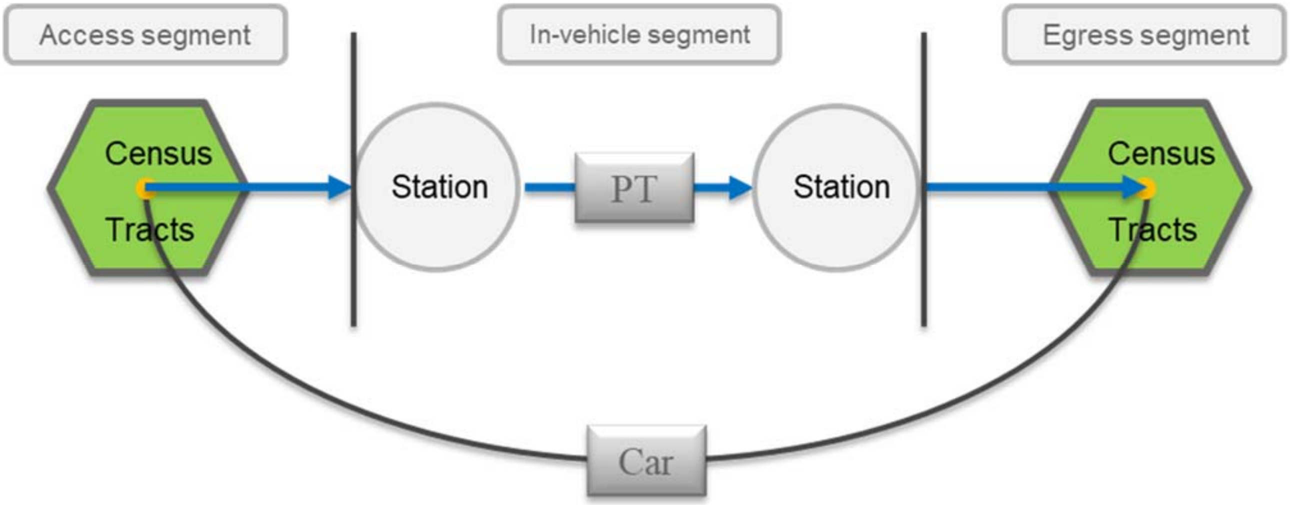


Figure 3.5 Different phases of LD trip by public transportation mode

According to Roman (Román, 2014), travel time is defined as the total time from origin to destination. In addition to total travel time, the waiting time plays a crucial role in mode competitiveness since this waiting time can significantly affect the total travel time specifically for short trips. Based on PT company websites (recommendations to travellers), the waiting time for the train mode and the bus mode was considered 30 and 15 minutes respectively.

The total travel time calculation for PT mode is estimated as Equation 1:

$$TT_{PT} = T_{acc} + T_T + T_{egr} + T_w$$ Equation 7

where the TT_{PT} is the total travel time with public transportation mode, T_{acc} is the access time, T_T is the travel time from station to station, T_{egr} is the egress time and T_w is the waiting time.

For the private car, a waiting time or a resting time is also considered because of tiredness while driving or the necessity of refuelling the car. For safety reasons, it is suggested to have 15 minutes break after two hours of driving (experts, 2021), so we added 15 minutes to a private vehicle mode for every two hours of travel time.

In addition, there is another consideration for PT trips. In the Montréal region, there are three bus stations and two train stations, and in Québec, there are two bus stations and two train statins. Therefore, for the calculation of public transportation travel times, several total travel times were calculated based on the composition of the stations and the availability of services between the OD pairs and the smallest travel time was selected as the total travel time by PT.

Equation 2 shows the estimation of total travel time for private car trips where the TT_{C} is the total travel time, T_{T} is the travel time from origin to destination and T_{R} is the combination of rest and refuelling time.

$$TT_C = T_T + T_R$$ Equation 8

Estimated travel time to and from the station was conducted during off-peak hours, and travel time by car from an origin to a destination was calculated under free-flow conditions. In the next section, the comparison of travel times using different modes is presented.

3.4.2 Study area

This study aims to conduct a mode competitiveness analysis covering the Census Agglomeration (CA) and Census Metropolitan Agglomeration (CMA) located in Québec; the census division of 2021 is used. For a more detailed analysis of travel time competitiveness, it is necessary to use a sufficient level of spatial resolution since the access and egress segments are crucial for using public transport (PT) options such as buses or trains. Therefore, the census tract (CT) level is used to estimate distributions of travel times between all CAs and CMAs. The access and egress components may not play a vital role in a small area but when it comes to large ones such as Montréal, it becomes a highly important element affecting total travel times between CTs.

Montréal, Québec, Trois-Rivières, Sherbrooke, Saguenay, Drummondville, Gatineau, and Granby are the cities selected in this study. They respectively consist of 970, 181, 40, 49, 45, 24, 73 and 20 census tracts, which amounts to a total of 1402 census tracts, resulting in more than 900,000 origin-destination (OD) pairs, Figure 3.6 shows the location of the cities which considered for mode competitiveness study on the map.

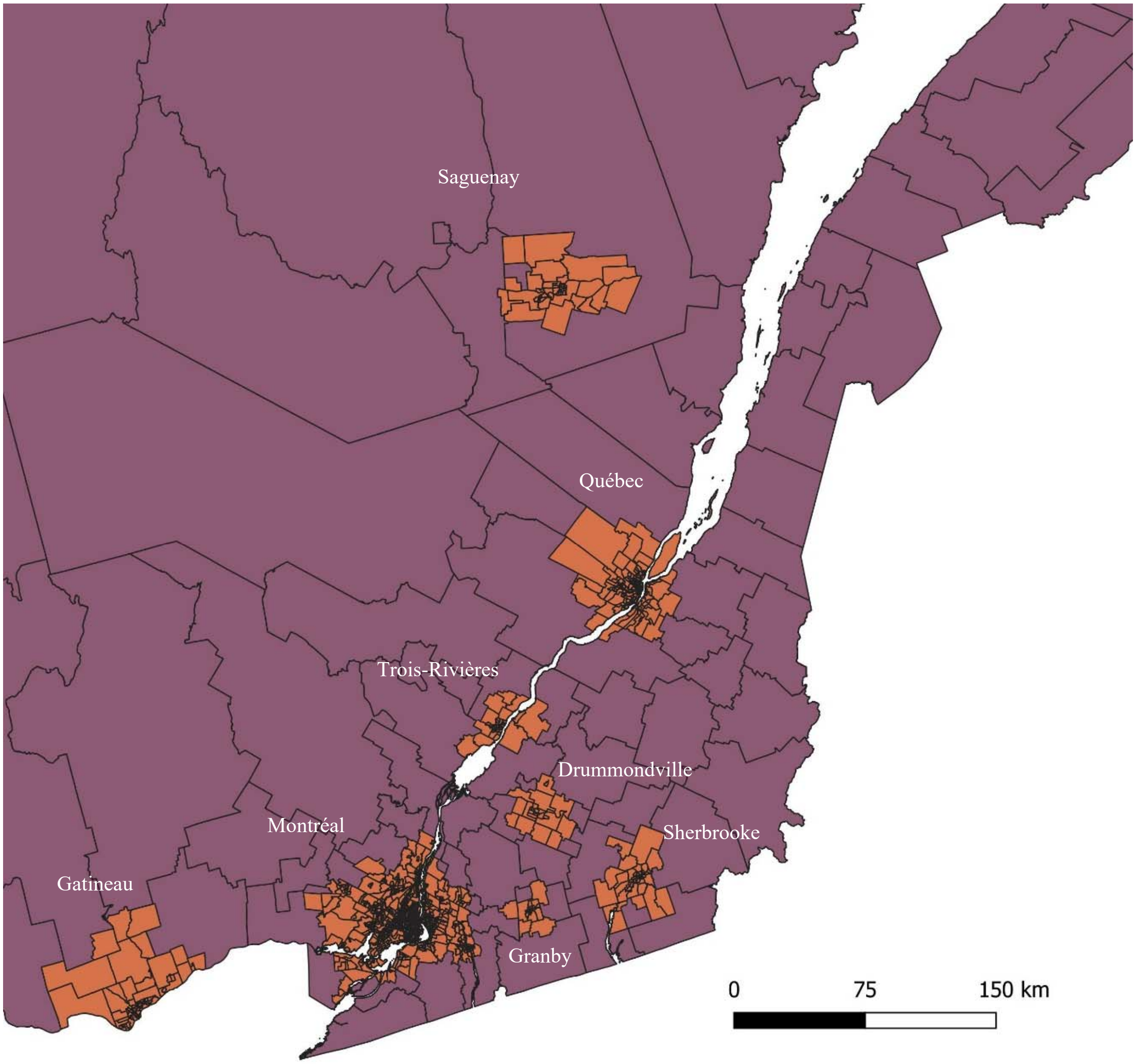


Figure 3.6 Location of CA and CMAs in this study

CHAPTER 4 RESULT

Chapter four presents the results coming from the application of the methodology in chapter three. This chapter starts with a descriptive analysis based on the Travel survey of residents in Canada (TSRC), then the corridor analysis finding is represented. Following, the trip generation model estimation results are shown, and at the end, the mode competitiveness outcomes are described.

4.1 Descriptive analysis result

The data and variables used in this research will be briefly described and their relations will be explored to assure that they support our general hypotheses in the trip generation models. The main source of data used in this study for primary analysis and the trip generation model estimation is the TSRC. This survey has been conducted since 2005 for domestic travel in Canada to provide information on volumes of trips on the demand side as well as expenditures during the trip. This survey includes:

“- The total volume of same-day and overnight trips taken by residents of Canada with destinations in Canada,

- same-day and overnight visits in Canada,

- the main purpose of the trip/key activities on trip,

- spending on same-day and overnight trips taken in Canada by Canadian residents in total and by category of expenditure,

- modes of transportation (main/other) used on the trip,

- person-visits, household-visits, spending in total and by expense categories for each location visited in Canada,

- person- and household-nights spent in each location visited in Canada, in total and by type of accommodation used,

- use of travel packages and associated spending and source of payment (household, government, private employer),

- demographics of adults that took or did not take trips, and

- travel party composition.” (Statistic Canada, Travel Survey of Residents of Canada (TSRC), 2022)

Table 4.1 No. of observations in TSRC data set by having or not having a long-distance trip

	No. of observations		Percentage			
	No Trip	Have Trip	No Trip	Have Trip	No Trip (Weighted)	Have Trip (Weighted)
Year						
2012	121,335	41,866	74.30%	25.70%	64.09%	35.91%
2013	101,772	37,408	73.10%	26.90%	63.87%	36.13%
2014	100,979	36,213	73.60%	26.40%	64.34%	35.66%
2015	99,655	34,843	74.10%	25.90%	64.25%	35.75%
2016	111,279	35,358	75.90%	24.10%	64.34%	35.66%
2017	104,948	36,992	73.90%	26.10%	65.53%	34.47%
Gender						
Male	290,135	101,040	74.20%	25.80%	63.43%	36.57%
Female	349,833	121,640	74.20%	25.80%	65.37%	34.63%
Age group						
18-24	39,327	17,593	69.10%	30.90%	60.27%	39.73%
24-34	88,663	37,577	70.20%	29.80%	61.23%	38.77%
35-44	98,788	38,912	71.70%	28.30%	63.45%	36.55%
45-54	112,067	39,717	73.80%	26.20%	64.78%	35.22%
55-64	125,604	44,811	73.70%	26.30%	64.05%	35.95%
65+	175,519	44,070	79.90%	20.10%	70.55%	29.45%
Employment						
Employed	356,439	147,629	70.70%	29.30%	61.14%	38.86%
Not Employed	283,529	75,051	79.10%	20.90%	70.16%	29.84%
Income						
Less than 50k	268,778	63,218	81.00%	19.00%	73.18%	26.82%
50k to 70k	85,809	32,632	72.40%	27.60%	63.85%	36.15%
70k to 100k	85,892	39,414	68.50%	31.50%	59.73%	40.27%
100k and over	113,274	65,487	63.40%	36.60%	53.59%	46.41%
N/A	13,742	3,729	78.70%	21.30%	71.42%	28.58%
Education						
Less than high school	122,249	20,941	85.40%	14.60%	64.41%	35.59%
High school certificate	131,927	38,805	77.30%	22.70%	77.47%	22.53%
University degree	121,431	63,469	65.70%	34.30%	57.59%	42.41%
Children						
No Child	474,494	157,307	75.10%	24.90%	64.89%	35.11%
Have Child	165,474	65,373	71.70%	28.30%	63.33%	36.67%

Table 4.1 presents a descriptive cross-tabulation of observations by having or not having a long-distance trip. This table contains some of the main variables included in the survey. This descriptive analysis of the variable provides valuable insights on having or not having a long-distance trip. Gender appears to not have a significant role in conducting a long-distance trip since the share for both classes is the same and around 36% for having a trip and 64% for not having a long-distance trip for men and women. However, the role of the dataset and the survey can play a vital role in the share of gender. Most of the trips in the survey are with the purpose of leisure or visiting family or friends, so these kinds of trips usually can be family trips or couple trips. If the dataset contains more business trips, the share of gender could be different. Reichert (2015) states that when it comes to LD travel with the purpose of business, men have a higher share of LD travel both in terms of frequency and distance travelled in all modes (Reichert, 2015).

The role of age in making a long-distance trip is not easy to assess since the proportion of respondents is not the same for each age group. As shown in the table, younger age group individuals are making more long-distance trips and the percentage starts to decrease with age and there is the least percentage for people aged 65 years and older. The role of age will be assessed more precisely in the trip generation modelling.

As for employment, it was expected that employed individuals have a higher probability to make a long-distance trip which is presented in

Table 4.1. Income is one of the most important variables, many studies state that higher income results in more trips. Besides, employment status and income can act as a proxy for each other because employed people have higher income which increases the odds of making long-distance trips.

The higher education cases make more long-distance trips. It can be seen in the table that higher educational level is increasing the percentage of individuals who make long-distance trips by an average of 6%. (Joyce M. Dargay, 2012) also claims that in her study higher education, car ownership and income increase the number of long-distance trips. The last variable is having children, and as of descriptive analysis having kids is related to a higher the number of people who make long-distance trips.

The trip rate for the study period is represented in Figure 4.1 for all trips performed in Canada. This graph shows how the number of trips per person is rising during the summertime and the end of each year which is a vacation period. The interesting finding of the graphs is how the trip rate is steady during the study period.

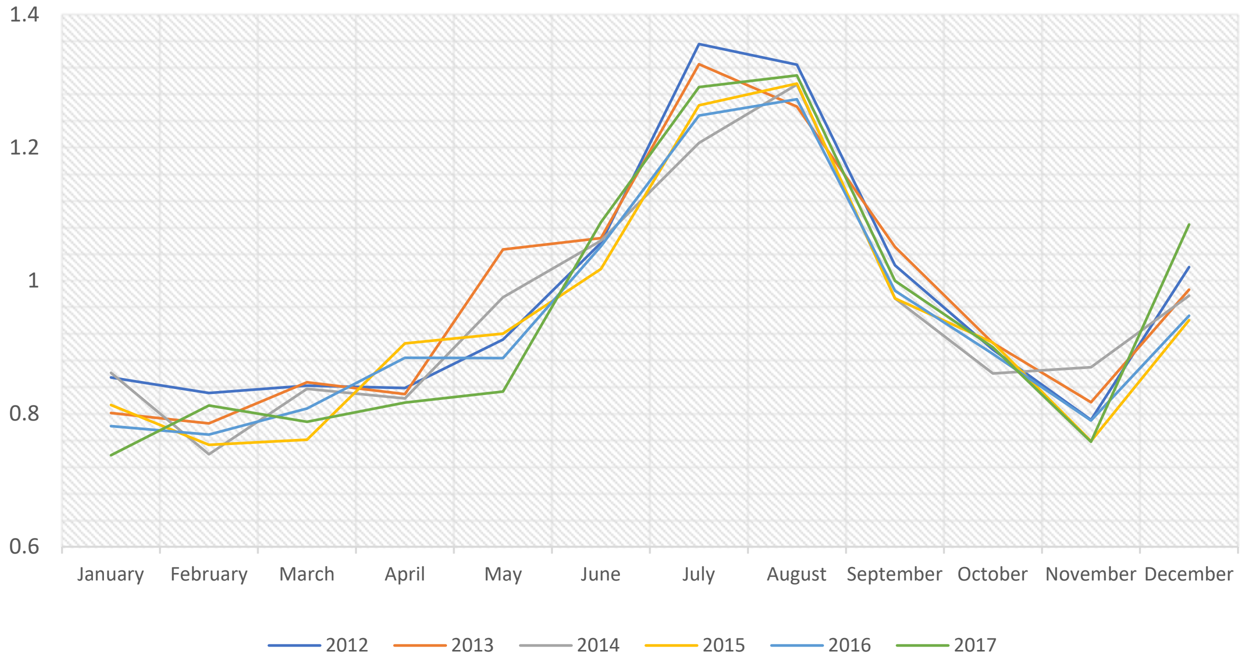


Figure 4.1 Number of trips per person per month for all trips performed in Canada

The same ratio of the number of trips per person per month is presented in Figure 4.2 for overnight and same-day trips for all trips in Canada. The result illustrates the same pattern for trip rate, however, in the overnight trips, there is a significant rise during summer since most of the overnight trips are with the purpose of leisure, recreation and visit friends and families and it is expected that these kinds of trips take place in the summertime. On the other hand, the TSRC survey is conducted with the purpose of tourism studies, so around 80 percent of trips have leisure and visit friends and family's purposes. The second graph shows, however, that the trip rate for same-day trips is higher during the holiday season, but it rises less than overnight trips.

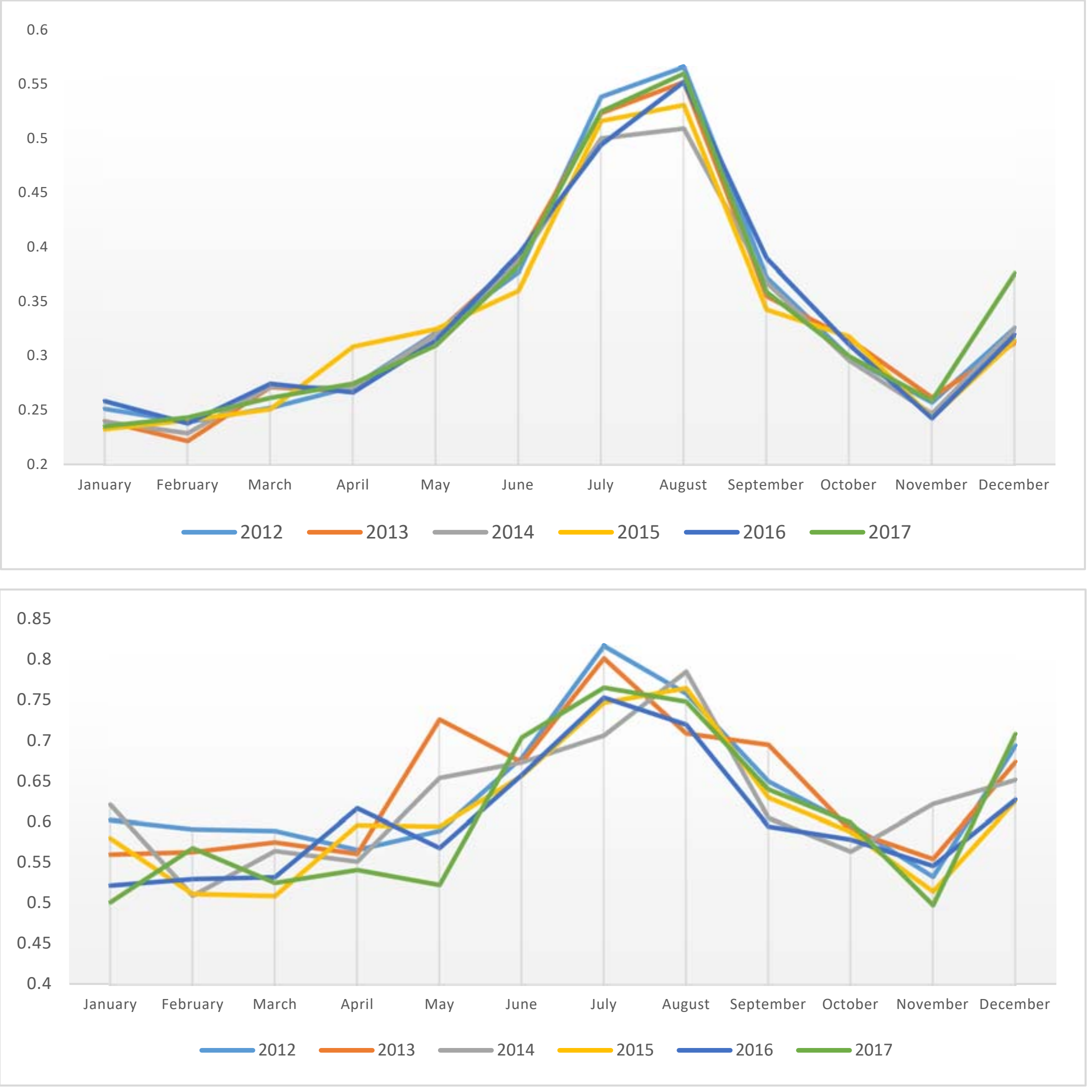


Figure 4.2 Number of trips per person per month for overnight(top) and same day (bottom) trips performed in Canada

The trip rates in Québec has minor differences. However, the trend for overnight trips in Québec is the same as for Canada; there is a peak during summer and at the end of the year but the highest trip rate in Québec is less than the one for Canada by 0.1 trip per person per month. Also, in Canada the highest trip rate in the year is three times more than the lowest trip rate for overnight trips while in Québec this number is close to two times.

In case of same-day trips, in 2013 the highest trip rate in Québec is 0.1 trip per person per month more than Canada. However, we do not see a similar trend in the other years. Also, the same day trip rate in Canada is between 0.5 to 0.8 while in Québec this rate is between 0.35 to 0.9 in different times of the years.

In both Québec and Canada, the trip rate for same-day trips is higher than overnight trips which can be a result of the type of survey, because as it is shown in Figure 4.3 around half of respondent are visiting their friends and families, this can be a reason why same day trip rates are higher than overnight trip rates.

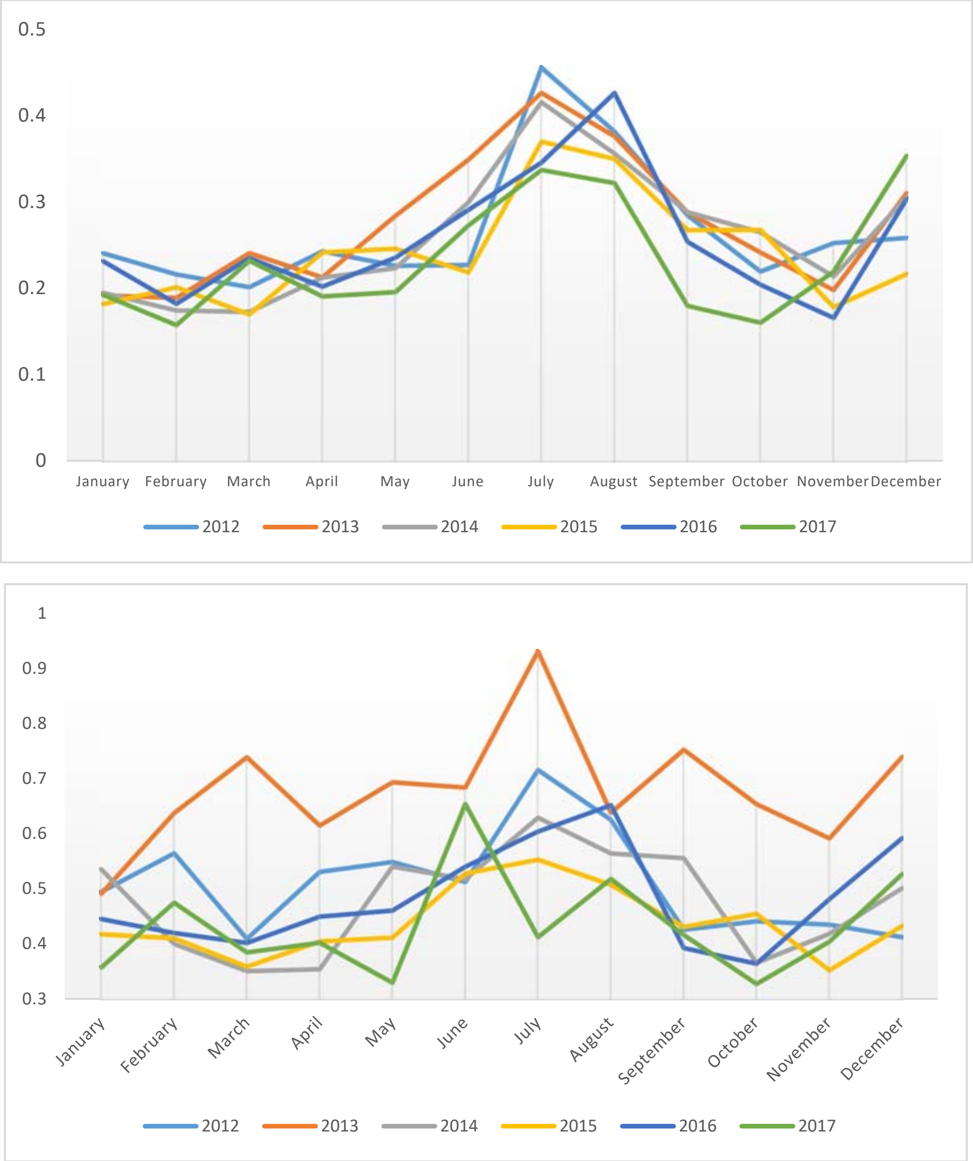


Figure 4.3 Number of trips per person per month for overnight trips (top) and same day (bottom) performed in Québec

As mentioned above, the data set extracted from a survey based on tourism studies, so as it is shown in Figure 4.4 around 44 percent of trips are for visiting friends and families, about 35 percent of trips are holiday and leisure trips, 8 percent are business trips and 12 percent of trips have other or not mentioned reasons. This graph can explain why the trip rate in holiday seasons is around two to three times higher than the rest of the year.

Besides seasonal effects which make the trip rate higher in summer, the trip purpose can be another reason for significant increase of the trip rate in summer, since most of the trips are for leisure and visit friends and family.

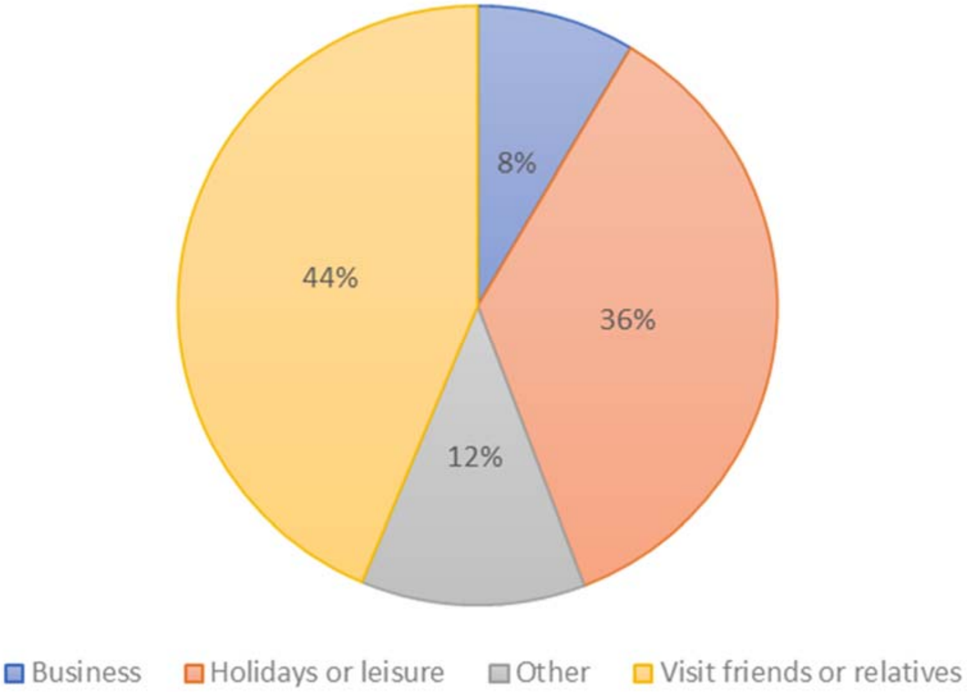


Figure 4.4 The proportion of trips by the purpose of travel for all trips in Québec during 2012 to 2017 (weighted)

To our knowledge, most studies in long-distance trip modelling are concerned with mode choice modelling. This study also had the objective of developing a mode choice model. However, Figure 4.5 shows that more than 92% of trips are done with the private car, so developing a mode choice model will plausibly result in all trips using the private vehicle. Because of this challenge a mode competitiveness study based on travel time point of view is conducted to evaluate how other modes can compete with private vehicles.

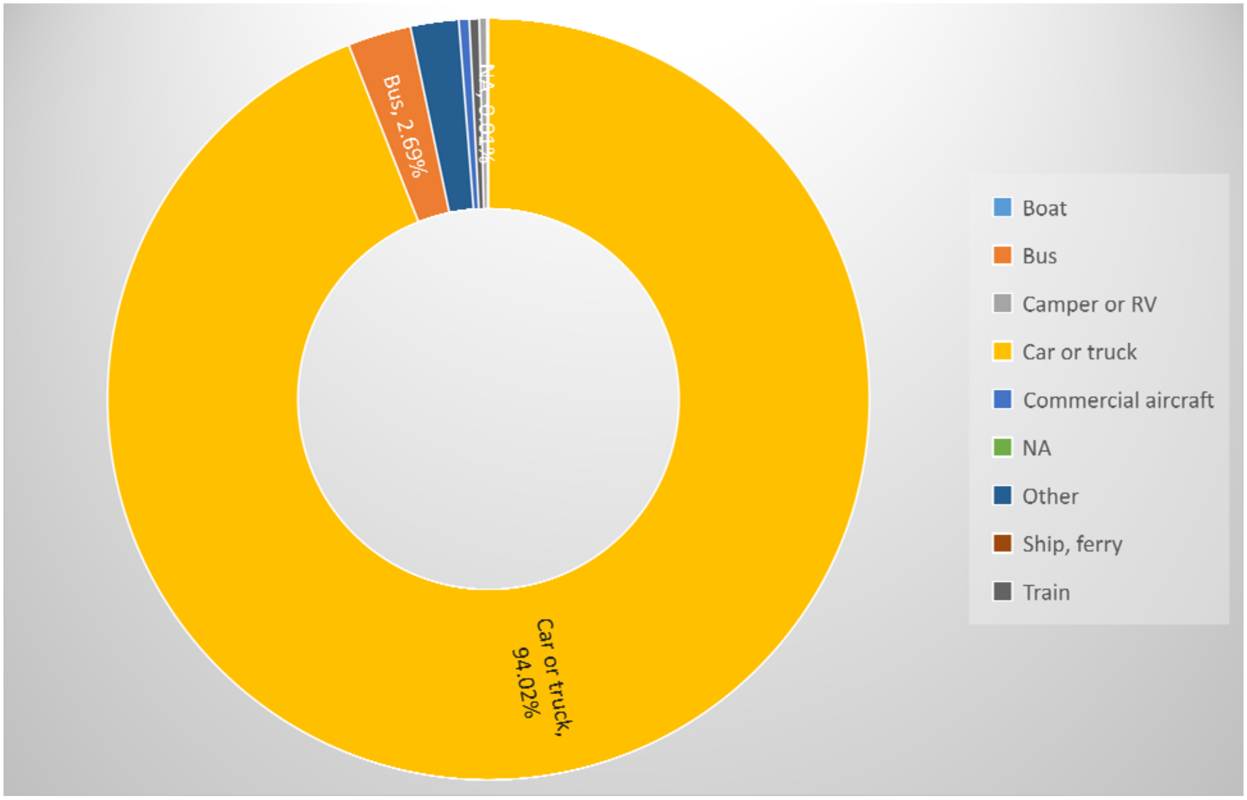


Figure 4.5 Mode share for all long-distance trips within Québec

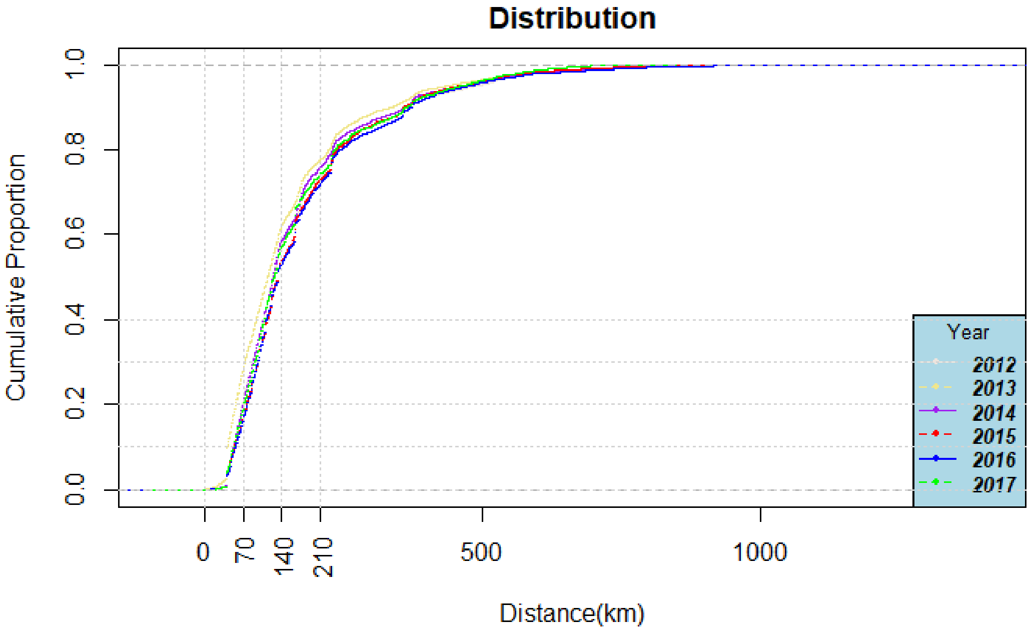


Figure 4.6 Cumulative distribution of travel distance for each year for all trip in Canada

One of the main boundaries to distinguish between long-distance trips and urban daily trips found in the literature is the distance travelled. Different studies based on the geographic area consider different distances. In the TSRC dataset there is no boundary in terms of distance while a long-distance trip is considered as a non-frequent day trip and overnight trip. Figure 4.6 shows the cumulative histogram of distances travelled in different years. This graph shows that long-distance trips start at 7 km and around 80% of these trips are less than 210 km. This results from the fact that origin and destination of most trips are within the same province.

Figure 4.7 presents the number of trips per month for all trips in Québec by gender. There is not a significant difference between men and women on making a trip, however, women are making slightly more LD trips than men.

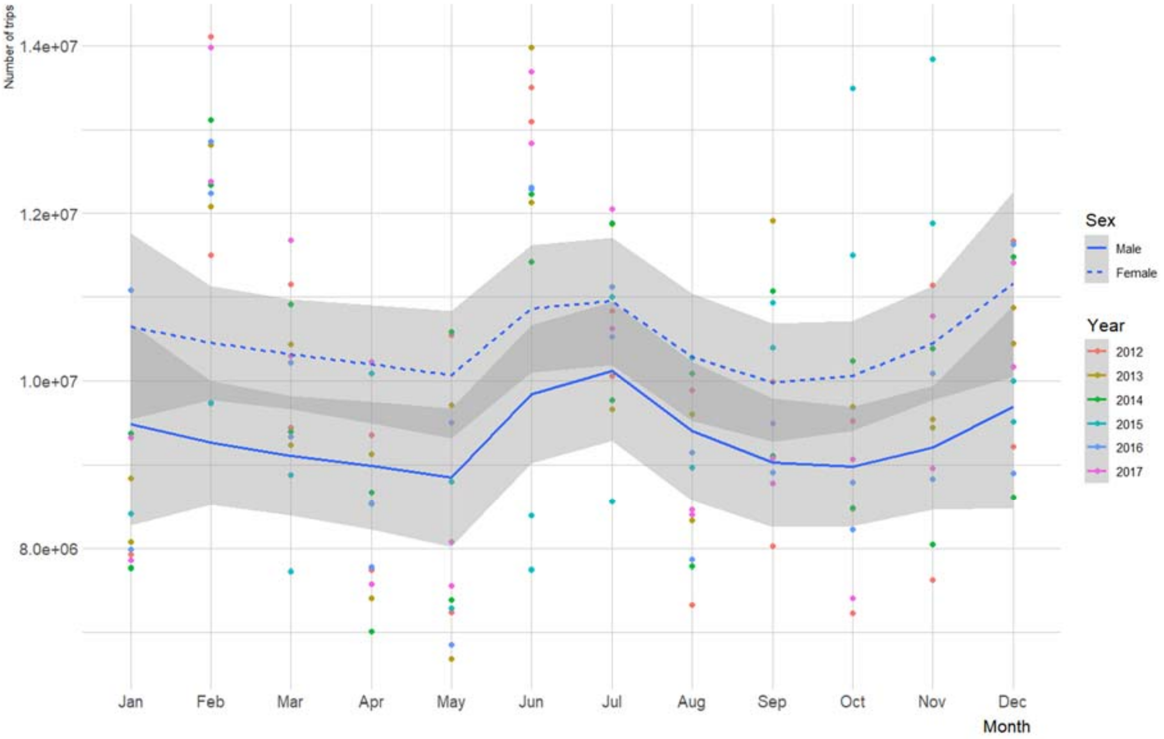


Figure 4.7 Number of weighted trips performed in Québec by gender

Figure 4.8 shows the number of trips for employed and unemployed people in Québec during the study period, a significant difference is observed between both groups, employed people make on average around three times more LD trips compared to unemployed people while their population is twice according to Table 4.2.

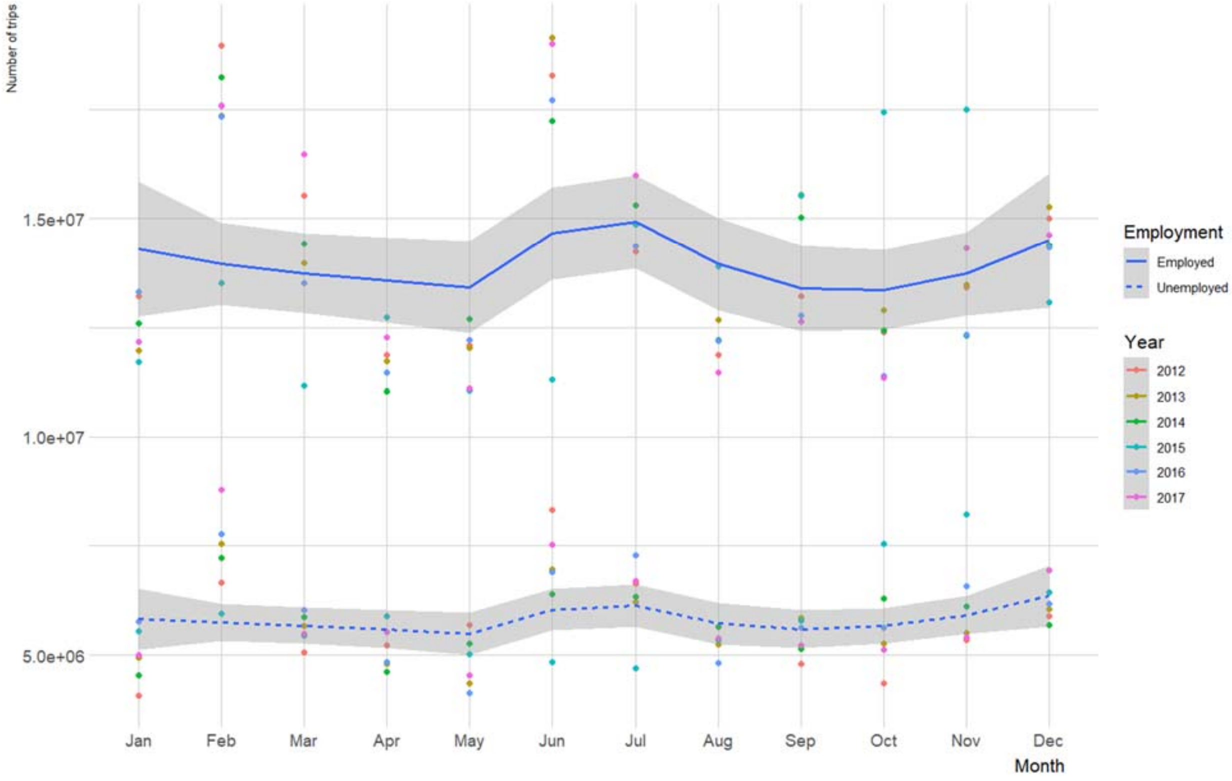


Figure 4.8 Number of weighted trips performed in Québec by labor force

The literature mentioned that income is the most important component influencing LD travel. Figure 4.9 presents the number of trips performed in Québec by different income groups. The highest number of trips is related to individuals with income between 50,000 $ to 70,000 $, and the lowest for people who did not reveal their income.

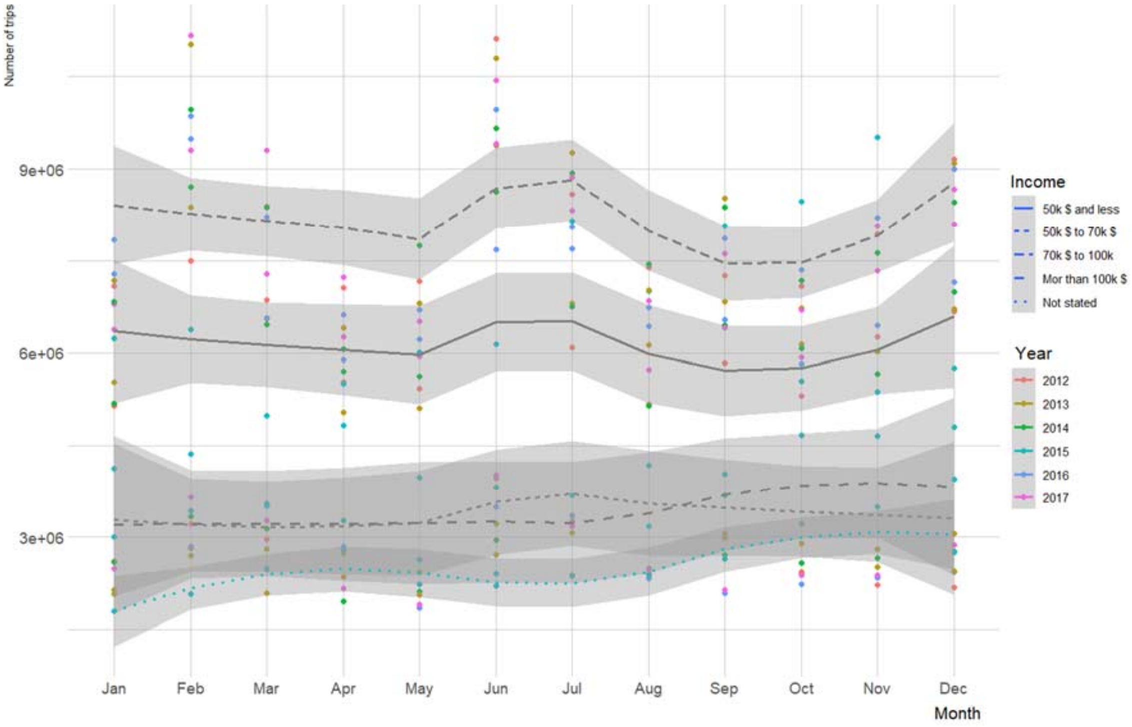


Figure 4.9 Number of weighted trips performed in Québec by income groups

Table 4.2 presents summary statistics of the number of observations in the TSRC dataset that is used to estimate the trip generation model in the following sections. The data and variables used in the model will be briefly described and their relations will be explored to assure that they support our key findings in the trip generation models. Some findings from Table 4.2 shows that most individuals (31%) of the sample have an income of less than 50,000 $ per year and the high-income group (more than 100,000 $ per year) has a share of around 27 % of the population. In terms of gender, not a big difference is seen, the data set includes around 49 % of men and 51 % of women. Employment is another important variable which has a significant role in making a long-distance trip specifically in trips with purpose of business. 64% of individuals are employed while around 36% of individuals are unemployed in the data set.

Table 4.2 Summary Statistics of the observation in the survey from 2012 to 2017

Variable	Category	Nb. Of observations	Percent	Percent (weighted)
Income	(-50k)	331,996	43.00%	31.03%
	(50k-70k)	118,441	15.30%	13.54%
	(70k-100k)	125,306	16.20%	15.83%
	(+100k)	178,761	23.20%	26.94%
	Not stated	17,471	2.30%	12.66%
Gender	Male	391,175	45.30%	49.20%
	Female	471,473	54.70%	50.80%
Employment	Employed	504,068	58.40%	63.68%
	Not employed	358,580	41.60%	36.32%
Education	Less than high school	143,190	16.60%	13.48%
	high school diploma	534,558	62.00%	60.75%
	University degree or more	184,900	21.40%	25.77%
Kids	No child	631,801	73.20%	69.26%
	Have child	230,847	26.80%	30.74%
Age group	18-24	56,920	6.60%	11.39%
	25-34	126,240	14.60%	17.56%
	35-44	137,700	16.00%	16.63%
	45-54	151,784	17.60%	18.32%
	55-64	170,415	19.80%	16.80%
	More than 65	219,589	25.50%	19.30%

Another variable used in the model is education which includes individuals who are educated less than high school, have a high school degree and a college or university degree or more with a share of around 17%, 62% and 21% respectively. It is expected that age plays a significant role in making a long-distance trip especially in this dataset since most trips are with purpose of leisure and visit family and friends. It is more likely that people at the young age group make more trips for leisure or to visit elderly people. People aged 65 years and older have the highest share in the sample with around 25% and except for the youngest group (18-24 years old) with around 6%, the other groups almost have an even portion in the sample as it is shown in Table 4.2.

To go further in details, it is interesting to know the role different variables have on making a LD trip. For example, if the share of employed people is higher than unemployed ones, is there any relation in making a trip and employment? To document this issue, it is important to find what percentage of employed and unemployed people are making a trip during the same period. Since the trip rate is constant during 2012 to 2017 but not in different months, data are aggregated based on months of the year.

Figure 4.10 characterizes the percentage of employed and unemployed individuals in terms of making at least one long-distance trip. A quick finding from the graph shows that more employed individuals are making a trip during all the months, specifically in July and August almost half of employed people are going to a trip while around 25% percent of unemployed people are making a trip during the same time. Nonetheless, in the months other than July and August this difference is not as significant as these two months.

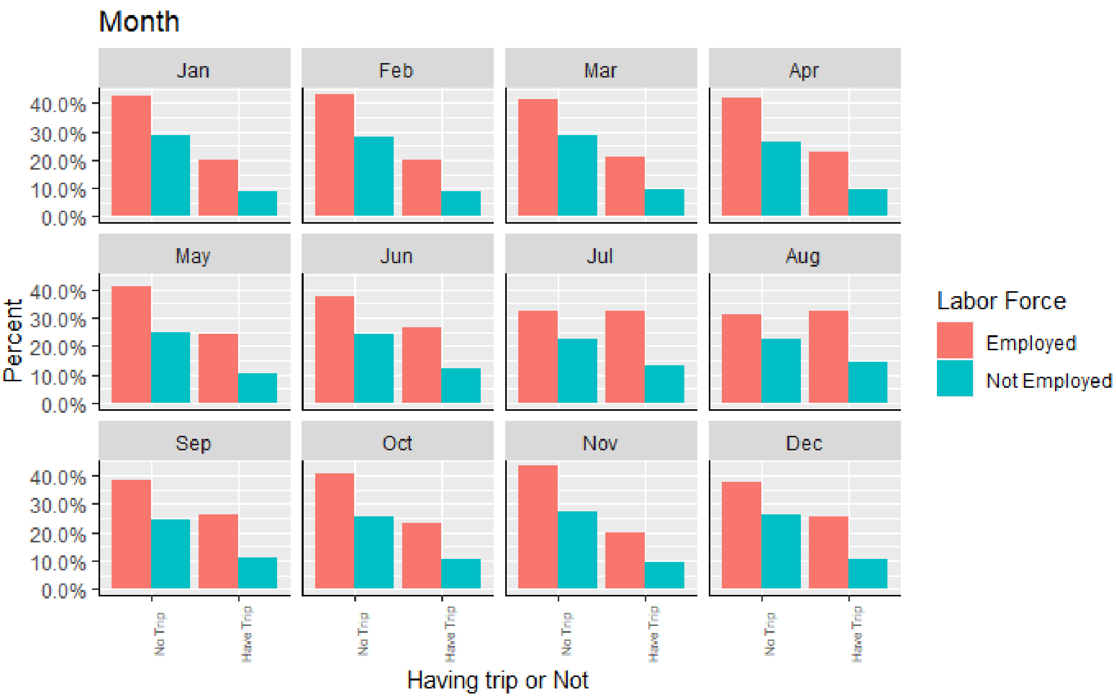


Figure 4.10. Proportion of people by employment status in terms of making a trip for 2012 to 2017 (Weighted)

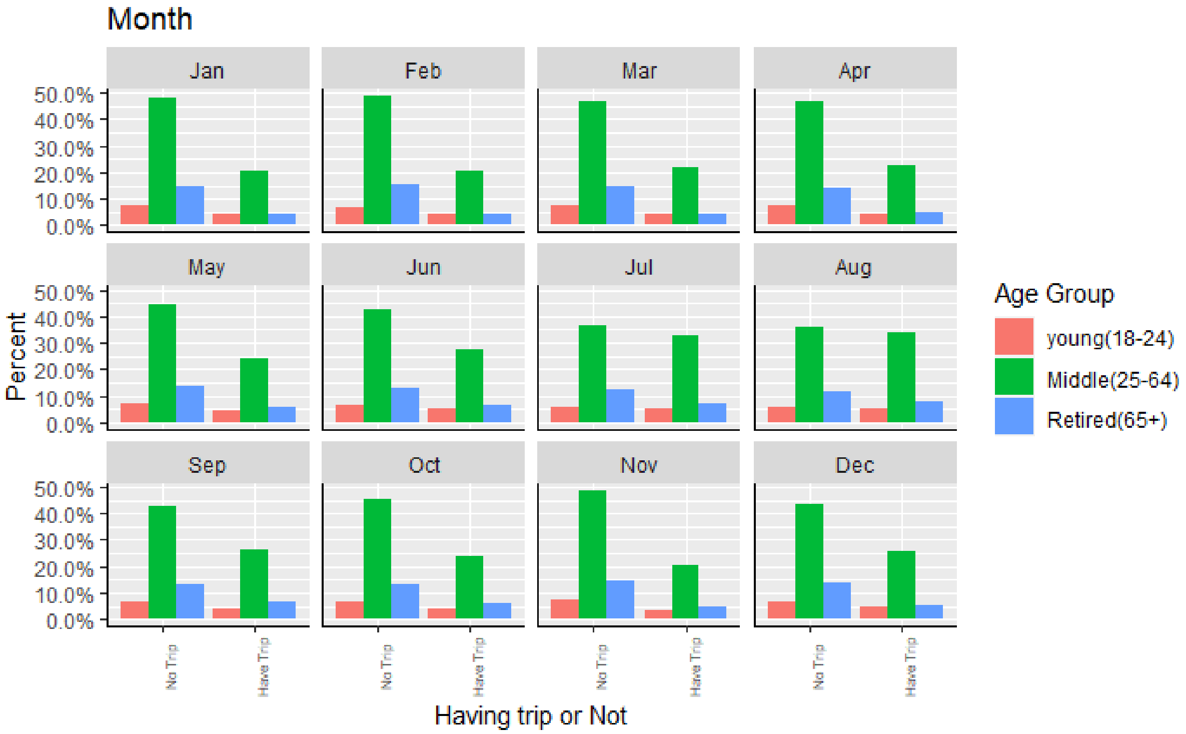


Figure 4.11. Proportion of people with different age groups in terms of making a trip for 2012 to 2017 (Weighted)

Regarding age, to have better visualization, the six different groups are aggregated to three of young, middle age and retired groups. However, as it is shown in Figure 4.11, the young age group has the lowest share in the dataset, but the proportion of young people who make a trip is higher than other groups. In July and August almost half of both young and middle age group are going to a trip while for retired groups, there is not a big difference between these two months and other months of the year.

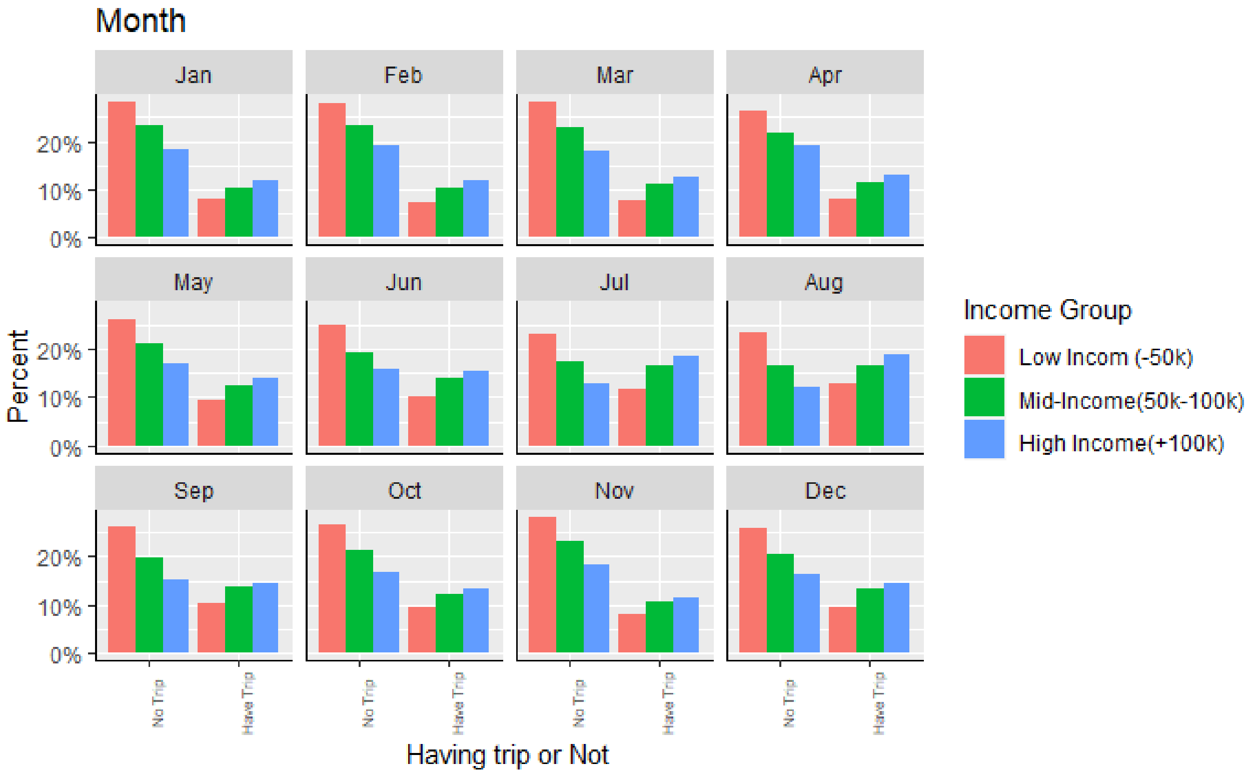


Figure 4.12. Proportion of people with different income levels in terms of making a trip for 2012 to 2017 (Weighted)

In the literature the most important variable in trip generation modelling for long-distance trips, is income. Higher-income individuals are more likely to go to a trip. Here also to have better visualization, the income groups are aggregated to three different classes: low income, mid-income, and high-income group. A key finding here also supports the idea that higher income results in opportunity to make a long-distance trip. As it is shown in Figure 4.12, during the most potential months in terms of long-distance trip (July and August), around 50% of mid-income and high-income group are conducting a long-distance trip while even in these two months the situation is not the same for low-income group. The low-income group has the least proportion of trip making in all the months while high-income group has the highest proportion of individuals who go to a long-distance trip. Surprisingly, in July and August for high-income group, percentage of individuals who are making a trip are higher than those who do not make a trip.

The mid-income group has an expected behavior, more percentage of trip makers in high-season months and fewer trip makers in the regular months.

Capturing a long-distance trip is not as easy as daily urban trip since the recall period is very important. On the other hand, as a long-distance trip is not a usual trip, even if they remember what they have done in the questionnaire period, did they have any long-distance trip or not? So, a long-distance trip can be a rare thing to happen to individuals.

4.2 Corridor analysis

In this section, we present the results for the first contribution of the thesis. This chapter started with identification of the cities that need to be considered as origin and destination of corridors. In this part, the cities with a population of more than 15,000 are selected for study. Also, three categories based on the population of cities were defined as core, mid-size and small cities. Cities with a population of more than 100,000 are considered as core and between 50,000 to 100,000 as mid-size and less than 50,000 considered as small cities.

Table 4.3 Categories of cities in the study based on the population

Core Cities		
1	Name	Population (2016)
2	Montréal	4,098,927
3	Ottawa - Gatineau	1,323,783
4	Québec	800,296
5	Sherbrooke	212,105
6	Saguenay	160,980
7	Trois-Rivières	156,042
Mid-size cities		
8	Drummondville	96,118
9	Granby	85,056
10	Saint-Hyacinthe	59,614
11	Rimouski	55,349
12	Shawinigan	54,181
Small cities		
13	Joliette	49,439
14	Victoriaville	49,151
15	Rouyn-Noranda	42,334
16	Sorel-Tracy	41,629
17	Salaberry-de-Valleyfield	40,745
18	Val-d'Or	33,871
19	Alma	32,849
20	Saint-Georges	32,513
21	Rivière-du-Loup	28,902
22	Sept-Îles	28,534
23	Thetford Mines	28,448
24	Baie-Comeau	27,692
25	Matane	17,926
26	Campbellton	15,746
27	Dolbeau-Mistassini	15,673
28	Gaspe	14,568

Table 4.3 expresses different categories of cities in the study. The city of Gaspé was added to the study manually since this city is a touristic city and expected to play a significant role in the intercity travel demand.

4.2.1 Defining the desire lines

Figure 4.13 illustrates the desire lines between selected cities. This figure includes 702 OD pairs and 351 desire lines. As already discussed in the methodology section, desire lines have been used for evaluation of demand between cities to avoid imposing any barriers for defining the corridor network.

Since there is no survey of demand between these cities, it was decided to use the population, the distance, the location of the city with regards to the Saint-Laurent River and passing a desire line from other cities as boundaries to find main corridors. In the desire line section, the first two variables are used to define the desire line and the two latter variables are used to move from a desire line to main corridors.

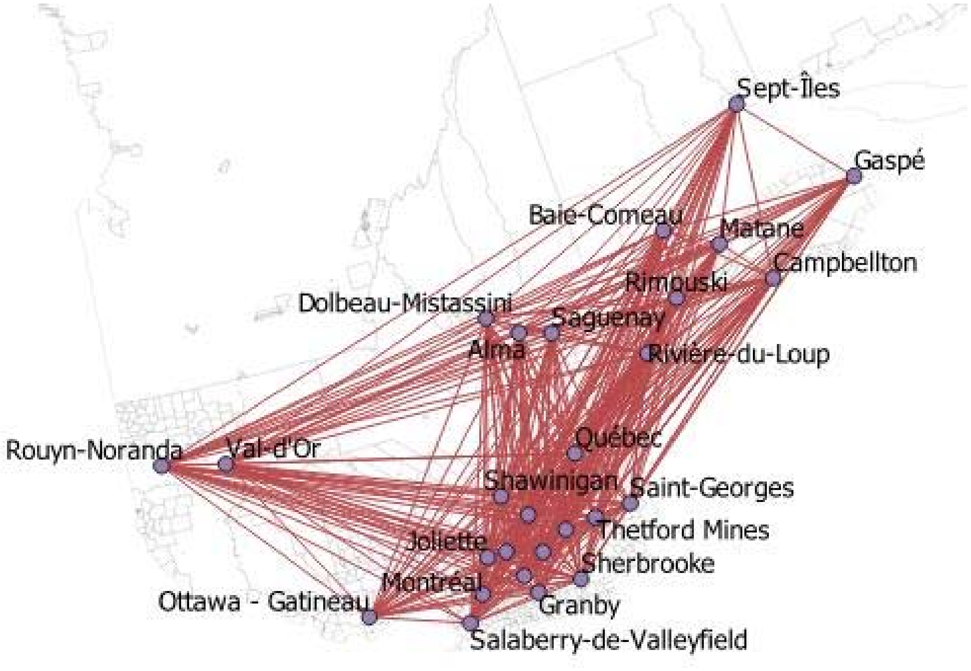


Figure 4.13 Desire line between OD pairs

In the first step, core cities were connected to each other, as it is shown in Figure 4.14. Main cities include seven cities with a population of more than 100,000. In this step, eight desire lines are passing the filters of the population and the direct distance of less than 300 km.

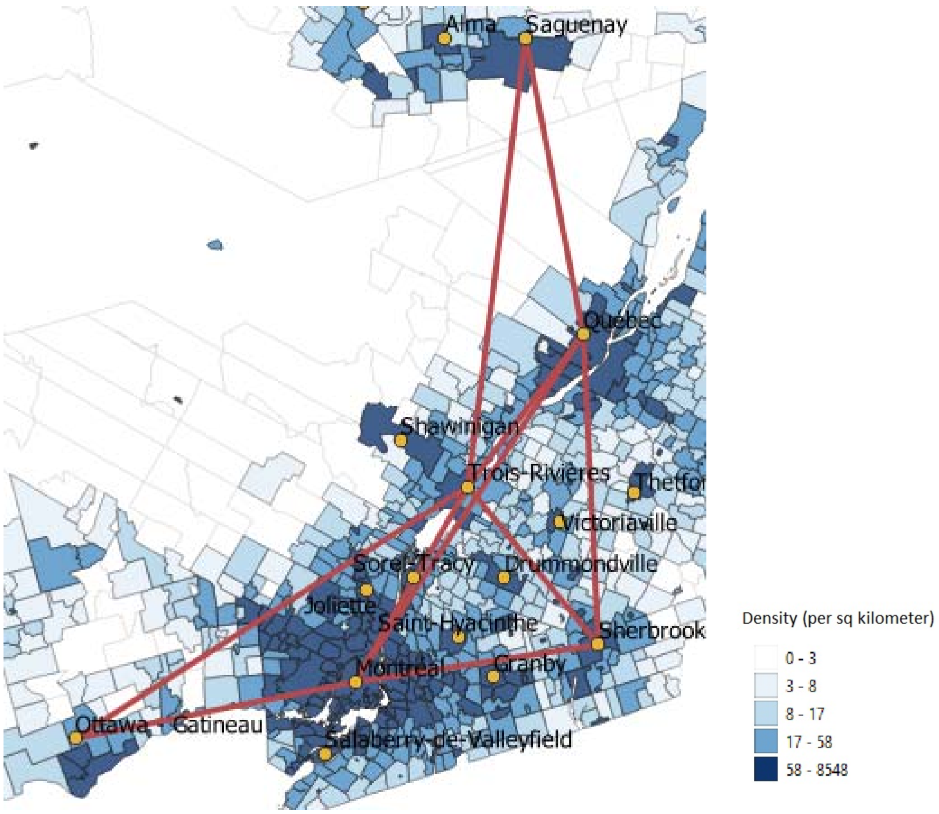


Figure 4.14 Desire line between core cities with population and distance boundaries

At a glance, in the figure, several points are noticed, several lines are passing from other core cities to connect two core cities, for example to have connection between Montréal and Québec the line is passing throughout Trois-Rivières, and it can be considered as a duplicate desire line. Also, in some cases, the desire line is passing throughout a very low-density area like Saguenay to Trois-Rivères, in these cases, the OD pair can be considered using a connection node, in this example the city of Québec can be the connection node.

In some parts, however, the line is not passing throughout any core city, but it is passing through a mid-size or small city which those cities can play a connection role. These points helped to move forward from the desire line to define corridors.

The next step is to connect core cities to mid-size cities. In this part, two main filters of distance and population are set. A distance of less than 100 km is chosen, and, regarding the population, the city of origin is set to cities with more than 100,000 and destination cities with populations between 50,000 and 100,000 people.

Figure 4.15 illustrates the desire lines between these two categories, as it can be deducted from the figure, there are many desire lines in this class that are passing through other cities.

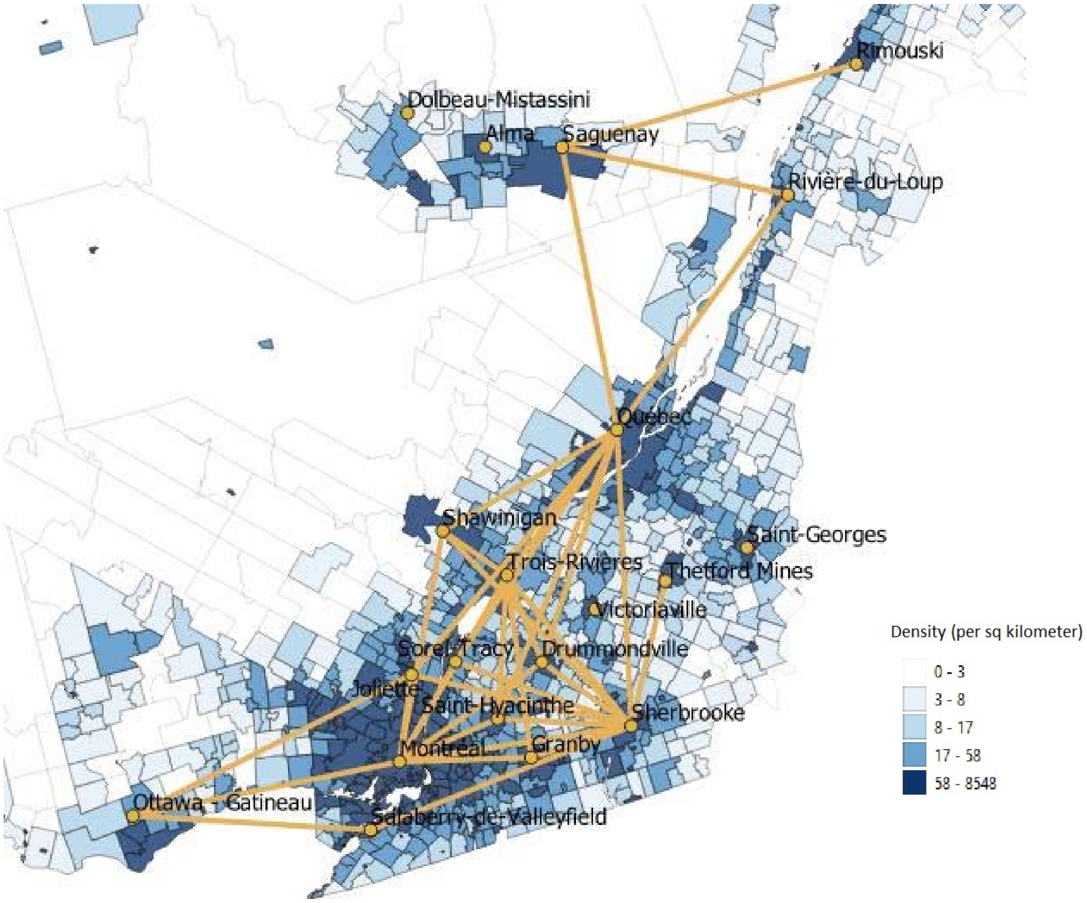


Figure 4.15 Defining of desire line between core to mid-size cities

On the other hand, there are many lines passing through rivers to connect cities (mainly the St. Lawrence). Also, some lines are passing through more than two CAs. For instance, Sherbrooke to Shawinigan is passing through Drummondville and Trois-Rivières.

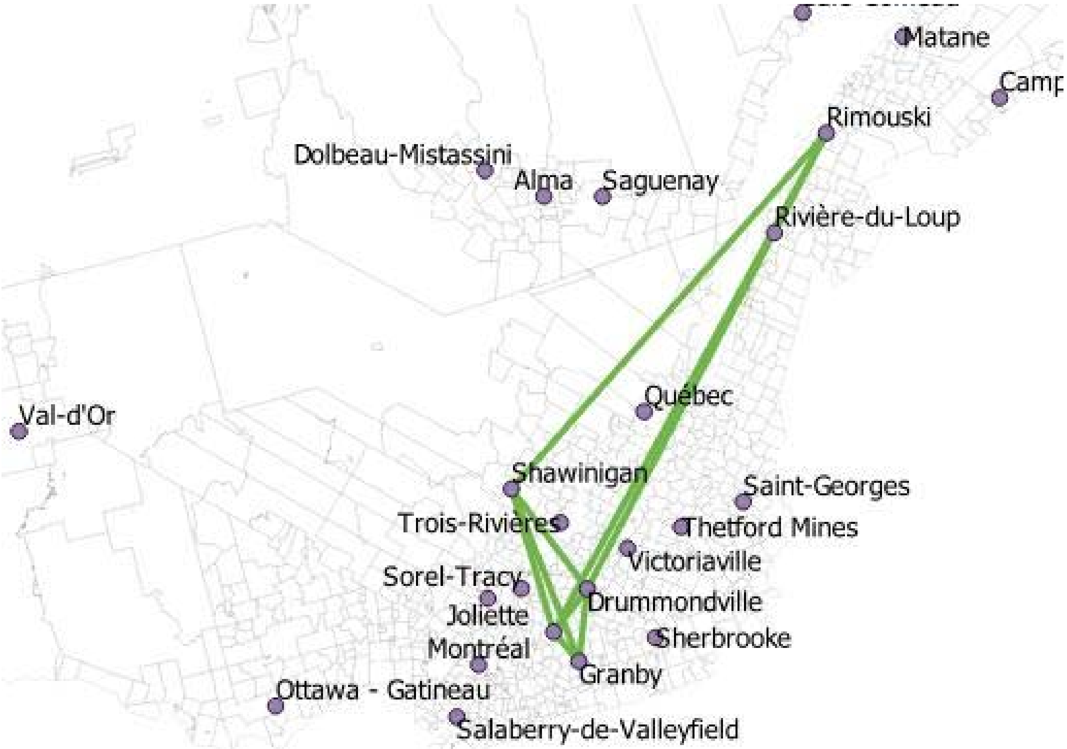


Figure 4.16 Defining of desire line between mid-size to mid-size cities

To connect the mid-size cities to each other, the direct distance was set to less than 350 km and population of origin and destination set to between 50,000 and 100,000 people. The results of mid-size to mid-size cities are shown in Figure 4.16. As it can be seen in the figure, only seven desire lines are passing the filters. Here also, lines are passing throughout other cities and some lines are passing the rivers to connect to cities that are far from each other. Rimouski to Saint-Hyacinthe is almost the same as Rimouski to Drummondville to Saint-Hyacinthe.

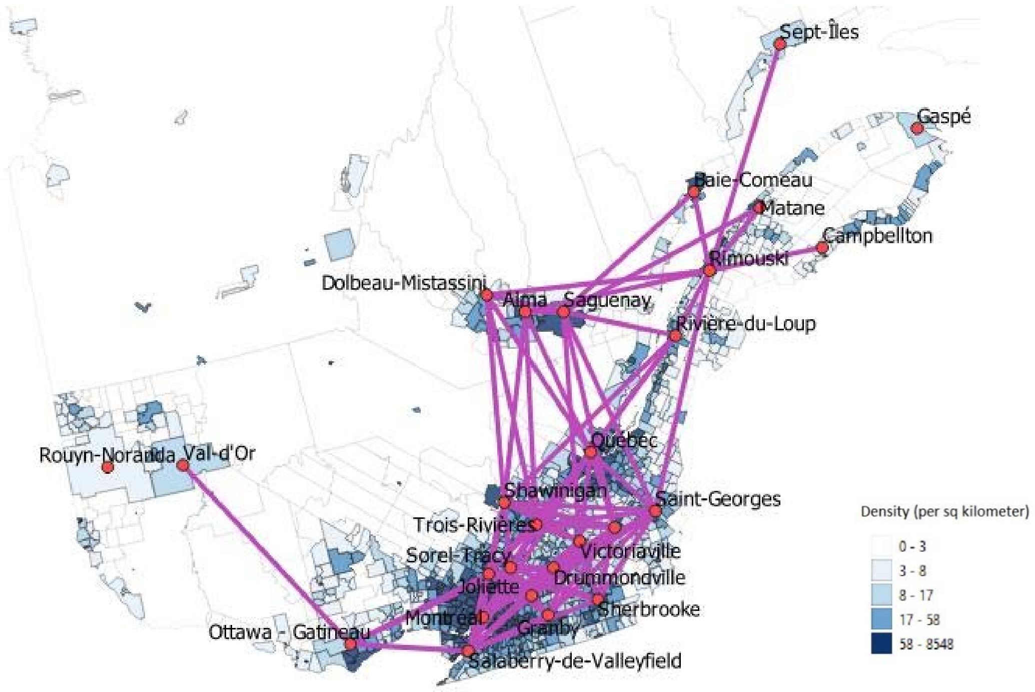


Figure 4.17 Defining of desire line between small to mid-size and core cities

In the next step, small cities were connected to mid-size or core cities. In this section, the population of origin is less than 50,000 and the population of the destination is more than 50,000. As of distance, since several small cities are far from others, the distance was set to less than 300 km.

It is noticed that there are small cities with no connection to any cities like Gaspé or Rouyn-Noranda, so it was important to keep connections of the type small-to-small cities. Here, the population of origin and destination is cities with less than 50,000 and the direct distance of less than 450 km. The distance set to 450 km to connect Val-d'Or to Dolbeau-Mistassini. Figure 4.18 illustrates the desire lines between small-to-small cities.

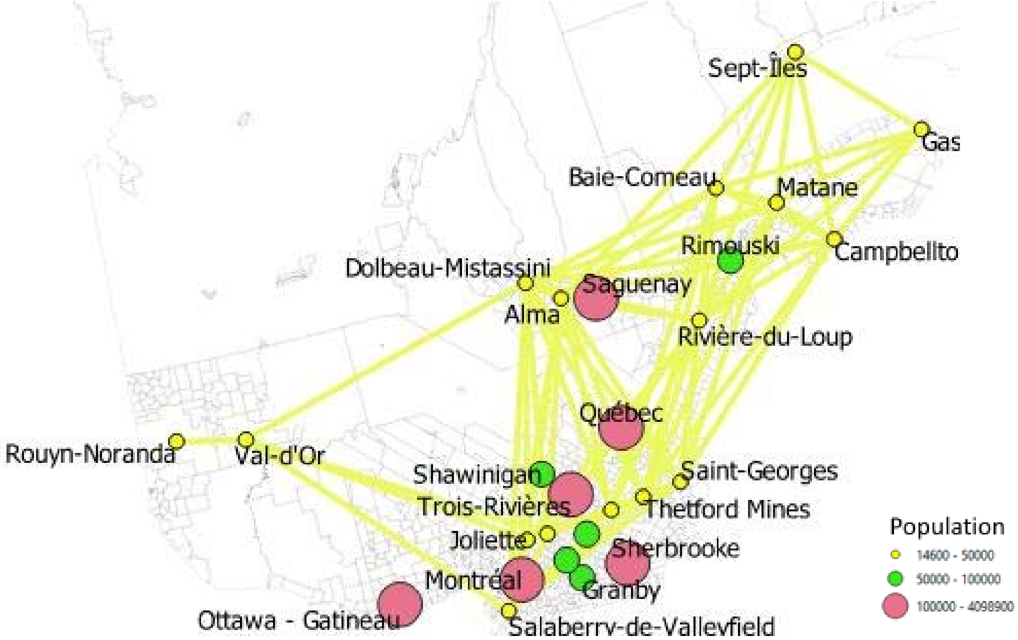


Figure 4.18 Defining of desire line between small-to-small cities

These five steps draw a preliminary concept of desire lines between the main cities in the study. By aggregation, the final map of desire lines is shown in Figure 4.19. This figure includes 122 OD pairs out of a total of 351 OD pairs. The key point mentioned above needs to be implemented to move from desire lines to the main corridors in the next section.

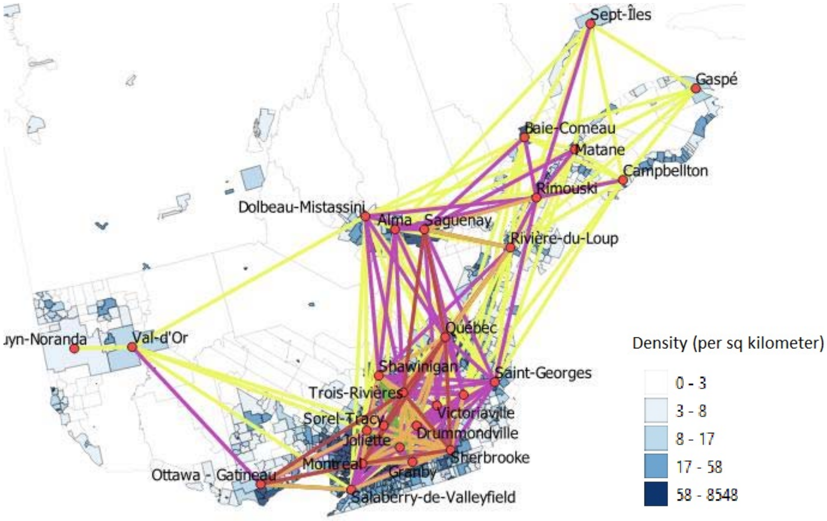


Figure 4.19 Final desire lines

4.2.2 Desire lines to main corridors

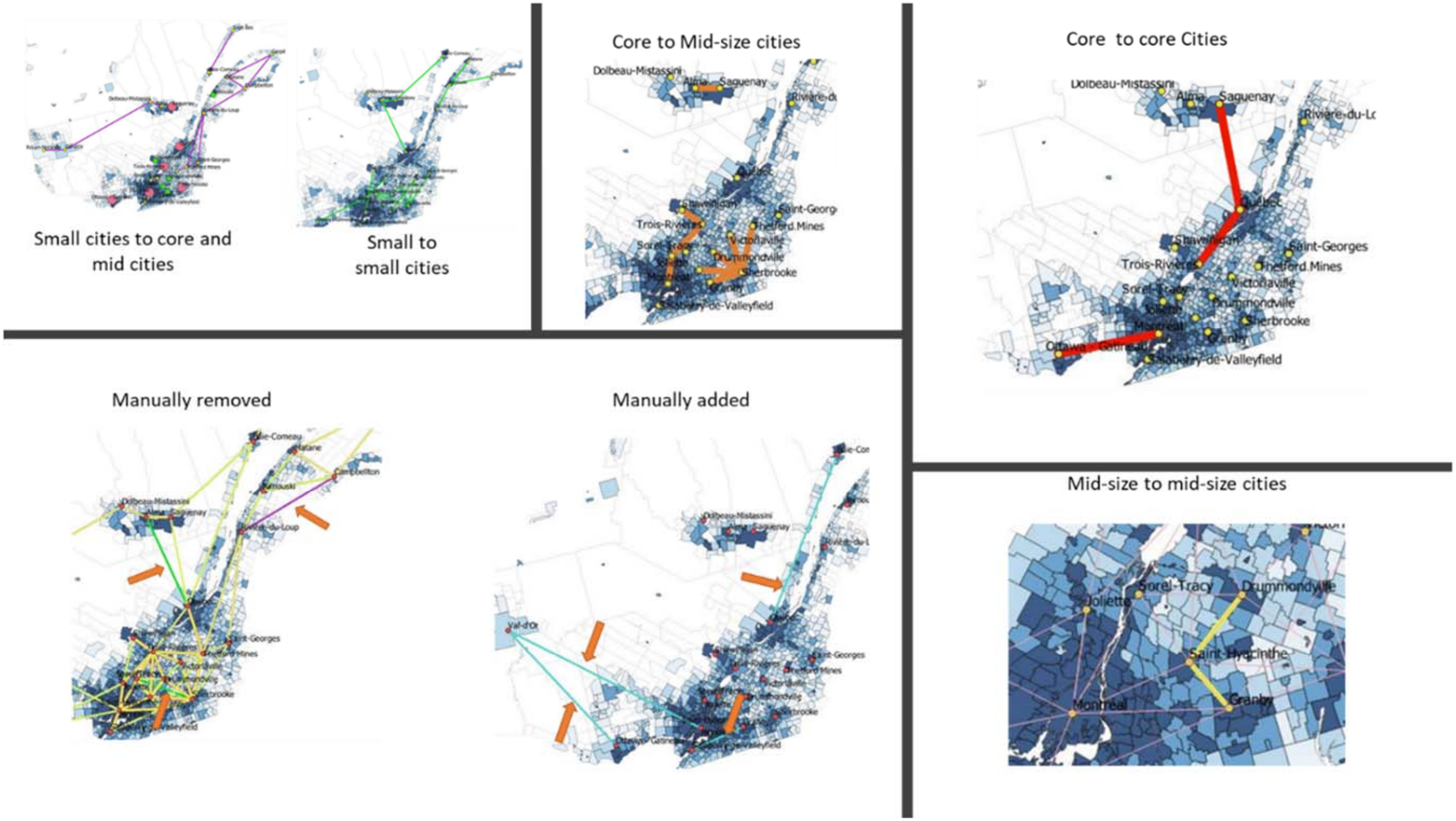
To achieve final corridors, some restrictions need to be tighter, and also new restriction defined. For connection of core-to-core cities, the distance of reduced to 250 km since there are connections between cities further than 250 km. Also, OD lines which pass over more than two CAs were excluded because those cities can play a connection role. In this step, being in the same side of a river was not considered, because core cities are important and there should be a high demand between two core cities.

In all other steps, two mains filter are used to avoid unnecessary OD lines. The first restriction is the passing over other cities: if a line is passing over more than two cities in the study, the line will be excluded. The next limitation is the location of the city in regard to the rivers. In this step, only the OD lines in the same side of rivers were considered to be in the main corridor and if a line was passing through the river and that line is not connecting two core cities, the line was removed.

After implementing the new restriction, it was found that there is not enough connection between two different sides of the river, so we implement a new step that considers the cities in different sides of the river.

In this step, those cities which are in different sides of the rivers and their distance are less than 100 km regardless of population were connected. In the last phase, it was found that there are some lines which are passing throughout a very low-density area or there is a line which can be combined by another OD line, these lines were removed manually. On the other hand, by taking a look on the final corridor and comparing with highway network of Québec, in some points there are some cities which need to be connected and they are excluded during the filtration process.

Figure 4.20 shows the process of defining the main corridors step by step, the manually removed pairs include: Rivière-du-Loup to Campbelltown, Québec to Alma and Sherbrooke to Sorel-Tracy and the manually added OD lines are Québec to Baie-Comeau, Val-d’Or to Montréal, Val-d'Or to Gatineau, and Granby to Salaberry-de-Valleyfield.



Small cities to core and mid cities

Small to small cities

Core to Mid-size cities

Core to core Cities

Manually removed

Manually added

Mid-size to mid-size cities

Figure 4.20 Steps taken from desire line to the main corridors

Figure 4.21 and Figure 4.22 illustrate the preliminary map of the final corridors and the current bus network in Québec. The final corridors consist of 50 OD pairs connecting 28 nodes (cities) to each other. As it is shown in the map, the defined corridors cover all the bus network within the area in the study.

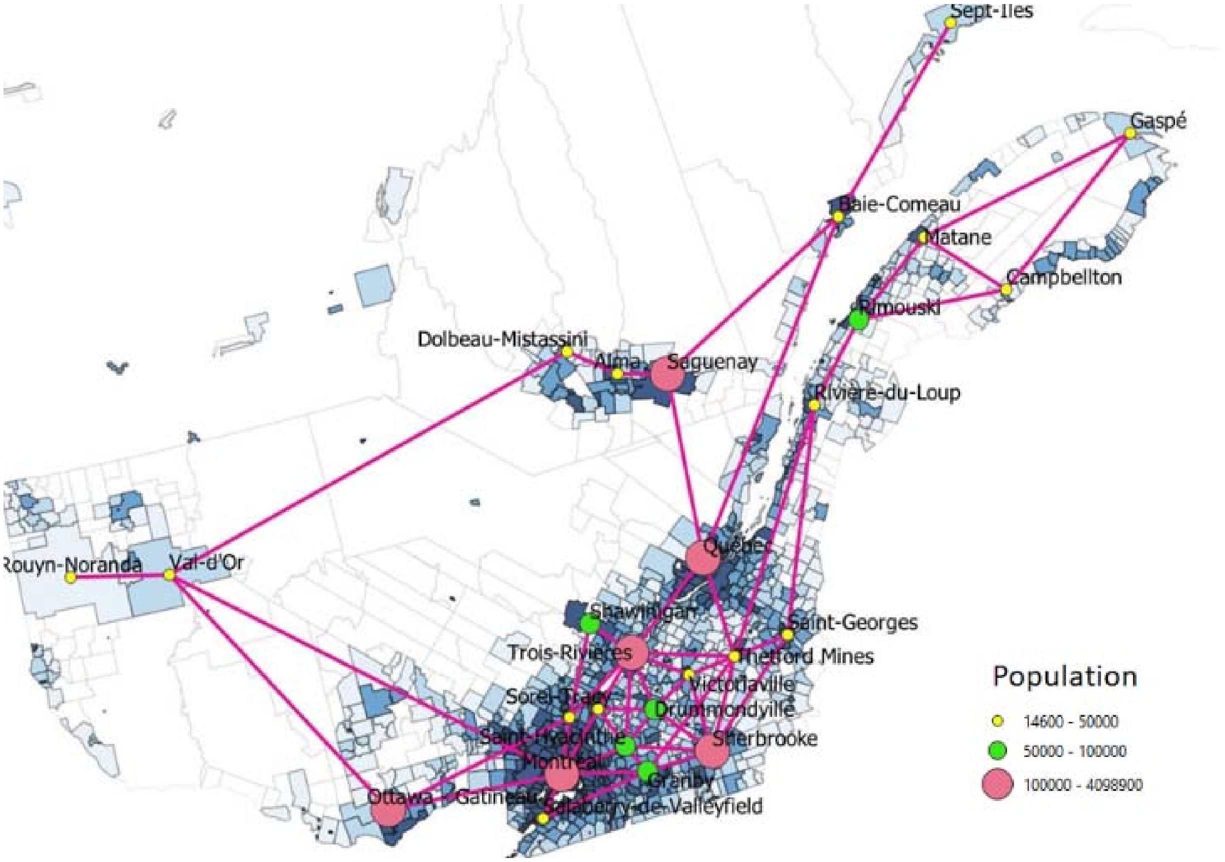
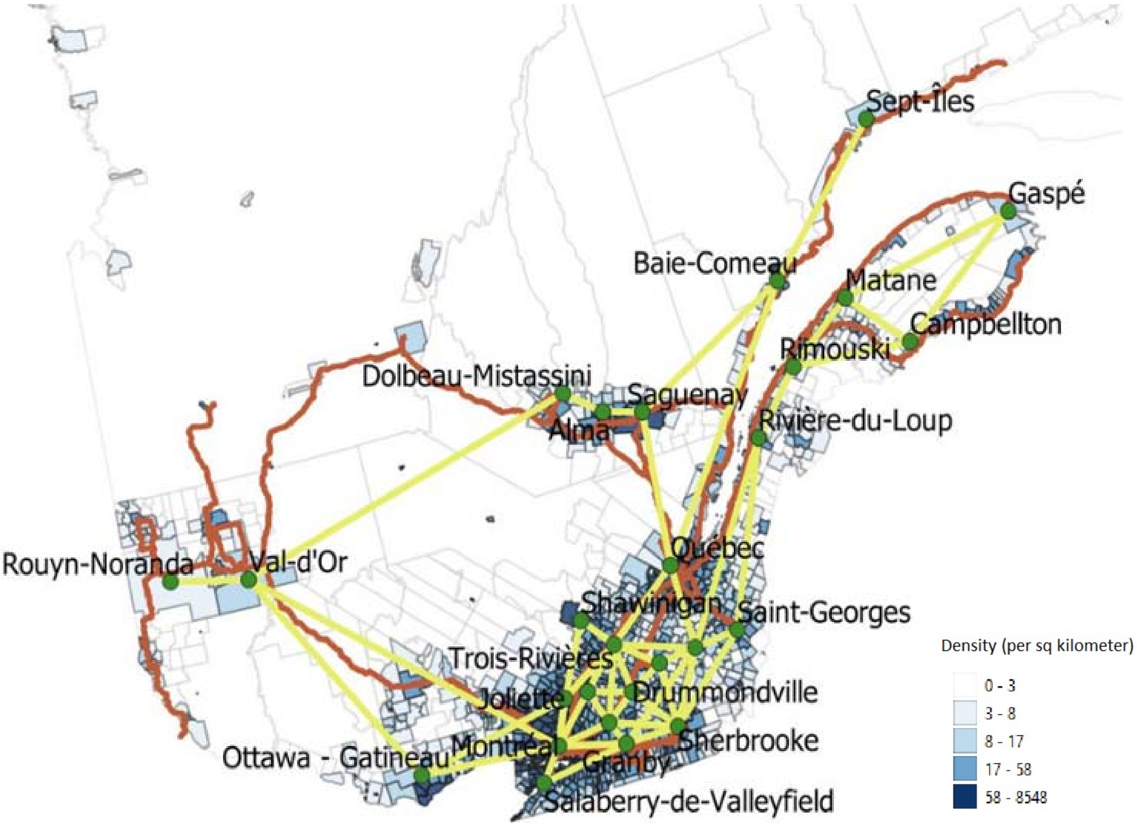


Figure 4.21 Final corridor

Figure 4.22 Final corridor and current bus network



To rank corridors, two different methods with several variables were employed. Figure 4.23 represents the map of ranked corridors using the sum of ranks of variables. In this method, the variables are primarily related to origin and destination, so OD routes with one or two ends to larger cities have a higher rank. The highest rank in a corridor demonstrates the highest potential demand in that corridor.

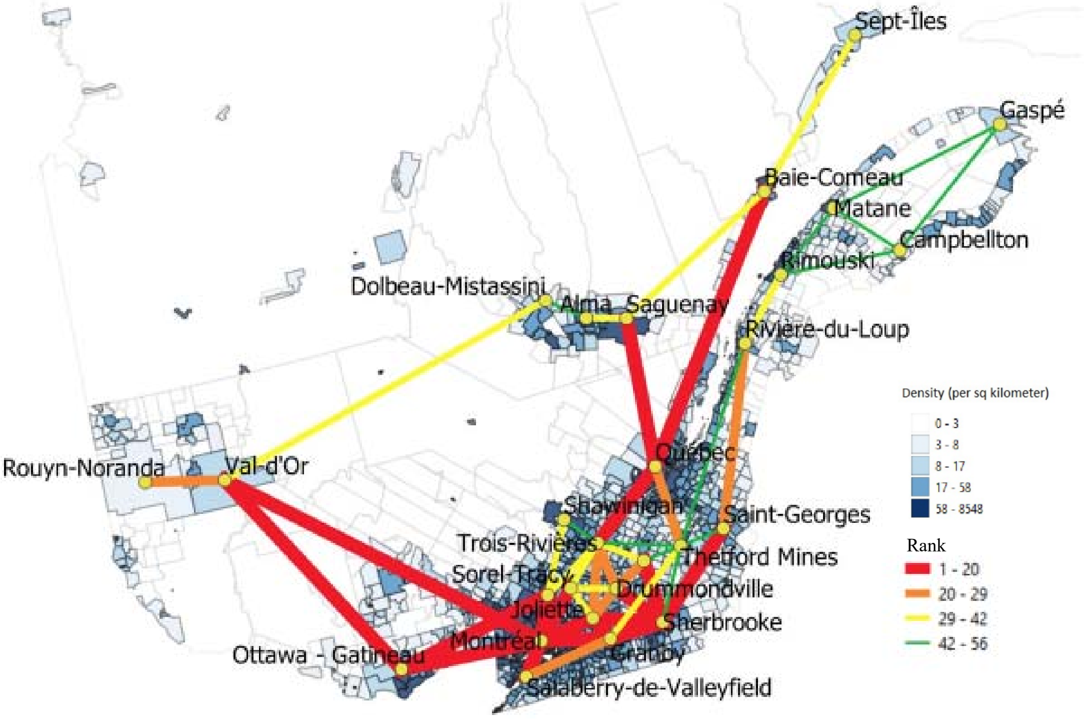


Figure 4.23 Ranking corridors using sum of rank of variables

In addition to a ranking method based on origin and destination, we must also take into account the areas along corridors. These areas may not be very important for travel with private vehicles, but at the same time they may be highly important for public transportation services. The PT companies need to propose the most optimized services to the passengers which is the fastest, safest and serving as many people as possible. So, the areas along corridors need to be considered for assessment of demand.

Figure 4.24 illustrates the map of the corridors based on rankings of the average of rank of variables. In this map, the OD pairs passing over denser areas have highest ranks. In general, the routes passing along low-dense areas have the lowest ranks, while the lines passing over the high-density area are potentially more important in terms of travel demand.

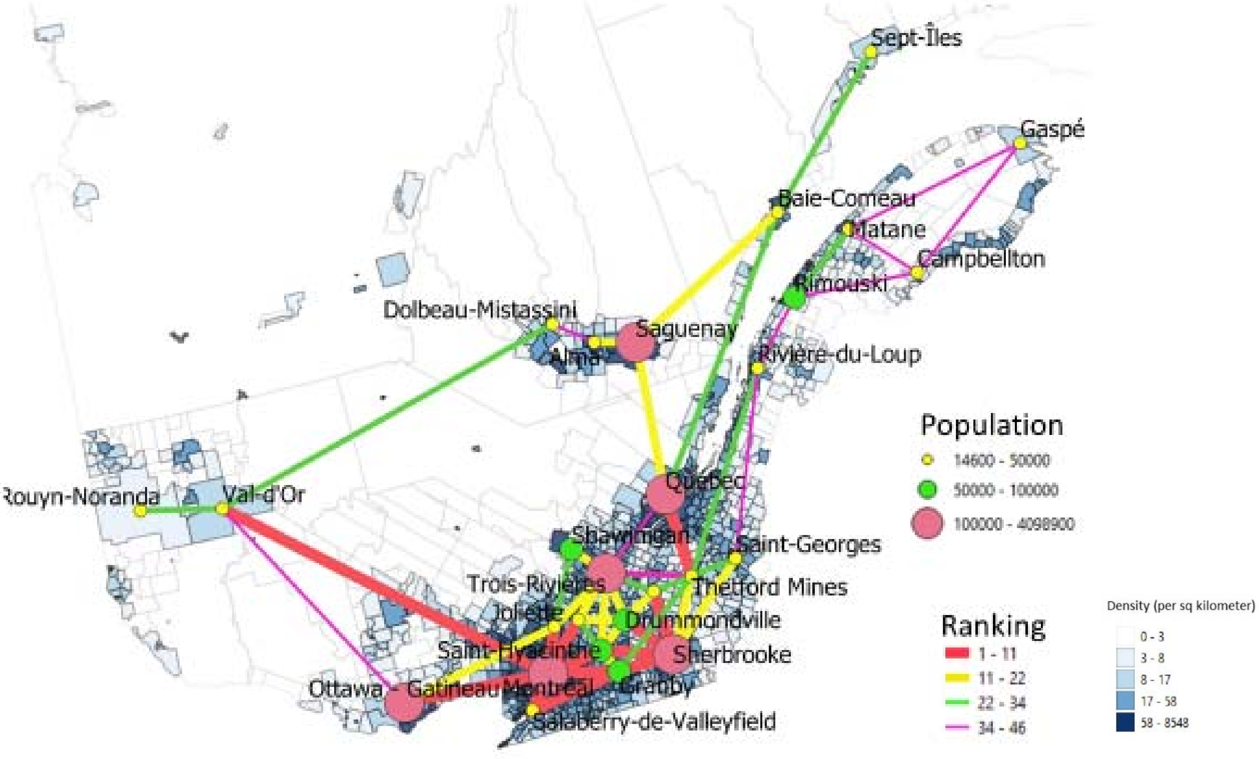


Figure 4.24 Ranking corridor using average of rank of variables

By comparing these two methods, it is found how important could be the areas along the OD pairs. So, it is vital to consider the aim of this study and the modes of transportation. When the focus is on the public transportation mode, the areas along the corridors become highly important, while, when the motivation of study is around the private vehicle, the importance of origin and destination needs to be considered firstly.

4.3 Model result

This section presents the results of the second contribution of this thesis.

4.3.1 Descriptive analysis

This study covers intercity or long-distance (LD) trips all over Canada. This country is composed of ten provinces and three territories, covering over 9 million square kilometres and a population of over 35 million people, according to the 2016 census. In this study, the intercity trip generation model considers all non-frequent day trips and overnight trips as reported in the Travel Survey for Residents in Canada (TSRC). The TSRC survey is designed with the purpose of supporting domestic and international tourism studies. The TSRC data collection is performed by phone for domestic trips and in-person for international travel; daily commute LD trips are excluded from the survey. The samples of the TSRC surveys are 88813, 75753, 74391, 67138, 65225, and 58361 people, respectively, for the 2012 to 2017 surveys.

Table 4.4 Demographic variable in the TSRC survey

Variable	Description	Coding of Input Value	Abbreviation (in the model)	Variable
Respondent gender	Male	1	SEX	Categorical
	Female	2		
Respondent educational level	Less than high school	1	EDLEVGR	Categorical
	High school certificate	2		
	Some post-secondary diploma	3		
	University degree	4		
Respondent age	18-24	1	AGE_GR2	Categorical
	24-34	2		
	35-44	3		
	45-54	4		
	55-64	5		
	65+	6		
Household size	Number of children	N/A	G_KIDS	Categorical
Income (CAD)	Less than 50k	1	INCOMGR2	Categorical
	50k to 70k	2		
	70k to 100k	3		
	100k and over	4		
	N/A	5		
Respondent employment	Employed	1	LFSSTATG	Categorical
	Unemployed	2		
CMA Level		1-34	RESCMA2	Ordinal

In the TSRC survey, all the respondent characteristics are categorical variables. In this study, for model estimation, all the variables are transformed into ordinal variables. To include the home location of the respondent in the model, the population of the census metropolitan area (CMA) of residence is included. The geography includes 33 census areas (CA) and CMA and one level representing the rest of Canada (all that is not included in any CA or CMA).

Figure 4.25 presents the share of people who made at least one LD trip during each month and year as the share of those who did not. This graph shows the observation of the survey (without any weighting factor). In the summer, where the LD trip rate is higher than for the rest of the year, only some 35% of respondents made at least one LD trip, and this share is below 20% for the other months.

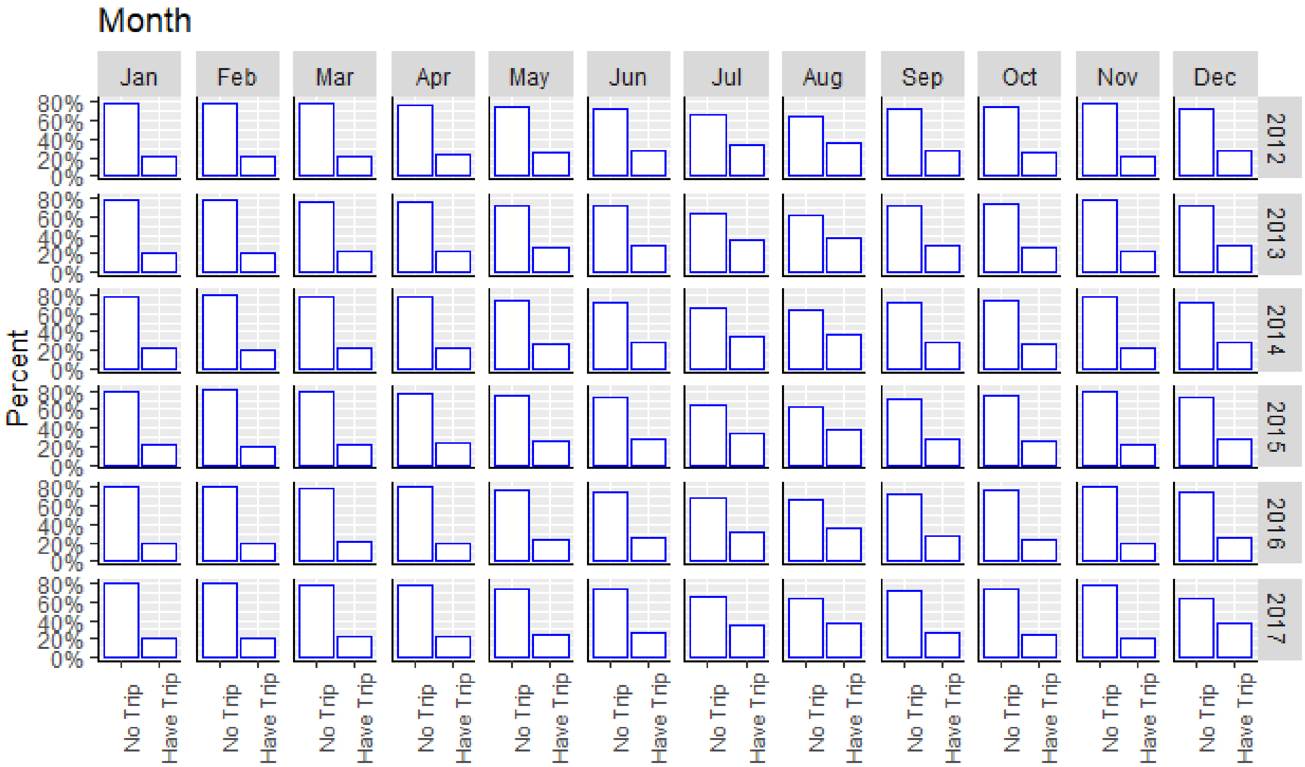


Figure 4.25 Proportion of people who did at least one LD trip during the month (event) and of those who did not (non-event) during the study period

4.3.2 Model results and comparison

In this study, four machine learning techniques are implemented to estimate trip generation models, and the performance of these models is compared. These methods are widely used for rare events or imbalanced data modelling in the literature (Ma, Lukas, 2021; Zhou, 2020). CART decision tree algorithm, CTree decision tree algorithm, Random Forest, and a generalized linear model are used.

For all techniques, 75% of the dataset is considered training, and the rest is used for validation. The dependent variable in the model is set as a binominal choice of whether an individual performs any long-distance trip during the study period or not, and socio-demographic features are used as independent variables. The "party" and "Rpart" package in R software is used for the estimation of DTs.

Several mentioned factors used to evaluate the performance of the model's estimation and the result of analysis are presented in the model's prediction ability.

To tackle the issue of imbalanced data, three different data balancing approaches are implemented before model estimation. Because the data set is imbalanced, the F-1 score is used to evaluate the performance of data preparation methods, as is shown in Table 4.5, the under-sampling method has the better result. So, this method was employed during the study for model performance evaluation.

Table 4.5 F-1 score for different methods of data preparation

Random Forest	Under-sampling	Oversampling	Synthetically oversampling
F1-score	0.457	0.413	0.426

Table 4.6 results demonstrate that Random Forest gives the highest performance in terms of research focus on positive class and CTree in terms of research interest on negative class. The overall accuracy is a factor stating the model's overall performance, which says that the CART model has the best performance overall but the least accuracy on prediction of the positive class. However, in the case of a rare event data set, other scores play an essential role in the model's prediction ability.

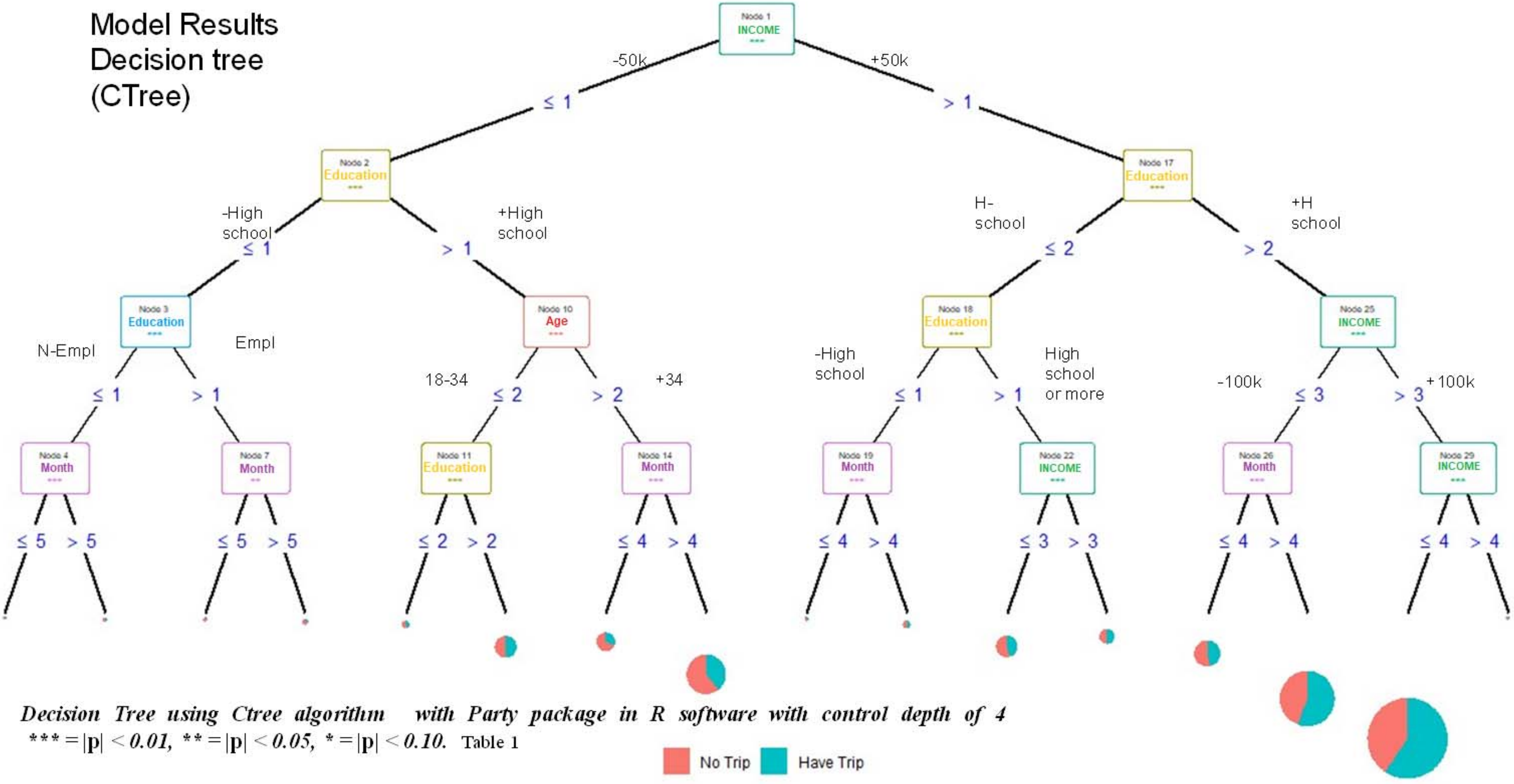


Figure 4.26 Decision Tree using CTree algorithm with Party package in R software with control depth of 4

Table 4.6 Model performance score

	CART	CTree	Random Forest	Logistic
Accuracy	0.642	0.612	0.588	0.597
Sensitivity	0.485	0.565	0.656	0.618
Specificity	0.704	0.630	0.563	0.542
Precision	0.392	0.375	0.351	0.358
F1-score	0.433	0.450	0.457	0.453

In this case study, since the positive class is a crucial event to predict the factors that affect people's LD trips, sensitivity plays a vital role in the determination of model performance. The Random Forest has the highest score through the models. Also, to avoid overestimating the positive class, F1-score is defined to ensure that the random forest model performs better than other models.

Figure 4.26 represents the CTree decision tree with a controlled depth of four levels. It shows that income has a high impact on the occurrence of a LD trip; also, educational level and the reference month are playing a vital role in LD travels. Higher academic level leads to more proportion of LD trips, and people tend to make LD trips in the summer because of vacation time and temperate weather conditions. Also, it shows that people with the lowest income and lower educational levels conduct significantly fewer trips. Interestingly, the results show that young people with the same income level make more LD trips if they have a higher level of education.

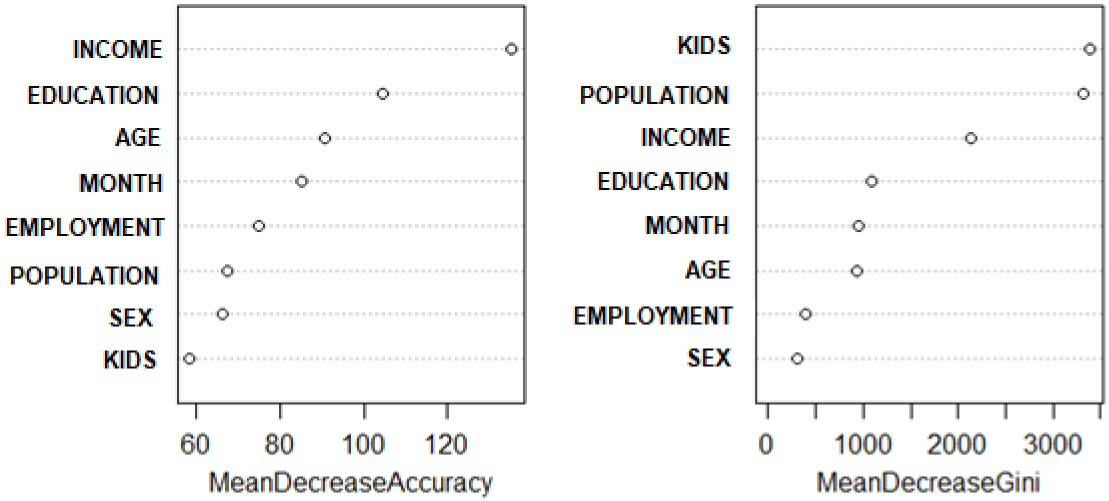


Figure 4.27 Variable importance using the Random Forest model

The interpretation of the variable importance in the random forest model can be made with two crucial graphs represented in

Figure 4.27. This outcome of the random forest model states how vital the variables are in classifying the data. In every tree of the random forest model, the prediction error of OOB data is recorded, and the prediction error of permuting each predictor variable is recorded as well. The average difference between these two errors is normalized by the standard deviation of the differences. This factor is called "MeanDecreaseAccuracy" (RColorBrewer, 2018).

In summary, this factor is a unitless factor that states how much the accuracy of the model depends on each variable. The variables are presented in descending order. The highest variable plays the essential role variable in the model. "MeanDecreaseGini" is the average of node impurities from splitting on a variable over all trees. In this case, a higher value results in a split with a purer node (RColorBrewer, 2018). In other words, a higher value for each variable states how much it contributes to the consistency of the nodes.

Interestingly, it was found here that income and educational levels are the most important variables in the model, which was also mentioned in the CTree model. Also, this was also found in the logit and CART models, which demonstrates that despite the accuracy of the models, all the methods agree on the contribution of the variables for LD trip generation. Having kids and the population of residence may not play a crucial role in the accuracy of the model, but they have a significant effect on having a purer node on splits.

The performance of the model also confirmed how necessary is the data collection procedure. Also, it confirmed the importance of improving survey methods with a focus on LD trip analysis. Better data collection helps to have more in-depth studies for LD travel and to find how individual characteristics affect their intention to make intercity trips.

4.4 Mode competitiveness

In this section, we present the results for the third contribution of the thesis.

4.4.1 Access time

The results from the combination of three different datasets show the competitiveness ratios of PT options vs car options. As stated earlier, the access component has a significant impact on mode competitiveness (based on total travel time) specifically when it comes to larger cities.

At first glance, the access component may not be considered an important factor in LD travel, since it may seem that an urban trip, compared to an intercity trip, is rather short. While it may be a short distance, it can still induce important travel times. For some shorter LD trips, the access component may even be as important as the station-to-station segment, hence doubling the total travel time. Figure 4.28 shows the relation between distance and travel time by car for access to bus (left graph) and train (right graph) station. The distance from the centroid of a census tract to the intercity station can be up to around 90 km. The longer access distances are related to the larger cities (Montréal, Québec City) which reveals the importance of the access time in travel by PT for intercity trips.

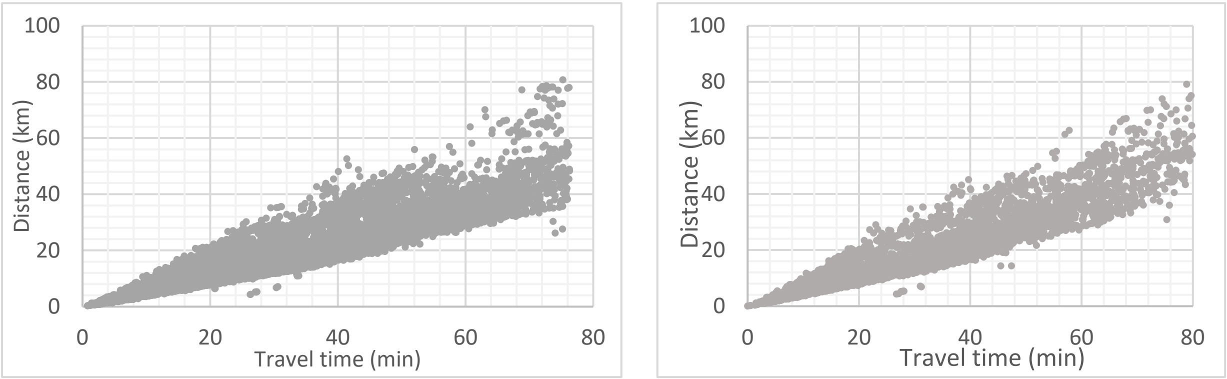


Figure 4.28 Relation of distance and travel time by car for access from the centroid of each CT to all bus (left) and train (right) stations in all CAs and CMAs.

As far as travel time is concerned, it can reach about 80 minutes; this number is even higher than the travel time from Montréal to Granby. Hence, this issue becomes even more intense when the station is not on the way to the destination.

To go further in detail regarding the access time, an investigation of access to Montréal's bus and train stations with different modes of travel was conducted. Figure 4.29 shows the proportion of access time from the centroid of each CT to different stations by walk, bike, PT, and car. The question then arises as to whether all individuals can afford to take a cab to the station? Or is it possible for everyone to use a private vehicle to access the station? On the other hand, from a sustainability perspective, it is more attractive to have a sustainable mode of transportation for all components of the trip.

A takeout from Figure 4.29 shows that when the travel time is less than about 30 minutes, there are alternatives to the car, but when the travel time is longer, there are very few alternatives to the car for getting to the stations. Moreover, the longest travel time for the car is more than 50 minutes in Montréal, while it can be as long as 120 minutes for the public transport options.

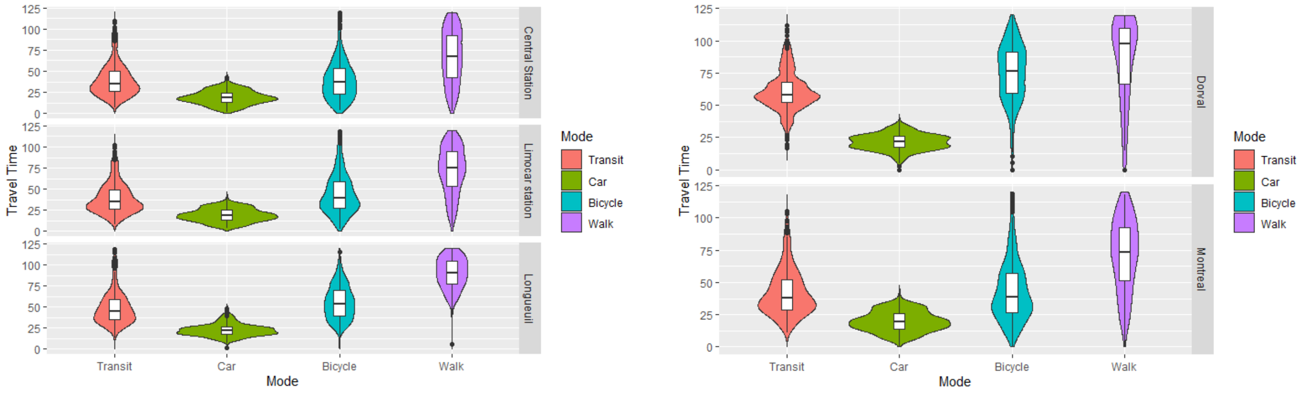


Figure 4.29 Proportion of access time for different modes from the centroid of CT to the central bus (left) and train (right) stations in Montréal Island

For a better comparison within the Montréal metropolitan area, Figure 4.30 and Figure 4.31 represent the ratio of minimum access time with transportation, walking, and biking versus driving for access to the central bus and train station in Montréal. The role of the metro is discretely significant in this figure with ratios lower in all CTs crossed by a metro line. Moreover, when a CT is outside the island of Montréal, this ratio is high, which is the result of less efficient transportation service for these areas to downtown Montréal. The ratio in some close CTs to the stations is also higher than expected which can be a result of waiting time for the bus or having a connection in PT mode to get to the station.

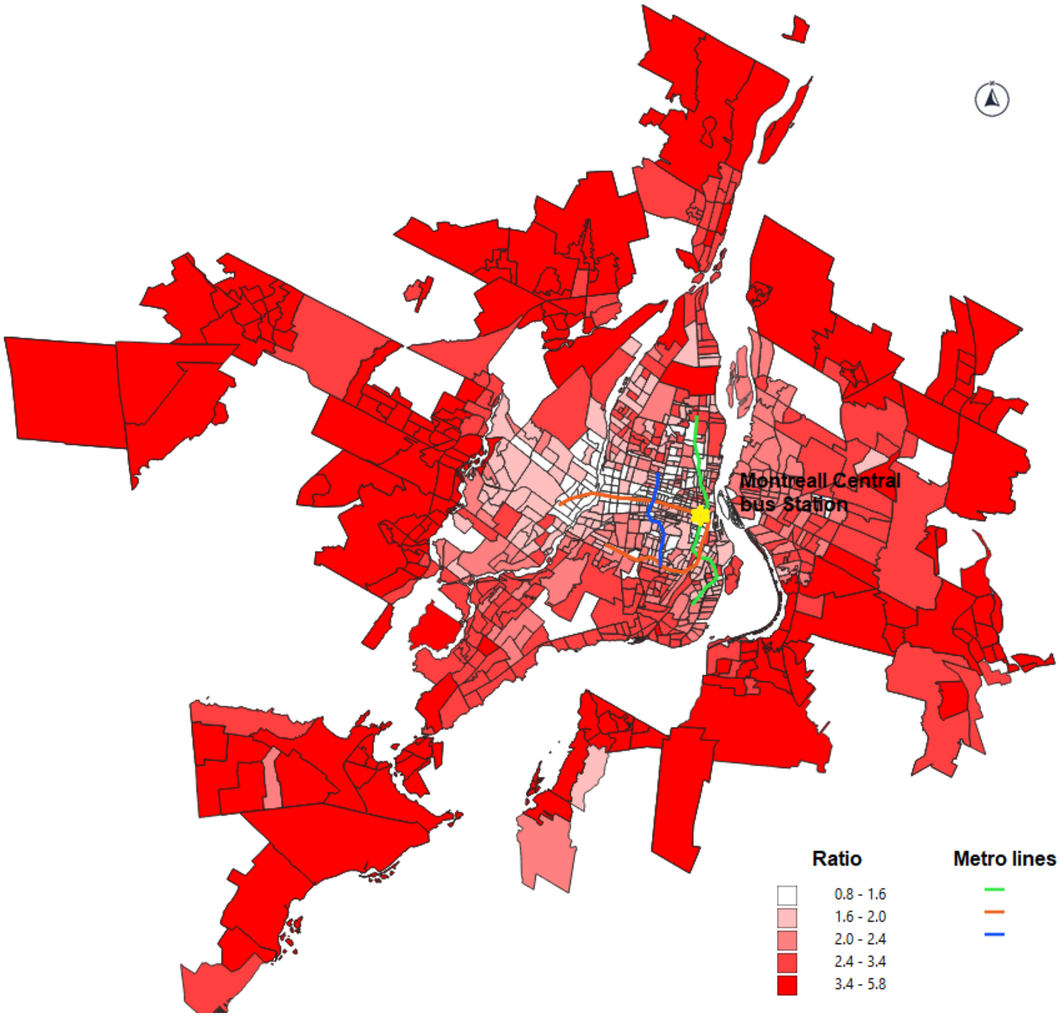


Figure 4.30 Gradient map of the ratio of minimum access time among public transportation, bike and walk to the private vehicle for access to the Montréal central bus station

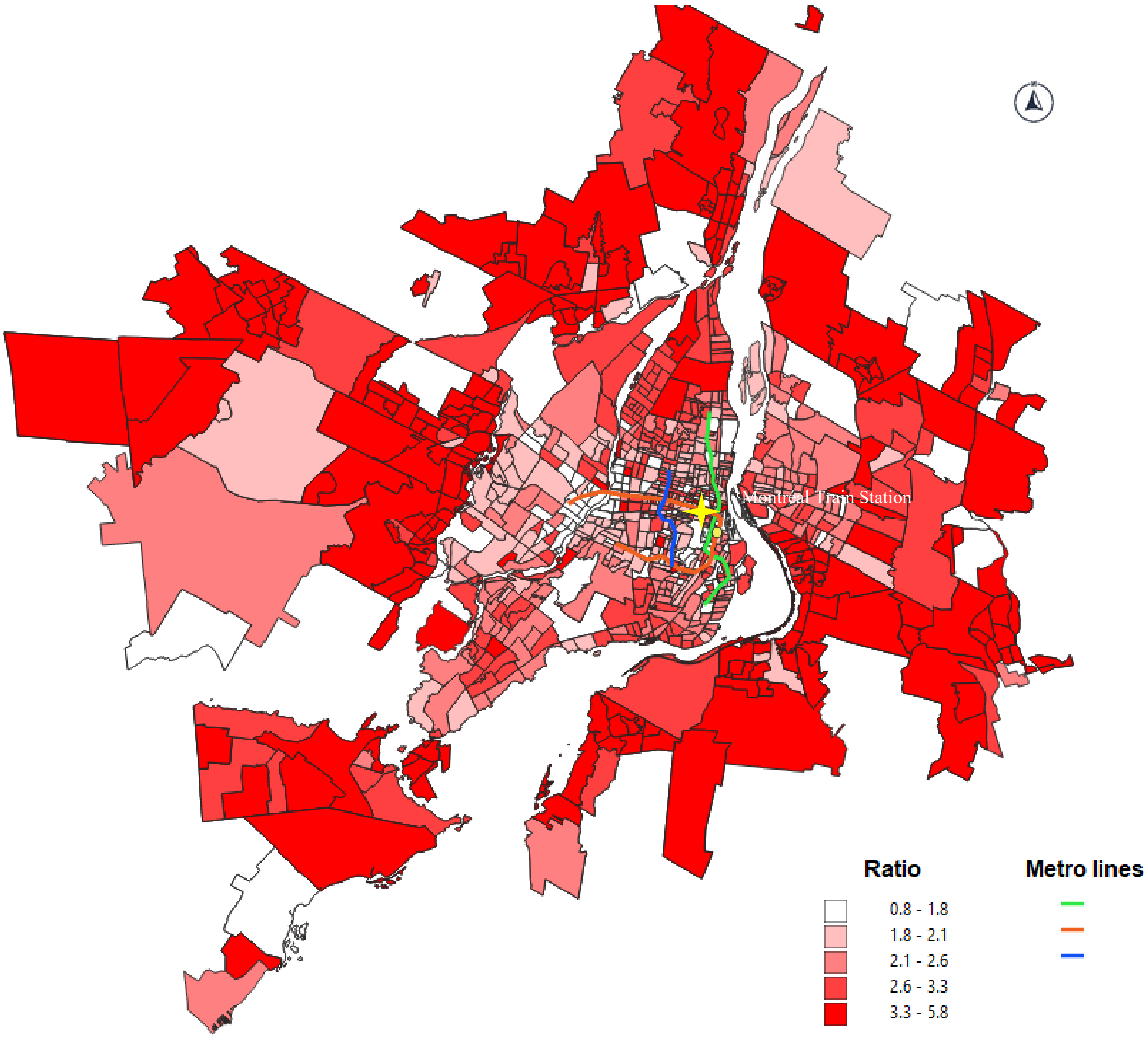


Figure 4.31 Gradient map of the ratio of minimum access time with public transportation, bike and walk to the private vehicle to the Montréal train station

Figure 4.32 represents the dispersion of travel time between OD pairs by car and bus without adding the waiting time to any of the modes. A quick takeout from the graph shows how travel time with the car is less than the bus in all OD pairs. This difference is more pronounced when the distance between origin and destination is smaller, for example, on average, the travel time by car between Drummondville and Sherbrooke is about half the one by bus, while for longer distances, such as Québec City to Saguenay, this difference is smaller.

4.4.2 Total travel time

The total travel time estimation between origin and destination consisting of access time, egress time, travel time and waiting or break time analysis are discussed in this section

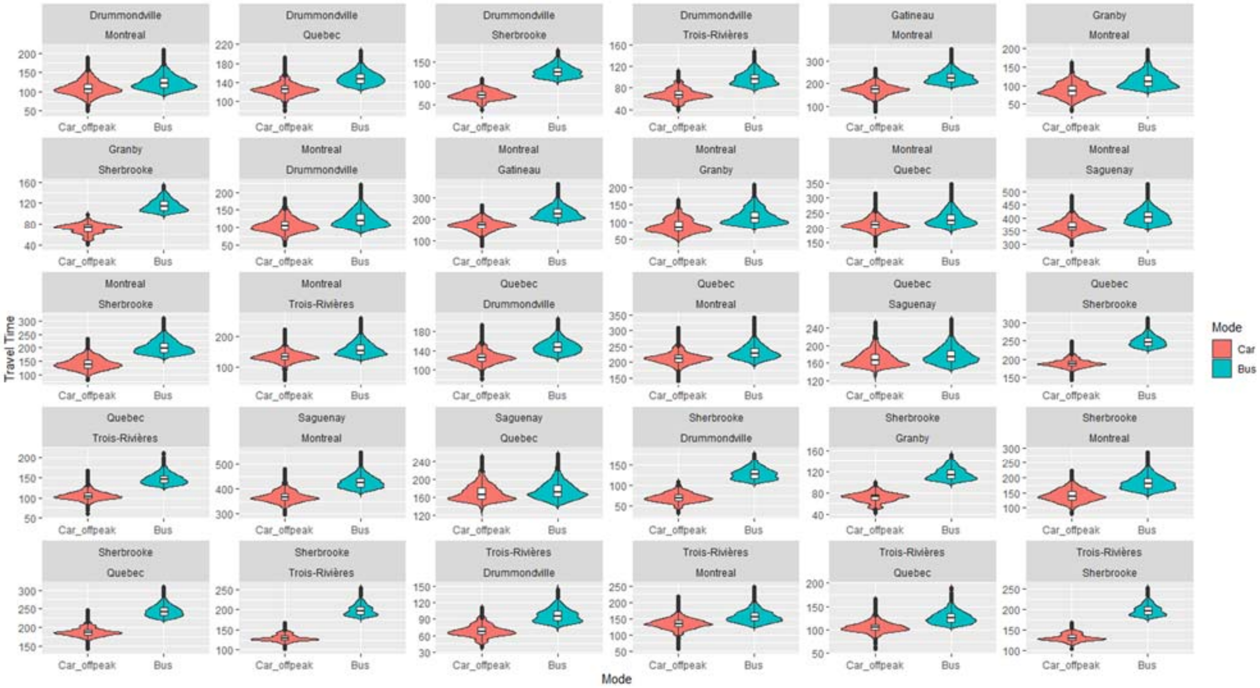


Figure 4.32 Travel time dispersion for car and bus travel

Figure 4.33 shows the dispersion of travel time between OD pairs by car and bus when the waiting time is added to the total travel time. A fraction can be observed in the car trip at the 120-minute point which corresponds to the 15-minute rest time. However, there is no fraction in the bus mode since the 15-minute wait time is added at the beginning of the trip. For OD pairs where most travel times between CTs are less than 120 minutes, the bus situation deteriorates further and the gap between the bus and car increases. For longer distances, such as Montréal-Saguenay, the gap decreases, because, for car trips, 45 minutes must be added as rest time while driving, while this figure remains the same for the bus mode.

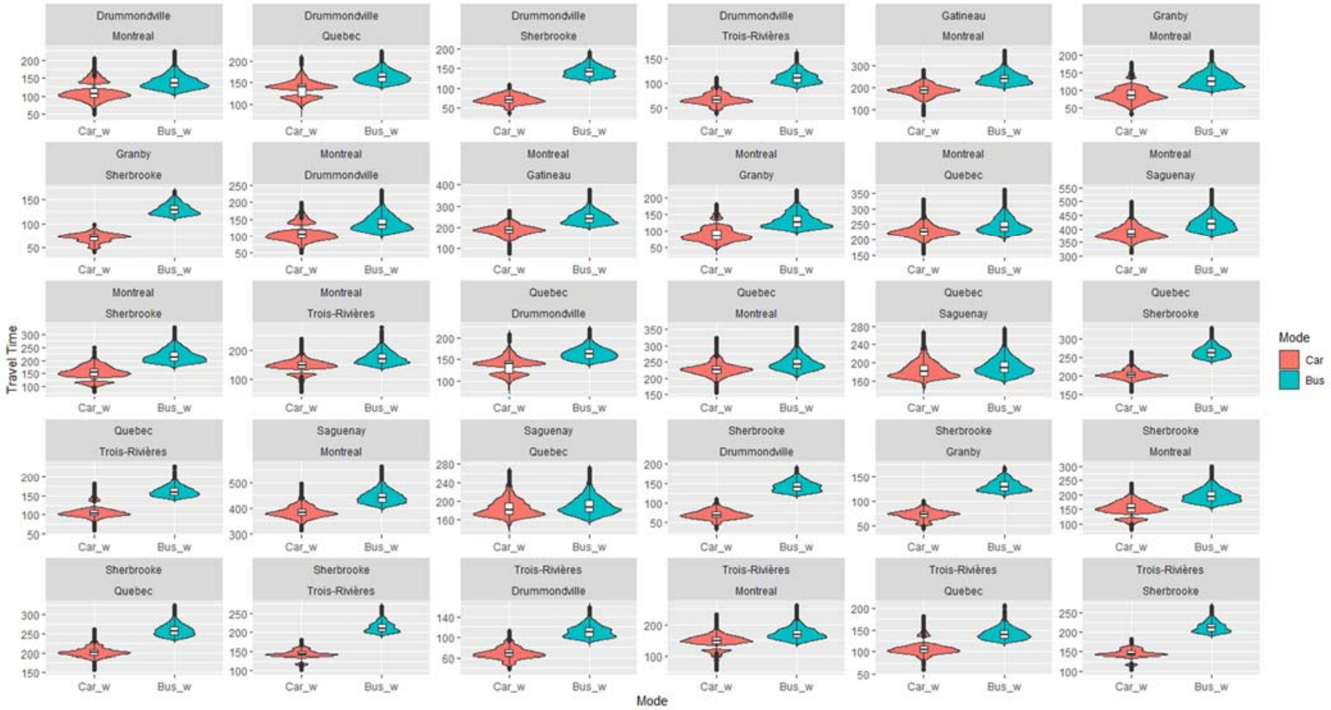


Figure 4.33 Travel time dispersion for car and bus travel with waiting time

Out of 56 OD pairs, only 8 ODs have available service by train. Figure 4.34 shows the travel time dispersion for the bus, the car and the train between CTs of origin and destination. Train service seems to be very slow when it comes to OD pairs with higher distances, while it is more competitive when the distance is smaller. In the OD pairs for Montréal-Saguenay, train travel time starts at 550 minutes for the shortest CT pair to about 650 minutes for the farthest distance, while for the same cities, car travel time in the farthest distance is about 500 minutes and on average train travel time is twice as long as the route by private car.

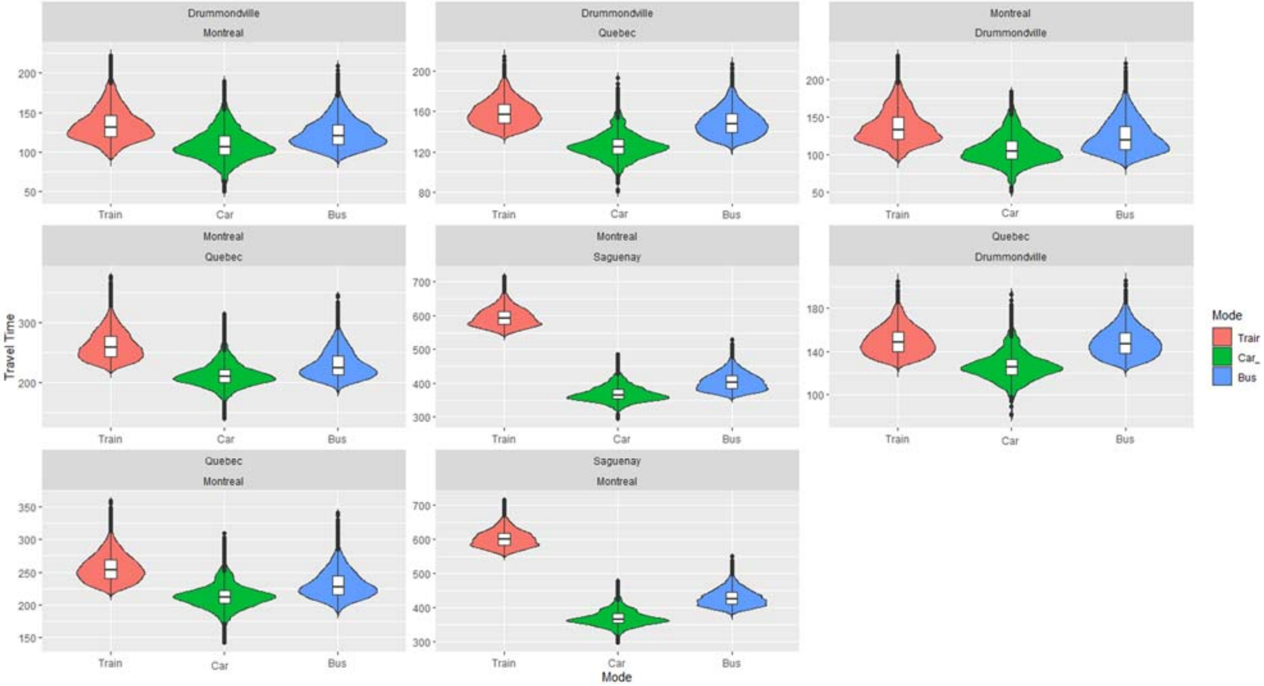


Figure 4.34 Travel time dispersion for bus, car, and train travel

The waiting time for the train is assumed to be 30 minutes, as Via Rail suggests to its passengers. Figure 4.35 shows the travel time dispersion for all three modes with added waiting time to the total travel time. In OD pairs with less than 120 minutes of travel, the train is even less competitive because there are 30 minutes of waiting time for the train and there is no waiting time for driving with the car. Even when driving for less than 240 minutes, the waiting time for the train is higher than the rest and refuelling time with the car. Just in the case of Montréal-Saguenay, the added time for the train is smaller than the added time for driving, however, in this case, while the travel time by train is much higher than by car, this added time does not make a significant change.

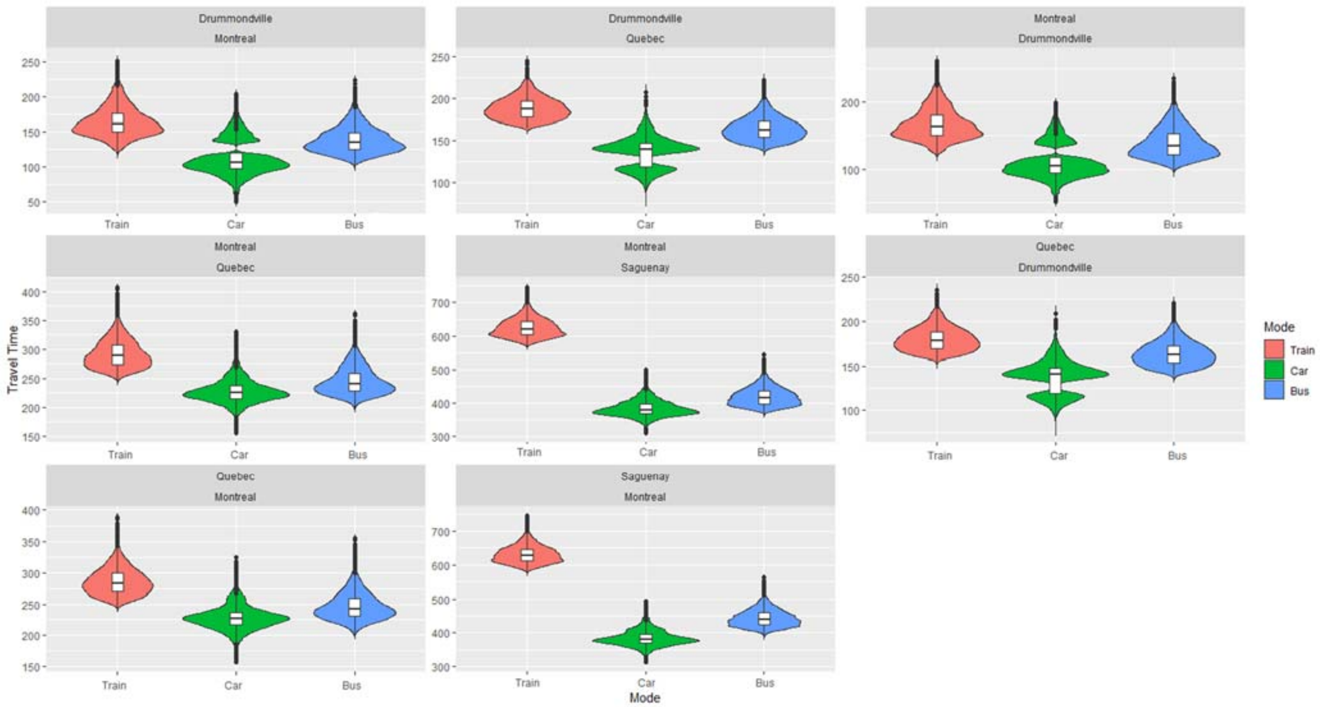


Figure 4.35 Travel time dispersion with waiting time for bus, car, and train travel

To have a better comparison between travel time by bus and car, a ratio of the bus to car travel time is calculated. Figure 4.36 illustrates the dispersion of the ratio of the bus to car travel time based on distance. This graph represents how this ratio varies between 1 and 5 depending on the distance of travel. However, this distance is based on the distance travelled by car. This ratio approves the above-mentioned assumption that when the distance is higher, the difference in travel times will increase between car and bus.

To our knowledge, there is no specific ratio to change the willingness of individuals to use PT instead of private cars. As it is shown for the distance of less than 100 km, this ratio is high which highlights the role of access and egress time and waiting time in shorter trips and maybe the type of service provided between OD pairs by bus whether it is an express or regular service. Meanwhile, when the distance is more than 200 in all cases, the ratio of the bus to car travel time is less than two.

The highest ratio of the same origin and destination city should have an end in a large city since the access time in the large city is high. Another explanation for those high ratios is that they may belong to CTs with a smaller distance in the same city since in some cases PT travellers travel a greater distance than car drivers, e.g., when an individual lives in the east end of Montréal and the destination is in southwestern Québec, there is a distance in both cities to go to the downtown area to access the bus station and return to the destination.

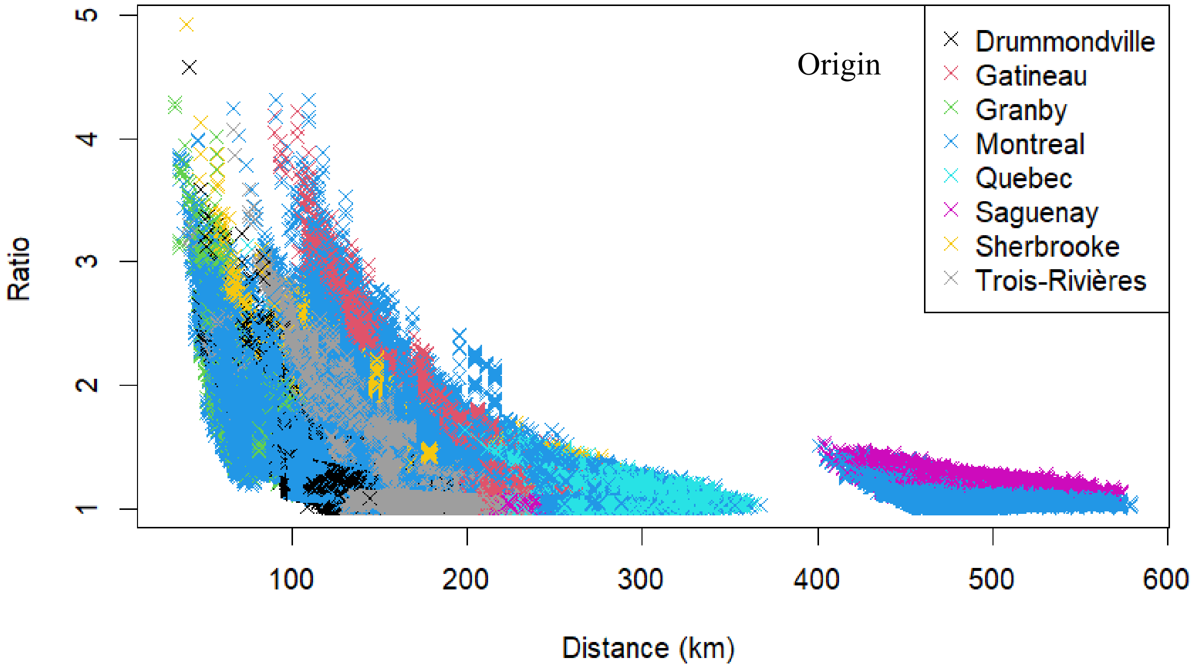


Figure 4.36 Dispersion of the ratio of the bus to car travel time based on the distance of the OD pair

The same analysis was conducted for travel by train versus car and the results are shown in Figure 4.37. For the train, the travel distance smaller than 100 km also has the highest ratio which can be the result of access time in larger cities. Still, the highest ratio for the train is smaller than the bus on shorter trips.

An interesting result in the train ratio appears when it comes to trips with a distance greater than 400 km. The ratio increases in this section. This increase can be explained by a longer travel time with the train when the origin and the destination are far from each other. This high ratio corresponds specifically to Montréal and Saguenay.

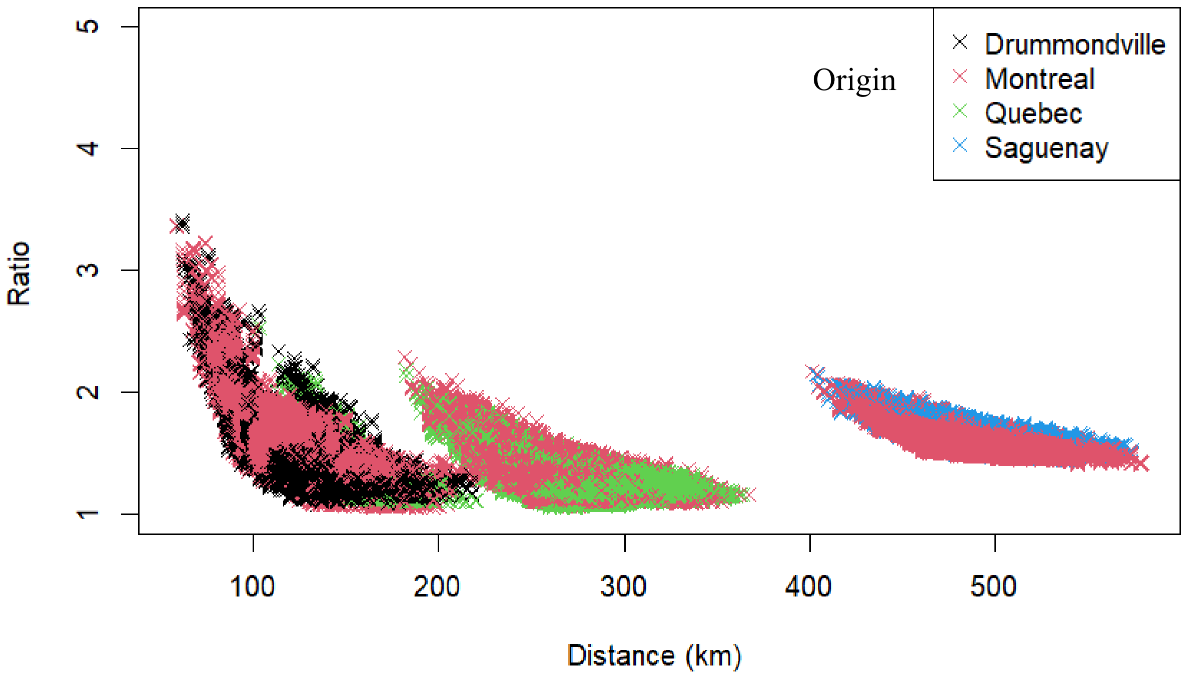


Figure 4.37 Dispersion of the ratio of the train to car travel time based on the distance of the OD pair

CHAPTER 5 CONCLUSION

5.1 Contributions

This research mainly consists of three main parts: 1) development of interregional corridors, 2) development of trip generation model for LD travel and 3) analysis of mode competitiveness on main interregional corridors.

Two main sources of data were employed in this study: census data for 2016 for corridor development, and data from Travel Survey for Residents in Canada (TSRC), covering the 2012 to 2017 period to develop a long-distance (LD) trip generation model.

For corridor analysis, 28 cities in Québec were classified into three categories based on population: cities with a population of more than 100,000 considered as core and between 50,000 and 100,000 as mid-size and less than 50,000 considered as small cities. Six steps were defined to achieve the main corridor. These corridors were ranked based on socio-economic and socio-demographic characteristics of the CSD along with an OD pair and cities of origin and destination.

Two ranking methods based on characteristics of CSDs along OD pairs and features of the city of origin and destination showed how different can be the importance of each OD pair. The presence of cities along a route becomes more important in the case of public transportation since this mode needs to serve those cities, while, when it comes to private vehicles, the cities of origin and destination become more vital.

It was found that, in both cases, the cities of origin and destination are highly important, the most important routes have one or two ends in the core cities. However, when the assessment is based on CSDs along the corridor, all the important pairs travel across highly dense areas.

In the next step, a trip generation model was developed. Because LD trips are relatively rare events and fewer individuals make LD trips, data is considered imbalanced. Several approaches were employed to tackle the imbalance distribution issue at the data preparation level. It was found that subsampling leads to better performance in the trip generation model. CART, CTree, Random Forest, and the generalized linear model are the four different methods used for model estimation to find the best-fitted model. To evaluate the models, overall accuracy, sensitivity, specificity, precision, and F1-score were measured. Findings from the models confirmed that the best overall accuracy was found in decision tree models, Random Forest, and logit models, respectively. In addition, as the model estimation results from imbalanced data, overall accuracy can be misleading because the model's accuracy might be biased by the majority class on the test data set. Therefore, sensitivity and F1-score are the two principal scores for the evaluation of the prediction performance of the model. It was found that the Random Forest model has the best performance in predicting the "having trip" class by individuals, while it is not the best model for the "no trip" class. In fact, the best model depends on the study's goal, whether both negative and positive classes have the same consequences, or one class plays a more crucial role in the expected use of study results.

Findings from this study state that income and educational level are the two most important variables affecting the LD trip occurrence. Higher income and higher academic levels lead to more LD trips. Also, it was found that young people with the same level of income make more LD trips if they have a higher level of education.

In the last section, this research aimed to conduct a mode competitiveness analysis in the main interregional corridors of Québec. Three different sources of data were used to estimate total travel time. OSRM was used to estimate travel time by car from each CT's centroid to the bus and train stations. In the case of Montréal, access using PT, walking and cycling were also estimated using the Transition platform. Hence, travel time between stations was extracted from the timetable of the bus and train companies. Additional times (waiting, resting + refuelling) were also added to improve the estimation of competitiveness ratios.

Regarding the spatial resolution, the research was conducted at the census tract level and travel time was estimated between all pairs of CTs of the CAs and CMAs located in Québec. Key findings from this section revealed that the car mode has a smaller travel time for LD trips in all OD pairs. However, for some pairs, this ratio is only a bit larger than one and it corresponds to the ODs with shorter access travel and travel distances ranging between around 120 to around 300 km.

Furthermore, the results show that for trips shorter than 100 km, the ratio of bus and train travel time to car travel time is higher, because the contribution of access time to total travel time is more important for short trips. However, for train trips, it seems that this ratio also increases for distances above 400 km, due to the quality of the train service in terms of travel time.

To make public transport more competitive concerning the private car, it is very important to facilitate and improve access and egress segments of long-distance PT options, especially when the destination or origin of the trip is not on the route. This problem is intensified in larger cities where access routes are longer.

5.2 Limitations

LD travel is a fairly challenging topic in transportation study both in terms of the complexity of LD travel and lack of data. LD trips are very different from urban commuting trips in terms of spatial and temporal characteristics, regularity, and frequency. In addition, LD trips are not easily captured in regular surveys because they are not typical trips, they are not necessarily made at the usual survey times (fall, weekdays), and they are less easy for respondents to remember because of their less usual occurrence. These characteristics make this type of trip more complex to observe and analyze. This thesis basically consists of three main topics, each part can have its perspectives, and the only common issue related to all of them is to have access to better data, developing and conducting a survey focusing on transportation planning study is the first step.

For the model estimation, as other studies have stated (Van Nostrand, 2013; Llorca, 2018), data-related issues are a fundamental challenge in LD travel demand modelling. In the TSRC survey, the only non-categorical variable is the population. Hence, factors like accessibility to the airport, bus, and train station in terms of travel distance and cost and land-use data might result in better model performance. The TSRC survey was mainly designed for tourism studies. This study proposed a relevant use of these data for travel demand forecasting and insists on improving data collection and survey methods to have opportunities to enhance the mode complexity and provide more insights into the factors having an incidence on LD travel behaviours.

The mode competitiveness analysis also has its limitation, the most important limit in this section was the use of different platforms to extract data for estimation of the total travel time for LD access, egress, car and PT mode which can bias the result. In addition, there are many factors influencing the LD trips, these factors include travel time, fare, distance, socio-demographic and socio-economic characteristics of travellers, etc. Also, because of the nature of the LD trip, the characteristic of the trip and the service quality has a significant impact on travellers choosing the PT mode over the private vehicle namely since PT options can allow valuing time while travelling.

These components increase the complexity of analyzing and modelling the mode choice process of travellers for LD trips. Also, there are other limitation in the mode competitiveness analysis, such as no accounting for congestion for travel time between cities, also it is not possible to estimate travel time by all mode for access to the station for all areas other than Montreal. Moreover, the departure time can play a vital role both in terms of choosing PT over private vehicle and the total travel time which is not considered.

5.3 Perspectives

Every part of this research project can be improved as discussed below. In the section on corridor analysis, it could be useful to use the current road network to analyze and rank the corridors. Even without having access to traffic data, by using socio-economic and socio-demographic data of cities along with the road network and corridors, a corridor analysis can be conducted to assess the potential demand through the current road network. The average annual traffic count data also can be applied for analysis of ranking the current road network which highest number of traffic count result in more important pair and least number of traffic count leads to least important OD pair. In addition, considering having a precise travel demand model, applying the model to corridor ranking could be interesting.

Modelling LD travel demand could be the most complex part of the LD travel study, this study used the machine learning method to develop a trip generation model, and it could be relevant to develop a mode choice model using the same data. In addition, by combining data on socioeconomic and sociodemographic characteristics of the origin and destination city, a destination choice model can also be developed. Moreover, it is strongly recommended to improve the trip generation model by using other methods such as statistical models or neural network methods or trying to access data from public transportation companies to have a more accurate model.

Since the mode used for travel in TSRC data was 95% of the time private vehicle, this research analyzes the mode competitiveness of PT modes with private vehicles, and it only compares the modes from the travel time point of view, while there are other features which are possible to analyze in mode competitiveness analysis. Fare, distance, and availability of alternatives can also be assessed, in addition, for PT service this study is investigating direct services between cities, while there are cities which are connected via a transfer service. And it does not take into account any improvement or additional services that could be added to the actual network. On the other hand, the combination of the modes could be included in the assessment. Moreover, adding new modes to this study can improve the mode competitiveness analysis such as carpooling and plane.

REFERENCES

Abdel-Aty, M. A. (1998). survey of the elderly: An assessment of their travel characteristics. *Transportation Research Board.*

Aguiléra, A. &. (2015). Socio-occupational and geographical determinants of the frequency of long-distance business travel in France. *Journal of Transport Geography*, 28-35.

Anderson, R. G. (2017). Long-Distance Smartphone-Based Travel Surveys in Ohio. Esterel, Canada: 11th International Conference on.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of economic theory* , 244-263.

Aultman, L. (2018). The implications of long-distance tour attributes for national travel data collection in the United States. *Transportation*, 875-903.

Aultman-Hall, L. H. (2018). The implications of long-distance tour attributes for national travel data collection in the United States. *Transportation, 45(3)*, pp. 875-903.

Berliner, R. M.-H. (2018). Exploring the Self-Reported Long-Distance Travel Frequency of Millennials and Generation X in California. *ransportation Research Record*, 208-218.

Biau, G. &. (2016). A random forest guided tour. *Test*, 197-227.

Bourdeau, J.-S. (2022, March). Personal communication.

Breiman, L. F. (2017). *Classification and regression trees. Routledge.* Routledge.

Canada, S. (n.d.). *Statistic Canada*. Retrieved Jan 2022, from https://www150.statcan.gc.ca/n1/en/catalogue/92-168-X

Chang, L.-Y. a.-C. (2005). Data mining of tree-based models to analyze freeway accident frequency. *Journal of safety research*, 365-375.

Chawla, D. A. (2008). *Learning Decision Tree for Unbalanced Data*. Retrieved July 2021, from University of Notre Dame: https://www3.nd.edu/~nchawla/papers/ECML08.pdf

Cieslak, D. A. (2008). Learning Decision Trees for Unbalanced Data. *Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg, Springer Berlin Heidelberg.*

Czepkiewicz, M. H. (2020). Who travels more, and why? A mixed-method study of urban dwellers' leisure travel. *Travel behaviour and society*, 67-81.

Dargay, J. M. (2012). The determinants of long distance travel in Great Britain. *ransportation Research Part A: Policy and Practice*, 576-587.

Dementias, S. A. (2017, October 06). *Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnagi.2017.00329/full

Dobilas, S. (2021, Jan). *CART: Classification and Regression Trees for Clean but Powerful Models*. Retrieved from towardsdatascience: https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powerful-models-cc89e60b7a85

donneesquebec. (2022, 01). *donneesquebec*. Retrieved from https://www.donneesquebec.ca/recherche/fr/dataset/stl-horaires-planifies-et-trajets-des-bus

Enterprise, S. (2018, Nov. 29). *SAS Institute Inc. Prior Probabilities*. Retrieved from SAS Help Center: http://documentation.sas.com/doc/en/emxndg/15.1/p1vqpbjwoo4bv7n1sw77e0z64xxs.htm

Enzler, H. B. (2017). Air travel for private purposes. An analysis of airport access, income and environmental concern in Switzerland. *Journal of Transport Geography*, 1-8.

EXO. (2022, 01). *EXO*. Retrieved from https://exo.quebec/en/about/open-data

experts, T. r. (2021, Nov). *The road trip experts*. Retrieved Jan 2022, from https://www.theroadtripexpert.com/do-cars-need-rest/#:~:text=How%20often%20should%20you%20let,a%20good%20state%20of%20repair.

Frei, A. T. (2010). *Long distance travel in Europe today: Experiences with a new survey*. Zurich: Arbeitsberichte Verkehrs-und Raumplanung 611.

Friedman, J. T. (2009). *The elements of statistical learning*. New York: Springer.

Georggi, N. L. (2000). *An analysis of long-distance travel behavior of the elderly and the low-income*. Florida: University of South Florida.

Gerike, R. a. (2018). Workshop synthesis: Surveys on long-distance travel and other rare events. *Transportation Research Procedia 32*, 535-541.

González-Savignat, M. (2004). Competition in air transport. *Journal of Transport Economics and Policy*, 77-107.

Guillemette, Y. (2015). *MIEUX COMPRENDRE L'OFFRE ET LA DEMANDE DE DÉPLACEMENTS.* Montreal: Polytechnique de Montreal.

Guillemette, Y. (2015). *MIEUX COMPRENDRE L'OFFRE ET LA DEMANDE DE DÉPLACEMENTS INTERURBAINS AU QUÉBEC.* Montreal: (Master thesis) Polytechnique de Montreal. Retrieved from https://publications.polymtl.ca/1829/

Haixiang, G. Y. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* , 220-239.

He, H. &. (2009). Learning from imbalanced data. *IEEE*, 1263-1284.

Hess, S. e. (2018). Analysis of mode choice for intercity travel: Application of a hybrid choice model to two distinct US corridors. *Transportation Research Part A: Policy and Practice* , 547-567.

Hothorn, T. H. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 651-674.

J.J. LaMondia, M. M.-H. (2015). Modeling intertrip time intervals between individuals' overnight long-distance trips. *Transportation Research Record* , 23-31.

Joyce M. Dargay, S. C. (2012). The determinants of long distance travel in Great Britain. *Transportation Research Part A: Policy and Practice, 46*(3), 576-587.

Kabir, E. S. (2018). Statistical modeling of tree failures during storms. *Reliability Engineering & System Safety* , 68-79.

Kuhnimhof, T. a. (2009). The path to better long-distance travel data in Europe–The potential of combining established household survey instruments and methodological innovations. Spain: First International Conference on the Measurement and Economic Analysis of Regional Tourism in San Sebastian.

LaMondia, J. J.-H. (2014). Long-distance work and leisure travel frequencies: Ordered probit analysis across non–distance-based definitions. *Transportation Research Record*, 1-12.

Lisa. (2021, November). *How Frequently Should You Take Breaks When Driving Long Distances?* Retrieved from theroadtripexpert: https://www.theroadtripexpert.com/how-frequently-should-you-take-breaks-when-driving-long-distances/

Llorca, C. M. (2018). Estimation of a long-distance travel demand model using trip surveys, location-based big data, and trip planning services. *Transportation Research Record*, 103-113.

Loh, W. Y. (1997). Split selection methods for classification trees. *Statistica sinica*, 815-840.

Lu, J. e. (2021). Modeling hesitancy in airport choice: A comparison of discrete choice and machine learning methods. *Transportation Research Part A: Policy and Practice*, 230-250.

Ma, Lukas. (2021). *Modelling rare events using non-parametric machine learning classifiers-Under what circumstances are support vector machines preferable to conventional parametric classifiers?* Göteborgs: Göteborgs universitet .

Mallett, W. J. (2001). Long-distance travel by low-income households. *E-C026*.

Menardi, G. &. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 92-122.

Miller, E. (2004). The Trouble with Intercity Travel Demand Models. *Transportation Research Board*, 94-101.

Mishra, S. D. (2003). Application of classification trees in the sensitivity analysis of probabilistic model results. *Reliability Engineering & System Safety*, 123-129.

Montreal, P. d. (n.d.). *Transition*. Retrieved Jan 2022, from http://transition.city/index_en.html#intro

Myers, R. H. (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, 274-291.

Outwater, M. B. (2015). *Foundational Knowledge to Support a Long-Distance Passenger Travel Demand Modeling Framework: Implementation Report.*

Patil, D. D. (2010). Evaluation of decision tree pruning algorithms for complexity and classification accuracy. *International Journal of Computer Applications*, 23-30.

Rachman, A. &. (2019). Machine learning approach for risk-based inspection screening assessment. *Reliability Engineering & System Safety*, 518-532.

RColorBrewer, S. &. (2018). *Package ‘randomForest'*. Berkeley: Berkeley, CA, USA: University of California.

Reichert, A. a.-R. (2015). Mode use in long-distance travel. *Journal of Transport and Land Use*, 87-105.

Rickard, J. M. (1988, May). Factors influencing long-distance rail passenger trip rates in Great Britain. *Journal of Transport Economics and policy*(1).

Román, C. e. (2014). Valuation of travel time savings for intercity travel: The Madrid-Barcelona corridor. *Transport Policy 36*, 105-117.

Rothengatter, W. (2010). 8 competition between airlines and high-speed rail. *Critical issues in air transport economics and business 84*, 319.

RTL. (2022, 01). *RTL*. Retrieved from https://www.rtl-longueuil.qc.ca/en-CA/open-data/gtfs-files/

Schlosser, L. H. (2019). The power of unbiased recursive partitioning: a unifying view of CTree, MOB, and GUIDE. *arXiv preprint arXiv*.

Shih, Y. S. (2004). A note on split selection bias in classification trees. *Computational statistics & data analysis*, 457-466.

Song, Y. Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 130.

Specification, G. T. (2022, 06). *General Transit Feed Specification*. Retrieved from General Transit Feed Specification: https://gtfs.org/

Statistic Canada. (2017). *Microdata User Guide Travel Survey of Residents of Canada 2017*. Statistic Canada.

Statistic Canada. (2022, 01). *Travel Survey of Residents of Canada (TSRC)*. Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3810

Strobl, C. B. (2007). Unbiased split selection for classification trees based on the gini index. *Comput. Stat. Data Anal*, 483-501.

Theofilatos, A. Y. (2016). Predicting road accidents: a rare-events modeling approach. *ransportation research procedia*, 3399-3405.

Van Can, V. (2013). Estimation of travel mode choice for domestic tourists to Nha Trang using the multinomial probit model. *Transportation Research Part A: Policy and Practice 49*, 149-159.

Van Nostrand, C. S. (2013). Analysis of Long-Distance Vacation Travel Demand in the United States: A Multiple Discrete–Continuous Choice Framework. Transportation. *Transportation 40*, pp. 151–171.

Vilaça, M. M. (2019). A rare event modelling approach to assess injury severity risk of vulnerable road users. *Safety*, 29.

Wang, Y. e. (2017). Influencing mechanism of potential factors on passengers' long-distance travel mode choices based on structural equation modeling. *Sustainability 9.11*.

Y. Wang, X. Y. (2015). Using AHP for evaluating travel mode competitiveness in long-distance travel. *2015 International Conference on Transportation Information and Safety (ICTIS)*, 213-218.

Yao, E. a. (2005). A study of on integrated intercity travel demand model. *Transportation Research Part A: Policy and Practice*, 367-381.

Zheng, Z. P. (2016). Decision tree approach to accident prediction for highway–rail grade crossings: Empirical analysis. *Transportation Research Record*, 115-122.

Zhou, X. L. (2020, Aug). Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. *Reliability Engineering & System Safety, 1*(200).