Titre: Title:	Assessing the Impact of Trip Chains on Travel Behavior
Auteur: Author:	Mohamad Abdul Majid Dabboussi
Date:	2023
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Dabboussi, M. A. M. (2023). Assessing the Impact of Trip Chains on Travel Behavior [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/53402/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/53402/
Directeurs de recherche: Advisors:	Catherine Morency, & Geneviève Boisjoly
Programme: Program:	Génie civil

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Assessing the impact of trip chains on travel behavior

MOHAMAD ABDUL MAJID DABBOUSSI

Département des génies civil, géologique et des mines

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

Génie Civil

Mai 2023

POLYTECHNIQUE MONTRÉAL

Affiliée à l'Université de Montréal

Ce mémoire intitulé :

Assessing the impact of trip chains on travel behavior

présenté par **Mohamad Abdul Majid DABBOUSSI**

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

Martin TRÉPANIER, président
Catherine MORENCY, membre et directrice de recherche
Geneviève BOISJOLY, membre et codirectrice de recherche
Éric Martel-Poliquin, membre

DEDICATION

To my beloved mother and sister,

To my supervisors, Catherine Morency and Geneviève Boisjoly

For their trust in me and their support for me through thick and thin

To all my family and friends

ACKNOWLEDGEMENTS

First and before all, I would like to express my eternal gratitude and thanks to my supervisors, Catherine Morency and Geneviève Boisjoly. These two believed in me and gave me chance after chance to prove myself. They gave me this opportunity in this country and were always supportive as I went through the darkest period of my life, always pushing me to do better while understanding what I was going through. They changed my life, and I don't think any other supervisors would've been as caring and supportive as these two were. Thank you from the bottom of my heart.

I would also like to thank Hubert Verreault, for always helping with whatever data I needed in my work and some coding processes.

A huge thanks to my family, who went through the toughest time but insisted on me staying here and continuing my dream and research.

Finally, a huge thanks to all my friends that stood by my side as I went through all of this and kept trying to motivate me all the time.

RÉSUMÉ

Comprendre les comportements de mobilité des personnes est une tâche complexe qui nécessite beaucoup d'études et de recherches. Au fil des années, plusieurs approches et méthodes ont été utilisées pour aborder ce problème. Le choix modal, c'est-à-dire le processus de choix du moyen de transport avec lequel une personne effectue ses déplacements, a toujours été l'un des aspects les plus importants et les plus marquants du comportement de mobilité. Les modèles de choix modal ont donc été et restent un élément essentiel de toute planification et politique de transport réalisée ou à réaliser.

Diverses variables et facteurs affectent ces modèles, tels que le temps de déplacement, la distance, le coût et bien d'autres. Un des facteurs qui est moins considéré dans la littérature est celui de la chaîne de déplacements (l'ensemble des déplacements, avec toutes les activités qui y sont associées, qui commencent et se terminent au domicile). Ceci se reflète par le manque de modèles de choix basés sur les chaînes (également connus sous le nom de modèles basés sur les tournées) par rapport aux modèles basés sur les déplacements qui considèrent chaque déplacement indépendamment sans le contexte des chaînes dont ils font partie. Un autre facteur qui affecte le modèle de choix modal est l'outil utilisé dans le processus. Ces dernières années, et avec les progrès de la puissance de traitement des données, les algorithmes d'apprentissage automatique ont gagné en popularité en tant que base des outils de modélisation du choix modal, comme la forêt aléatoire qui a donné des résultats impressionnants. Cependant, ces algorithmes ont rarement, voire jamais, été utilisés dans le contexte des modèles basés sur les chaînes de déplacement.

Cette recherche vise donc à évaluer l'impact des chaînes de déplacements sur le comportement des personnes, en particulier sur le choix modal. Le but est de voir comment un modèle de chaînes de déplacements basé sur une forêt aléatoire peut être performant sur la base des données de la Grande Région de Montréal et comment les caractéristiques de la chaîne évoluent et affectent le comportement de mobilité. Pour y parvenir, trois objectifs principaux doivent être atteints:

1) Mener une analyse exploratoire des tendances des chaînes de déplacements au fil des ans.

2) Identifier les déterminants du choix modal des chaînes de déplacements et explorer la relation entre les variables explicatives et le choix modal des chaînes.

3) Tester l'efficacité de l'algorithme d'apprentissage automatique " forêt aléatoire " comme outil de modélisation du choix modal au niveau de la chaîne de déplacement.

La revue de la littérature présente les différents déterminants du choix modal, en particulier lorsqu'on considère le niveau de la chaîne de déplacement, et plonge dans les différentes typologies et définitions de chaînes proposées dans la littérature. L'augmentation récente de l'utilisation d'algorithmes d'apprentissage automatique, et en particulier de la forêt aléatoire, comme outil de modélisation du choix modal a également été examinée, montrant comment ce type de modèles donne généralement de bons résultats en termes de pouvoir prédictif. Les différents déterminants du choix modal sont également discutés, et sont généralement divisés en quatre groupes : les variables sociodémographiques, les variables liées aux ménages, les variables liées au milieu bâti et les variables liées aux chaînes.

Dans la méthodologie, la préparation des variables basée sur les données des enquêtes Origine-Destination de la Grande Région de Montréal en 1998, 2008 et 2018, a été expliquée. Une méthode de catégorisation du choix modal de la chaîne utilisée, mais sans la génération d'un trop grand nombre d'alternatives possibles pour les chaînes complexes a été créée. Il a été constaté qu'environ 90% des chaînes totales étaient réalisées en utilisant un seul mode pour l'ensemble de la chaîne. Cela permet de créer cinq alternatives uniques pour les chaînes : Auto-Conducteur seulement, Auto-Passager seulement, Transport en Commun seulement, Vélo/Marche à pied seulement et Autres modes seulement. Pour les autres chaînes, trois alternatives ont été créées : Bimodale (pour Kiss-and-Ride et Park-and-Ride), Mixte avec voiture (où différents modes sont utilisés, incluant la voiture) et Mixte sans voiture (différents modes sont utilisés mais sans la voiture). De la même manière, les systèmes d'activités (le nombre et le type de chaînes effectuées par une personne au cours d'une journée) ont été calculés en observant que quatre systèmes d'activités particuliers forment 96% de tous les systèmes possibles : une chaîne simple par jour, deux chaînes simples par jour, une chaîne complexe par jour, une chaîne simple et une chaîne complexe par jour. Tous les autres systèmes d'activités possibles ont été qualifiés d'"autres". D'autres variables ont également été construites sur la base du travail effectué par les recherches précédentes menées à Montréal à l'aide des données des enquêtes Origine-Destination.

Dans l'analyse exploratoire, plusieurs tendances du comportement de chaînage des déplacements de la population de Montréal ont été examinées. L'évolution de la complexité des chaînes, du nombre de chaînes par jour, de la durée, de la distance et des motifs a été montrée de 1998 à 2018. De plus, la répartition modale des chaînes et l'évolution du système d'activités ont été présentées au fil des ans. Les tendances observées montrent une augmentation de la complexité des chaînes, en particulier pour les femmes, et une augmentation globale de la longueur totale des chaînes si l'on compare 2018 aux années précédentes, ainsi qu'un changement dans les systèmes d'activité, où l'on observe une diminution de l'activité de deux chaînes simples par jour et une augmentation de l'activité d'une chaîne complexe par jour.

La relation entre différentes variables de différents types (socio-démographiques, environnement bâti, ménage et chaîne) et le choix modal de la chaîne a ensuite été observée. Les résultats montrent que des variables telles que la distance entre le ménage et le Centre-Ville et le motif de la chaîne sont étroitement liées au choix modal de la chaîne.

Dans la partie modélisation, des classes ont été créées pour les variables catégorielles, car elles sont essentielles pour le processus de division effectué dans la forêt aléatoire. Pour certaines variables numériques, des classes ont été créées afin de mieux interpréter l'importance de leurs caractéristiques, comme la création de groupes d'âge. Plusieurs modèles ont été testés avec différentes hypothèses et variables. Le modèle final réalisé, qui incorporait les systèmes d'activités d'une personne comme variable indépendante, a donné les meilleurs résultats en termes de précision globale, soit 79 %. Tous les modèles ont donné les meilleurs résultats dans la prédiction des chaînes d'Auto-Conducteurs seulement, mais ont souffert dans la prédiction du mode Bimodal, moins représenté. Un modèle prédictif a également été réalisé, sans utiliser de variables de chaîne, qui a montré une précision de 60%, démontrant la pertinence de considérer la chaîne de déplacements dans la modélisation du choix modal.

En ce qui concerne l'importance des caractéristiques, la longueur des chaînes s'est avérée être la variable la plus décisive pour les modèles de forêt aléatoire, tant pour la diminution moyenne du Gini (mesure de la contribution d'une variable à l'homogénéité des nœuds et des feuilles de la forêt aléatoire) que pour la diminution moyenne de la précision. L'accès à une voiture, la distance par rapport au centre-ville, le motif de la chaîne et le système d'activités se sont également avérés

importants après la longueur de la chaîne. Trois variables utilisées se sont avérées ne pas avoir une importance significative, à savoir la présence d'une gare, le statut de la personne et la durée de l'activité. Nous constatons toutefois que le statut est plus important dans le modèle prédictif lorsque le motif de la chaîne n'est pas présent, ce qui suggère une forte corrélation entre ces deux variables.

La catégorisation simplifiée des combinaisons de modes (lorsque plusieurs modes sont utilisés au sein d'une même chaîne) en mixte avec voiture et mixte sans voiture est l'une des principales limites de cette étude. Bien qu'elles représentent moins de 10% de toutes les chaînes, le fait de ne pas connaître les modes exacts utilisés ou l'ordre dans lequel ils sont utilisés rend difficile l'interprétation des résultats pour ces deux catégories. Le développement d'un moyen de mieux saisir les alternatives où différents modes sont combinés, et l'intégration d'un plus grand nombre de variables de milieu bâti pourraient fournir un meilleur aperçu et des résultats plus performants à l'avenir.

ABSTRACT

Understanding travel behavior of people is a complex matter that requires a lot of studies and research. Throughout the years several approaches and methods were used to tackle this problem. Mode choice, the process by which a person selects their transportation mode, has always been one of the most important and prominent parts of travel behavior. Mode choice models, therefore, were and remain an essential part of any transportation planning and policies made or to be made.

Various variables and factors affect these models such as travel time, distance, cost, and many others. One type of factors that is less represented in the literature is that of the trip chain (the sequence of trips, with all its associated activities that ends and begins at home). This is extremely evident by the lack of chain-based (also known as tour-based) choice models when compared to trip-based models that consider each trip independently without the context of the chains they are embedded in. Another factor that affects the mode choice model is the tool used in the process. In recent years, and with the advancement of computational power, machine learning algorithms have gained a lot of popularity as the basis of mode choice modeling tools, such as the random forests which showed impressive results with respect to their predictive capabilities. However, the use of these algorithms has rarely if ever been used in the context of trip-chain-based models.

As such this research aims to assess the impact of trip chains on travel behavior, especially the modal choice. The goal is to see how well a random forest trip-chain-based model can perform based on data from the Greater Montreal Area and how chain variables are evolving and affecting travel behavior. To achieve this three main objectives need to be attained:

1) Conduct exploratory analysis of trip chaining trends over the years.
2) Identify the determinants of trip chain mode choice and explore the relationship between the explanatory variables and chain mode choice.
3) Test how well the machine learning algorithm “random forest” functions as a tool to model mode choice at the trip chain level.

The literature review presents the different determinants of mode choice, especially when considering the trip chain level, and dives into different typologies and chain definitions proposed in the literature. The recent rise in the use of machine learning algorithms and especially the random

forest as a tool to model mode choice was also examined, showing how these models are generally providing good results in terms of predictive ability. The different determinants of mode choice are also discussed, where generally they were divided into four groups: socio-demographic, household, built environment and chain variables.

In the methodology, the preparation of the variables to be incorporated in the random forest model, based on the Origin-Destination survey of the Greater Montreal Area in 1998, 2008 and 2018, was explained. A method for categorizing the modes used in the chain, but without the generation of too many possible alternatives for complex chains were created. It was found that about 90% of total chains were conducted using a single mode for the whole chain. This allows the creation of five unique alternatives for chains: Car Driving only, Car Passenger only, Public Transit only, Cycling/Walking only and Others unique. For the remaining chains, three alternatives were created: Bimodal (for Kiss-and-Ride and Park-and-Ride), Mixed with car (where different modes are used including the car) and Mixed without car (different modes are used but without the car). In a similar way, activity systems (the number and type of chains done by a person through a single day) were calculated by observing that four particular activity systems form 96% of all possible systems: one simple chain per day, two simple chains per day, one complex chain per day, and one simple and one complex chain per day. All the remaining possible activity systems were labelled as “others”. Other variables were constructed as well based on the work done by previous research conducted in Montreal using the Origin-Destination survey.

In the exploratory analysis, several trends of the trip chaining behavior for the people of Montreal were examined. The evolution of chain complexity, chains per day, duration, distance, and purpose were shown from 1998 to 2018. Furthermore, the chain modal split and activity system evolution were shown throughout the years. Takeaways of the trends observed showed an increase of the complexity of chains especially for women, and an overall increase in chain total distance when comparing 2018 with the previous years. We also observe a decrease in the two simple chains per day activity and an increase in the one complex chain per day. The relationship between different variables of different types (socio-demographic, built environment, household, and chain) and the chain mode choice was then observed. Results show that variables such as the household distance to Central Business District and chain purpose are closely related to the chain mode choice.

In the modeling part, classes were created for the categorical variables as it is essential for the splitting process done in the random forest. For some numerical variables, classes were created to have better interpretation of their feature importance such as the creation of age groups. Several models were tested with different assumptions and variables. The final model which incorporated the activity systems of a person as an independent variable performed the best in terms of overall accuracy at 79%. All the models performed exceptionally well in the prediction of Car Driving only chains but suffered in the less represented Bimodal mode prediction. A predictive model was done as well where no chain variables were considered, which had an accuracy of 60%.

In the feature importance, chain distance was shown to be the most decisive variable for the random forest models in both Mean Decrease Gini and Mean Decrease Accuracy. Access to a car, distance to Central Business District, chain purpose and activity system were also found to be significant, ranking after chain distance in terms of importance. Three variables used were found to not have a significant importance which was the presence of a train station, people main occupation and duration of the activity. Although, we see that the main occupation is more significant in the predictive model where the chain purpose is not present, which suggests a strong correlation between these two variables.

The simplified categorization of mode combinations (where several modes are used within the same chain) to Mixed with car and Mixed without car is one of the main limitations of this study. Although they form less than 10% of all chains, not knowing the exact modes used or the order in which they are used makes it difficult to draw conclusions regarding these two categories. Developing a way to better capture the alternatives where different modes are combined, and the integration of more built-in environment variables could provide greater insight and better performing results in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CBD Central business district

CD Car driver

CP Car passenger

DA Dissemination area

GMA Greater Montreal Area

GTFS General Transit Feed Specification

K&R Kiss-and-Ride

MNL Multinomial logit

ODS Origin-Destination survey

P&R Park-and-Ride

PT Public transit

RF Random Forest

STM Société de transport de Montréal

CHAPTER 1 INTRODUCTION

Sustainable development, an approach which ensures meeting the current human needs without sacrificing or endangering the future of others on three main levels (economic, environmental, and social), is one of the main goals for all major and leading cities around the world. In Montreal, achieving this goal requires many strategies, with the reduction of greenhouse gases and car use being primary examples. The Metropolitan Plan for the Planning and Development of Greater Montreal (PMAD) has set the objective of increasing the modal share of public transportation in peak hours in the morning, currently at 26% according to the 2018 Origin-destination survey (ODS), to 35% by 2031 (Montréal, 2019). And while this goal could be unattainable due to the still largely unknown COVID-19 pandemic effects, the goal to achieve better transit mode share in order to reduce car use remains a long-term goal in Montreal (Montréal, 2019).

In this context, having the right planning and modelling tools is essential to test and predict which variables and elements play key parts in the mode choice decisions travellers make. Mode choice models therefore are a critical step in transportation planning (Sicotte, 2014).

1.1 Context

Mode choice models have been used for quite some time. However, many details define the models and make their performance and results differ. First, mode choice models are usually based on either: 1) individual trips or 2) trip chains (tours) (Hasnine & Nurul Habib, 2021). Although trip chain-based models have shown promising results when compared to trip-based models, trip-based models remain by far more common in the literature and in practice as they are less complex and easier to formulate.

Second, the type and the approach of the model play a big role in its performance. There are dozens of different types used all with their pros and cons such as the multinomial logit model (MNL), probit models, generalized extreme value models, and nested logit models. In the last few years, machine learning mode choice models such as Support Vector Machine (SVM) and random forests have also been developed with very impressive results (Cheng, Chen, De Vos, Lai, & Witlox, 2019; Hagenauer & Helbich, 2017). However, there has been very few attempts to utilize these machine learning algorithms in a context to model mode choice at the chain level, with recent studies

mentioning the importance of incorporating trip chaining variables in the model and calling for such application in future studies (K. Kim, Kwon, & Horner, 2021).

Finally, the independent variables used play a huge role in the performance of mode choice models. The variables can be divided into many levels (socio-demographic, built environment, household, etc..) and be included in many types (nominal, ordinal etc..). Variables that consider the type and number of chains a person will do in their day are rarely proposed in the literature, although there is strong indications that show the interdependency between trip chaining and mode choice (Schneider et al., 2021).

1.2 Research objectives

Given the previous context, the main objective of this research is to a have a better understanding of mode choice by considering trip chaining. Within the main objective of the research project, there are different objectives:

1) Conduct an exploratory analysis of trip chaining trends over the years 1998, 2008 and 2018 in Montreal.
2) Identify the determinants of trip chain mode choice and explore the relationship between the explanatory variables and chain mode choice.
3) Test how well the machine learning algorithm "random forest" functions as a tool to model mode choice at the trip chain level.

1.3 Document structure

After the introduction, the document is divided into five main parts. First, the literature review where some of the previous works on trip chain modeling and definitions will be exposed. Furthermore, a section will review the mode choice determinants belonging to the four different levels (socio-demographic, household, built environment and chain). Then, previous studies relevant to this research will be reviewed as well. Their main topics will focus on previous applications of machine learning algorithms as a tool to model mode choice, and the work done concerning trip chaining in the Greater Montreal Area (GMA). The second section will present the methodology, in which the data and assumptions made in the research are further explained and presented along with data preparation. The following section contains the exploratory analysis for the trip chain trends over the years in the GMA and then the relation between the independent variables and chain mode choice will be examined. Then the next part is about the Random Forest

model that will be used to predict the mode choice at the chain level. The results will be discussed based on the accuracy-performance and the feature importance of the variables used in the random forest decision trees. The final section of this document will return on the major conclusions of this work and present the limitations, along with future methodological approaches that could help further improved mode choice modeling while considering trip chains.

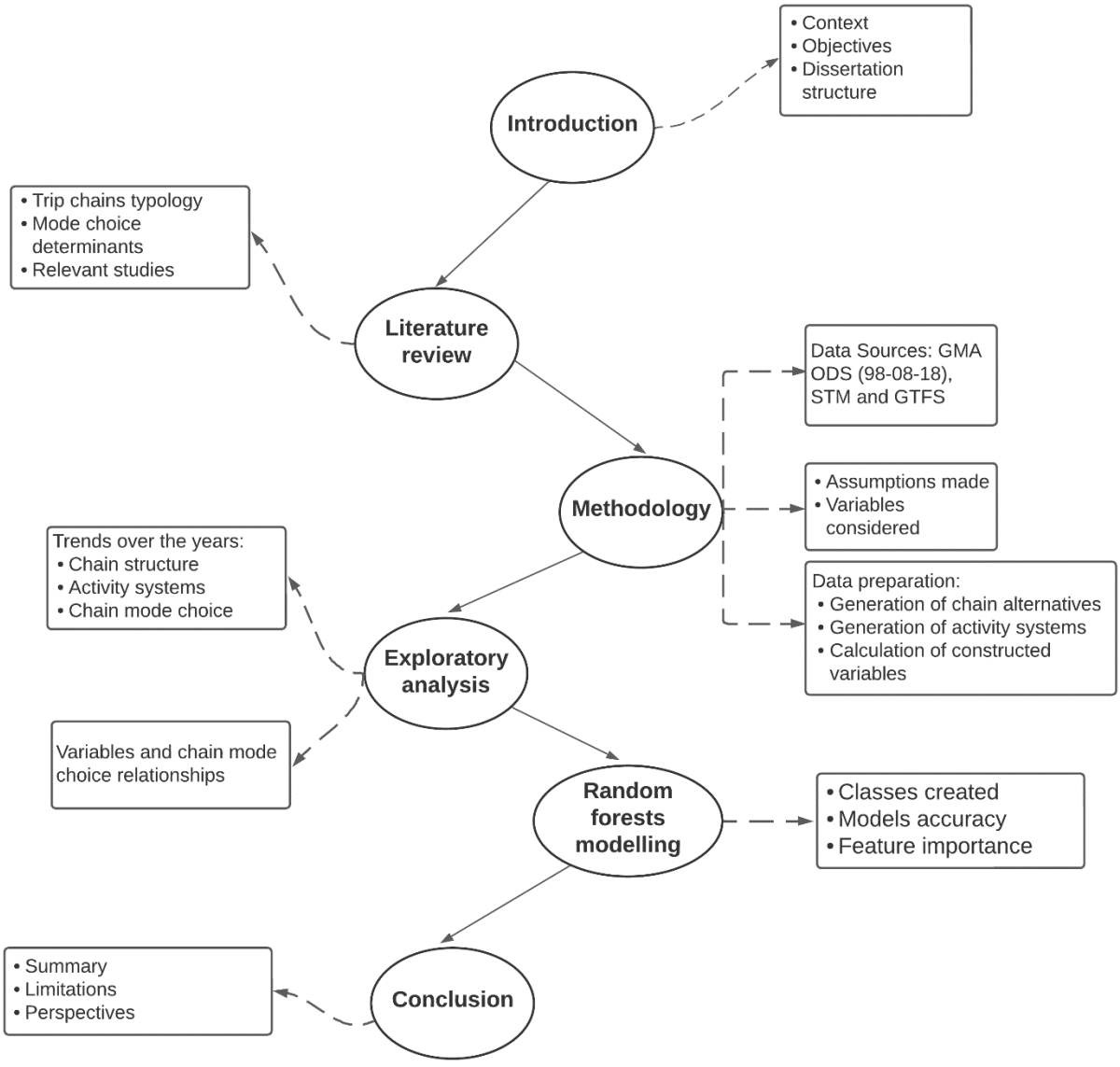

Figure 1-1 Document structure

CHAPTER 2 LITERATURE REVIEW

2.1 Definition of a trip chain

This section will present the various definitions of a trip chain (referred as a tour as well) used in previous studies, from the early first simple definitions, to the one adopted by this work introduced by Valiquette (2010). The underlying concepts introduced by the used definition will also be discussed briefly.

2.1.1 Definitions' evolution

When working at the trip chain level, the way an author defines a chain plays an important role in the eventual studies and results. Throughout the years, the concept of a trip chain has evolved heavily ranging from very simple definitions to very complex ones. Holzapfel (1986) was the first to mention the idea of a trip chain in the literature. He proposed the home of a person as an anchor point (a location with a great importance and influence on all trips within a chain). According to this definition, a chain consists of a sequence of location changes. Nevertheless, for this author, a chain requires a minimum of three segments to be considered as such. Thus, one of the most typical chain structures: the simple chain (Ex: Home-Work-Home) would not be considered a chain for this author.

Another definition followed, by Thill and Thomas (1987), in which the duration of the activities played a big role. For this author, the chain is a function of the duration of the activities involved and not their number, as long as their duration exceeds a predetermined set time. The chain was defined as a sequence of trips between different locations over a given period.

Goulias and Kitamura (1991) were the first authors to introduce the idea of the trip chain purposes as part of the chain definition. They defined it in terms of the number of trips while taking into account the frequency, timing, and purpose of these trips. Chains were then defined based on mandatory purposes (work, study) and discretionary purposes (shopping, leisure …). Anchor points were defined based on the locations of the mandatory activities (home, school, and work), and the chain is thus defined by the sequence of trips between these points.

McGuckin and Murakami (1999) went on to propose a definition in which work, and home are predetermined anchor points. As such, any sequence of trips that starts and ends at one of these

anchor points is considered a trip chain. As such, the sequence of Home-Work-Home for example (which is the most popular type of all chains in the ODS of 1998 to 2018), would consist of two chains: Home-Work and then Work-Home.

In the 2000s, the definition of a chain became more and more consensual between authors. Vleugels, Steenbergen, Vande Walle, and Cornélis (2005) defined a trip chain as a loop starting from and eventually returning to the home and including all trips with their different destinations along the way (e.g., Home-School-Work-Shopping-School-Sitter-Home).

The definition of the trip chain provided by Primerano, Taylor, Pitaksringkarn, and Tisato (2008) is based on a number of ideas mentioned by earlier authors. They defined it as the sequence of primary and secondary trips made between departure and return home, a definition that is most frequently used in recent literature. To support their selected definition, the authors provided three separate justifications. First, the majority of first daily trips begin at home, and the majority of last daily trips end at home, according to the 1999 Metropolitan Adelaide Household Survey. Second, a chain beginning from home means that various choices must be taken in advance, and that those will have an impact on the entire chain until the return home. After the initial trip, the available modal options will unavoidably shift and, more than likely, get reduced as the chain develops. Third, this definition of the trip chain places more emphasis on the trips that connect the activities than the trips themselves. This is in line with how demand forecasting activity models define the trip as a demand derived from an activity.

Several recent pieces of literature considering trip chains still use Primerano et al. (2008) definition, with some recent examples being the work by Schneider et al. (2021) who worked on complex chains and their relation to the daily mobility patterns of people, and the work of Huang, Gao, Ni, and Liu (2021) that analyzed the relation between trip chain patterns and travel mode choice in a case study in Shanghai, China.

The definition adopted in this work is the one proposed by Valiquette (2010), and later adopted in the works of Sicotte (2014). It is not so different and is inspired by the definition proposed by Primerano et al. (2008). The major difference between the two typologies is that Valiquette (2010) typology does not consider the order in which the secondary activities occur within a complex chain, while Primerano et al. (2008) had 5 different types of complex chains depending on the order of the secondary activity within it. An example to show the differences is two complex chains:

Home-Shopping-Work-Home, and Home-Work-Shopping-Home. For Valiquette (2010) these two chains are defined as complex chains and are no different in their characterization (both three trips complex chains with work as primary activity), while for Primerano et al. (2008) these two chains are different and labelled as: complex with secondary activity before primary one and complex with secondary activity after primary one.

2.1.2 Underlying concepts of the chain typology

From the different definitions collected in the literature, several concepts overlap. They are key concepts used by Valiquette (2010), M. Islam (2010) and Primerano et al. (2008) among others, to establish a typology of trip chains. These different concepts are the primary and secondary activity, the anchor point, and the loop.

2.1.2.1 Primary and secondary activities

The primary activity can be defined as the main reason for the trip chain and for any traveller to leave home. Working and studying are usually considered as the primary activities, however when these activities are absent other activities with the longest duration will be considered as the main activity. Secondary activities are then defined as any activities that are performed within the same chain of another primary activity (Valiquette, 2010).

2.1.2.2 Anchor point

According to Valiquette (2010), the anchor point is generally the location of the primary activity in addition to the home location. The most frequent anchor points are the places of residence, work and study.

Since the starting point of a chain is the home and a person will have to return to it since they reside there, it is considered as an anchor point. The location of work and study are also considered as anchor points since these activities are considered necessary and mandatory for most of the population, hence their locations are predetermined within a chain i.e., a person does not have a choice in the location of these activities. The locations of these primary activities offer the chance to perform a multi-loop chain as explained later.

2.1.2.3 The loop

With the definition of anchor points, the concept of the chain loops is created. Chain loops are defined as trips starting from an anchor point and returning to that same point. Loops are differentiated from chains as they are based on any anchor point and not just the home location, even if both concepts are very similar. This is how a chain can contain more than one loop (Valiquette, 2010).

2.1.3 Chain types

Based on all the above-mentioned concepts, several chain types arise and are considered in this research.

2.1.3.1 Simple, complex and open chains

The complexity of the chain is defined by the number of trips within a chain. This variable allows us to split the chains into three main categories. Figure 2-1, Figure 2-2 and Figure 2-3 show the different types of chains, inspired by the work of (Valiquette, 2010).

2.1.3.1.1 Simple chain

These chains consist of only two trips: one to any given activity, starting from home, and then another to return home.

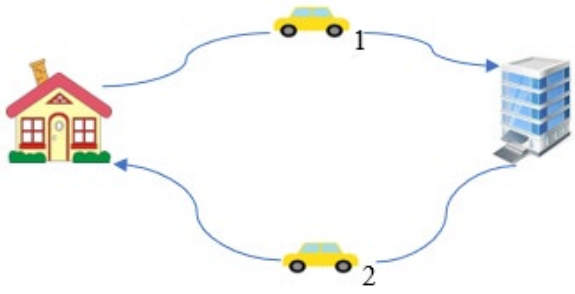

Figure 2-1 Example of a simple chain

2.1.3.1.2 Complex chain

These chains include all chains that consist of more than two trips, i.e., more than one activity. In this case, we find a primary activity and one or more secondary activities to complete the chain.

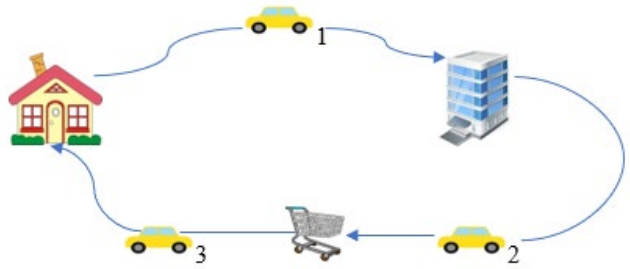

Figure 2-2 Example of a complex chain

2.1.3.1.3 Open chain

Open chains are sequence of trips that either do not start or end at the home location making them outside the typical definition of chains considered in this study. These chains are rarely considered in existing studies of trip chains, and the general focal points of previous studies generally focus on home-based chains, i.e., chains that start and end at home (Primerano et al., 2008).

2.1.3.2 Mono-loop and multi-loop chains

The anchor point consists of the location of the primary activity and the home. The loop corresponds to the set of trips between the departure and returns to the same anchor point. This variable is used to divide complex chains into two categories: mono-loop and multi-loop.

2.1.3.2.1 Mono-loop chain

Mono-loop chains are chains where a person does not return to the same anchor point - home excluded - a second time through their activities making the trip contain one loop only. All simple chains are considered mono-loop chains, while complex chains could be mono-loop or multi-loop. Figure 2-1 and Figure 2-2 are examples of mono-loop chains.

2.1.3.2.2 Multi-loop chain

Multi-loop chains consist of more than one loop and are exclusively complex chains composed of at least 4 trips (2 trips at least per loop). A common example is an individual who leaves home to

go to work and leaves the workplace at lunchtime to go out to eat, then returns to work later and finally returns home at the end of the day.

Figure 2-3 Example of a multi-loop chain

2.1.3.3 Constrained and unconstrained chains

Depending on the primary pattern of the chain, it is possible to group chains into two further subcategories.

2.1.3.3.1 Constrained chain

A chain is considered constrained when the primary activity of the chain is either: “work”, “study” or “driving/picking up someone”. The constraint of these purposes comes from two main points:

1) They are mandatory for a large part of the population.
2) The location of the activities is, for the most part, predetermined, making people have no choice about their trip distance and often their timing as well, unlike for other purposes like “Shopping” and “Leisure” where people can choose the location and timing they want. Driving/picking up someone is also considered constraining as it is the car passenger’s activity that controls the features of the trip.

2.1.3.3.2 Unconstrained chain

A chain is unconstrained when the primary purpose involves a trip that is considered to be non-compulsory. Thus, when the primary reason for a chain is "leisure", "shopping" or "other", the chain is unconstrained as people have control over the location and timing of such chains.

2.2 Explanatory variables of the chain mode choice

Throughout the literature, identifying the explanatory variables or determinants of an individual's mode choice has always been one of the main tasks when attempting any kind of mode choice modeling. These determinants vary depending on the location of the study, time, and between every single individual. Typically, these variables are grouped into four main categories: socio-demographic, household, built environment, and trip/chain variables. The following section describes the effect of the most important variables identified in the literature on mode choice.

2.2.1 Socio-demographic variables

A person's travel behavior is heavily affected by the socio-demographic variables that define them. In the work of Martel Poliquin (2012), where he focused on identifying and better understanding the determinants of mode choice using the date from the Greater Montreal Area Origin-Destination Survey of 2008, he identified two sets of socio-demographic variables: intrinsic such as gender and age, and extrinsic such as main occupation and ownership of a driving license. The alternatives available for any given trip can differ heavily according to these variables. For example, people under 16 cannot possess a driving license so car driving is not an option, and women, in general, could be less tempted to make trips at night via public transit (PT).

2.2.1.1 Gender

In the literature, gender is usually an ever-present socio-demographic variable when considering the impact it might have on the mode choice people make and is almost always used as an explanatory variable in any modeling process. Some key findings in previous works show that men are more likely to use cars and motorcycles, and less likely to use public transport to go to work (Godefroy, 2011; Roorda, Passmore, & Miller, 2009; Xianyu, 2013).

Women were found to conduct more of their trips by walking then men, while also making more child-serving stops in their trip chains, affecting their travel behavior, and performing overall more complex chains (Elias, Benjamin, & Shiftan, 2015).

When compared to men, women performed more shopping trips, and women with children in the household chained these trips before or after work trips (McGuckin & Murakami, 1999).

Another study found that gender did not affect the number of chains performed, but had strong effects on the complexity of the chains (Jing Ma, Mitchell, & Heppenstall, 2014).

Other findings suggest that women in general are more inclined to use pro-environmental modes than men (Briscoe, Givens, Hazboun, & Krannich, 2019; Taru, Kr, & Rao, 2021).

2.2.1.2 Age

Like gender, age is another socio-demographic variable whose impact is heavily studied in the literature on modal choice. With different ages, people's travel behavior is heavily affected. An example of that is how according to Xianyu (2013) the older people get the less likely they are to use public transit. This could be due to several elements such as how physically their energy level lessens with various pains and disabilities or due to how older people tend to go to new destinations at different times, where public transit might not be the best alternative, as transit tends to be the most efficient during peak work hours. A more recent study by Li (2021) shows how higher age can affect the trip chain patterns for people. Indeed, more than half of the elderly in the U.S conduct only one trip chain a day, with more than half of those being simple chains. Otherwise, people aged 14 to 19 tend to use public transport more. Factors such as lack of a transit pass, inaccessibility to a car, and lack of financial resources may explain this observation (Roorda et al., 2009). Furthermore, different ages imply usually different activities and purposes that affect people's travel behavior (Martel Poliquin, 2012).

2.2.1.3 Ownership of a driving license

Owning a driving license is a key individual variable as it is, along with age, a determinant to whether the car driving alternative is available for a person, which is in return the most common mode choice for all trips and chains in Montreal (Xianyu, 2013).

The impact of owning a driving license on trip chaining is a subject of contradictory results. While some studies found that owning a driving license increases trip chaining (conducting more complex chains instead of simple ones) (A. J. Horowitz, 1982; Liu, Sun, Chen, & Susilo, 2018), others found that it decreases it (Lu & Pas, 1999).

Telecommuters conducted more complex trip chains and with higher complexity compared to those who do not telecommute when both own a driving license. However, for those who do not own a driving license there was no significant difference in the trip chaining patterns between

telecommuters and non-telecommuters, which is relevant when considering the increase of telecommuting after the pandemic (Zhu & Guo, 2022).

2.2.1.4 Main occupation

The main occupation of a person heavily affects their travel behavior. It is a strong indicator of a person's primary activity during their day. For example, a full-time worker is most definitely heading to work on the weekdays and a student is going to school. Thus, a person's travel destination and schedule are correlated with their main occupation. Hilgert, von Behren, Eisenmann, and Vortisch (2018) show that a person's main occupation can help predict the variability of their activities, and a change in occupation leads to a strong change in activity patterns. Several previous works show that students are more likely to walk and less likely to drive even when possessing a driving license. Also, being a full time-worker increases the attractiveness of public transit modes (Cirillo & Axhausen, 2002; M. T. Islam & Habib, 2012; Roorda et al., 2009).

2.2.2 Household variables

Features of a household have an important influence on the mobility patterns of the people living in it. Some of these variables include the number of people in the household, presence of children, number of cars owned, income and distance to CBD.

2.2.2.1 Size of the household

The size of the household is defined as the number of people living within the household. It plays an important role in determining the dynamics of the household and its typology. The car driving/ car passenger modal choices seem to increase with the increase in the number of adults in the household (Cirillo & Axhausen, 2002; Sicotte, 2014; Yun, Liu, & Yang, 2011).

2.2.2.2 Number of cars owned

According to several studies, the greater the number of cars in the household the greater the chance an individual will use this mode, this same logic is used for bicycles (Sicotte, 2014; Xianyu, 2013; X. Ye, Pendyala, & Gottardi, 2007).

Households where higher number of cars are owned were found to have a larger chance to perform more complex trip chains, and also performed more complex chains when work is the primary activity (Shiftan, 1998).

2.2.2.3 Presence of children

The presence of children under 16, and especially six and under is a key feature in a household. Children do not have independent mobility and therefore require to be driven to certain places, especially at younger ages when they cannot use public transit yet. Therefore, more car usage is observed in households where children who are 6 and under are present (X. Ye et al., 2007). A study showed that the presence of children in the household has three main effects: more ownership of private cars, a slightly higher use of car as the preferred mode of travel and huge differences in the travel behavior patterns between weekends and weekdays (N. Ye, Gao, Juan, & Ni, 2018).

2.2.2.4 Distance to CBD

The distance to CBD implies usually easier access to several key transit modes such as the metro, along with more bus stops and shorter walking time to access key activity hubs. A study done in Washington D.C showed the distance to CBD as the explanatory variable with the highest relative importance in the machine learning algorithm (gradient-boosting decision trees) mode choice model used (Ding, Cao, & Wang, 2018).

2.2.2.5 Income

Households with low to medium income were shown to be less likely to use car than households with high income, which can be explained by how the ownership of cars is correlated with income. Similarly, households with higher income are less likely to walk or use public transit (Martel Poliquin, 2012; Xianyu, 2013).

2.2.3 Built environment variables

2.2.3.1 Population density

Population density is defined by the number of people per unit area. The effect of density on trip chaining is not universally agreed. While some studies found that land use and in particular high density lead to an increase in trip chaining and chain complexity (Frank, Bradley, Kavage,

Chapman, & Lawton, 2008; Maat & Timmermans, 2006), others found that lower densities could lead to more trip chaining, especially when considering people above 65 (Noland & Thomas, 2007; Schmöcker, Su, & Noland, 2010). High density areas were found to be associated with higher use of cycling and walking, and less usage of the car (Chica-Olmo & Lizárraga, 2022).

2.2.3.2 Distance to nearest metro station

The access distance to the nearest metro station is an important built environment variable. It is usually defined as the walking distance to the nearest access point of the metro. In the work done by Cheng et al. (2019) the distance to metro was the 4th most important variable in the random forest model used. Furthermore, it was shown that people living within 800m of access to public transit have a higher transit modal share than those who do not (Litman, 2012).

There is no agreed upon walking access distance for the metro. However, the range in the literature varies between 400 m-1400 m. In this study, 700 m is considered which falls between the average of 608 m average walking distance to the metro by (Olszewski & Wibowo, 2005) and the 756 m found in a survey done by Schlossberg, Agrawal, Irvin, and Bekkouch (2007) for North America and Europe.

2.2.3.3 Distance to nearest train station

The distance to the nearest train station is calculated using the car driving distance from the household to the nearest train station available. The importance of this distance lies within it falling in the catchment area of the train stations particularly the “Park-and-Ride” catchment area. The calculation of this area is a rigorous and complex process as each train station has different specifications and variables. In fact, several studies have tried to develop a method to calculate the catchment areas of train stations (Duncan & Christensen, 2013; Lin et al., 2016; Ortega, Hamadneh, Esztergár-Kiss, & Tóth, 2020).

Limtanakool, Dijst, and Schwanen (2006) found that the presence of a train station was associated with higher use of commute and less usage of car. This was especially evident for work purpose trips where the modal split for car was 91.8% vs 8.2% for commute when no train station was available, while these values were 65.1% for car and 34.9% for commute with the presence of a train station. Leisure trips were found to be less effected by the presence of a train station than work trips. Another study conducted in Chicago by Lindsey, Schofer, Durango-Cohen, and Gray

(2010) found that closer proximities to train stations increased the use of the train, bus and walking modes while it decreased the use of privately owned vehicles.

It was found that the most important factor to start a chain that includes the train is its distance from the home of the traveler (Souza, Bodmer, Zuidgeest, Brussel, & Amer, 2010). In the debate of what comes first: trip chaining or mode choice, it was found that for great distances to train stations the process of picking a station comes first before the decision of mode choice, effectively making the chaining of the trip coming before the mode choice decision (Chakour & Eluru, 2014).

Compare to people living within 500 m radius of a train station, people living within 500-1000 m used the services 20% less, while people within 1 km to 3.5 km used it 30% less. For people living more than 3.5 km away, the usage of the services was 50% less than for those living within 500 m (Keijer & Rietveld, 2000).

2.2.3.4 Intensity of transit service in 24h

The intensity of transit service is defined as the summation of transit stops made by metro, buses, and trains within 500 m of a household in 24 h. In the work of Martel Poliquin (2012) this variable was significant in the decision tree model as one of the main splitting criteria. This variable is a strong indicator of accessibility to transit from a given household.

Higher transit service was shown to increase the use of transit and walking, especially when considering non-work chains (Cervero & Gorham, 1995; Lee, He, & Sohn, 2017).

Areas with higher transit service appeared to have high “efficient trip chaining” which means complex trip chains in which a trip is done by walking to connect two close activities (Sabouri, 2021).

2.2.4 Trip chain variables

When considering the variables that have the greatest influence on an individual’s mode choice, the ones associated with the trip and the chain are likely to be the most important ones. When considering the trip level only, variables such as travel time, trip distance, purpose, and parking are some of the prime examples of the most discussed variables in the literature. However, when considering the mode choice of the entire chain, some other variables that are normally not considered at the trip level play an important role in mode choice decision-making. One such

example is variables related to the chain structure such as chain complexity, while other chain-related variables are constructed based on calculations of trip-level variables such as the chain distance. Since this study focuses on mode choice at the chain level, below are some of the most important chain variables discussed in the literature.

2.2.4.1 Trip chain distance

Chain distance is defined as the total distance travelled in any given chain. It is the summation of all individual trips' distances in one chain. This variable is one of the most important when considering the effect it has on mode choice. Xianyu (2013) showed that cycling appears the most for chains less than 1.5 km long. According to the work of Sicotte (2014) in the GMA based on the 2008 ODS, shorter distances of the chain (0-2 km) were associated with much higher use of walking/cycling and lower use of car, while the use of PT is almost completely absent. It is only after the chain total distance is more than 4 km that we begin to see a an increase in the modal share of PT and car. Valiquette (2010) on the other hand showed how other variables affect the chain total distance. Men in general were found to preform longer chains, while the possession of two or more cars also increased the chain distance. Being a student or a part time worker seemed to lead to a decrease in chain distances. It is important to note that the exact chain distance varies depending on the mode used for each trip, as different modes have different routes with different distances.

Another study found that short distance chains (less than 3.2 km) that used transit were more likely to be complex when compared to ones that did not have transit as one of the modes (Lee et al., 2017).

Longer chain distances were found to affect the vehicle choice, and be associated with higher complexity chains, and ones with multi-passengers (Paleti, Pendyala, Bhat, & Konduri, 2011).

2.2.4.2 Trip chain duration

On the individual trip levels, the duration of travel or travel time has always been one of the most influential criteria for mode choice decision, and is one the most commonly used variables to express non-monetary cost in generalized cost models. Usually, lower travel times are considered more attractive while higher travel times are perceived negatively by the travelers (Bhat & Sardesai, 2006; Limtanakool et al., 2006), although there is specific cases where higher travel times

where shown to have more attractiveness mainly for long distance trips using the train (Malichová, Cornet, & Hudák, 2022).

When considering the chain level, chain duration is composed of two main components: the total duration of all trips within the chains and the total duration of all the activities within the chain. It is basically the time spent from the moment a person leaves their house for a chain till the moment they come back. Therefore, this variable cannot be used in the same vein as in the trip level, as it is not reflecting solely the attractiveness of the mode choice. A suggested way to utilize this variable is done by Sicotte (2014), where the chain duration was split into travel time by each alternative and the duration of the activities. The time of travel for each possible alternative for the chain is then calculated. The difficulty of this method comes from the generation of all possible alternatives for each chain where 1000s of alternatives could be possible. (Sicotte, 2014) tackled this problem by considering simple chains only, reducing the number of possible alternatives greatly.

The chain mode choice seemed to be effected the most for chains lasting between 6 and 9 h, where a noticeable drop in car driving and higher use of public transit is observed (Valiquette, 2010). Other studies tried to understand the relationship between travel time and activity duration and see which effect they have on each other. They found that higher travel times resulted in higher activity times as well, prolonging the entire chain duration (Lu & Pas, 1999; Zhu & Guo, 2022).

When studying the effects of past trips on current trips within a chain, a study found that the shorter a leisure trip duration was, the longer the trip after would be (June Ma & Goulias, 1998).

2.2.4.3 Duration of activities

The duration of activities corresponds to the time spent at a given activity in a chain after completing a trip, and before starting the next trip in the chain. This duration can play an important part in the chain mode choice decision. An example for that could be a woman that takes the car in the afternoon if she knows the activity duration is long and will continue until late at night, making the public transit less attractive (Gardner, Cui, & Coiacetto, 2017; S. Kim, Ulfarsson, & Todd Hennessy, 2007). It was also found that a shorter duration for activities other than “work” and “study” seem to favor the use of public transit and car passenger modes, and that a longer activity duration lead to the Bimodal mode being less attractive (Sicotte, 2014).

Women were found to have on average 1.32 times longer duration of activities that involve shopping then men, and also to have longer secondary activity duration within a complex chain where the main activity is work (Niemeier & Morita, 1996).

2.2.4.4 Trip chain complexity

The complexity of a trip chain refers to the number of trips performed within a chain. Several chain-based mode choice models have used this variable as a determinant of mode choice (Martel Poliquin, 2012). However, its use remains questionable as the dilemma of whether trip chaining precedes mode choice, or mode choice precedes trip chaining remains discussed and argued about in the literature. Recent findings seem to agree that it is in fact trip chain structure that comes first and not the other way around. That means that people are likely to decide which mode to take depending on the trip chain they have in mind (Krygsman, Arentze, & Timmermans, 2007; Schneider et al., 2021; X. Ye et al., 2007). Although some other authors remain adamant that these two could be simultaneous decisions and in some cases mode choice will precede trip chaining (M. T. Islam & Habib, 2012).

Complex trip chains were found to increase peak time demand, while also encourage the usage of a car due to the flexibility it could offer during chains with many stops (X. Ye et al., 2007).

Another study compared the spatial context of various regions/cities (Canada, United States, Netherlands, United Kingdom and Japan) and their effects on a chain's complexity. The spatial context seemed to have little to no effect on the chain complexity, with the exception of suburban areas having a very slightly higher tendency to chain more trips together and conduct chains with higher complexity (Timmermans et al., 2003).

It was indicated that the more a chain is complex the more likely an individual is to choose a car and less likely to choose public transit for their trips (Ho & Mulley, 2013).

2.2.4.5 Trip chain purpose

The chain purpose is defined by the primary activity present within a chain, as defined earlier. Even if several other trips with other activities are present, the primary activity is always considered as the chain purpose (Valiquette, 2010).

Chains with the purpose of work and shopping were found to be less complex than ones that had other purposes (Krizek, 2003).

When comparing work chains (where the main activity is work) vs non-work chains, it was found that mode choice prediction is a lot more inaccurate for non-work chains, suggesting that non-work chains have a higher diversity in mode choices (Yun, Chen, & Liu, 2014). In the GMA in 2008, chains where the main purpose was driving/picking up someone and work showed the highest modal split for car driving, while the study purpose showed the highest use of the PT mode (Sicotte, 2014). However, modal shares largely differ between cities where different factors are present, such as Amsterdam where the modal share for cycling is highest for work trips, and Tokyo where PT has the higher modal share for work (Rodrigue, 2020).

2.2.4.6 Activity systems

Activity systems are defined as the number and type of chains done by a person in one day. In his work, Valiquette (2010) studied the activity systems of people aged between 25-44 in the GMA based on the OD of 2003. His findings showed that the one chain only per day system is the one with the highest use of public transit and the lowest use of car driving. Valiquette also showed how several socio-demographic and household variables can affect the activity system of a person. Part-time workers were found less likely to do one simple chain per day than full time workers and that presence of children between 0-16 years old in the household decreased the chance of that activity system as well.

2.3 Relevant previous work

It is clear that throughout the literature, many previous works focused on mode choice modeling through various scopes and approaches. However, the implementation of trip chains in the models remains rare, especially when considering complex chains and their complex alternatives. In this section, previous works and studies that focused on the use of machine learning algorithm as a mode choice modeling tool are reviewed and discussed. Finally, recent works on trip chains that were conducted in the GMA are discussed. This helps give more context to this work as it is set in the same area and utilizes the same ODS data.

2.3.1 Machine learning algorithms as mode choice modelling tool

Throughout the years, many tools have been used to model the mode choice process of travelers. Random utility maximization, a theory that considers the traveler as an agent that considers his actions and the surrounding variables in order to produce the decision with maximum utility, is one of the most used approaches in mode choice modelling. MNL models are an example of said theory, and have been extensively used as one of the main ways to model mode choice ever since they were introduced in the works of (Stopher, 1969) and (McFadden, 1974). Other models such as probit and nested logit are also used in predicting travel demand and are based around random utility maximization theory (Hensher & Ton, 2000; J. L. Horowitz, 1993; Koppelman & Bhat, 2006; Van Can, 2013).

In recent years and with the increase in computational power, there has been remarkable emergence of machine learning algorithms as a new tool to model mode choice. Karlaftis and Vlahogianni (2011) explored the idea of statistical methods vs neural networks in transportation research. They concluded that in general, neural networks offer more flexibility especially when dealing with complex datasets when compared to statistical methods, and although they are harder to interpret, they usually offer better predictive powers.

Several studies tried to compare several machine learning algorithm with statistical methods, mainly MNL. Sekhar, Minal, and Madhu (2016) compared MNL and RF in a context of mode choice analysis based on 5000 travel diary samples from Delhi. The results showed that RF had a prediction of 98% compared to 77% for MNL. Similarly, Cheng et al. (2019) found random forests performing the best when it comes to both prediction accuracy and computational runtime when compared to three other models: MNL, SVM, and adaptive boosting (AdaBoost). The relative importance of each variable was also evaluated and showed that travel time (10.1% relative importance) is considered the most important variable among 20 variables used, while the presence of a child in the household was least important (0.8%). Other studies showed similar results when it comes to the prediction power of RF, but found that variables such as travel time and waiting time have a higher importance (Mohd Ali, Mohd Sadullah, Abdul Majeed, Mohd Razman, & Musa, 2022). Hagenauer and Helbich (2017) compared six different machine learning classifiers along with the MNL and found that RF is also the highest performing classifier in terms of accuracy while trip distance was the most important variable for all classifiers (travel time was not

considered as a variable due to the same causes mentioned later in this study). For all the previously mentioned studies, only trips at the individual level were considered. There have been extremely rare attempts to utilize machine learning algorithms and specifically RF at the trip chain level. K. Kim et al. (2021) utilized RF to determine the effects of built environment variables on mode choice and showed that they had the most effect on the decision process of elderly people when they want to take the subway. At the end of the study, it was suggested that trip chaining variables can also play an important role in the mode choice process.

2.3.2 Works in the GMA

Several previous works conducted in the context of the GMA serve an important role in the buildup for this study. Martel Poliquin (2012) examined the relationship between independent variables on all levels and mode choice. He mostly used the data from the 2008 OD survey and utilized it to build several constructed variables that were critical for the study. Of his constructed variables, two are used and evaluated using the same method as in this study: access to a car, and the intensity of transit services over 24 h. Furthermore, he used decision trees to model the mode choice at the trip level using many determinants, where the results found showed impressive accuracy of almost 75%. In many ways, this research continues the work of Martel Poliquin (2012) by utilizing some of the variables that were found to be important in his study. Another important study relevant to this one is the one presented by Sicotte (2014). In his work, he also adopted the typology of trip chains suggested by Valiquette (2010). Explanatory variables were categorised into four main categories: personal, household, trips, and the built environment. These variables were characterized based on their effects on mode choice through descriptive analysis and then later used in a multinomial logit model, where a trip-based model was compared against one that considers trip chains. A typology of mode choice was adopted that splits different modes into two main groups: with an anchor and without an anchor. This allowed for the generation of different alternatives depending on the presence or absence of anchors. The MNL model results showed that the model considering the trip chains produced a better prediction than the one considering trips alone (72.15% simulated were observed vs 12.74%), especially for the car driving case (89% vs 11%). For the models developed in this research, only simple trip chains were considered, and the study was done using the trips of the Vaudreuil-Hudson commuter train line in the Greater Montreal Area.

2.3.3 Research gaps

Throughout the literature, two main gaps are identified that this research tries to tackle. First, while the usage of machine learning algorithms and especially RF is gaining popularity and showing impressive predictive results in mode choice modelling, there is still very few attempts if any to utilize this tool at the chain level. Second, throughout the work of Sicotte (2014), complex trip chains were not considered. Other gaps include the lack of trip chaining variables in the previous works. This research proposes to address these gaps by utilizing the RF tool for trip chain modelling in the GMA, while also incorporating trip chaining variables such as chain complexity and activity systems.

CHAPTER 3 METHODOLOGY

This section's objective is to present the data used in the study, the preparation process, some assumptions taken and to explain the overall methodology applied leading to the development of a random forest mode choice model incorporating trip chains.

3.1 General methodology

For this study, data from the ODS of the GMA from 1998, 2008 and 2018 are used along with some data from the STM and Statistics Canada. Several variables are then created based on the existing data through some manipulations and calculations. A generalized categorization for the different chain mode alternatives and people's activity systems is established.

Next, descriptive analysis is used to show the trip chaining trends throughout the years, along with the relationship between chain structure variables and chain mode choice. The incidence of the various variables on the chain mode choice is observed in order to have a general understanding of how each of these elements interact with one another.

Classes are then created within several variables and used in several chain mode choice models powered by the random forest machine learning algorithm. The models are then compared and the relevancy of each variable is viewed through MeanDecreaseAccuracy and MeanDecreaseGini feature importance. MeanDecreaseGini measures the average gain of purity by splits of a given variable. It is calculated based on the mean decrease in Gini coefficient (the degree of variation or inequality represented in the dataset) after each split in the decision tree. The importance of the variables is measured by which ones tend to split mixed labeled nodes into pure class nodes the most. MeanDecreaseAccuracy measures the increase in the model's prediction error after permuting a variable. A variable is considered of more importance if shuffling its values increases the model out of bag error.

3.2 Data used

The data used in this research comes from several main sources: the GMA origin-destination surveys, the Canadian census, and STM data.

3.2.1 Origin-Destination surveys

The data used is primarily from the 1998, 2008, and 2018 GMA Origin-Destination Surveys. Every five years since 1970, the ODS has been conducted among households to collect the travel habits of GMA residents. The ODS, conducted by the transportation authorities in the Montreal region, provides details of all trips made by about 5% of the GMA population, in addition to place of residence, origin, and destination of trips. The origin-destination (OD) matrix, developed from this survey, allows, among other things, to simulate the travel time required from the various existing modes of transportation. The goal of using several data throughout the years is to observe the evolution of trip-chaining trends in the GMA, while comparing several key variables.

Through the work of Valiquette (2010) a code was created that permits the calculation of several key trip chains data, and the creation of trip chains typology discussed earlier in the research. The ODS data used in this study had the code already applied in them. The code was originally made for the 2008 ODS, and then re-applied for several past years and following years. This code allows the creation of several trip chains variables such as chain purpose, chain duration, number of places visited in a chain, type of chain (simple, complex), and whether the chain is constrained or not among other things. The process of generating the variables for this study is explained a bit later in this chapter. Furthermore, the trip travel times were also provided by the Chaire Mobilité research team.

In this work, some trips were omitted from the study. First, all open and undetermined chains were removed as they can prove to be a problem in the modeling process, since they do not fall under the established definition of a trip chain used in this study, that states that a trip chain must start and end at home.

Also, this research only considers data for people aged between 15 and 64, as they represent the portion of the population that has the most trips (71.5% of all trips) and could offer more flexibility to change modes and travel behavior.

Table 3-1 below shows key insights and a summary for the ODS data observed in 1998, 2008 and 2018. The total number of trips observed signifies the number of trips as reported by the people responding. It is important to note that people who complete several trips or chains within a day are not counted twice. For each individual, a weighting factor is assigned in ODS to help generalize the ODS to the whole population of the region. The people's weighting factor is based on many

variables, such as age and gender. The total number of trips, chains, people, and households (weighted) are higher for the latest ODS in 2018, mostly due to the increase in population observed in the GMA over time.

Interesting initial observation shows us that the average age of the respondents is higher throughout the years, along with the number of cars owned per household, which increased from 1.39 to 1.60 between 1998 and 2018.

Table 3-1 Initial observations of ODS data from 1998, 2008 and 2018

	1998	2008	2018
Number of trips	290,786	237,718	255,817
Number of trips (weighted)	6,086,581	6,056,021	6,916,834
Number of chains	127,010	106,913	111,674
Number of chains (weighted)	2,660,637	2,720,620	3,026,743
Number of people	93,820	86,188	90,536
Number of people (weighted)	1,967,488	2,189,175	2,453,525
Number of households	52,112	48,053	50,490
Number of households (weighted)	1,103,744	1,206,732	1,298,037
People average age	37.17	38.8	39.73
Average number of cars owned per household	1.31	1.43	1.55
Average number of people per household	2.60	2.61	2.64

3.2.2 Census data

The main other sources of data for this research come from the Canadian census. The national census in Canada is conducted by Statistics Canada every five years. It is usually conducted in May. The Census provides demographic and social information that is used widely throughout the country in various research and studies in order to make important decision-making and perform various analyses. When collecting the census, there are several geographical levels divided into two main types of areas. For this study, the dissemination areas known as Das are used. A dissemination area (DA) is a small, geographically defined area used by Statistics Canada to collect and disseminate statistical information. Dissemination areas are the smallest standard geographic areas for which census data are produced. Dissemination areas are used to provide detailed information about the population and housing characteristics of small geographic areas, such as neighbourhoods, small towns, and rural communities. For this study, the population density was

collected at the DA level and then each household was assigned its corresponding population density based on the DA that it is in, this step is further explained later.

3.2.3 STM data

The STM (Société de transport Montreal) is the main enterprise that integrates and operates the transit network in Montreal. Created in 2002 to replace the previous STCUM (Société de transport de la Communauté Urbaine de Montréal), it operates four metro lines with 68 stations and the bus network going through the Montréal Island. The STM data used in this study are those of the coordinates of metro stations and train stations across the territory. Also, the GTFS (General Transit Feed Specification) files provided by the STM were used. GTFS shows, among other things, the routes, schedule and stops of the transit network in a given area or city (Fortin, Morency, & Trépanier, 2016).

3.3 Explanatory variables

The random forest and the explanatory analysis done later in this study all use several key important explanatory variables of different types. Table 3-2 below highlights these variables, their type, description, and the source of each variable. Variables are grouped into four different groups: socio-demographic, household, built environment and trip variables. All the sources for the variables in question are mentioned below. An important remark to be made is the omission of the total travel time of chain. Due to the nature of Random Forests prediction, providing the travel times in the training datasets is already implicative of the mode choice used, i.e. since the mode of travel is already considered to calculate travel times putting it in the model beats the purpose of predicting the mode choice based on other variables (exogeneous), hence its use is not recommended.

Table 3-2 Explanatory variables selected for the study

Type	Variable	Description	Source
Socio-Demographic	Age	The age of the person doing the trip chain	ODS (extracted directly)
	Gender	The gender of the person doing the trip chain	ODS (extracted directly)
	Occupation	The occupation of the person doing the trip chain	ODS (extracted directly)

	Possession of a driving license	Whether the person doing the trip chain owns a driving license or not	ODS (extracted directly)
Household	Number of people in the household	The number of people in the household of the person doing the trip chain	ODS (extracted directly)
	Presence of children	How many children are in the household of the person doing the trip chain	ODS (calculated)
	Distance to CBD	The distance between home location and the central business district of the person doing the chain	ODS (calculated)
	Number of cars per licensed person	The number of cars per licensed person in the household of the person doing the trip chain	ODS (calculated)
Built environment	Presence of metro station	The presence of a metro station within a 700 m walking distance from the home location of the household	STM ODS (calculated)
	Presence of train station	The presence of a train station within a 2 km driving distance from the home location of the household	STM ODS (calculated)
	Population density	The population density / km^2 within the DA of the home location of the household	Census 2016 (calculated)
	Intensity of transit service	The intensity of bus transit service / 24h in the CT of home location of the household	STM GTFS (calculated)
Chain	Chain distance	The total distance of the entire chain	ODS (calculated)
	Chain purpose	The main purpose of the chain	ODS (calculated)
	Chain modes	The modes used throughout the chain	ODS (calculated)
	Activity system	The number and type of chains done by the person throughout the day	ODS (calculated)
	Chain Complexity	The number of activities within a chain	ODS (calculated)
	Activity Duration	The duration spent at all activities within a chain	ODS (calculated)

3.3.1 Data preparation

The process of data preparation for each different variable is mentioned in this section. While some data were used directly as they were presented in their source, others required data manipulation.

Most notably, the preparation of the chain mode choice and the different activity systems was very important for the model, and several assumptions and coding processes were used.

3.3.1.1 Chain mode choice

The chain mode choice is the most important variable in the study, as it is the dependent variable. Typically, when considering trips individually, the number of alternatives available for a given trip is limited and easily countable. In the case of the ODS 2018, 11 different mode alternatives were observed for the selected trips. However, the issue of alternatives arises when considering trips at the chain level. In his work, Sicotte (2014) tackled the possible alternatives for trip chains by considering simple trip chains only made of 2 trips (11*11=121 possible alternatives maximum), and by using the concept of anchor mode to eliminate an important number of possible alternatives (i.e., when a person drives in their first trip, they must eventually return home with their car). However, in this work even complex chains are considered in the modeling process which complicates the problem of possible alternatives. Theoretically, a complex trip chain made of five trips could have 11*11*11*11*11= 11^5 = 161051 different alternatives, and even more complex trip chains are observed in the ODS. Therefore, a set of simplified alternatives was suggested to tackle this problem. It was observed that for all chains, 91% are conducted using one mode for the whole chain, while only in 9% of chains we do see a mixture of modes throughout the chain. From this point, five classes were created that represent the use of one given mode "only", as it is the only one used for the entirety of the chain and three others that represent a combination of modes. Given that the end goal of the study is to move towards more sustainable transportation, the mixed alternatives were grouped based on whether there is any usage of a private car within a chain or not. The alternatives were created using the data from the 2018 ODS. A python code was created that first grouped all the modes of one chain done by a person together such as: "CD-CD-CD" or "PT-Cycling/Walking-PT" (thus firstly creating all the possible alternatives observed) and then labeling each one according to the assumption above based on the presence of unique modes or mixed modes. This code was later used on the 1998 and 2008 ODS data as well.

The final classes for the chain mode choice can be grouped as followed:

- CD only: A chain in which all the trips were done by car driving only
- PT only: A chain in which all the trips were done by public transit only

- CP only: A chain in which all the trips were done as being a car passenger only
- Cycle/Walk: A chain in which all the trips were done by either cycling only, walking only, or a combination of the two
- Bimodal: A chain in which all trips were done by Kiss-and-Ride or Park-and-Ride or a combination of those modes only
- Mix with car: A chain in which the trips were done by a combination of modes, one or more of which include the use of CD or CP
- Mix without a car: A chain in which all the trips were done by a combination of modes, without the usage of a car (CD or CP, Bimodal included)
- Others Unique: A chain in which all the trips were done by a single mode only that has not been included in previous groups (taxi, school bus, etc. …)

3.3.1.2 Activity systems

Activity systems are an important variable in this study as they are defined as the number and type of chains done by a person throughout the day. This variable is important as it reflects the structure of the chain, and thus when used in the model later, we can see how integrating this variable helps the model performance, and the feature importance of this variable. Like in the case of mode choice, a person can have many possible activity systems, especially when making many chains per day. While not as many as the possible combinations for mode choice, the number of possible activity systems is still a lot, and thus it was decided to only consider five different activity systems. This was done as it turns out that four activity systems form about 96% of all observed ones and the other 4% are spread around many other systems in the 2018 ODS. Again, this process was first done on the 2018 ODS data using python. A code that counts how many chains a person did, and then labels the chain type based on the number of trips within each (2 trips → simple chain, 3+ → complex chains) was used. After this process, the code allocates each person one of the following activity systems based on the results. The five activity systems are: one simple chain, one complex chain, two simple chains, one simple and one complex chain, and “others” which includes all other possible activity systems combinations.

3.3.1.3 Chain Distance

Chain distance was also calculated using the straight-line distance. Each individual trip distance was calculated and then trips belonging to the same chain were summed to have the total chain distance.

3.3.1.4 Population Density

To prepare the population density (people/km^2) for each household's zone of residence, the population density associated with the DA in which the household is located was considered. This was done by getting the population density files for the GMA from the census of 2016, and then importing them along with the GMA Das shapefile and the household coordinates file into QGIS. Based on the ID of the Das, a population density is associated with each DA, and then by intersecting the household information with the GMA shapefile, we get the associated population density for each household. Figure 3-1 shows the different population density throughout the GMA. (Note that due to some households being situated outside the CMA of Montreal, the DA from the province of Quebec and their shapefile were used).

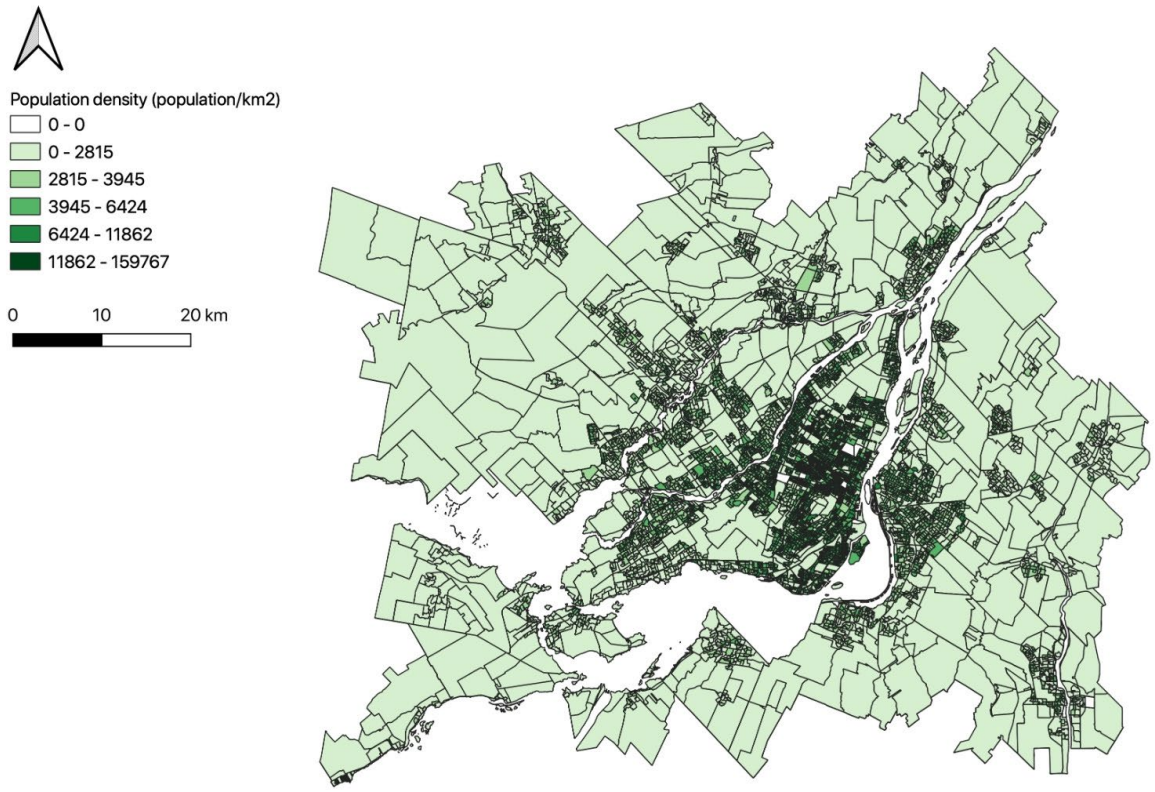

Figure 3-1 Population density based on DA in the GMA

3.3.1.5 Metro stations and train stations

After obtaining the coordinates of the metro stations and train stations from the GTFS files, OpenStreetMaps was used to integrate the GMA road network, and then OpenTripPlanner (OTP) was used to calculate the corresponding distances from the closest station to each household. 700 m was considered an acceptable walking distance for metro access as mentioned in the literature review section, so this was calculated by using the "walk" mode in OTP and calculating the distances between households coordinates and those of the metro stations via a package in R. The presence of metro station within a 700 m distance would be considered as a dummy variable, as to imply that people have access to the metro or not.

Identifying a catchment area for train stations was a bit trickier, as the acceptable distance differs in the literature depending on the city and several assumptions. For this study, a driving distance of 2.5 km from the house to the train station was considered as we are mostly concerned about the

P&R opportunities presented. This was also calculated using the same process but this time by selecting the" Car" mode. Again, a dummy was also used to indicate whether a train station is accessible or not to the person.

3.3.1.6 Distance to CBD

For this variable, the straight-line distance was assumed in the study as people can use different routes and thus different distances, depending on the mode they choose in their chains. Distance from CBD was considered as the distance from the household to the intersection of Peel and Sainte-Catherine streets in downtown Montréal, to keep in line with previous related studies that considered this point as the CBD of Montreal (Sicotte, 2014; Valiquette, 2010). The equation of the distance is as follows:

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

Where $x1, y1$ are the coordinates of a given household, and $x2, y2$ are the coordinates of the CBD.

3.3.1.7 Presence of children

For the presence of children in the household, a code was used that searched each household ID to look for children 6 and under and for under 16 and 3 classes were created: with children 6 and under, with children between 7-15 and without children. If a house had 2 or more children from different classes, the younger class was considered. Younger children are considered more dependant on their parents for any kind of trips they must do, while older children can have a bit more freedom for their mode choice like taking public transit for school for example. Based on this assumption, younger children (6 and under) are more likely to affect the travel behavior and mode choice of their parents' trips than older ones (between 7-15). Therefore, the younger class was considered if two children from different age groups were present.

3.3.1.8 Access to car

For this variable, the same method adopted by Martel Poliquin (2012) was used, where using the ODS data the ratio of cars owned/ licensed drivers was calculated.

Three classes are then created depending on the obtained ratio:

- "Full access" is when the ratio is equal or greater to 1.

- “Partial access” is when the ratio is between 0 and 1.
- “No access” is when the ratio is equal to 0 (no car).

3.3.1.9 Intensity of transit service over 24 h

This is another determinant where the methodology adopted by Martel Poliquin (2012) is used. The intensity of transit service over 24 h is calculated using the GTFS data for the day of October 6, 2018, as it coincides with the ODS done in 2018. It is done in three main steps:

1) Calculating the sum of all transit passages and stops done by all transit modes (metro, bus, train) within 24 h. This is done regardless of directions and different lines at each station, or bus stop.
2) Creating a 500 m radius buffer zone for each household in the 2018 ODS.
3) Intersect all stops and stations within each household’s buffer zone and summing all transit stops done in that zone and assigning it to each household. This will mean including the same bus if it passes in several stops in the same buffer zone.

3.3.1.10 Duration of activities

The duration of each activity was calculated using two steps:

1) Calculating the difference between the times of departure for consecutive trips in each chain. This time corresponds to the total duration of the trip + activity.
2) Then subtracting the travel time of each trip from the time we got in step 1, which gives the duration of each activity within a chain.

   For any given chain, the duration of all activities is simply the summation of the time of all activities done in the chain.

CHAPTER 4 EXPLORATORY ANALYSIS

In this section the aim is to explore the trends of trip chaining behavior in the last 20 years according to the ODS. Exploratory analysis will be done on several key variables that relate to trip chains such as chain complexity, duration, and distance. The chain structure relationship with the chain mode choice will also be examined. The relationship of the explanatory variables mentioned in Chapter 3 with trip chain mode choice will also be evaluated.

4.1 Trends over the years

While it is already known that trip chaining influences mode choice, it is still not known exactly how each trip chain variable contributes to the decision process of mode choice. Given the importance of trip chaining, it is important to examine what kinds of trends are observed throughout the years in the GMA. Table 4-1 shows the different types of chains and their evolution throughout the years. The biggest takeaway from the table is that the percentage of simple chains is lowest in the latest ODS of 2018 at 81% with the highest percentage of complex chains at 19%.

Table 4-1 Chain trends for years 1998, 2008 and 2018

	1998	2008	2018
Total Chains Observed	127,011	106,913	111,674
Total Chains (Weighted)	2,660,658	2,723,690	3,019,465
Simple Chains (% of total chains)	104,299 (82%)	90,663 (85%)	90,144 (81%)
Complex Chains (% of total chains)	22,701 (18%)	16,242 (15%)	21,530 (19%)
Mono-looped chains (% of complex chains)	17,128 (75%)	12,516 (77%)	17,497 (81%)
Multi-looped Chains (% of complex chains)	5,573 (25%)	3,726 (23%)	4,033 (19%)
Constrained Chains (% of total chains)	86,179 (68%)	78,180 (73%)	82,081 (74%)
Non-Constrained Chains (% of total chains)	40,833 (32%)	28,733 (27%)	29,594 (26%)

4.1.1 Chains per person

The first trend to look at is that of chains per person as shown in Figure 4-1. This variable shows the average number of trip chains done by person per day for men and women across different age groups. At first glance, it is obvious that people performed more chains across all age groups in 1998 compared to 2008 and 2018. This could be the result of two elements: 1) more accessibility to services throughout the years making trips not as necessary (it is easier to get to one mall having all kind of services in 2008 and 2018 than it was in 1998, plus the possibility of doing shopping and obtaining other services online) 2) people are leaning more towards making less chains per day in favor of making more complex ones where they get everything done in one chain, as could be seen later on for the 2018 increase in chain complexity. Another reason could be the increase in duration for main activities and travel time, leaving less time to do other activities. For the 2008 and 2018 years, we can see that the two have similar chains per person throughout the age groups with the younger and older groups having slightly more chains per day in 2008 and the middle age groups having slightly more in 2018. When comparing the results at the gender level, we see a similar pattern for all years. Women tend to perform more chains per day than men at younger (15-24) and older (50-64) ages, but perform overall less chains per day in the middle age groups (25-49). This could be since women in the middle age groups tend to perform the most complex chains, making them perform less chains per day in favor of more activities per chain.

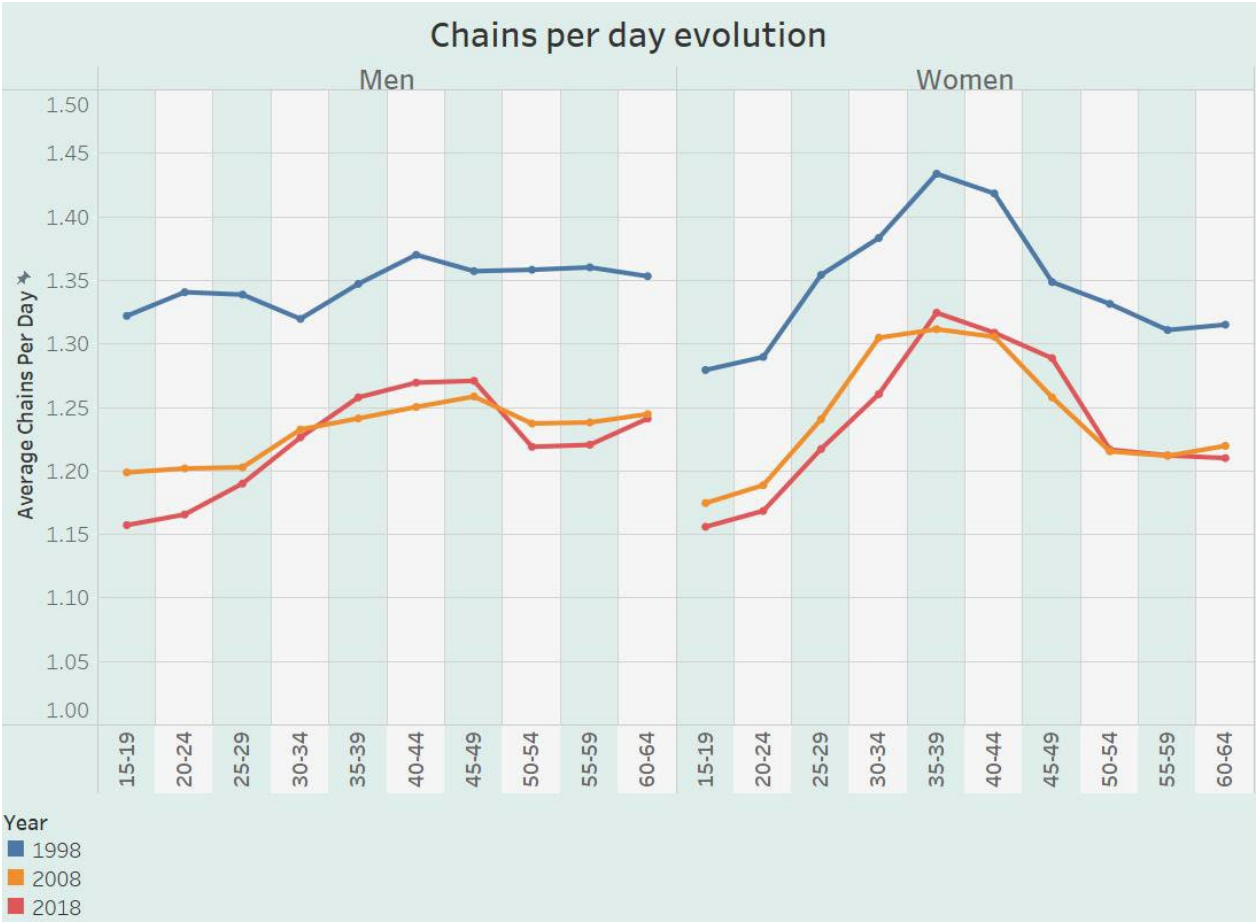

Figure 4-1 Evolution of the number of chains per person per day (1998-2018)

4.1.2 Chain complexity

Chain complexity is one of the most important variables when it comes to chain structure and its effect on mode choice. Chain complexity indicates the number of trips within a given chain. In general, complex chains tend to lead to higher use of the car and lower use of active transportation (Currie & Delbosc, 2011). One main reason is that when conducting a complex trip chain, the alternatives become extremely limited when starting your trip with an anchor point (example: when using a car) as you must make your last trip with it as well. Another explanation can be that car typically brings more flexibility in the stops you can make in a chain. In Figure 4-2, we can see that while 2008 saw a significant drop in chain complexity compared to 1998, 2018 saw a jump in chain complexity especially for women. For men, the average chain complexity is highest in 2018 for the ages 30-44. Women on the other hand are doing more and more complex chains in 2018 than ever before averaging 2.45 trips per chain for ages between 30 and 44. It is also worth

mentioning that women across all years and age groups tend to complete more complex chains than men.

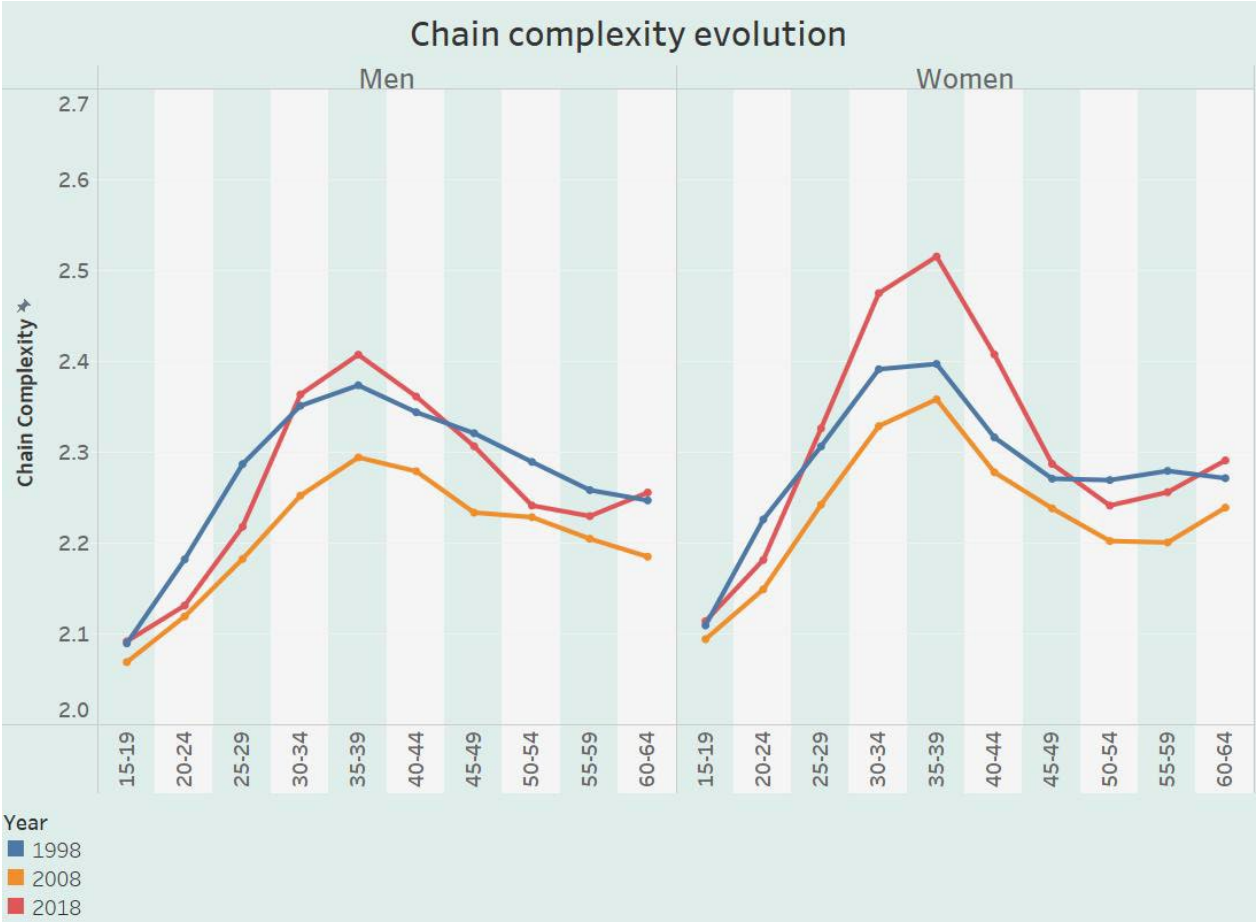

Figure 4-2 Chain complexity evolution (1998-2018)

4.1.3 Chain duration

Figure 4-3 shows the average chain duration across age groups for men and women in 1998, 2008 and 2018. The observed results show that the average chain duration increased from 1998, to a similar level between 2008 and 2018. The highest chain duration average is observed between 25-34 years old for men and between 15 and 24 years old for women. For all years, a sharp decrease in chain duration begins to show after the age of 54. In general, women tend to have shorter average chain duration compared to men, except for the 15-19 years old where they show similar durations, this is likely due to both genders having the same primary activity of going to school.

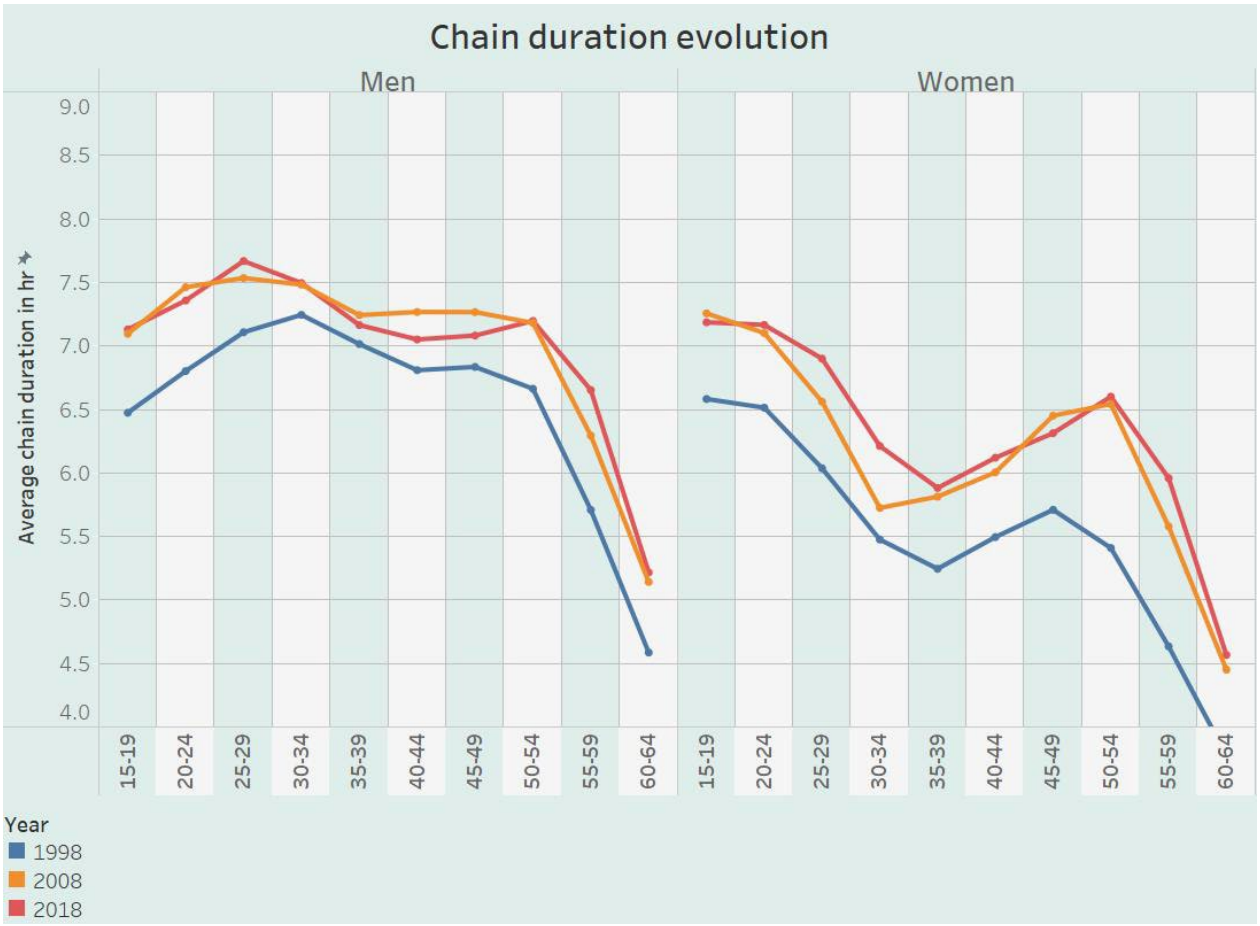

Figure 4-3 Chain average duration evolution (1998-2018)

4.1.4 Chain distance

Chain distance is one of the most important variables when considering mode choice decision. In Figure 4-4, we can see the average chain distance evolution throughout the years with respect to the different age groups and gender. Two main trends can be observed: one is that men continue to do longer trip chains on average across all years for all age groups when compared to women. Second, we can observe that while years 1998 and 2008 have somewhat similar average chain distances, 2018 have higher distances across all age groups for both men and women. One possible explanation can be that of the growth in population in the suburban areas (Institut de la Statistique du Québec, 2022), where longer chain distances are required to get to different services, job locations and to CBD.

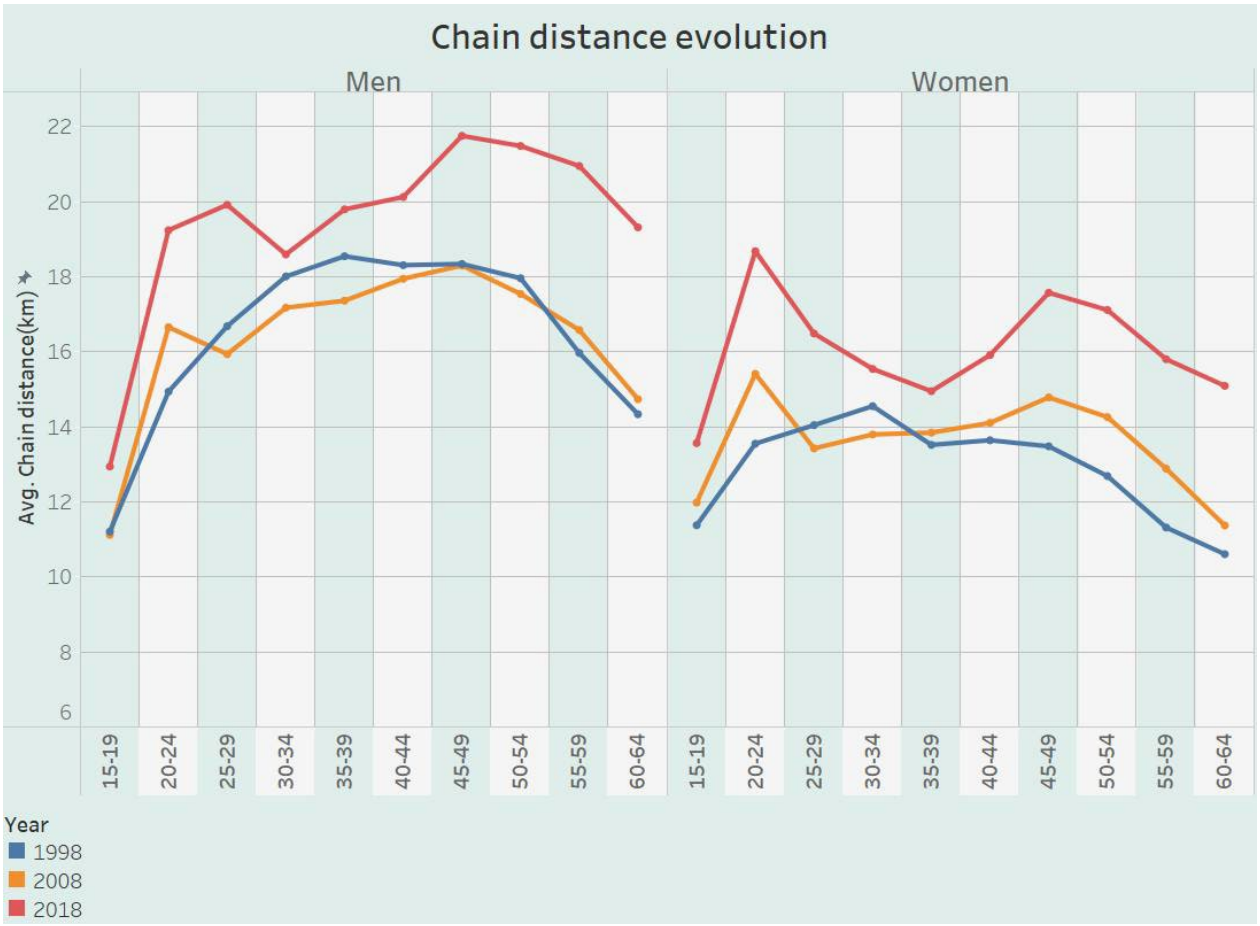

Figure 4-4 Chain average distance evolution (1998-2018)

4.1.5 Chain purpose

The evolution of the trip chains' main purpose throughout the years is shown in Figure 4-5. The chain main purpose is determined based on the primary activity of the chain. A steady increase in the percentage of chains where the main purpose is work is observed, where the value increases from 50% in 1998 to 55% in 2018. This can be due to the increase in complex chains observed before, and that work will always be the primary activity of a trip meaning that complex trips where work and another purpose are present are still labelled as work motivated trips. With the increase of work chains throughout the years, a decrease in shopping trips is observed from 14% in 1998 to 10% in 2018. Several factors can be assumed as the reason such as the previously mentioned increase in complex trips where work and study are always considered main purposes when paired with other purposes and the increase of online shopping options. The increase of women's labour force is perhaps another reason of the increase in work purpose chains. Other purposes do not show

noticeable trends, especially for the purpose “driving/picking up someone” where the percentages are identical at 8.5% for all years.

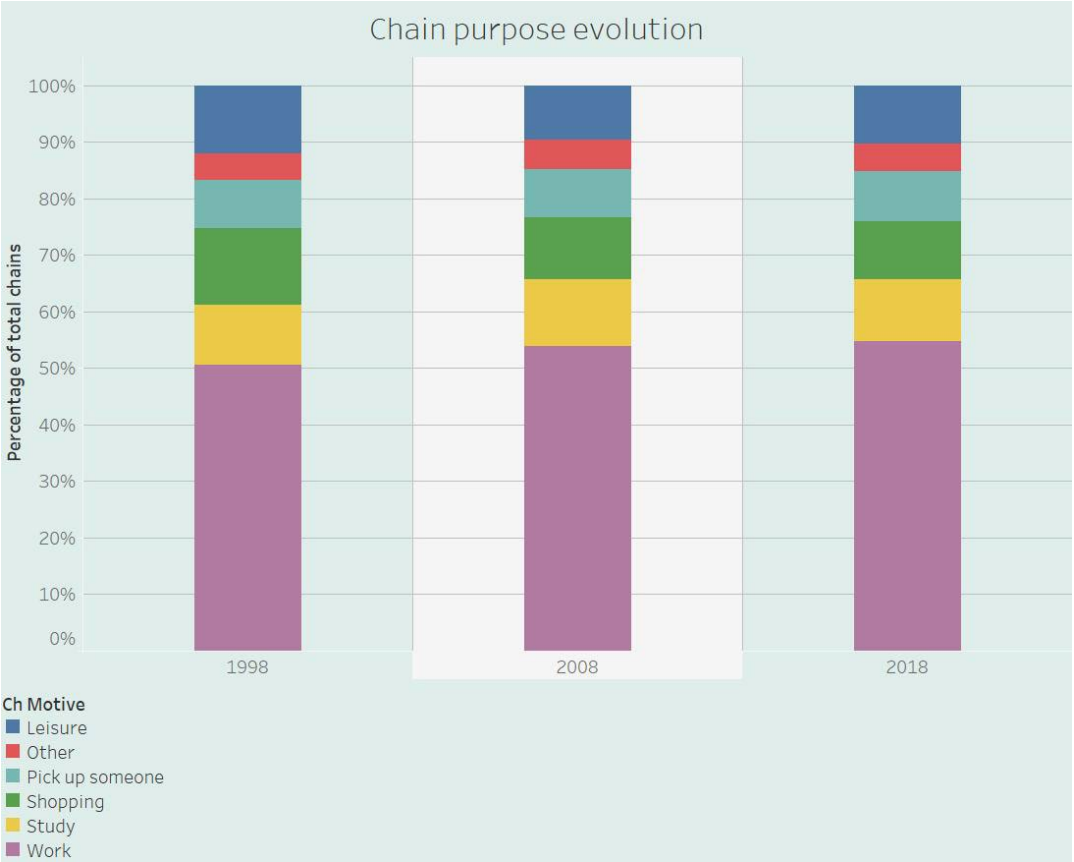

Figure 4-5 Chain purpose evolution (1998-2018)

4.1.6 Chain structure and mode choice trends

With the goal of exploring the relationship between trip chaining and mode choice, Table 4-2, Table 4-3 and Table 4-4, show the relationship between the structure of the chain in terms of complexity and the chain modal split in the years 1998, 2008 and 2018. An obvious trend can be observed in all years, where the more complex a chain is in terms of loops, the more the car is used. This increase in car usage is accompanied with a steep decrease in PT, and Cycling/Walking modal split. In fact, the CD only chain modal split increases throughout the years by an average of 8% for complex mono-looped chains and by 15% for multi-looped chains when compared to simple chains, while the average modal split for modes in which a car is used (CD only + CP only + Mixed with car+Bimodal) increases by 15% for complex mono-looped and by 23% for complex multi-looped chains when compared to simple chains. The modal split for PT only chains, decreases on

average by 11% when going from simple chains to mono-loop chains and by 14% when taking complex-multi loop chains. One positive trend observed in the year 2018 compared to 1998 and 2008, when considering the decrease of car use, is the increase in the Mixed without car modal split. This increase is noticeable in both complex mono and multi-loop chains. This could indicate that people are becoming more willing to explore combining different active modes without the usage of cars when considering complex chains.

Table 4-2 Chain modal split with respect to chain structure (1998)

Mode choice	Simple (104,298)	Complex mono-loop (17,138)	Complex multi-loop (5,574)	Total (127,010)
CD only	59.64%	68.67%	71.47%	61.38%
CP only	9.90%	8.45%	3.53%	9.42%
PT only	13.81%	4.48%	1.38%	12.01%
Cycling/Walking	9.82%	3.65%	2.01%	8.65%
Mixed with car	2.26%	10.11%	17.39%	3.98%
Mixed without car	1.20%	4.57%	4.14%	1.78%
Others unique	2.06%	0.08%	0.07%	1.70%
Bimodal	1.31%	0.00%	0.00%	1.07%

Table 4-3 Chain modal split with respect to chain structure (2008)

Mode choice	Simple (90,663)	Complex mono-loop (12,516)	Complex multi-loop (3,726)	Total (106,913)
CD only	56.20%	66.13%	75.65%	58.04%
CP only	8.00%	6.66%	2.82%	7.60%
PT only	17.66%	6.15%	1.74%	15.76%
Cycling/Walking	10.48%	4.50%	1.99%	9.48%
Mixed with car	2.07%	10.59%	14.12%	3.46%
Mixed without car	1.00%	6.08%	3.65%	1.82%
Others unique	2.40%	0.1%	0.03%	2.05%
Bimodal	2.10%	0.09%	0.00%	1.78%

Table 4-4 Chain modal split with respect to chain structure (2018)

Mode choice	Simple (90,144)	Complex mono-loop (17,495)	Complex multi-loop (4,035)	Total (111,674)
CD only	59.11%	63.54%	71.00%	60.24%
CP only	6.25%	5.30%	2.33%	5.96%
PT only	16.83%	5.24%	1.44%	14.46%
Cycling/Walking	9.66%	5.54%	3.12%	8.77%
Mixed with car	2.00%	11.28%	15.24%	3.60%
Mixed without car	1.66%	9.03%	6.79%	3.33%
Others unique	2.02%	0.07%	0.05%	1.64%
Bimodal	2.48%	0.00%	0.00%	2.00%

4.1.7 Activity systems

Activity systems are an important calculated variable in this work. Individuals are associated to an activity system that shows the number and the type of chains they made throughout the day as explained in section 3.2.1.2. Table 4-5 shows the general evolution of activity systems through the years while Figure 4-6 and Figure 4-7 show the evolution of activity systems for each age group throughout the years 1998 to 2018 for both genders. For both men and women, a general trend can be seen across all age groups: a decrease in the two simple chains only per day and an increase in the one complex chain only activity system. This trend could indicate that people are now preferring to chain their trips especially two simple ones into one complex chain. This trend is extremely evident for women in the year of 2018, where there is a noticeable increase in the complex chain only activity system and a decrease of the one and two simple chains only. In fact, women aged between 35 and 39 are the only group to have one simple chain only forming less than 50% of their activity systems. Another worthy note is the decrease in the "others" activity system from 1998 to 2008 and 2018, where people are now more likely to stick to one of the four most common activity systems.

Table 4-5 Activity systems composition evolution (1998-2018)

Activity systems	1998	2008	2018
1 simple chain	57.08%	66.61%	63.23%
1 complex chain	14.1%	12.93%	16.81%
2 simple chains	17.24%	13.46%	12.27%
1 simple + 1 complex chain	5.56%	3.57%	4.34%
Others	6.01%	3.43%	3.35%

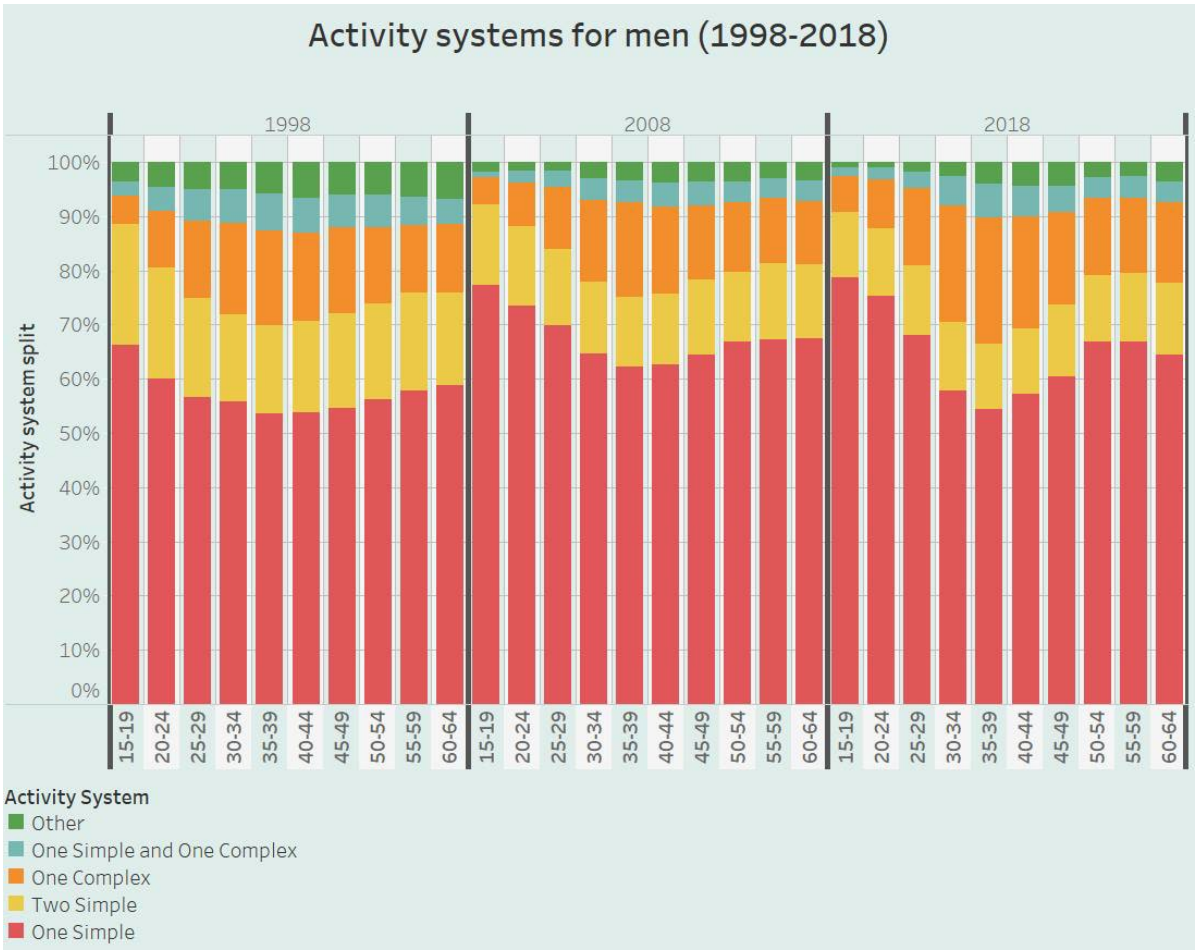

Figure 4-6 Activity systems evolution for men (1998-2018)

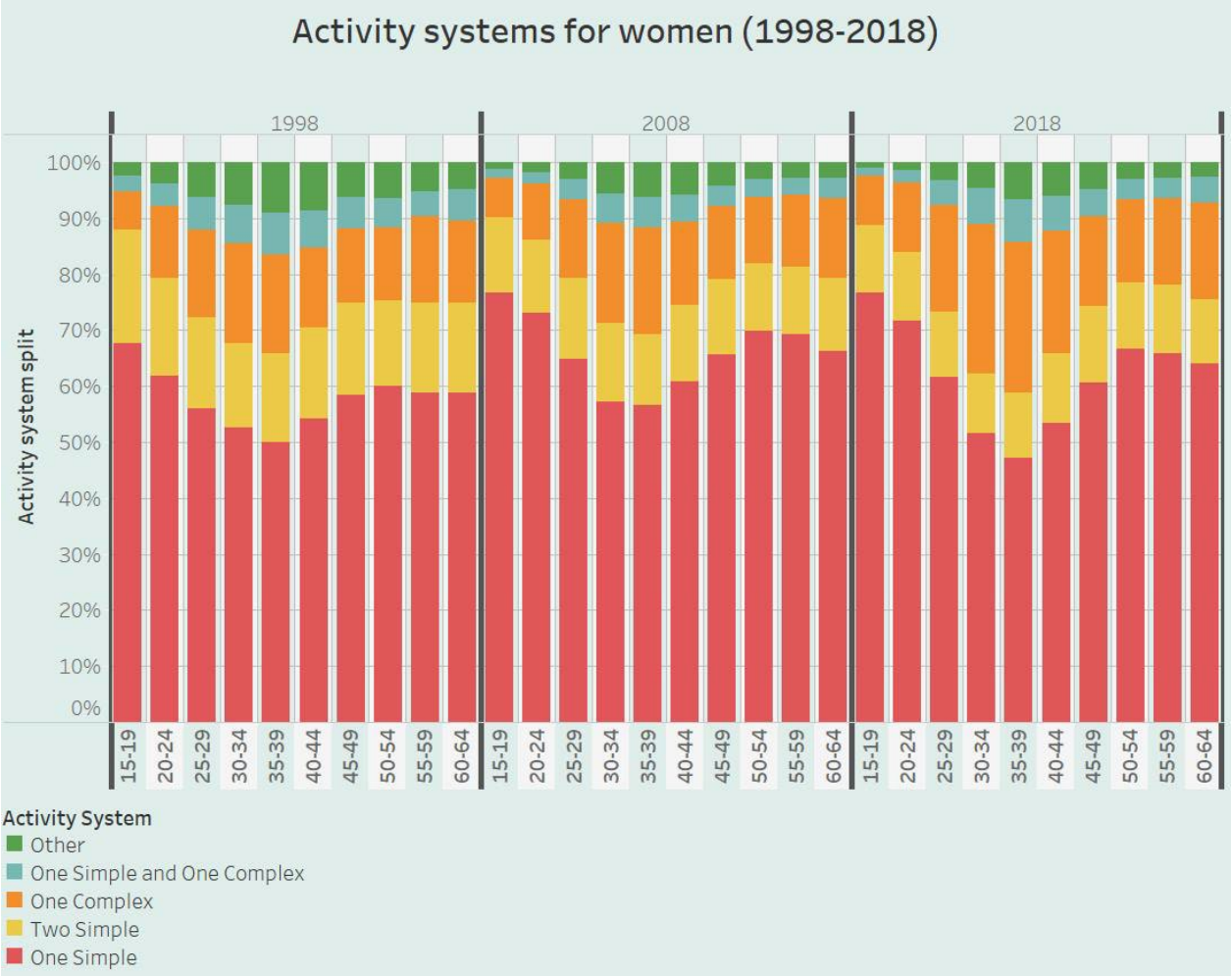

Figure 4-7 Activity systems evolution for women (1998-2018)

4.2 Relation between trip chain mode choice and independent variables

To later integrate the different independent variables in the random forest chain mode choice model, it is important to analyze the relationship each one has with the chain mode choice. To keep in mind is that all these variables are calculated from the latest ODS of 2018, which will be the basis for the modelling process later.

4.2.1 Socio-demographic variables

4.2.1.1 Age and gender

Figure 4-8 shows the chain modal split with respect to age and gender. A lot of trends can be observed considering the mode choice for various age and gender groups. The first noticeable difference is that men tend to use CD only as their mode choice more frequently than women, especially after the age of 49. Women start to use CD only less with the advancement of age, while men use it more. The highest modal split of CD only is 68% for women aged 44-49, while for men it is 75% for the 55 to 59 years old. Another insight is that women use the CP only mode more often than men in all age groups, especially in the older age groups, where women between 60 and 64 have a chain modal split of 15.1% compared to men, at 4.2%, for the same age group. The use of PT only as the chain mode choice follows the same trend for both men and women, as it decreases with the advancement of age. However, it is interesting to note that between the ages of 15 and 39, men have a higher modal split for PT only than women, while from 39 onwards women tend to use the PT only more than men. This could be explained by the increase of CD only for men when they grow older, in contrary of women who tend to use CD only less in favor of other modes like PT and CP. As for walking and cycling only, the modal split is consistent throughout the years for both men and women, except for young ages between 15 and 25 where men have a slightly higher share for active modes.

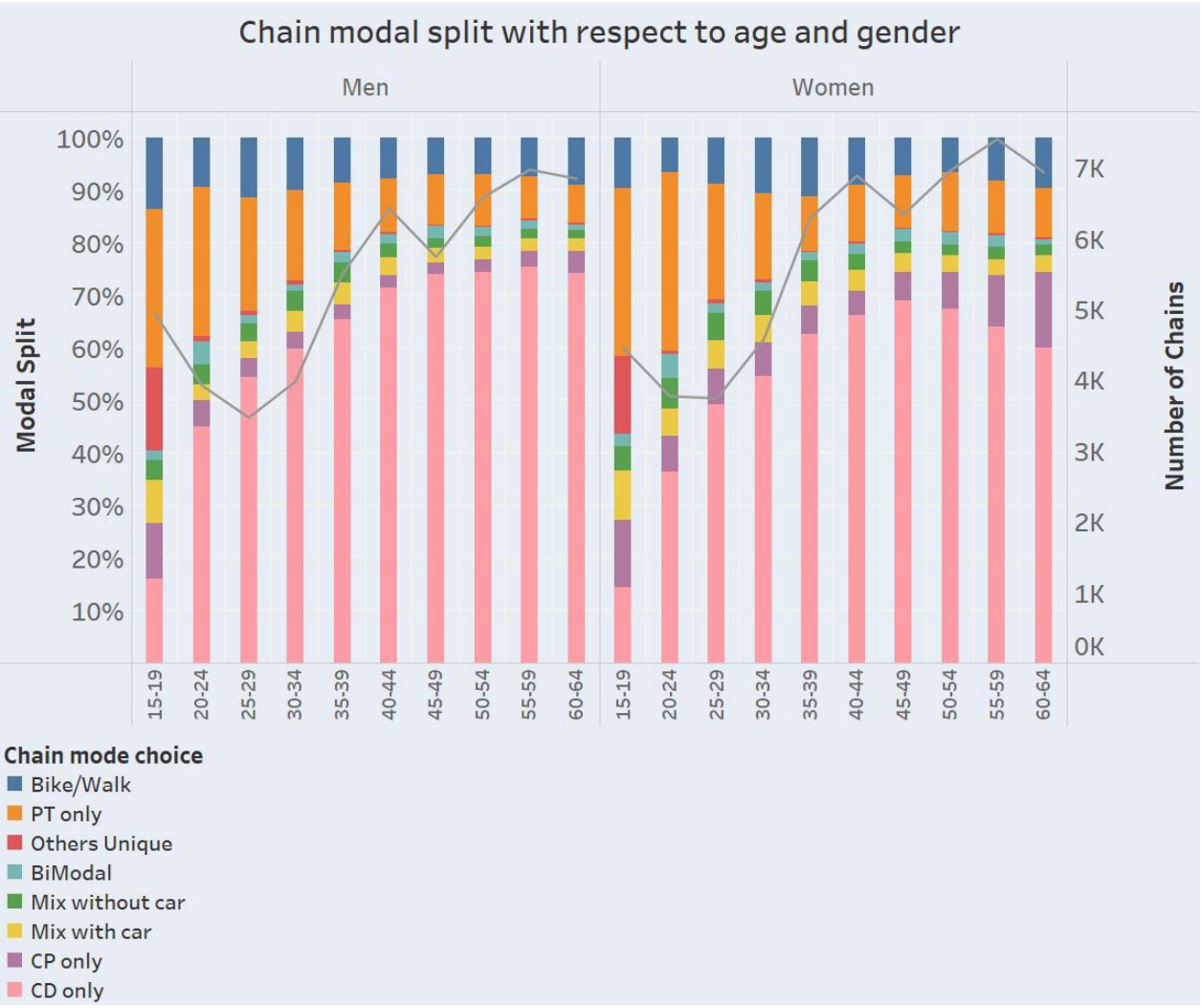

Figure 4-8 Chain modal split with respect to age and gender (2018)

4.2.1.2 Ownership of a driving license

It is with no surprise that the possession of a driving license changes drastically the chain modal share for people. Figure 4-9 shows that people without a license tend to use PT only as their main chain mode (48% modal split) while people with license have CD only as the dominant mode (68% modal share). It is notable to see that people without license tend to use every other alternative more than CD only more often than those with license, even ones such as Mix with car, where the usage of a car is involved. This implies that people with license are unlikely to mix other alternatives in their chain when they are driving themselves, while people without license tend to do it more even if they are being driven around as car passengers. The no license group does not include people under 17, who legally are not allowed to drive. For these people the most common

mode is others unique, which most likely due to school bus being in that category. Also of interest is the large difference of Cycling/Walking only between people with and without driving (5.8% vs 17.2%). People without a license tend to complete chains using Cycling/Walking only more than 3 times as much than people with one. While these differences could be due to constraints rather than choice, as non-license owner have fewer choices than those with licence, it still poses the question of how many trips done by CD only can be easily substituted with ones by Cycling/Walking given the difference between the two, and also the difference of perception of walking distance accessibility for people with and without driving licenses.

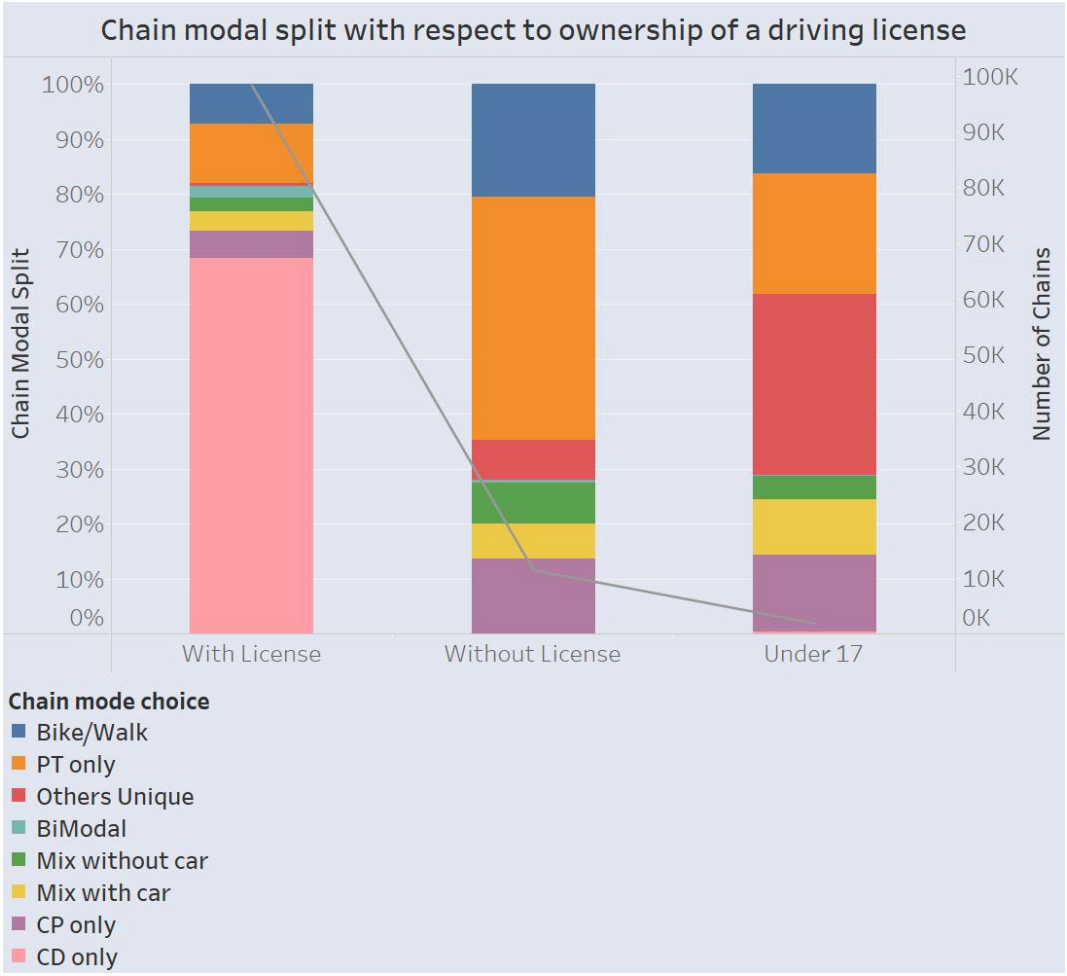

Figure 4-9 Chain modal split with respect to ownership of driving license

4.2.1.3 Main occupation

Figure 4-10 shows that the main occupation of the person is correlated with the chain mode choice.

The students show a particularly different distribution. Indeed, while for the non-student population, the CD only option is the most important mode, students make greater use of public transport and active transport. This could be due to both legal constraint for students under 17, or economical constraints where students cannot afford car ownership so they tend to use other modes. Also, active transport is used less among full-time workers than among other categories.

It is worth noting that although they have a low total modal share, Bimodal only sequences are only present among workers and students, which could be explained by the fact that the other categories have less need to make long-distance chains. Another reason is that Bimodal trips are often based on train usage which is typically more convenient during peak hours, where most trips are work or study related.

Finally, CP only trips tend to be much less for workers (full time and part time) than any other occupation, and highest for people with the main occupation “at house”. People “at house” tend to also use PT the least for their chains and CD almost as high as full time workers. They are the category that uses car in their chains the most (sum of CD, CP, Bimodal and Mixed with car) with 81.2% of their chains involving the usage of car. This could be due to these people making a lot of shopping trips for the house, and picking up/ driving someone, where the usage of car offers more flexibility.

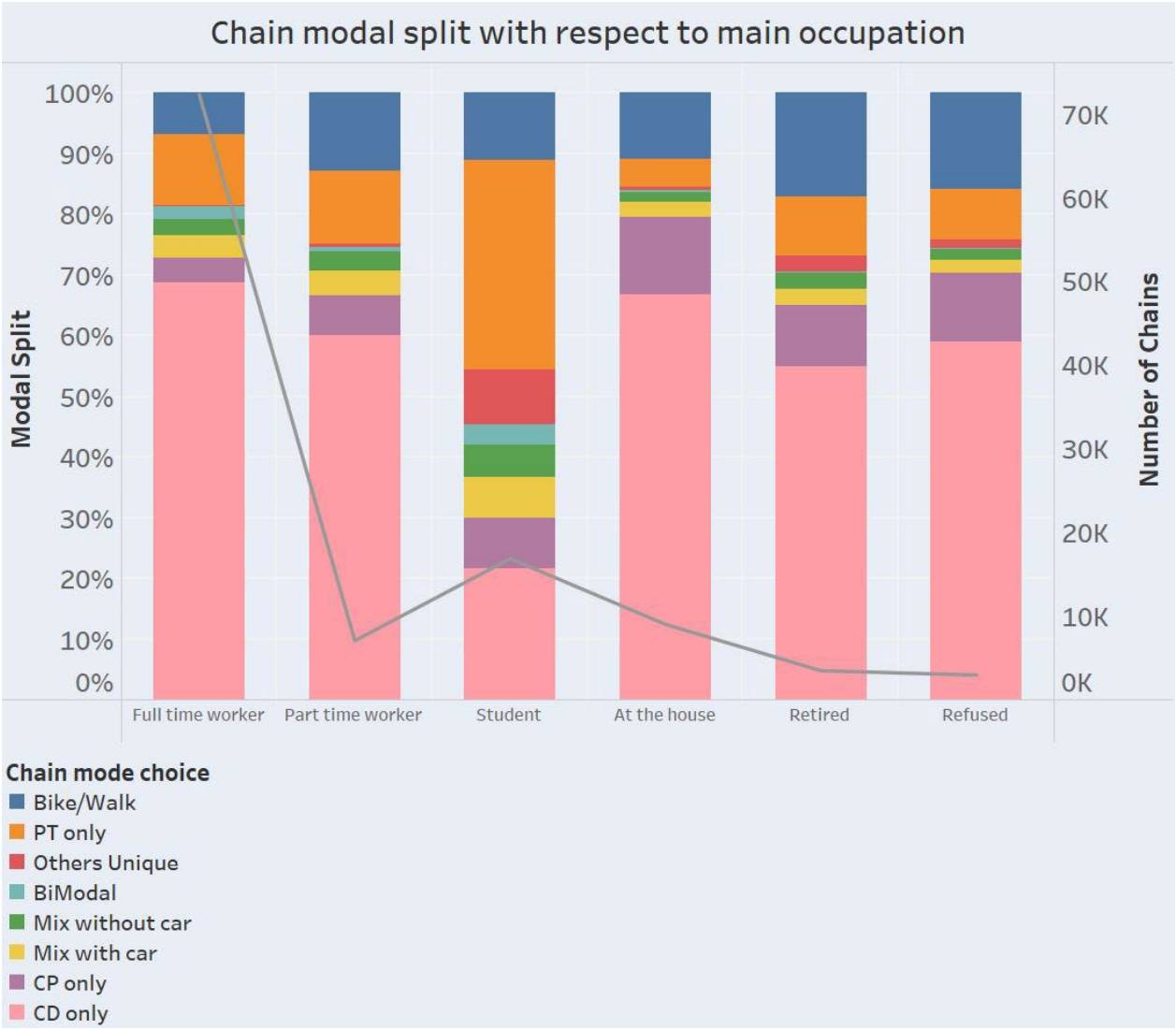

Figure 4-10 Chain modal split with respect to main occupation

4.2.2 Household variables

4.2.2.1 Size of the household

Figure 4-11 shows the chain modal split with respect to the number of people in the household. It is observed that the CD only alternative is the most used for all categories, and it is mostly used for the 2-4 people households with a modal share of about 61%. It is noticeable that the usage of CD only starts to decrease with the increase of the household size after the count of four people.

PT only is mostly used by single person household, while all other sizes share about an equal use of this alternative. As for CP only, households of two people have the highest chain modal share.

Active transport is highest for households with 1 person only, and so is the use of PT only. This could be due to students who live alone forming a good portion of these people.

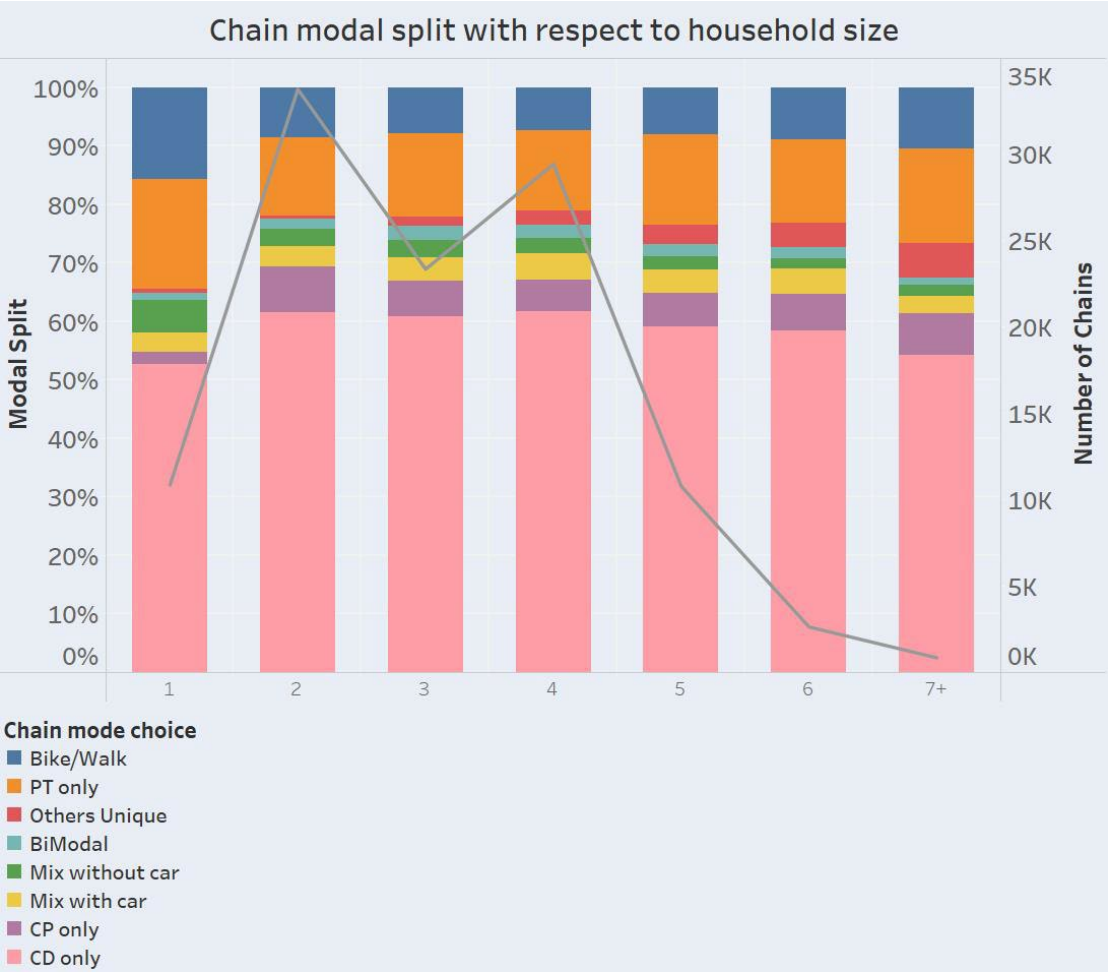

Figure 4-11 Chain modal split with respect to household size (2018)

4.2.2.2 Number of cars owned in the household

Figure 4-12 shows the chain modal split with respect to the number of cars owned in the household. Briefly, we can see how owning one car is enough to change drastically the way people choose their chain mode. In fact, owning one car is enough to reduce PT only and active transportation by about 52.3% of all chains.

Another big change is seen when comparing households with one car against households with two. The CD only modal split jumps from 50.3% to 72.2%, while the modal split for PT only drops from 20.4% to 7.7%. This clearly shows that owning more than one car has a huge impact on choosing PT as the chain mode and shifting towards car-based alternatives a lot more. It is

interesting to see that owning more than two cars does not show the same kind of impact but rather just small increases in the usage of CD only.

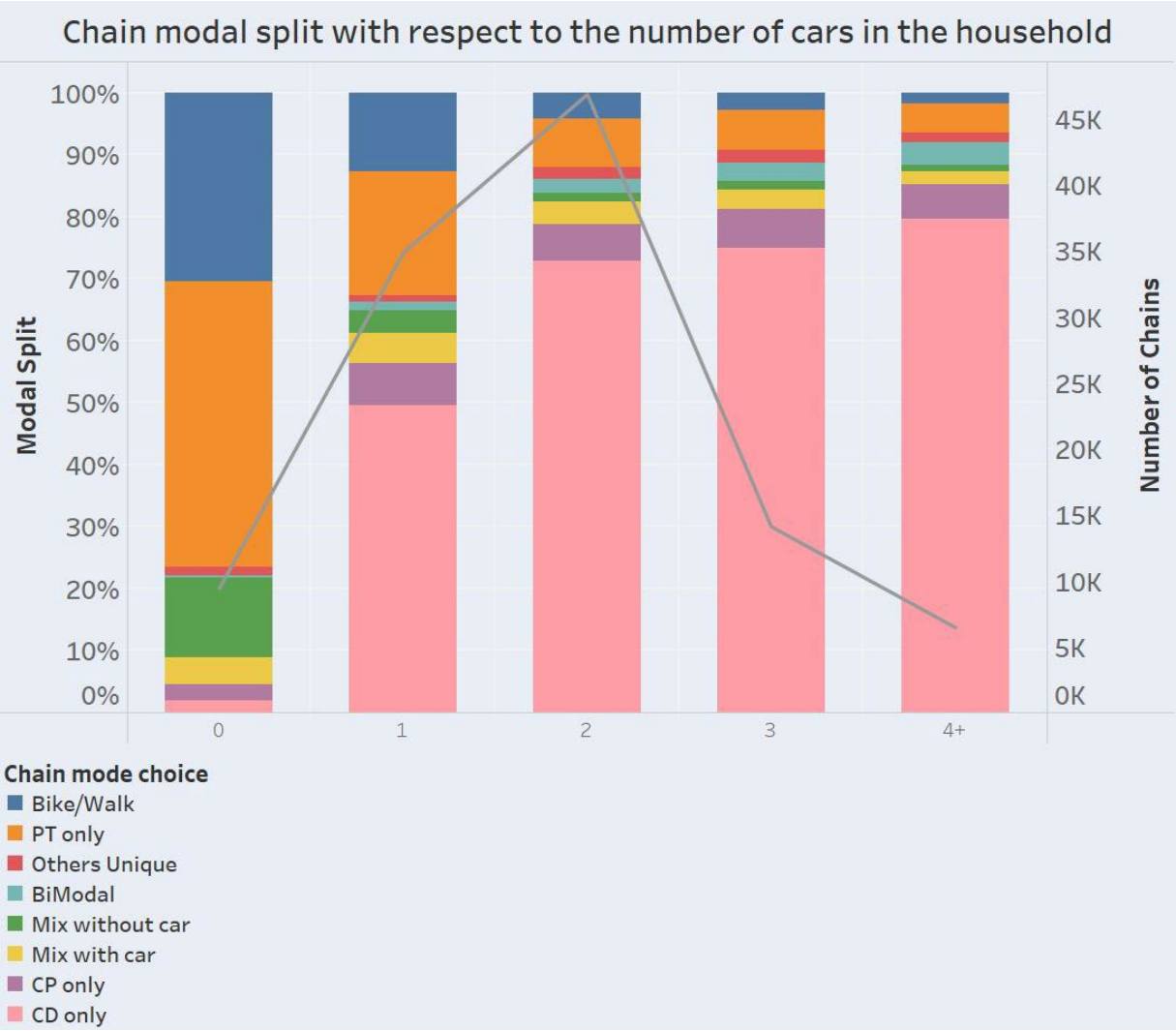

Figure 4-12 Chain modal split with respect to the number of cars owned per household (2018)

4.2.2.3 Presence of children in the household

Figure 4-13 shows the modal split with respect to the presence of children in the household. It is noticeable how having children 6 and under leads to more CD only trip chains and overall, a lot less use of PT only chain. The PT only modal share is 10% for households with children of 6 and under, 13% for households with children between 5 and 15 and 16% for households without children. It is also very evident how the others unique chain mode is by far more noticeable for households with children between 6 and 15. This is most likely due to the school bus alternative

that is considered one of the others unique modes. Finally, the increase of the CP only modal split increases when there are no children/ or the children are older than 6. This is mostly explained by the fact that households composed of two people (mostly couples) make the most CP only trip chains.

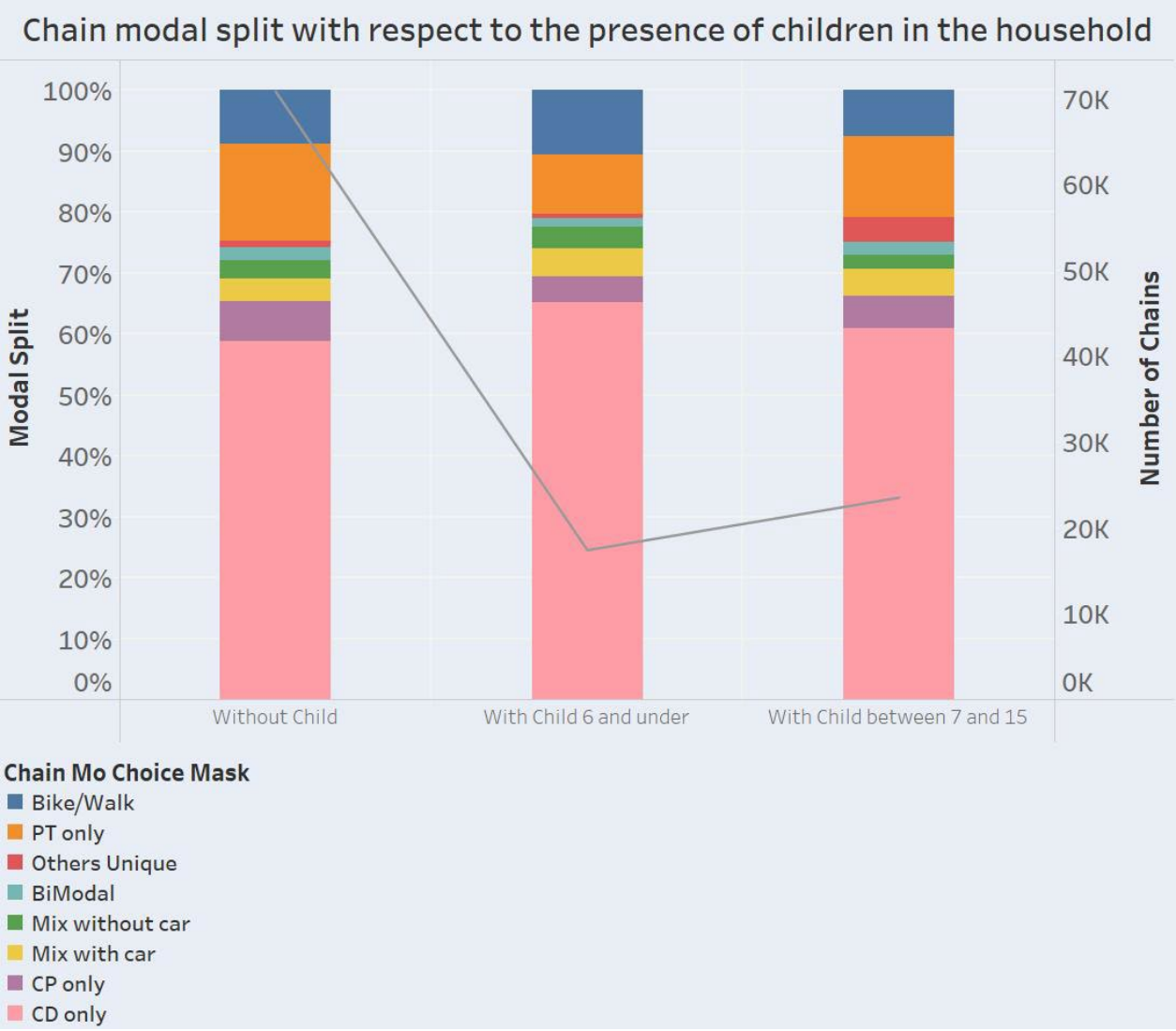

Figure 4-13 Chain modal split with respect to the presence of children in the household (2018)

4.2.2.4 Distance to CBD

This indicator has a significant relationship with the modal choice of individuals. Indeed, we can see in Figure 4-14 that the closer a household is to downtown, the greater the use of PT only is.

It is obvious that the trend is rather the opposite with respect to the use of the CD only, given that the availability of public transit services decreases the further away from the city center one is which led to the increased use of CD only the furthest we go from CBD.

Households situated within 5 km of the city center offer a very insightful indication of how accessibility to public transit and attraction points can limit the use of cars. The sum of modal split for alternatives in which a car is used (CD only, AP only, Bimodal and Mix with car) is 44.1%, while that of active transport and PT alternatives form 55.9% of the modal share. This is the only range in which the car is used in less than half of all chains.

The modal split for PT only, as well as that of Bimodal only sequences, is practically zero for homes located more than 50 km from the city center, because after that point public transport services are less existent. Also, note that the modal share of active transport is significant for chains whose home is less than 15 km away and seems to decrease with distance from the city center. Nevertheless, it remains more important than that of PT, as these modes remain an interesting option for completing what is a short chain.

Finally, it is worth noting that the obvious increase in the use of CD only chain happens when going from the 0-5 km range to 5-10 km range (from 35% to 55%) and from the 5-10 km range to 10-15 km (from 55% to 68%). Above the 15 km range the increase becomes minimal, which could suggest that these three ranges could serve as important buffer ranges in which there is a large impact to accessibility and transit supply. Given this information, it appeared like examining the relationship between the chain mode choice and a smaller scale distance to CBD would give better insights. In Figure 4-15, we see that most chains between 0 and 10 km happen at the distances of 3 and 4 km. Moreover, with each km increase, we see a steady pattern of increase in CD only chains corresponding with a decrease in Cycling/Walking only chains and PT only chains.

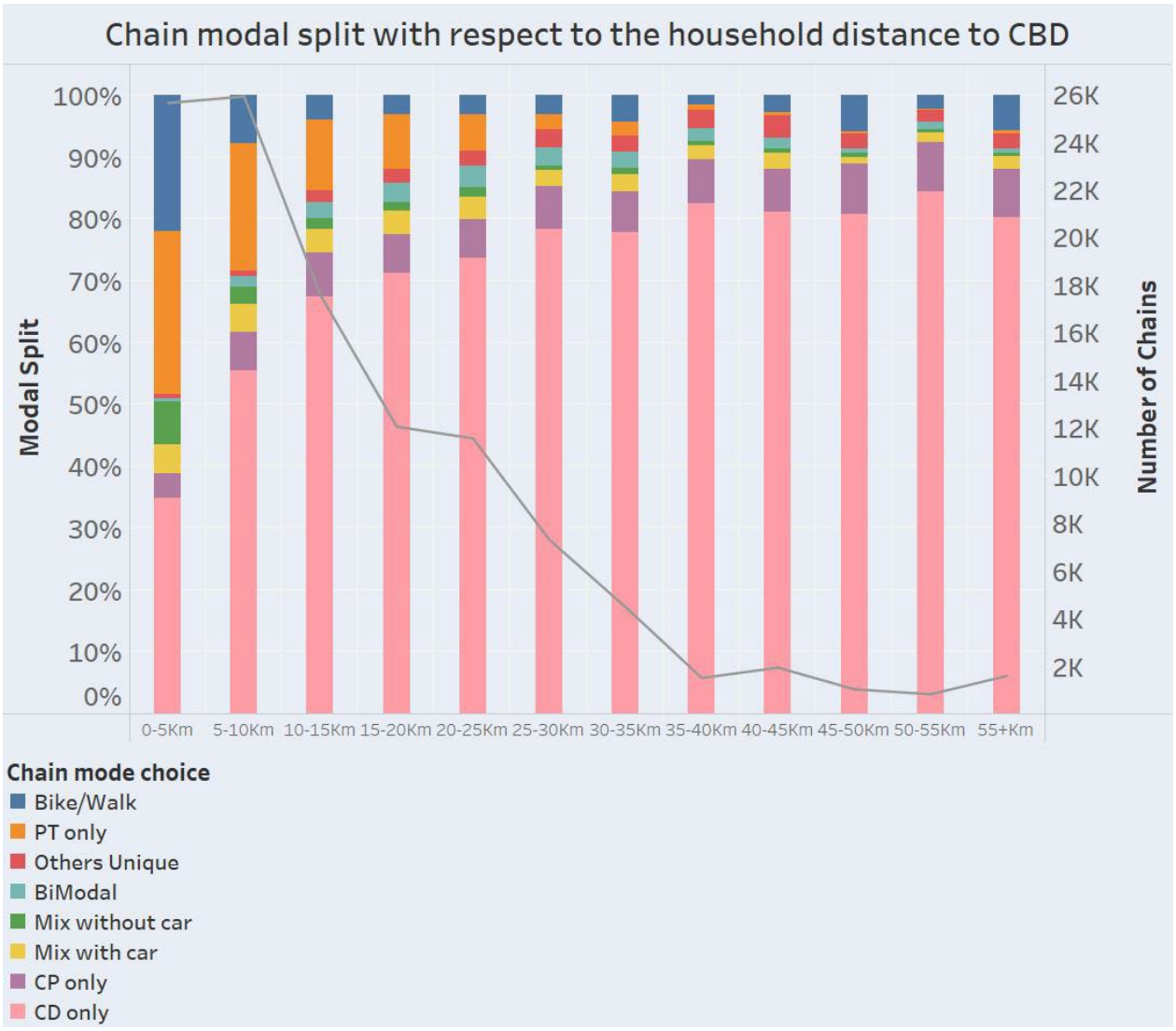

Figure 4-14 Chain modal split with respect to the household distance from CBD (2018)

Figure 4-15 Chain modal split with respect to the household distance from CBD 0-10 km range (2018)

4.2.3 Built environment variables

4.2.3.1 Population density

Figure 4-16 shows the chain modal split with respect to different population density in km^2. Household in DAs between 0-1000 people/km^2 from the largest sample of observed chains where we see the highest modal split for CD only chains and the lowest for PT only. We observe gradual decline in CD only chains with the increase of population density. With the decrease in CD only

in higher densities we see an increase in PT only mode and in Cycling/Walking. It is noticeable that at a density of 20k+ people/km^2 the Cycling/Walking chain mode has the highest modal split, but only a small sample size of the chains is observed in these densities.

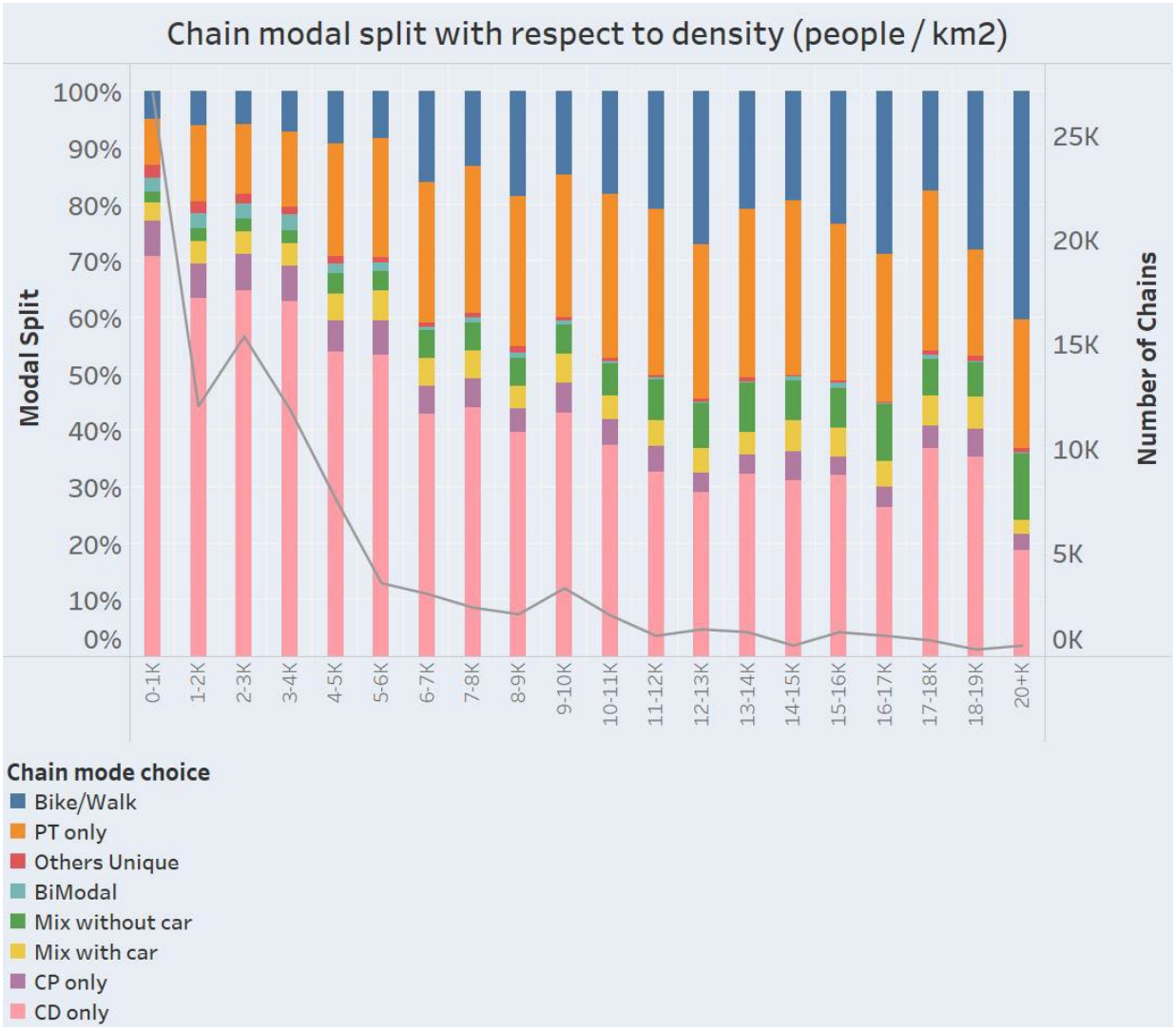

Figure 4-16 Chain modal split with respect to population density (people per km^2) (2018)

4.2.3.2 Distance to closest metro station

The initial chain modal split with respect to the nearest metro station from the household showed a large difference between the 0-2 km walking distance and every other value as shown in Figure 4-17. In fact, CD only chains form about 50% of the chain modal split in the 0-2 km but more than 70% for every other distance greater than 2 km. With that increase of CD only chains, we see a decrease in the PT only chains. While this is likely correlated with the fact that households near

metros have in general higher accessibility to transit services and other attraction points, it is still interesting to see how the chain modal split is divided within these 2 km walking distances. Figure 4-18 shows the chain modal split where the nearest metro station walking distance is between 0 and 1800 m from the household. The results show that CD only chain mode choice is lowest between the 0 and 600 m threshold, and then increases gradually until the 1200 m value where it becomes constant. The use of PT only has a reverse pattern where the use is maximal between 0-600 m and then decreases to be stable after the 1200 m value. This could indicate that the catchment area for metro walking distance in Montreal is about 600 m, while distances between 600-1200 m are still within the acceptable range for some. It is after the 1200 m that we could see a pattern that resembles the walking distances above 2 km.

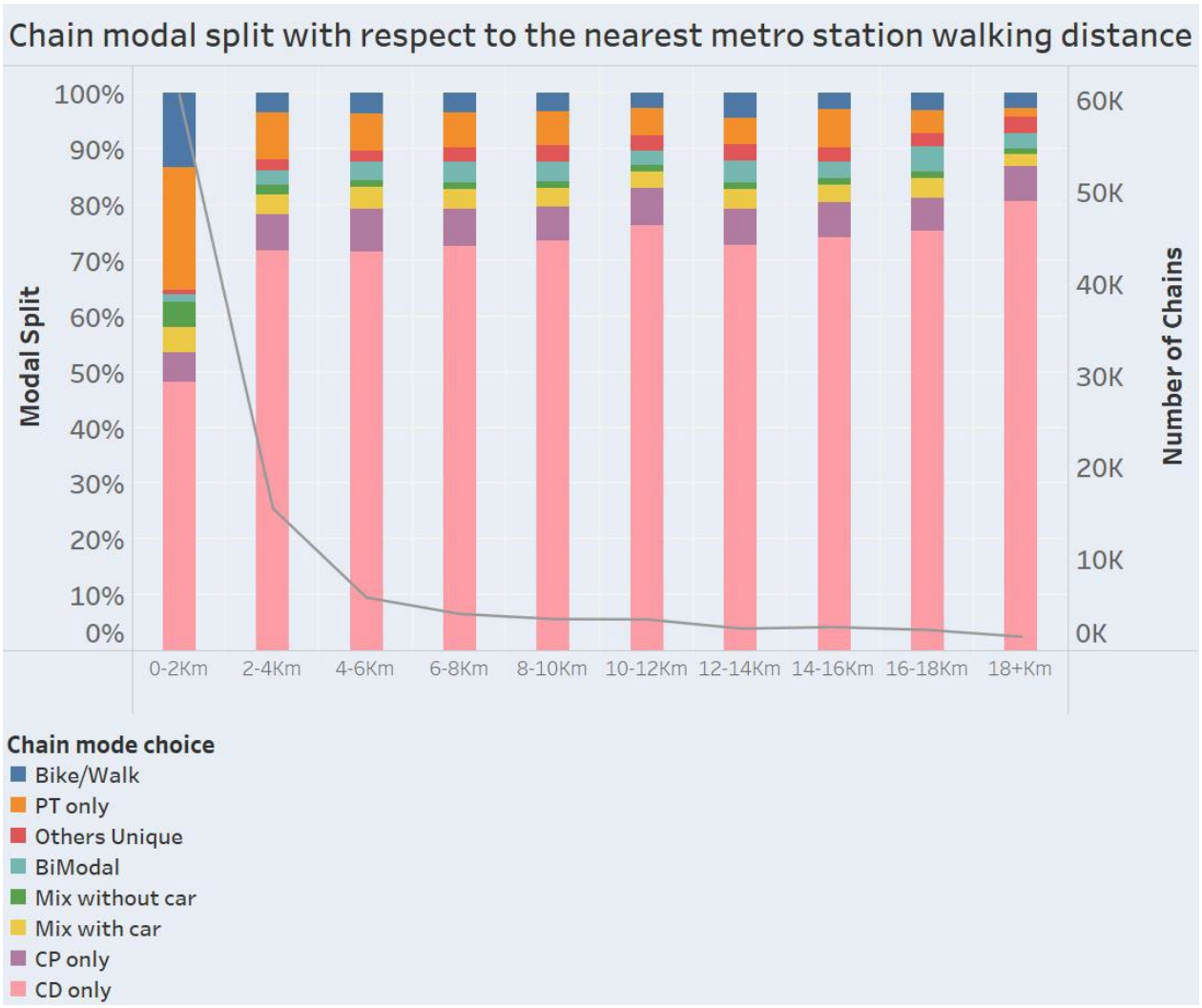

Figure 4-17 Chain modal split with respect to the nearest metro station walking distance (2018)

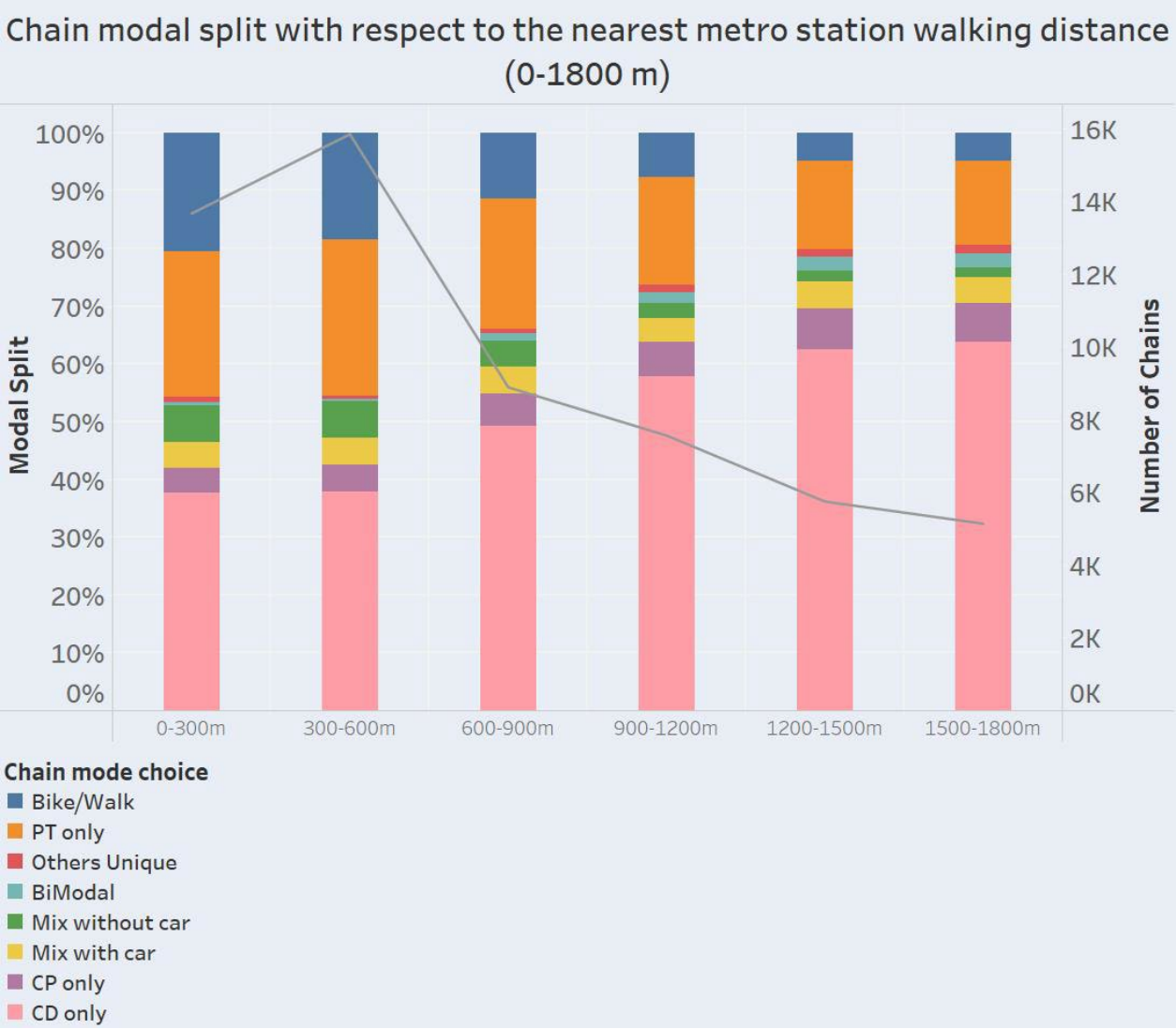

Figure 4-18 Chain modal split with respect to the nearest metro station walking distance 0-1800m range (2018)

4.2.3.3 Distance to closest train station

When considering the distance to the nearest train station, it is the catchment area that we are most interested about especially for P&R and K&R opportunities, so all distances correspond to the driving distance from the household to the nearest train station. In Figure 4-19, we see that the use of PT only chains is highest for the distances between 0-3 km and is almost completely nonexistent for distances above 6 km, which could be due to limited presence of transit service available in these areas. Interestingly, the Bimodal (K&R and P&R) chain modal split remains constant for distances between 0-6 km and then decreases for higher values which could mean that the

catchment area for this mode is around 6 km for the GMA, although this remains a very difficult analysis as each station has different catchment area depending on many variables. Like for the distance to the nearest metro, it is interesting to see the chain modal split for the distances on a smaller scale, so a plot showing the chain modal split between 0 and 4 km was made as shown in Figure 4-20. PT only decreases slightly with increasing distances between 0 and 3 km and then suffers a steeper decrease above this value. The Bimodal only on the other hand shows no decrease from 0-4 km.

Figure 4-19 Chain modal split with respect to the nearest train station driving distance (2018)

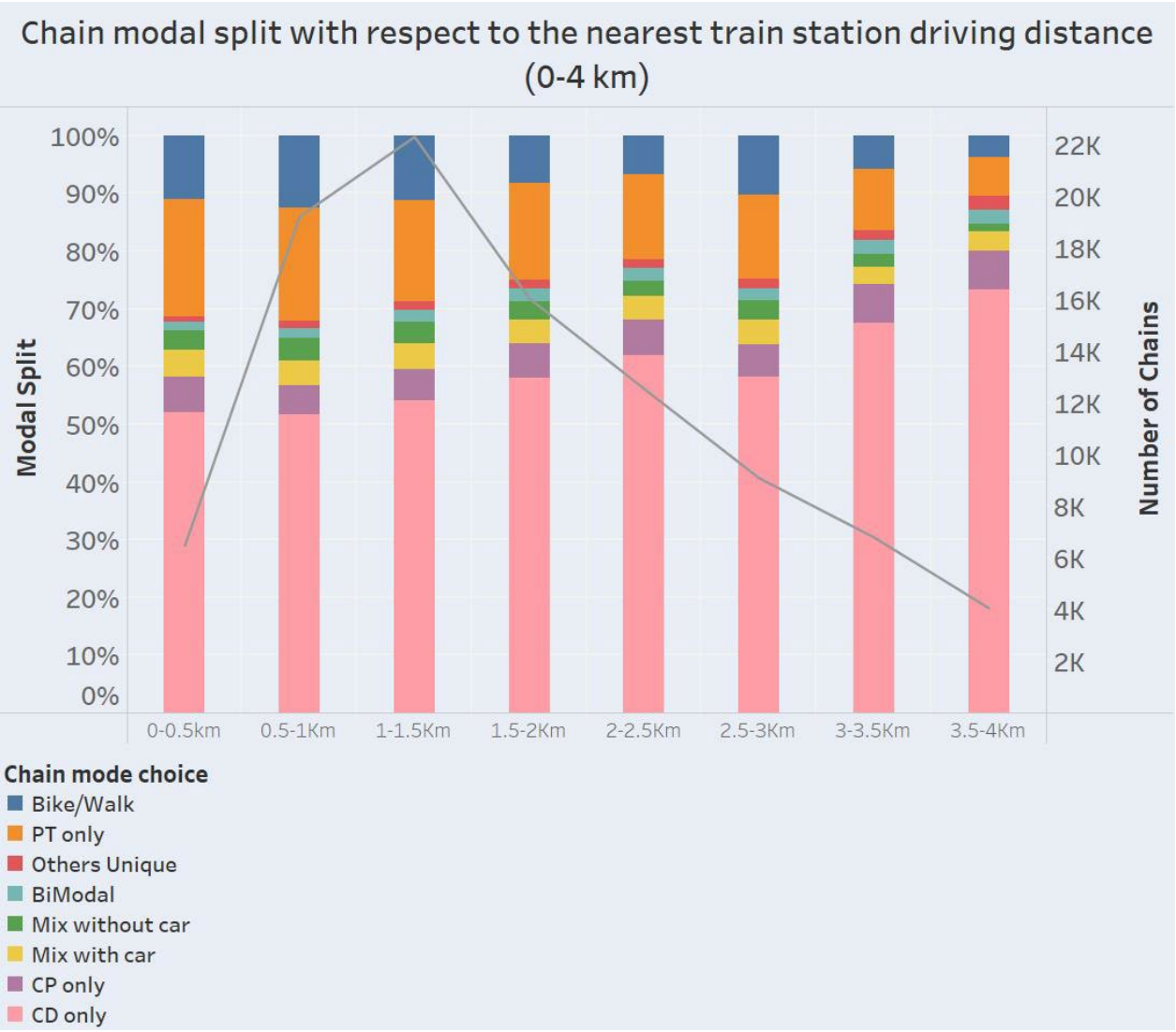

Figure 4-20 Chain modal split with respect to the nearest train station driving distance 0-4 km range (2018)

4.2.4 Chain Variables

4.2.4.1 Chain purpose

Figure 4-21 shows the chain modal split with respect to the chain main purpose. The first observable insight is that driving / picking up someone is the purpose where the use of the car is the highest as expected (78.3%). It is observed that for non-study chains and chains where the purpose is work have the highest modal split for PT only (16%) and the second highest for CD only

(67%). This implies that people going to work usually chose one of these 2 alternatives to do their chains. Study is the purpose where we see the bulk of the mode Others unique as it includes school bus. For the non-constraint chains (where the purposes are shopping, leisure and others) the use of the CD only is highest for the shopping trips, while the PT only alternative is highest for chains where the purpose is "other". The chains observed are mostly those done for work purpose, while other purposes have around the same number of chains split between them.

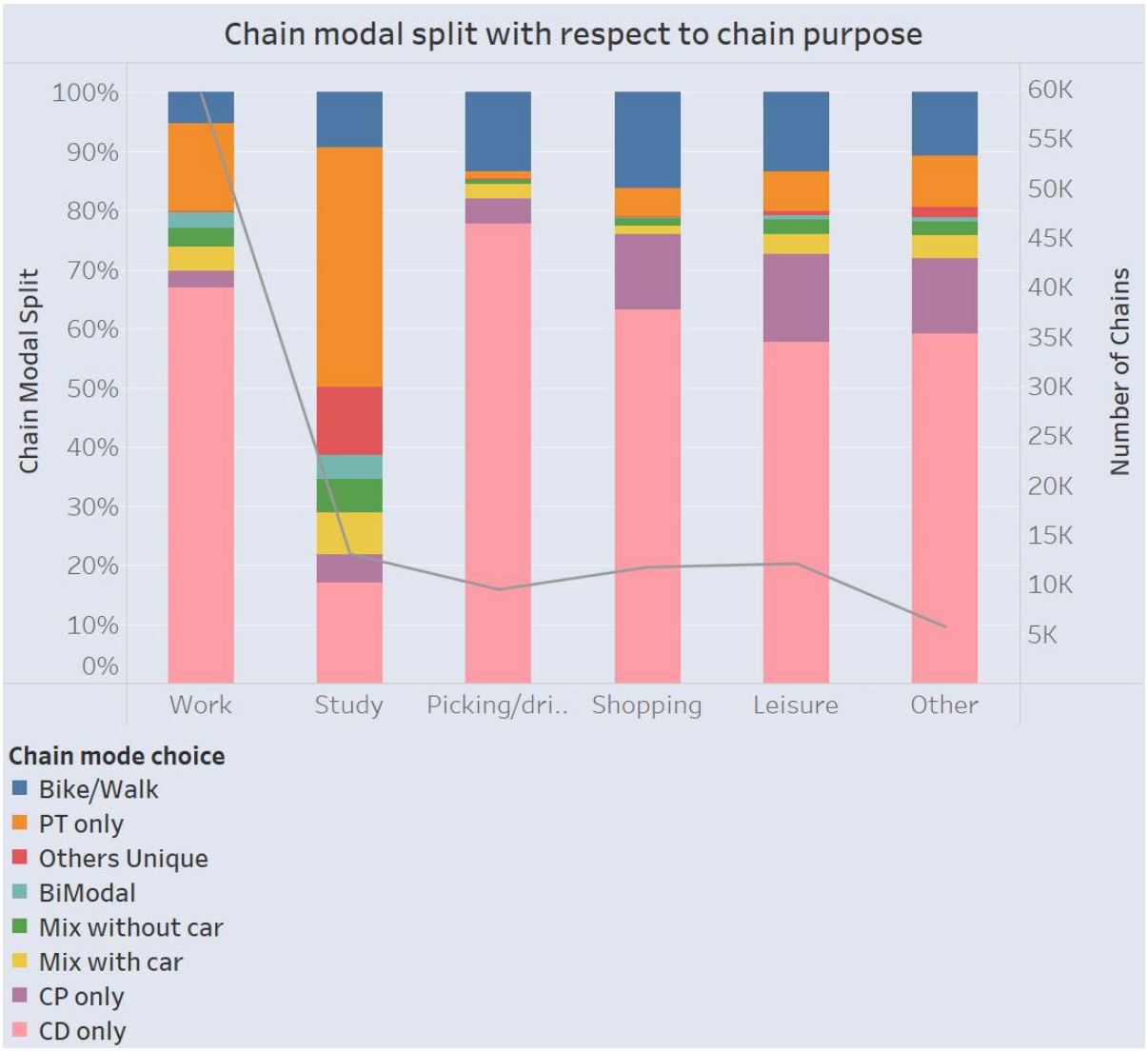

Figure 4-21 Chain modal split with respect to the chain main purpose (2018)

4.2.4.2 Chain complexity

Chain complexity offers a good insight on the effect of the chain structure on the chain mode choice as shown in Figure 4-22. Two things are most noticeable when jumping from a 2-trip chain (a

simple chain) to a 3-trip chain (complex chain). First a large decrease in the chains done using PT only is observed, dropping from 17% for 2-trip chains to 6% for 3-trip chains and then even less for 4 and 5+ trip chains. Second, a notable increase in CD only chains, and an even more notable increase in Mix with car chains is observed. In fact, the use of the Mix with car alternative increases more than 5 times from 1.7% to 10% when jumping from a simple chain to a complex one. All these observations could imply that people are less likely to consider taking only PT as their alternative for a complex chain and are very likely to use the car in their chains either by CD only or by mixing with other modes.

Figure 4-22 Chain modal split with respect to chain complexity (2018)

4.2.4.3 Chain duration

Although this variable is not used in the modelling process presented in the next chapter, it is still interesting to see the relationship of chain duration and chain mode choice as shown in Figure 4-23. It is observed that chain durations that are associated with going to school (where the chain time is between 7h30 - 8h00) are the least used chains for Car Driving only. This is where we observe an increase in others mode (corresponding to school bus) and in public transit as well. Furthermore, chains with durations between 9 to 9 and half hours are the most common, as they correspond to the typical work commuting + activity time for a work chain. An interesting observation is that generally the usage of CD only decreases slightly with the increase of trip chain duration from 0-30 mins until the 7h-7h30 chains, where it spikes back up and follows an increase again. Finally, it is worth noting how we see spikes around chain durations with exact timings (3h exactly, 4h exactly, etc..) which is most likely due to the declaration bias of people to say o'clock times, for trip start times, when responding to a survey.

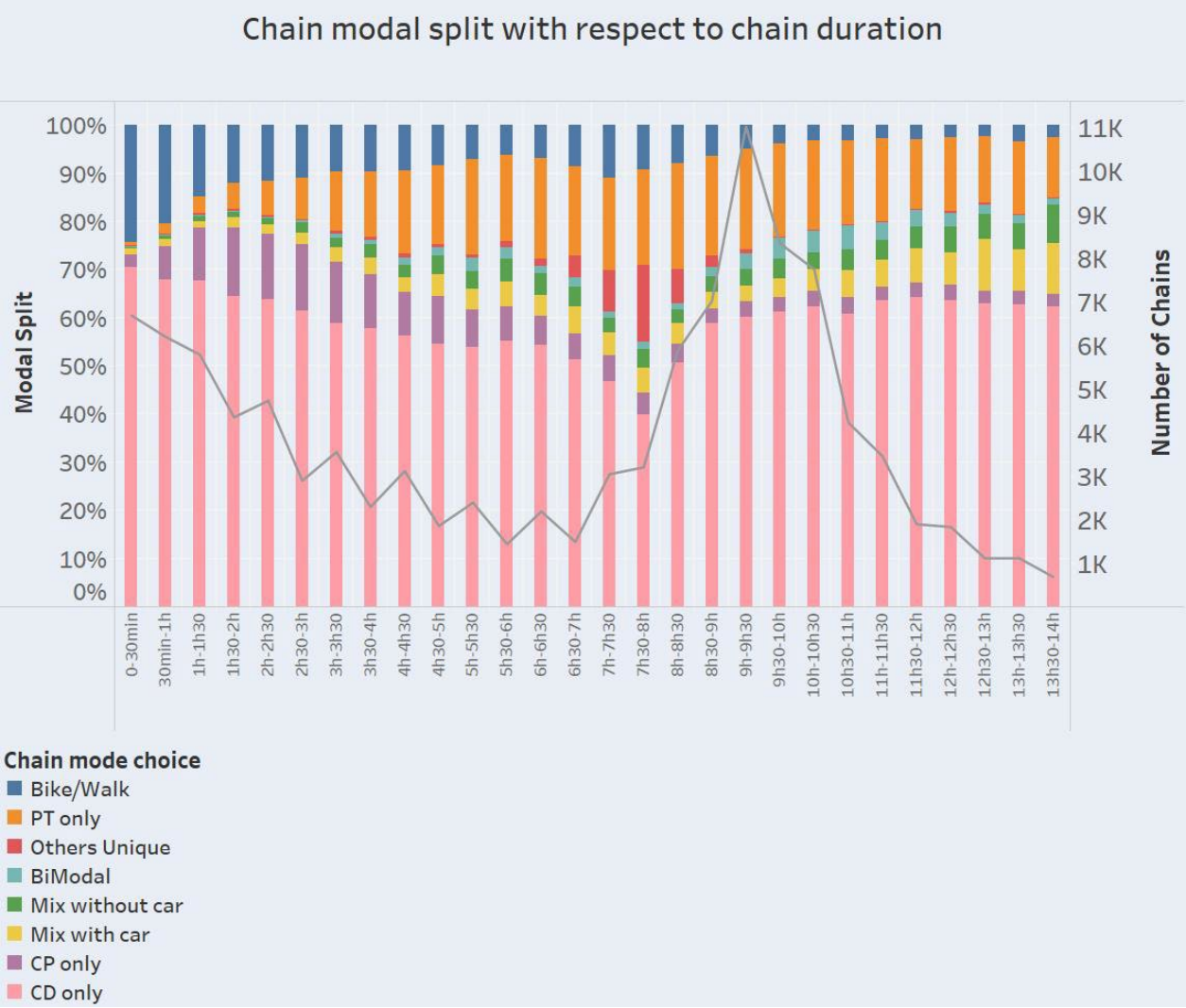

Figure 4-23Chain modal split with respect to the chain duration (2018)

4.2.4.4 Chain distance

Chain distance is one of the most important variables when considering chain mode choice. Figure 4-24 shows the chain modal split with respect to chain distance. The data shows that with the increase of chain distance the usage of CD only chain increases as well. For the PT only chains, the usage is minimal when the distance is between 0-5 km but increases for distances between 5-15 km. After 15 km we see the PT only chain mode choice decreases with the increase of chain distance. Also, worth mentioning that the Bimodal only mode choice increases with the increase of chain distance, which could be due to the long P&R trips that implies longer chains. Since most of the chains observed happen between the distances between 0 and 15 km, Figure 4-25 was plotted

to see the chain modal split in more details between 0-15 km chain distance range. It is seen that the Cycling/Walking modal split is highest between 0-4 km range, where it has more than 25% of share. Interestingly, for these distances we see a decrease in the PT only chains, while the car usage mostly remains the same. This could imply that people are more willing to substitute the usage of public transit in short chains in favor of Cycling/Walking but are less willing to substitute CD only chains with Cycling/Walking.

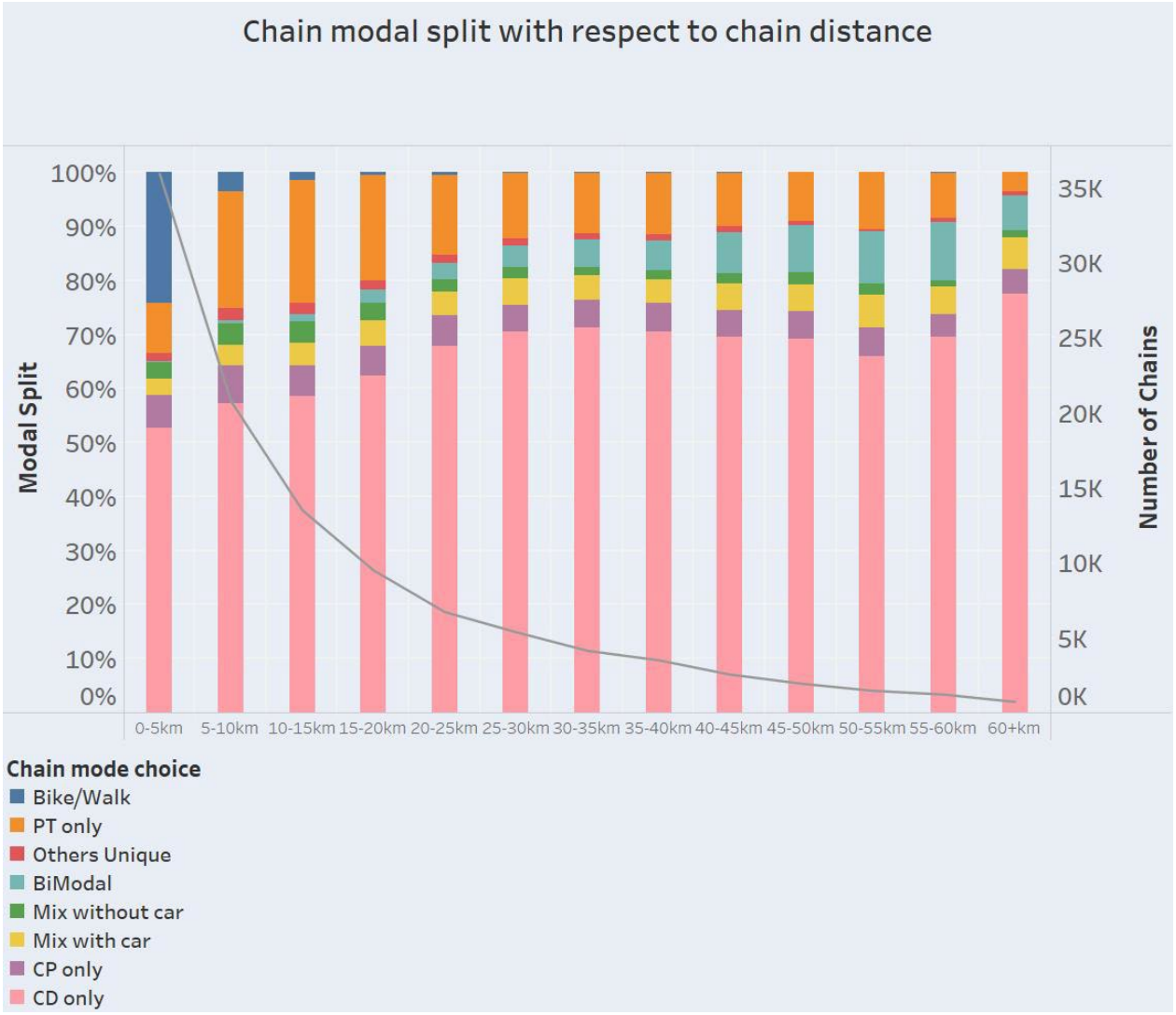

Figure 4-24 Chain modal split with respect to the chain distance (2018)

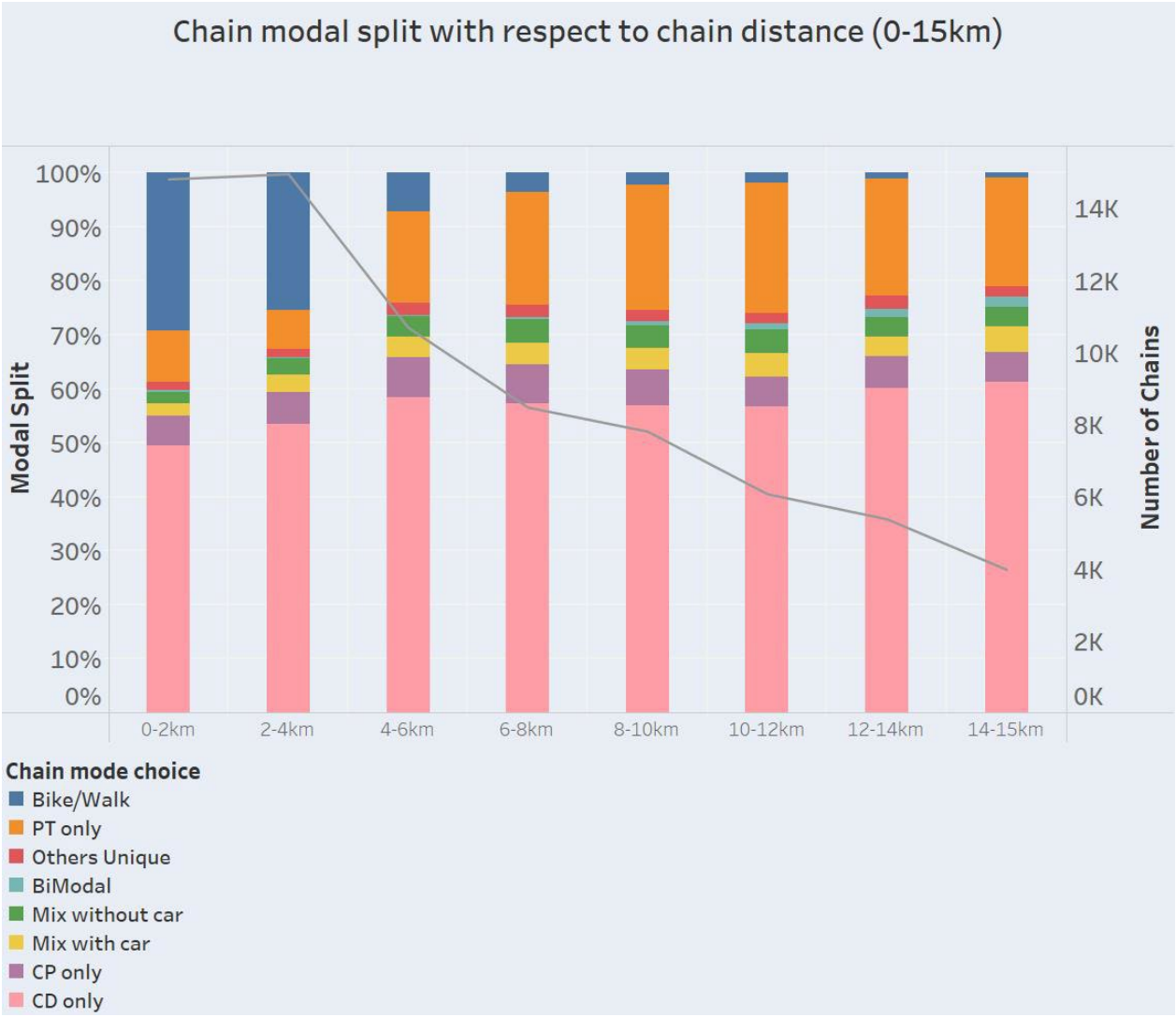

Figure 4-25 Chain modal split with respect to the chain distance 0-15 km range (2018)

4.2.4.5 Activity systems

The chain modal split and activity system relationship is shown in Figure 4-26. People that do two simple chains per day have the highest modal split for Cycling/Walking between all other activity systems. As expected, and shown before, complex chains encourage the use of car and decreases the use of PT as seen for the people who do “One complex chain” or “1 simple and 1 complex chain” only per day. We see an important increase in the modal split for “mixed” chain mode choices in the one complex chain only activity system for both Mixed with car and without.

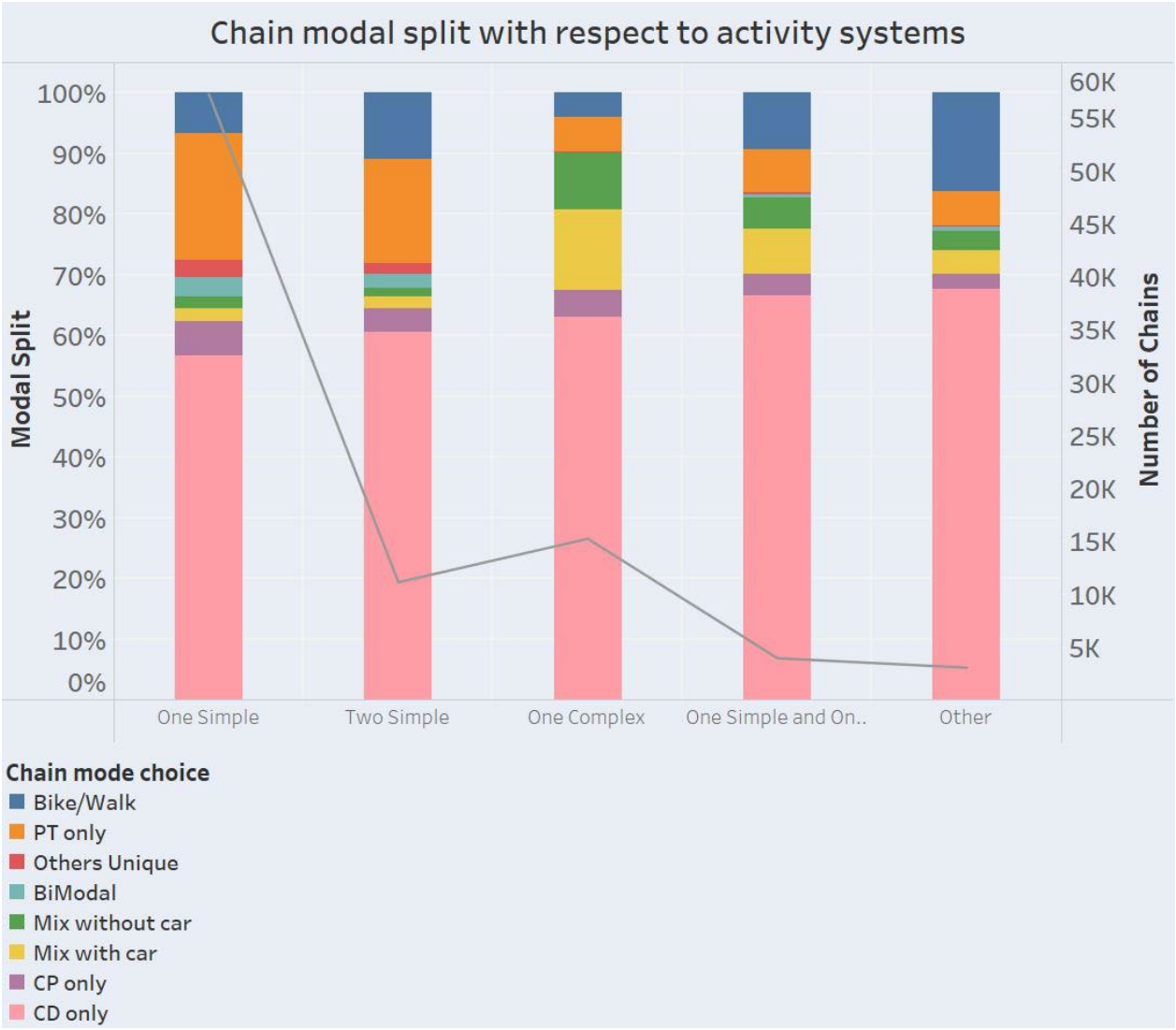

Figure 4-26 Chain modal split with respect to the activity system of a person (2018)

4.3 Descriptive analysis synthesis

The results of the descriptive analysis done for the trends over the years and for the relationship between chain mode choice and the independent variables offer various insights and important takeaways. This section shows the most important conclusions for sections 4.1 and 4.2.

4.3.1 Trends conclusion

All in all, the exploratory analysis conducted on the trends shows a general direction of moving towards more complexity in chains when compared to previous years, especially for women. This could prove problematic as we can see how complex chains and multi-loop ones lead to a higher

usage of car in the chain modal split and a steep decrease in public transit. The takeaways of the trip chaining evolution can be summarized as follows and as shown in Table 4-6:

- People are doing less chains per day and increasing the complexity of their chains, meaning they are probably fitting more trips in a single chain. This also shows in the activity systems where we see the decrease of the two simple chains system in favor of the one complex chain.
- Women seem to conduct more complex chains then men in 2018, especially for the middle age groups. It is especially obvious in their activity systems where middle-aged groups show the largest share of activity systems composed of complex chains.
- Both chain duration and chain distance are on the rise when compared to previous years.
- In all years, the increase in complex chains was seemingly accompanied by an increase in car use in the modal split and a steep decrease in public transit use.
- In 2018, there was a remarkable increase in the Mixed without car mode choice, which could mean people are more willing to try combining active modes for complex chains than before.
- The work chain purpose continues to increase while that of shopping continues to decrease.

Table 4-6 Chain trends takeaways

Trends	Men	Women
Chains per person	↓ when compared to 1998 = for 2008 and 2018	↓ when compared to 1998 = for 2008 and 2018
Chain complexity	↑ when compared to 2008	↑↑ when compared to 2008
Chain duration	↑ when compared to 1998 = for 2008 and 2018	↑ when compared to 1998 = for 2008 and 2018
Chain distance	↑ ↑ when compared to 1998 and 2008	↑ ↑ when compared to 1998 and 2008
Chain purpose	↑ for work ↓ for shopping	
Chain mode choice	↑ for Mixed without car	
Activity systems	↓ for 2 simple chains ↑ for 1 complex chain	↓ for 2 simple chains ↑↑ for 1 complex chain

4.3.2 Chain mode choice relationship with independent variables conclusion

The chain mode choice for the latest ODS data of 2018 was shown to be affected differently by each variable. While some variables were shown to have a stronger relationship with chain mode choice than others, all variables were shown to influence mode choice to an extent. The takeaways from the descriptive analysis of chain mode choice can be summarized as follows:

- For the socio-demographic variables, CD only chains increased and PT only chains decreased with the increase of age for men. This pattern is also observed for women, but only until the age of 49 as after that CD only chains decrease, and PT only ones increase slightly. CP only chains were significantly higher for women in the higher age groups when compared to men. The ownership of a driving license seemed to heavily reduce Walking/Cycling chains and PT only chains when compared to people who do not own a license. As for main occupation, full-time workers tend to use CD only chains the most and

have the least use of active transport, while students were found to use PT only chains the most.

- For the household variables, the size of the household was shown to have a noticeable relationship with chain mode choice when comparing one person household with two people or more household. One person household performed less CD only chains and more PT only chains than most other sizes. Also of note is the decrease of CD only chains when the size of the household becomes greater than four. When comparing households that own one car with household that own two cars or more a significant increase was observed in CD only chains accompanied by a decrease of PT only chains. Having kids seemed to increase the use of CD only and decrease the use of PT only chains in the household especially for households with younger children of six or under. Greater distances from the household to the CBD led to higher use of CD only chains and less use of PT only and Walking/Cycling chains. Distances between 0-10 kms were shown to have higher use of active transport and less use of the car.
- For the build environment variables, higher population densities seemed to be directly correlated with a decrease in CD only chains and an increase in PT only and Walking/Cycling. The distance to the closest metro was shown to have a noticeable effect on chain mode choice when considering distances less than 900 m, where the use of PT only is highest and that of CD only is lowest.
- For the chain variables, complex chains of 3 or more trips were found to have a higher modal split for CD only chains and for Mixed with car chains when compared to simple chains. Short chain distances between 0-4 km have the highest Cycling/Walking modal split while having the lowest use of PT only chains. A notable observation is that CD only chains does not suffer the same type of decrease as PT only chains when considering very small distances (0-4 km) with higher chain distances (5-60 km). This could imply that people are more willing to substitute the usage of public transit in short chains in favor of Cycling/Walking but are less willing to substitute CD only chains with Cycling/Walking. People's activity systems showed that people who make simple chains (one or two) uses PT only chains way more often than people who make a complex chain in their day. Also

the modal split for Mix with car was considerably high for people who make one complex chain per day.

CHAPTER 5 CHAIN MODE CHOICE MODELING

In this part, the random forest models are shown along with the results they provide and the variables' importance for each model. The goal of the models is to predict the mode choice of the chain, based on the different modes presented in section 3.3.1.1, by utilizing the various independent variables discussed in the rest of chapter 3.For each model, two types of feature importance are presented: Mean Decrease in Impurity (Gini), Mean Decrease in Accuracy. The definition of these features and how they are evaluated are explained below. Several models were tested based on different assumptions and different variables, with the objective of seeing how well the random forest perform as a tool for trip chain mode choice modeling. It also allows to see how each variable performed according to its feature importance rank. The most recent data from ODS was used, for the year 2018. The models used did not account for weights. Using unweighted data eases the processing time and computational power needed to run the models however it can have some impacts such as misrepresentation of some population segments or some mode choices.

5.1 Random Forests

Random forest (RF) is a machine learning algorithm that is widely used as regression and classification model. It is composed of several decision-trees that all conduct the classification process for a given variable (Breiman, 2001). Random forest reduces the variance between each tree by adapting the bootstrap aggregation also known as bagging. Bagging is the process that allows each tree to randomly select a subset of the training dataset as a sample to utilize in the decision process. This allows each tree to have a different random sample from the training set, which is then split at each node into categories by the decision tree according to the input explanatory variables. Thus, the number of observations corresponding to different classes decreases after each node split. The splitting criteria at each node is decided based on "purity", which means maximizing having the most amount of sampled data belonging to the same class after a split, a method also referred to as maximizing information gain. In other words, variables that can help the model split the dataset into homogenous subsets are chosen in the splitting process of each node. The output of the class is eventually decided by the majority voting done by all trees i.e., the most selected class by the ensemble of decision trees.

The performance of the RF models is evaluated by checking the overall accuracy of the model. However, this could be misleading for cases where there are many classes that have different representation and weight. Therefore, for all the models the overall accuracy of the model is displayed, along with the precision and recall value for each chain mode choice. The explanation of precision and recall is provided below as:

Precision is described as the number of positive class predictions that belong to the positive class where: $$Precision = \frac{True\ positive}{Predicted\ Positive} \quad (2)$$

For a specific mode, True positive corresponds to all the chains where that mode was correctly predicted.

For a specific mode, predicted positive corresponds to all the chains where that mode was predicted.

For example, the precision for CD only chains would be all the chains where CD only was predicted and observed in the dataset divided by all the chains where CD only was predicted (regardless of being observed or not).

Recall can be defined as the number of positive class predictions made out of all positive examples in the dataset where: $$Recall = \frac{True\ Positive}{Actual\ Positive} \quad (3)$$

For a specific mode, actual positive is defined as all the chains where a certain mode was observed in the dataset.

For example, the recall for CD only chains would be all the chains where CD only was predicted by the model and observed in the dataset divided by all the chains where CD only was observed in the dataset.

Although mode choice models usually focus on the prediction more than interpretations (K. Kim et al., 2021), RF models offer the chance to view the relative influence of each independent variables on the model through feature importance. Two main types of feature importance are shown for each model:

1) Mean Decrease Gini
2) Mean Decrease Accuracy

Mean Decrease Gini is a measure used in random forest algorithms to quantify the importance of a particular feature in making predictions. It is calculated by measuring the total decrease in impurity (measured using Gini coefficient) across all decision trees in the forest when a particular feature is used for splitting the data. The intuition behind mean decrease Gini is that a feature that results in a large decrease in impurity when used for splitting the data is likely to be more important for making accurate predictions. Conversely, a feature that does not significantly decrease impurity when used for splitting is likely to be less important. To calculate mean decrease impurity, the algorithm first computes the baseline impurity of the data using the Gini coefficient. It then considers each feature and calculates the total decrease in impurity that results from using that feature to split the data in each decision tree in the forest. The mean decrease impurity for a feature is then calculated as the average of the total impurity decrease across all trees in the forest. In practice, mean decrease impurity can be used to identify the most important features in a random forest model, which can help in feature selection and in understanding the underlying patterns in the data. One potential issue with Mean Decrease Gini is that it can be sensitive to correlated features. When two or more features are highly correlated, it may give similar or equal importance scores to both, even if one of the features is redundant or less informative. This can make it difficult to interpret the relative importance of each feature.

Mean decrease accuracy is another measure used to evaluate the importance of features in a Random Forest model. The algorithm creates many decision trees, each trained on a random subset of the data and a random subset of the features. The importance of a feature is calculated by measuring how much the accuracy of the model decreases when that feature is excluded. Mean decrease accuracy is the average reduction in accuracy of the Random Forest model across all decision trees when a particular feature is removed from the data. This measure helps identify which features are the most important for making accurate predictions.

To determine the best number of trees used in each model (also known as n-estimators), a different set of trees was used from “100”, “200”, “300”, “400”, etc. In theory, increasing the number of trees will increase the accuracy, however, after a certain point the improvement seen in the results is negligible and increasing the number of trees will only increase the computation time. Some studies found that the optimal number of trees has a strong correlation with the number of observations in the dataset used (Płoński, 2020). For this study, the number of optimal trees was

chosen when the improvement in precision and recall observed in any of the classes was less than 0.01.

For each model, the random forest model firstly uses two subsets of data: a training subset which is used to learn and discover pattern, and a testing subset used to test the accuracy of the model. The data chosen by the random forest is random for both subsets, but in all models 80% of the observations were used as a training subset while the remaining 20% are used for validation and testing, which is a commonly used split for training/validation datasets for machine learning algorithms.

5.1.1 Variables' classes

For each categorical variable, the creation of classes is necessary as the random forest model utilizes the different classes as a splitting criterion for each node. For some numerical continuous variables such as age classes were created to see the effect a particular age group has on the mode decision process. Finally, some variables such as gender were treated as dummy variables.

For the socio-demographic variables, classes were created for both age and main occupation while gender and possession of driver's license were treated as dummy variables. As for the main occupation of the person, 6 classes were created in accordance with those present in the 2018 ODS data. For the age, 4 bins were created that felt would classify people with similar activity and trip patterns. The rationale for the bins is as follow, these classes have been selected arbitrarily and their relevance for modelling would need to be validated in a further research:

1. 15-19: Adolescence is a period of significant transition in terms of travel behavior as individuals are transitioning from being chauffeured by parents to independent travel. This age group often involves more non-motorized travel modes such as walking and cycling, as well as public transport use.

2. 20-39: This age group consists mostly of young adults/millennials. This group also tends to be more mobile than the 15-19 age group, often owning a car or using ride-sharing services. Plenty of studies investigated the travel behavior of this age group, where it was found that they were more multimodal than others (Circella et al., 2017), and display more openness towards different modes as long as it suits their needs (Delbosc & Nakanishi, 2017).

3. 40-54: This age group consists mostly of Generation X individuals, who are in their prime working years and are typically more established in their careers. This group may have different travel patterns based on their work schedules, with more trips during the week and fewer on the weekends. When compared to millennials, this age group was found to use the car more in both higher and lower density areas (Wang, 2019). They may also engage in more travel for household-related activities such as grocery shopping and errands (Krygsman, 2004).

4. 50-64: This age group consists of individuals before generation X, also known as baby boomers. They are approaching retirement age, with many still working or engaged in volunteer activities. This group may have different travel patterns, such as more flexible schedules, which can influence their travel behavior. People from this group may also begin to experience physical limitations, which can impact their travel behavior. A study found that it is in this age group that we begin observing a drop in the vehicle miles travelled (Zhang & Li, 2022).

For household variables, classes were created for presence of children and access to car. For the presence of children, 3 classes were assumed: without children, with children between 7 and 15 and with children 6 or under. This was done as children 6 or under are considered fully dependent on their parents to make their trip chains, while children between 7 and 15 can take alternative modes (school bus, walking, etc.) but are still generally dependent on their parents for long trips.

For built-environment variables no classes were created but presence of metro station and presence of train station were considered as dummy variables based on the criteria explained in chapter 3.

For chain variables, the main purposes for the chain mentioned in the ODS were used as classes. Figure 5-1 shows all the different classes created for the random forest.

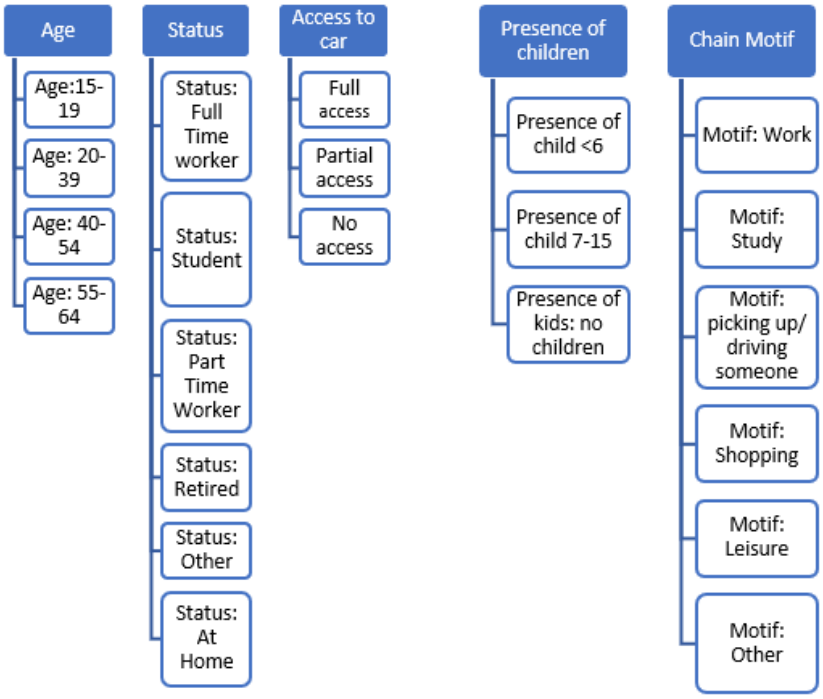

Figure 5-1 Classes created for independent variables

For all variables not mentioned, no classes were created, and the variables were considered as continuous numerical ones.

5.1.1.1 Software and hardware

All the random forests models were done using the Scikit- learn machine learning library on python. The models were computed on a 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz 2.80 GHz processor using a 16 GB RAM. The computation time for all models were relatively short, even when using a higher number of trees.

5.2 Models

The different models conducted in the study are shown here, each with the different variables used in each and tackling a different assumption. All the models are based on the data of the year 2018. The 1st model considers 16 different explanatory variables. Initially, activity systems were not considered in the model. The 2nd and 3rd models have different approaches regarding the

correlation of some variables shown in the section below. The idea is to see how removing these variables and keeping one as a proxy for the others would:

1) Affect the model's overall precision.
2) Influence each variable's importance in the mean decrease Gini given how it can be affected by correlation between variables.

For the 2nd model, distance to CBD was considered as a proxy for the other highly correlated variables: distance to metro, intensity of transit service and population density. For the 3rd model distance to CBD was removed while the other variables were kept. For the 4th model, people's daily activity systems were added as an extra variable to the 1st model. Finally, a predictive model that does not take into consideration chain specific variables is evaluated.

5.2.1 Correlation between variables

As with all studies that consider several independent variables, correlation between them is highly probable and could affect the performance of any given model. Although according to the algorithm of the random forest, high correlation between variables does not affect the overall precision and accuracy of the model when using many trees and nodes like in this case, correlation can still lead to different learning patterns for the model, and specifically in different feature importance especially using Mean Decrease Gini. Table 5-1 shows the correlation between the different variables. Given that typical Pearson correlation is not appropriate for dummies, the values used for "Presence of metro", "Presence of train" and "Access to car" are those corresponding to their numerical values before being put into different classes. For presence of metro and presence of train the distance to nearest metro and distance to nearest train were used, while for access to car the number of cars available per licensed person was used.

We can see that distance from CBD is highly correlated with population density, presence of metro, and intensity of transit stops.

Table 5-1 Correlation between variables

	Household size	*Access to car*	*Age*	*Distance from CBD*	*Trip distance Total (km)*	*Distance-metro(m)*	*Distance-train(m)*	*Population-density*	*Chain Complexity*	*Intensity of transit service*
Household size	1.00									
Access to car	0.36	1.00								
Age	-0.31	-0.04	1.00							
Distance from CBD	0.11	0.37	0.03	1.00						
Trip distance Total (km)	0.01	0.03	0.00	0.02	1.00					
Distance-metro(m)	0.07	0.28	0.02	0.88	0.01	1.00				
Distance-train(m)	0.03	0.18	0.02	0.45	0.00	0.40	1.00			
Population-density	-0.11	-0.29	-0.04	-0.44	-0.01	-0.32	-0.19	1.00		
Chain Complexity	-0.01	-0.06	0.02	-0.04	0.00	-0.03	-0.01	0.03	1.00	
Intensity of transit service	-0.09	-0.30	-0.04	-0.55	0.07	-0.68	-0.32	0.35	-0.09	1.00

5.2.2 1st model

The initial model suggested for this study uses 16 different independent variables (mentioned in chapter 3.2). Activity systems are not used in this model, as we plan to see what effect implementing them in the model will have on its performance. Table 5-2 shows a summary of the variables used in the 1st model.

Table 5-2 Variables used for 1st model

Socio-demographic variables	Household variables	Built environment variables	Chain/activities variables
Age	Number of people in the household	Presence of metro station	Chain Complexity
Gender	Presence of children	Presence of train station	Chain distance
Possession of driver's license	Distance to CBD	Population density	Chain purpose
Main occupation	Access to car	Intensity of transit service	Duration of activities

The primary goal of this model was to see how effective the random forest mode choice modelling is on the trip chain level.

5.2.2.1 Results

Model training accuracy: 85.3% (done on 80% of the data (88,234 chains)

Model testing accuracy: 75.6% (done one 20% of the data as validation (22,336 chains)

Number of trees used: 500

The results of the initial model show good overall performance in terms of prediction. The overall accuracy of the model is 75.6% in the testing process. Table 5-3 shows that the model excels in predicting CD only chains, with a precision of 0.86 and a recall of 0.88. PT only and Cycling/Walking also showed good precisions at 0.73 and 0.63 respectively. The chain mode choice that the model struggled the most in predicting is the Bimodal one. This could be due to the low sample size for the Bimodal chain mode, and for the simplified way the catchment area of train station was considered due to the complexity of that variable. For other modes, the results were less in the precision and the recall section, indicating that the model struggles to predict

chain mode choices that are less used, which could be due to the mode share being lower making this equivalent to rare events.

Table 5-3 Precision and Recall results for the 1st model.

Chain mode	precision	recall	observations
Cd only	0.86	0.88	13516
PT only	0.73	0.62	3238
Cycling/Walking	0.63	0.54	1900
CP only	0.43	0.31	1364
Bimodal	0.09	0.03	436
Others unique	0.57	0.44	360
Mixed with car	0.43	0.47	842
Mixed without car	0.45	0.36	680

5.2.2.2 Feature importance

For the feature importance when considering Mean Decrease Gini, Chain distance was the variable with the most importance followed by the distance to CBD as shown in Figure 5-2. Full access to car and no access to car were the 3rd and 4th, and we see after that the importance of the variables becomes very close. From the chain purpose classes picking up someone/driving someone and study were the most important which is logical since knowing the former implies heavy usage of car, and the latter greater use of active modes as shown in section 4.1.5. For age, only the class of 15-19 was found to be relevant in the node splitting criteria, meaning that this age group is the most that helped the model achieve pure nodes. This could be because it helps the model predict that the trip will mostly be done without CD only for most of these ages. For the presence of children classes, only households where there are children six and under were found to be of importance. This could be due to how younger children can only rely on their parents for any activities they need to do, and that mostly imply more use of the car. Chain complexity, intensity of transit service and gender were also found to be of some importance. Some variables and classes did not have a significant feature importance score such as the presence of train station, age classes between 20 and 64, chain purposes other than “study” and “driving/ picking someone” and the person’s main occupation. These variables were therefore included in the figures below. For train station presence, the lack of importance could be due to that variable affecting mostly the Bimodal choice, which in return does not have many observations and thus will not affect the overall accuracy of the model

as much. Finally, it is noticeable that ownership of a driving license was not significant, which can be mainly due to the "access to car" variable which incorporates having no license in the "no access" class.

When considering Mean Decrease Accuracy, the calculation process involves fully removing a variable from the dataset (with all its corresponding classes if applicable). This makes it easier to compare the importance of variables that were assigned different classes with those that were not. Figure 5-3 shows that chain distance remains the variable with the most importance as in the Mean Decrease Gini. However, we start to see some changes in other ranks. Access to car is the second most important variable in Mean Decrease Accuracy (removing this variable would result in the second highest drop of accuracy for the model), mostly because it is so important to know the level of car access a person have as it makes the CD only option a possibility. The main purpose of the trip becomes ranked 3rd, which highlights how important activities are in determining the chain mode choice. Elsewhere, the variables are ranked in similar ways as in the Mean Decrease Gini.

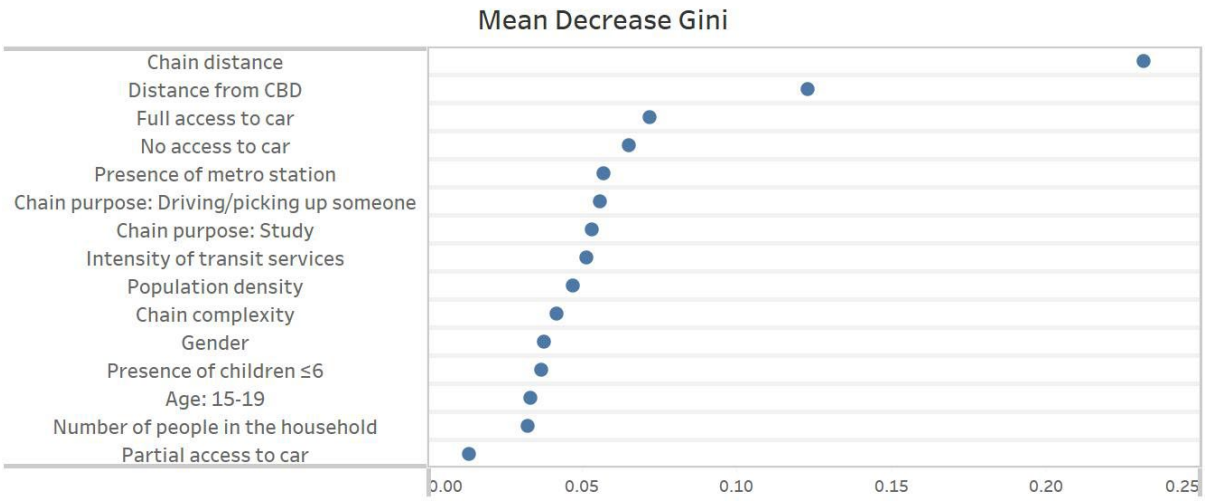

Figure 5-2 Feature importance: Mean Decrease Gini for the 1st model

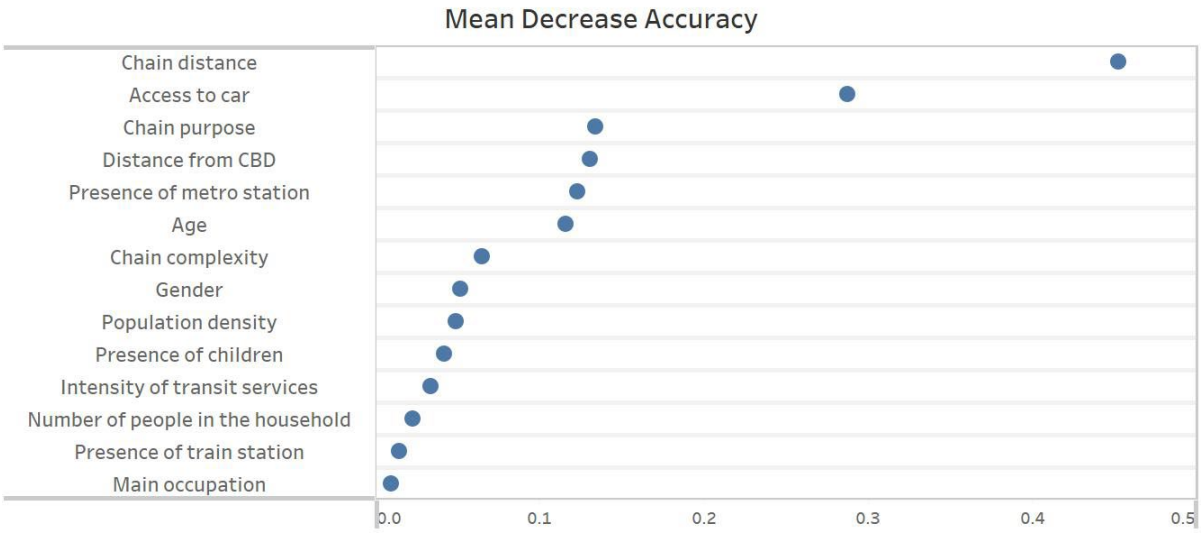

Figure 5-3 Feature importance: Mean Decrease Accuracy for the 1st model

5.2.2.3 Confusion matrix

To help better understand the limits of the models, a confusion matrix is shown after each one. A confusion matrix is a table that is often used to evaluate the performance of a machine learning model, particularly for classification problems. It compares the predicted class labels of the model to the actual class labels of the data, and displays the number of true positives, true negatives, false positives, and false negatives.

These terms are defined as follow:

- True Positive (TP): Represents the number of instances where the model predicted a certain mode choice (example: CD only), and the actual mode choice observed is correct (CD only). True positive can be summarized as when the model achieves a hit (correct prediction of the mode).
- False Positive (FP): Represents the number of instances where the model predicted a certain mode choice (example: CD only), but the actual mode choice observed is different (example: PT only). False positive can be summarized as when the model overestimates a certain mode.
- False Negative (FN): Represents the number of instances where the model predicted a negative class (PT only), but the actual class was positive (CD only). False negative can be summarized as when the model achieves a miss, and an underestimation.

- True Negative (TN): This cell represents the number of instances where the model predicted a negative class (PT only), and the actual class was also negative (PT only). True negative can be summarized as when the model has a correct rejection for a given mode choice.

Table 5-4 shows the confusion matrix of the first model. When considering the CD only option we see that the TP instances were 86.1%. We see that the instances for FN where the actual mode was CD only are split between other modes with Cycling/Walking having the highest value at 3.2%. This means that the model is mistakenly predicting CD only chains as Cycling/Walking the most, although the difference between all other modes is not great. For CP only chains, 43.2% were correctly predicted as TP. The FN instances that were CP only are split between the other modes with PT only having the highest value at 18.3% followed by CD only at 12.4% and Others unique at 11.9%. It's important to note that the FN instances that were Mix without car and Bimodal form a low 0.7% of all instances, a trend that is observed for almost all other modes. For Cycling/Walking the TP instances form 63.1%. From all FN instances, CD only and PT only form most of the instances at 23.2% and 9.3% respectively. This means that the model is mostly confusing the Cycling/Walking modes with these other modes, which can maybe be explained by the fact that chain distance is the main indicator of this mode, yet there are still very short chains that are done by CD only and to a lower extent by PT only. For PT only the TP instances were at 73.2%, while the FN were mostly split between Cycling/Walking (9.3%), CD only (6.4%) and CP only (6.2%). The Cycling/Walking FN instances could be down to these two modes being the most common for people with "no access to car" making them prime choices at that splitting criterion in the model. For the Mix with car we see a TP instances of 43.4%. Most of the FN instances are CD only (24.9%) and CP only (13.9%), which makes sense as these two modes form at least a part of the Mix with car mode, meaning that the model is likely having trouble knowing exactly when another mode is paired with one of these two modes. Notably, this is the only mode where we see a significant FN instance for Bimodal, which could be due to the fact that Bimodal utilizes the car in the first half of a trip. For Mix without car we see a TP value of 45.1%, while the FN instances are split almost equally between CD only, Cycling/Walking and PT only. Bimodal remains the mode with the least instance of TP at 9.0%. Most of the FN instances are identified as CD only at 42.2%. Finally for the Others unique chain mode choice, we see that the TP instances are at 56.8%. The FN instances are mostly Cycling/Walking or PT only which makes sense given that the

majority of the Others unique trips are actually school buses, where some students might cycle/walk there or take public transit.

Table 5-4 Confusion matrix for the 1st model

	Simulated choice							
Observed choice	**CD only**	**CP only**	**Cycling/Walking**	**PT only**	**Mix with car**	**Mix without car**	**Bimodal**	**Others Unique**
CD only	86.1%	2.4%	3.2%	2.1%	2.1%	1.2%	1.2%	1.7%
CP only	12.4%	43.2%	7.3%	18.3%	6.2%	0.5%	0.2%	11.9%
Cycling/Walking	23.2%	0.4%	63.1%	9.3%	0.3%	0.4%	0.8%	2.5%
PT only	6.4%	6.2%	9.3%	73.2%	1.3%	1.5%	0.6%	1.5%
Mix with car	24.9%	13.9%	4.6%	2.5%	43.4%	0.5%	8.3%	1.9%
Mix without car	13.4%	6.8%	12.8%	14.2%	4.2%	45.1%	1.0%	2.7%
Bimodal	42.2%	6.9%	6.3%	23.9%	7.7%	1.1%	9.0%	2.7%
Others Unique	7.6%	6.7%	10.3%	12.3%	2.0%	2.4%	1.9%	56.8%

5.2.3 2nd model: Distance to CBD as proxy for population density, presence of metro and intensity of transit stops

In this model the population density; presence of metro and intensity of transit stops variables were removed from the model. The idea is to see if the use of the distance to CBD by itself is enough to capture the effect of built environment features without decreasing the accuracy of the model, while also seeing how the feature importance changes for each variable. Table 5-5 summarizes the variables used in the 2nd model.

Table 5-5 Variables used for 2nd model

Personal variables	Household variables	Built environment variables	Trips/activities variable
Age	Number of people in the household	Presence of train station	Chain distance
Gender	Presence of children		Chain purpose
Possession of driver's license	Distance to CBD		Duration of activities
Main occupation	Access to car		Chain Complexity

5.2.3.1 Results

Model training accuracy: 82.5% (done on 80% of the data (88,234 chains)

Model testing accuracy: 72.1% (done one 20% of the data as validation (22,336 chains)

Number of trees used: 500

Although heavily correlated, we can see that the overall accuracy of the model drops from 75% to 72% when removing population density, presence of metro and intensity of transit stops and keeping distance to CBD only. For almost all modes, the precision of the model dropped a bit, especially for CD only chains as shown in Table 5-6.

Table 5-6 Precision and Recall results for the 1st model

Chain mode	precision	recall	observations
Cd only	0.82	0.84	13516
PT only	0.70	0.61	3238
Cycling/Walking	0.59	0.63	1900
CP only	0.42	0.36	1364
Bimodal	0.08	0.02	436
Others unique	0.58	0.34	360
Mixed with car	0.44	0.23	842
Mixed without car	0.41	0.21	680

5.2.3.2 Feature importance

When considering Mean Decrease Gini, the ranking of distance to CBD remains second in terms of importance, however its value increases from 0.12 in the first model to 0.16 in the second as shown in Figure 5-4. The largest takeaway we can see in feature importance happens in the Mean Decrease Accuracy. The distance to CBD, previously ranked 4th in the initial model, becomes the second most important variable as shown in Figure 5-5. This indicates that the model accuracy will be more affected by the removal of distance to CBD, if its correlated variables are no longer used. For other variables, few differences are observed.

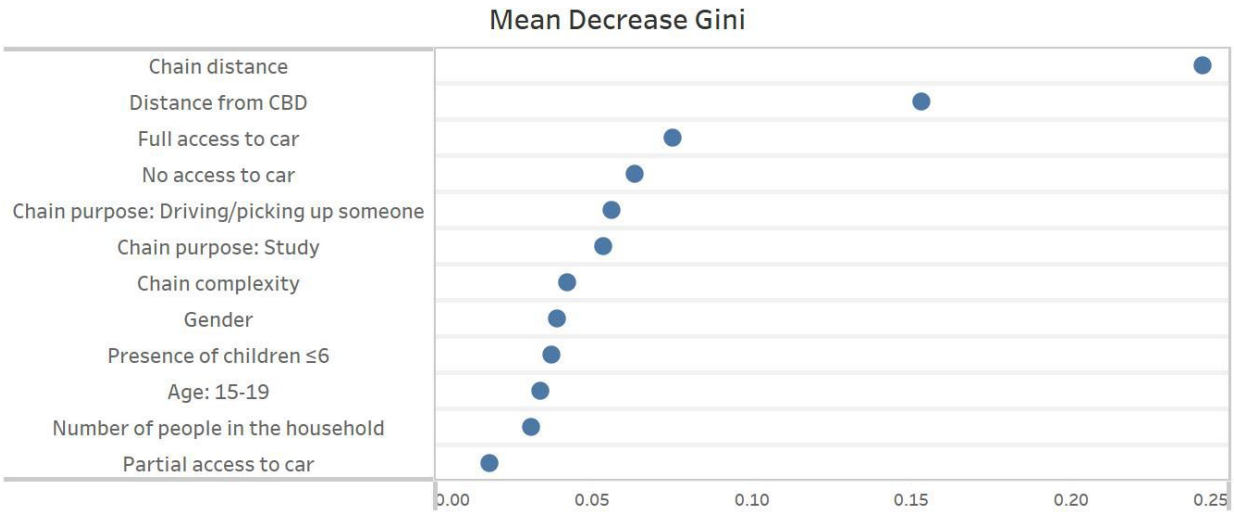

Figure 5-4 Feature importance: Mean Decrease Gini for the 2nd model

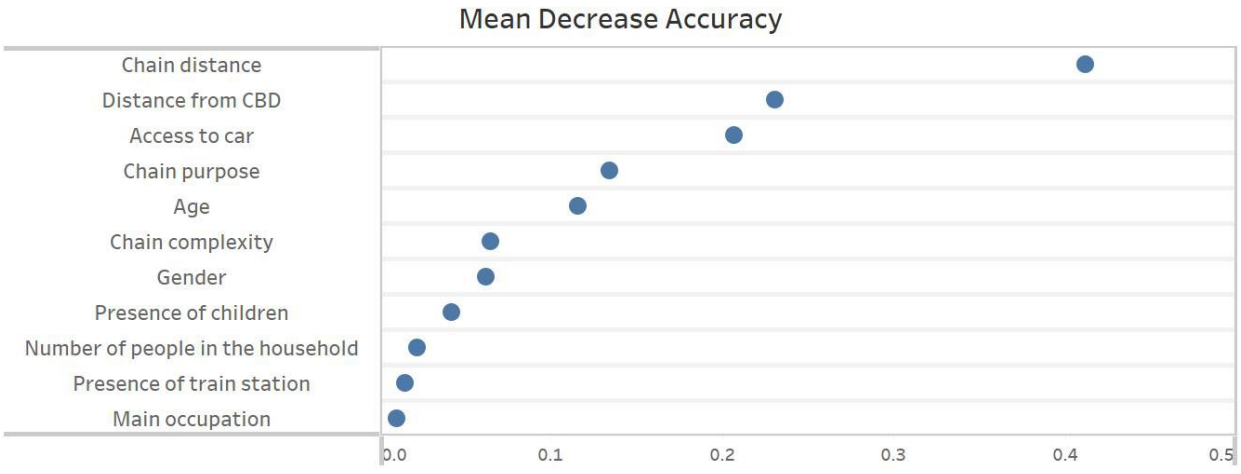

Figure 5-5 Feature importance: Mean Decrease Accuracy for the 2nd model

5.2.3.3 Confusion matrix

Table 5-7 shows the confusion matrix for the 2nd model. When compared to the first model every TP instance dropped except for Others unique and Mix with car that showed a slight increase. Overall, the results between the confusion matrix of the first and second model are not so different with some few notable differences. The FN instances for CD only were split mostly between Cycling/Walking and PT only instead of CP only as it is the case in the first model. The instances for FN for CP only increased for Mix with car from 6.2% to 13.2% while it dropped for Others unique from 11.9% to 5.2%.

Table 5-7 Confusion matrix for 2nd model

	Simulated choice							
Observed choice	**CD only**	**CP only**	**Cycling/Walking**	**PT only**	**Mix with car**	**Mix without car**	**Bimodal**	**Others Unique**
CD only	82.3%	1.2%	6.1%	5.1%	1.9%	0.9%	1.2%	1.3%
CP only	13.5%	42.3%	8.1%	16.3%	13.2%	0.6%	0.8%	5.2%
Cycling/Walking	21.2%	0.9%	59.3%	12.8%	0.3%	0.4%	0.8%	4.3%
PT only	10.2%	7.4%	5.7%	70.2%	1.3%	2.1%	0.6%	2.5%
Mix with car	23.7%	14.1%	4.5%	2.5%	44.4%	1.2%	7.9%	1.7%
Mix without car	13.4%	6.8%	15.7%	14.6%	4.2%	41.8%	0.8%	2.7%
Bimodal	43.3%	6.9%	6.3%	23.9%	7.7%	1.3%	8.4%	2.2%
Others Unique	10.6%	6.7%	9.3%	8.9%	2.0%	2.4%	1.9%	58.2%

5.2.4 3rd model: Distance to CBD not used

In this third model, distance to CBD is removed to see the effect on the model and how the feature importance for its correlated variables is influenced. Table 5-8 shows the variables used in the 3rd model.

Table 5-8 Variables used for 3rd model

Personal variables	Household variables	Built environment variables	Chain/activities variable
Age	Number of people in the household	Presence of metro station	Chain Complexity
Gender	Presence of children	Presence of train station	Chain distance
Possession of driver's license	Access to car	Population density	Chain purpose
Main occupation		Intensity of transit service	Duration of activities

5.2.4.1 Results

Model training accuracy: 83.2% (done on 80% of the data (88,234 chains)

Model testing accuracy: 72.7% (done on 20% of the data as validation (22,336 chains)

Number of trees used: 500

Although not by a significant margin, the accuracy of the model is a bit higher in this model when compared to just using the distance CBD as a proxy for these variables, but still less than the 1st model where all the variables are included. In terms of specific mode, the only noticeable difference in precision is observed for Mixed without car where the 3rd model has a 0.46 precision compared to 0.41 for the 2nd, however this difference is hard to interpret since this mode have few observations and the difference is not that important. The complete precision and recall values for each mode is shown in Table 5-9.

Table 5-9 Precision and recall results for 3rd model

Chain mode	precision	recall	observations
Cd only	0.83	0.87	13516
PT only	0.71	0.62	3238
Cycling/Walking	0.59	0.57	1900
CP only	0.41	0.44	1364
Bimodal	0.07	0.08	436
Others unique	0.57	0.34	360
Mixed with car	0.43	0.21	842
Mixed without car	0.46	0.25	680

5.2.4.2 Variables' importance

The difference in feature importance can be seen in the Mean Decrease Gini in Figure 5-6. Intensity of transit services and the presence of metro station became the 2nd and 3rd in rank. This means that the model replaces the splitting criteria found in the distance to CBD by utilizing these 2 variables more. In the Mean Decrease Accuracy (Figure 5-7), we see a slight increase in the importance of presence of metro, intensity of transit services and population density when compared to the 1st model.

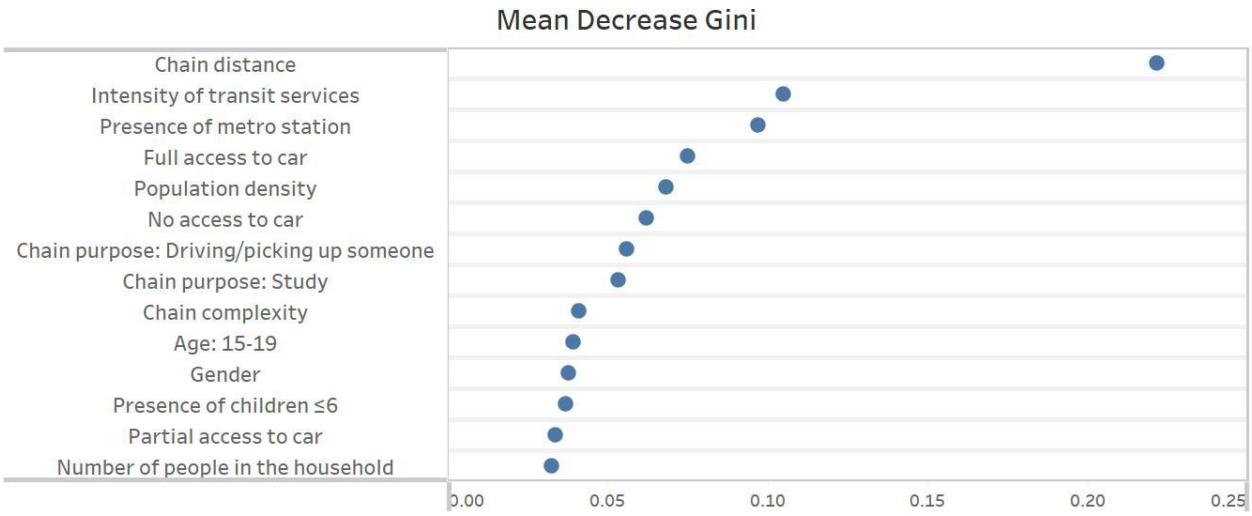

Figure 5-6 Feature importance: Mean Decrease Gini for the 3rd model

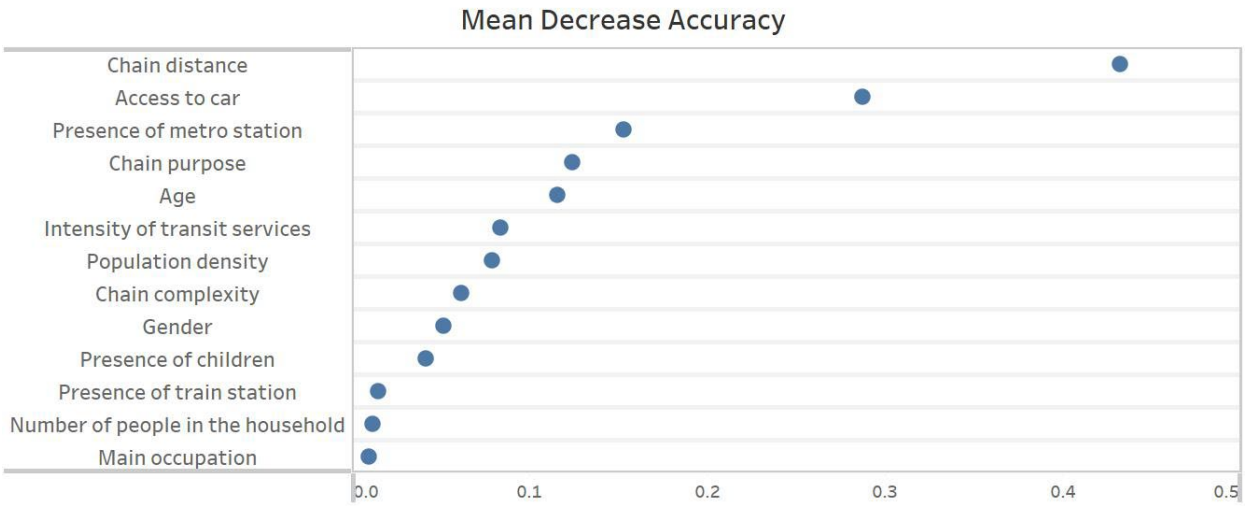

Figure 5-7 Feature importance: Mean Decrease Accuracy for the 3rd model

5.2.4.3 Confusion matrix

Table 5-10 shows the confusion matrix for the 3rd model. Overall, only one noticeable difference is seen when comparing the confusion matrix of the 2nd and 3rd model and that is the increase in TP instances of Mix without car which increased from 41.8% to 46.4%. This was accompanied by less FN instances for Cycling/Walking only, which could mean that the variable distance to CBD leads the model to predict more instances of Walking/Cycling only predictions when considering complex chains. Beyond that other values are basically the same with no differences above 1%.

Table 5-10 Confusion matrix for 3rd model

	Simulated choice							
Observed choice	**CD only**	**CP only**	**Cycling/Walking**	**PT only**	**Mix with car**	**Mix without car**	**Bimodal**	**Others Unique**
CD only	83.2%	1.1%	5.4%	5.4%	1.4%	1.1%	1.0%	1.4%
CP only	13.7%	41.4%	8.0%	15.6%	14.3%	0.8%	0.9%	5.3%
Cycling/Walking	21.1%	1.0%	59.2%	12.9%	0.6%	0.7%	0.8%	3.7%
PT only	9.9%	7.1%	5.2%	70.9%	1.1%	2.3%	0.6%	2.9%
Mix with car	23.9%	14.6%	4.3%	2.9%	43.1%	1.4%	7.9%	1.9%
Mix without car	12.4%	6.6%	12.3%	14.6%	4.3%	46.4%	0.8%	2.6%
Bimodal	43.1%	7.7%	6.3%	24.0%	7.7%	1.6%	7.4%	2.2%
Others Unique	11.2%	6.5%	9.5%	8.8%	2.2%	2.7%	1.9%	57.2%

5.2.5 4th model: Use of person's activity system as an independent variable

Given the recent literature that supports the idea that it is in fact the trip chaining that comes before mode choice in most cases, and the direction of this study to show how chain structure affects the chain mode choice, activity systems were added to the initial model to see:

1) If the model accuracy improves with the addition of this variable;
2) The relevancy of this variable in the feature importance.

Activity systems show the structure and the number of chains a person does during their day as discussed in chapter 2 and 3.

Table 5-11 Variables used in 4th model

Personal variables	Household variables	Built environment variables	Chain/activities variable
Age	Number of people in the household	Presence of metro station	Chain Complexity
Gender	Presence of children	Presence of train station	Chain distance
Possession of driver's license	Distance to CBD	Population density	Chain purpose
Main occupation	Access to car	Intensity of transit service	Duration of activities
			Activity systems

5.2.5.1 Results

Model training accuracy: 86.1% (done on 80% of the data (88,234 chains)

Model testing accuracy: 79.1% (done one 20% of the data as validation (22,336 chains)

Number of trees used: 600

The model training and testing results are both the highest in this model. The overall testing accuracy is 79.1%. When compared to the initial model, the prediction accuracy raises or stay the same for all possible chain mode choices. Both CD only and Mix with car showed significance improvement with the addition of activity system as a new variable, with the latter being in accordance with the relationship viewed earlier between the chain structure and the increase in the use of Mix with car mode choice (shown in section 4.2). This may explain why the model improved its precision the most for this particular mode as shown in Table 5-12.

Table 5-12 Precision and Recall results for 4th model

Chain mode	precision	recall	observations
Cd only	0.91	0.9	13516
PT only	0.76	0.62	3238
Cycling/Walking	0.65	0.53	1900
CP only	0.43	0.37	1364
Bimodal	0.1	0.08	436
Others unique	0.56	0.58	360
Mixed with car	0.51	0.47	842
Mixed without car	0.48	0.37	680

5.2.5.2 Variable's importance

As for feature importance, activity systems is the 3rd most significant variable in the Mean Decrease Gini shown in Figure 5-8. It is also notable how the chain complexity importance increases as well when the model uses activity system, which can imply that the model can learn more from the chain complexity if it also has the data of a person's activity system. For Mean Decrease Accuracy, activity systems was ranked 7th while chain complexity ranked 8th among all variables in terms of importance as seen in Figure 5-9.

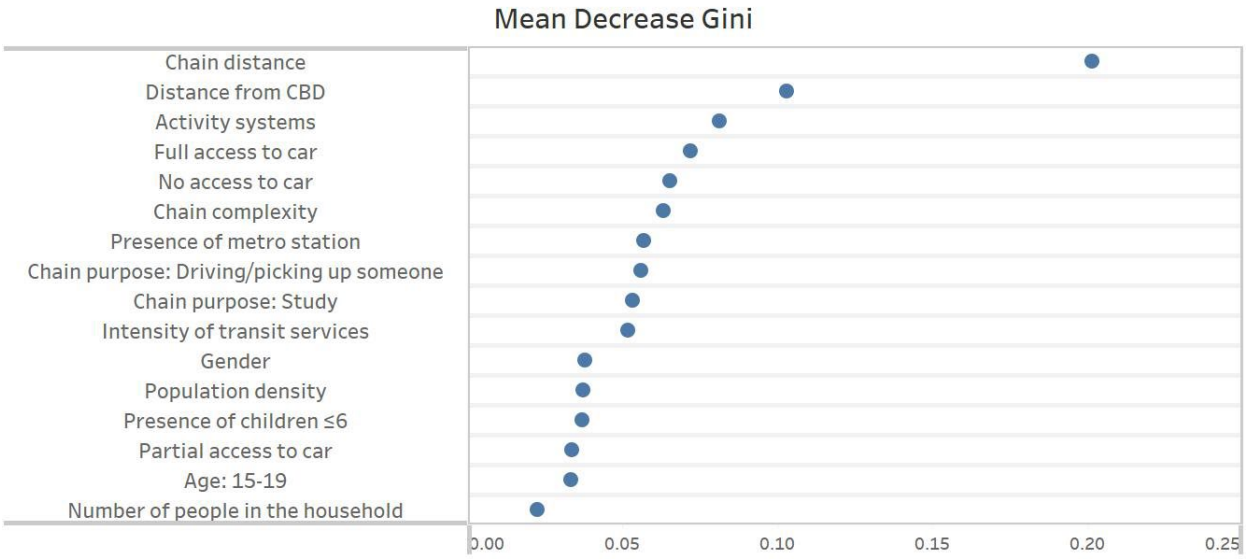

Figure 5-8 Feature importance: Mean Decrease Gini for the 4th model

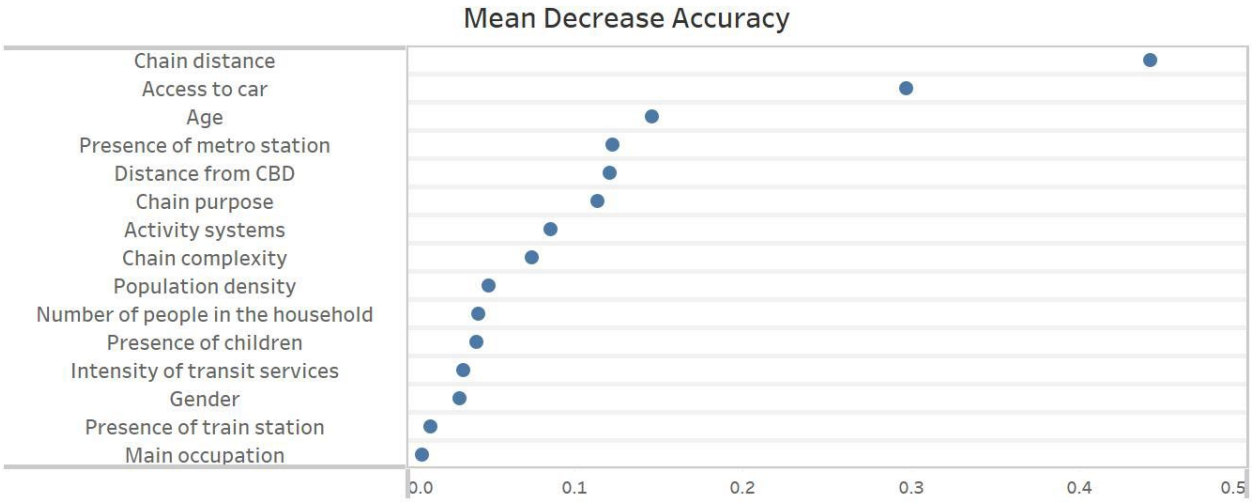

Figure 5-9 Feature importance: Mean Decrease Accuracy for the 4th model

5.2.5.3 Confusion matrix

The confusion matrix shown in Table 5-13represents the results of the 4th model. When compared to the first model the TP instances increased for all modes after the addition of the activity systems as a new variable in the model. No real noticeable differences for FN instances, but overall the model tends to have less FN instances for CD only predictions than the 1st model.

Table 5-13 Confusion matrix for 4th model

	Simulated choice							
Observed choice	**CD only**	**CP only**	**Cycling/Walking**	**PT only**	**Mix with car**	**Mix without car**	**Bimodal**	**Others Unique**
CD only	91.2%	2.1%	0.3%	2.1%	1.1%	1.2%	0.9%	1.1%
CP only	11.9%	43.2%	7.5%	18.5%	13.2%	0.6%	0.3%	4.8%
Cycling/Walking	22.8%	0.3%	65.1%	9.3%	0.2%	0.3%	0.1%	1.9%
PT only	3.2%	11.4%	4.3%	76.5%	1.3%	1.2%	0.7%	1.4%
Mix with car	23.1%	11.3%	2.7%	2.3%	51.3%	0.2%	8.0%	1.1%
Mix without car	11.4%	4.8%	12.8%	18.2%	2.2%	48.3%	0.2%	2.1%
Bimodal	43.2%	5.1%	6.3%	23.9%	8.0%	1.2%	10.0%	2.3%
Others Unique	12.5%	6.7%	9.3%	8.3%	2.0%	2.2%	2.1%	56.9%

5.2.6 5th model: predictive model

For the chain-based models tested above, chain variables such as chain distance and chain complexity play a big role in predicting the chain mode choice of the traveler. In reality though, when trying to predict randomly the travel behavior and mode choice of a given person at any given time, such variables are not known. This is where it was interesting to see how the random forest model can perform as a predictive model based on solely using the variables about the person, their household and built environment without using any trip / chain variables, even if removing such variables will almost undoubtedly lower the accuracy of the model. The variables used in this model are shown in Table 5-14.

Table 5-14 Variables used in predictive model

Personal variables	Household variables	Built environment variables
Age	Number of people in the household	Presence of metro station
Gender	Presence of children	Presence of train station
Possession of driver's license	Distance to CBD	Population density
Main occupation	Number of cars per licensed person	Intensity of transit service

5.2.6.1 Results

Model training accuracy: 71.3% (done on 80% of the data (88,234 chains)

Model testing accuracy: 65.6% (done one 20% of the data as validation (22,336 chains)

Number of trees used: 400

The results show the model provides a testing accuracy of 65.6%, which is impressive for a predictive model, especially given the under-sampling in less used chain mode alternatives. The model shows good predictions for the usage of CD only mode choices at 72% precision, and decent precision for PT only and Cycling/Walking only and Others unique. However it does not perform

well for the other choices as shown in Table 5-15. Still, such predictive model can help identify somehow accurately when a person is likely to make a CD only chain.

Table 5-15 Precision and recall results for the predictive model

Chain mode	precision	recall	observations
Cd only	0.72	0.68	13516
PT only	0.58	0.47	3238
Cycling/Walking	0.54	0.46	1900
CP only	0.31	0.1	1364
Bimodal	0.07	0.04	436
Others unique	0.55	0.49	360
Mixed with car	0.29	0.17	842
Mixed without car	0.21	0.15	680

5.2.6.2 Variables' importance

Interestingly, removing the chain variables helps to shed some light on how important other variables are. One of the most notable changes is that of the "Main occupation" variable, which was irrelevant in all previous models. However, when removing chain purpose this variable becomes of significant importance as shown in Figure 5-10. This is probably due to the main occupation being a good indicator of the main activity the person is going to do in their chains. As such, the model can use this variable as a replacement in the node splitting process. Access to car becomes the most important variable both in Mean Decrease Gini and Mean Decrease Accuracy (Figure 5-11), while gender and presence of children become more important in the "Gini" feature importance when compared to previous models.

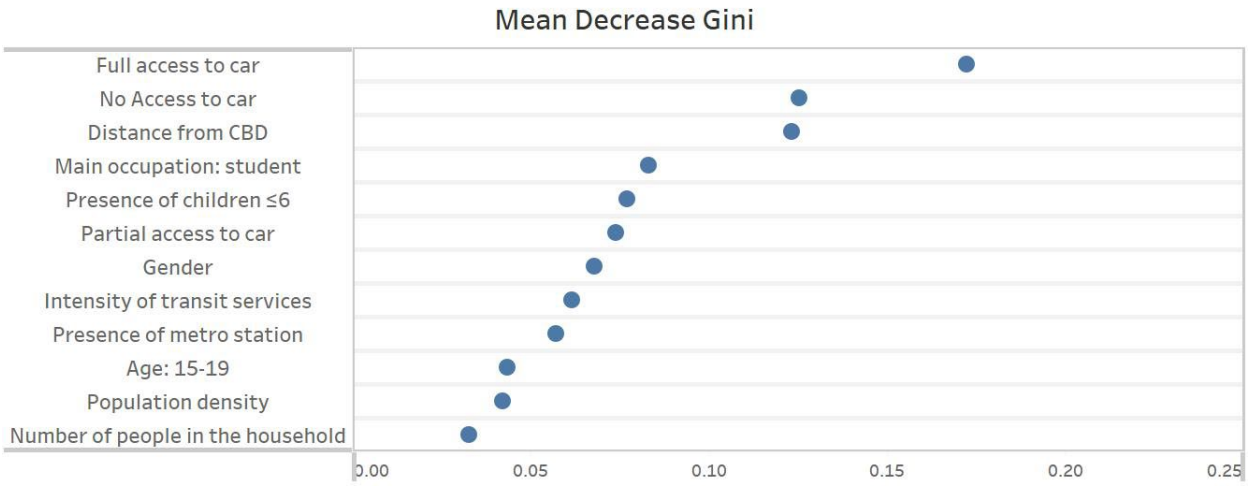

Figure 5-10 Feature importance: Mean Decrease Gini for the 5th model

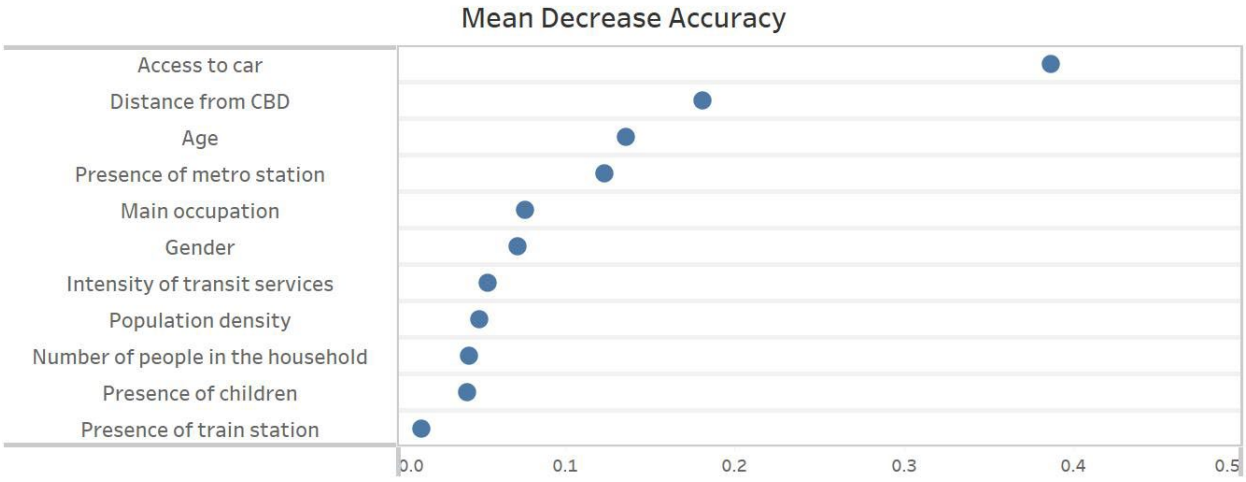

Figure 5-11 Feature importance: Mean Decrease Accuracy for the 5th model

5.2.6.3 Confusion matrix

When analyzing the confusion matrix for the predictive model, we can see how the model predictions are affected by the chain variables as shown in Table 5-16. The TP instances dropped moderately for CD only, CP only, PT only and Cycling/Walking with a drop around 15% for each, while those for Mix with car and Mix without car dropped significantly at around 25%. This means that the chains variables are important for the model to predict complex chains where more than one mode was used. Furthermore, we see no noticeable drop for Others unique, which can be since

this mode is mostly made of school buses trips, which can be predicted in the model via the age variable.

Table 5-16 Confusion matrix for predictive model

	Simulated choice							
Observed choice	**CD only**	**CP only**	**Cycling/Walking**	**PT only**	**Mix with car**	**Mix without car**	**Bimodal**	**Others Unique**
CD only	72.3%	1.2%	12.1%	7.9%	1.9%	0.9%	1.2%	2.5%
CP only	21.5%	31.4%	8.1%	18.3%	13.5%	0.7%	0.6%	5.9%
Cycling/Walking	24.2%	1.9%	53.9%	14.2%	0.3%	0.4%	0.8%	4.3%
PT only	16.4%	8.1%	9.9%	58.2%	1.5%	2.4%	0.7%	2.8%
Mix with car	33.7%	14.6%	7.8%	3.5%	29.6%	1.2%	7.9%	1.7%
Mix without car	19.0%	7.5%	19.7%	24.7%	4.2%	21.4%	0.8%	2.7%
Bimodal	48.3%	14.9%	2.1%	15.9%	7.7%	1.4%	7.1%	2.6%
Others Unique	10.9%	9.6%	9.5%	9.2%	2.5%	2.2%	1.1%	55.0%

5.2.7 Comparison between models

A comparison between the accuracy of all tested models is shown in Table 5-17.

Overall, all models considered performed exceptionally well in the accuracy of predicting the CD only chain mode choice. The 4th model where activity systems were implemented showed an accuracy of 91% which was the highest accuracy for any mode choice across all models. The models also showed good accuracy in the prediction of PT only, Cycling/Walking modes and Others unique. It is important to note that the relatively good prediction of Others unique is heavily affected by how the school bus choice makes the most of the observations for this mode, and how perhaps it is easy for the model to split this node using the "Age:15-19" class. Elsewhere all models performed on average around 43% accuracy for modes Mixed with car, Mixed without car and CP only. Bimodal mode remains problematic to predict across all models, which opens the perspective on how better model this choice.

Table 5-17 Precision comparison between all models

Chain mode	1st model	2nd model	3rd model (no	4th model	5th model
Cd only	86.1%	82.3%	83.2%	91.2%	72.3%
PT only	73.2%	70.2%	70.9%	76.5%	58.2%
Cycling/Wal	63.1%	59.3%	59.2%	65.1%	53.9%
CP only	43.2%	42.3%	41.4%	43.2%	31.4%
Bimodal	9.0%	7.9%	7.4%	10.0%	7.1%
Others unique	56.8%	58.2%	57.2%	56.9%	55.0%
Mixed with	43.4%	44.4%	43.1%	51.3%	29.6%
Mixed	45.1%	41.8%	46.4%	48.3%	21.4%
Total	75.6%	72.1%	72.7%	79.1%	65.6%

CHAPTER 6 CONCLUSION

6.1 Contributions

This research project offers some contributions on different levels. In terms of literature review, this research presents the different determinants of mode choice, especially when considering the trip chain level, and dives into different typologies and chain definitions defined through the literature. The recent rise in the use of machine learning algorithms and especially the random forest as a tool to model mode choice was also examined, showing how these models are generally providing impressive results.

In the methodology, the preparation of the variables based on the latest ODS of the GMA to be incorporated in the random forest model was explained. A method of identifying the chain mode choice observed in the data was implemented where different modes were considered but without the generation of too many alternatives for complex chains. Other variables were also constructed based on previous works done in the GMA.

In the exploratory analysis, several trends of the trip-chaining behavior for the people of Montreal were examined. The evolution of chain complexity, chains per day, duration, distance, and purpose were shown from 1998 to 2018. Furthermore, the chain modal split and activity system evolution were shown throughout the years. Takeaways of the trends observed showed an increase in the complexity of chains especially for women, and an overall increase in chain total distance.

The relationship between different variables of different types (socio-demographic, built environment, household, and chain) and the chain mode choice was observed for the latest travel data of the ODS of GMA in 2018.

In terms of modeling, few works are done that consider modelling on the trip-chain level, especially for the city of Montreal where the latest attempt was done in 2014 albeit without the consideration of complex chains. This research project contributes with a trip chain-based mode choice model utilizing the random forest tool which is growing in popularity. The models showed extremely impressive results in terms of accuracy, especially for the prediction of Car Driving only chains.

6.2 Limitations

Some limitations are to be considered when discussing this research project. First, while the alternatives used to summarize the large number of alternatives possible for trip chains are effective, especially for chains conducted with 1 mode only, they still have their drawbacks especially when considering the Mixed with car and Mixed without car mode choices. Indeed, while these chains only form less than 10% of all chains, they represent the possibilities of mode combinations within the chain. As such, not knowing the exact alternatives used or the order they are used in negatively affects how well we can draw some conclusions and interpretations. The same line of thought is true but to a lesser extent for activity system, where the system "others" hold various possibilities and thus cannot be interpretated correctly.

Other limitations lie within the variables of the built environment. Several other variables not mentioned in the study such as land use mix and availability of parking could be of importance, but such variables are harder to obtain or develop with good quality. Also, while considering built environment variable, the distances considered for the access distance for metro station and especially train station are still only rough estimation based on the literature, while in reality each metro station and train station could have a different access distance depending on many different and complex variables such as population density, land use patterns, transportation infrastructure, and the availability of parking facilities (Chakour & Eluru, 2014; Martínez, Moyano, Coronado, & Garmendia, 2016). Train stations in particular are less explored in the literature and have a far wider range of catchment areas, especially when considering the fact that these stations can have both "an active access range" and "car access range". These ranges were not differentiated in this study and could lead to better significance for the presence of a train station.

The method of calculating the transit service per 24h is also limited in the fact that it counts for the same bus passing in several stops within the buffer zone, while in reality a person is most likely to look for the bus in one bus stop from their home, most likely the one closest to their house. This method also serves to increase the service for households located within proximities of several bus stops where a single route can serve multiple stops.

In the modeling process, the most important limitation faced is that of the modeling of lower share modes such as Bimodal, Mix with car and Mix without car. These rare events led to fewer training

attempts for the model when compared to other modes, and thus this could explain why the prediction of these modes were less impressive in the models.

Finally, while the correlation between variables does not affect the accuracy of the random forest models it could still heavily affect the feature importance rank of these variables. This was observed in the predictive model where the variable “main occupation” became significant after the removal of the chain purpose variable.

6.3 Perspectives

Several parts of this research can be improved by incorporating some few recommendations.

It is interesting figure out a way that allows the use of the exact alternatives for the Mixed without car and Mixed with car mode choices. This would offer more insights about possible mode combinations and which of these combinations are easier to predict for the model and are used the most.

Incorporating more built-environment variables might also be a nice addition and could lead to better models. Along with the development of more accurate ways that would help establish the access distances for metros and train stations for Montreal. The addition of latent variables such as travel pattern behaviors or personal attitudes and perceptions could also help better capture the complex decision making that goes into chaining trips.

Furthermore, some methods are present in the literature that can deal with problems of rare events modeling by either over-sampling these events randomly or under sampling the majority, which could offer a better result when it comes to the modelling of less represented modes such as Bimodal.

REFERENCES

Bhat, C. R., & Sardesai, R. (2006). The impact of stop-making and travel time reliability on commute mode choice. *Transportation Research Part B: Methodological, 40*(9), 709-730. doi:https://doi.org/10.1016/j.trb.2005.09.008

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Briscoe, M. D., Givens, J. E., Hazboun, S. O., & Krannich, R. S. (2019). At home, in public, and in between: gender differences in public, private and transportation pro-environmental behaviors in the US Intermountain West. *Environmental Sociology, 5*(4), 374-392. doi:10.1080/23251042.2019.1628333

Cervero, R., & Gorham, R. (1995). Commuting in Transit Versus Automobile Neighborhoods. *Journal of the American planning association, 61*(2), 210-225. doi:10.1080/01944369508975634

Chakour, V., & Eluru, N. (2014). Analyzing commuter train user behavior: a decision framework for access mode and station choice. *Transportation, 41*(1), 211-228. doi:10.1007/s11116-013-9509-y

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society, 14*, 1-10. doi:https://doi.org/10.1016/j.tbs.2018.09.002

Chica-Olmo, J., & Lizárraga, C. (2022). Effect of Interaction between Distance and Travel Times on Travel Mode Choice when Escorting Children to and from School. *Journal of Urban Planning and Development, 148*(1), 05021055. doi:doi:10.1061/(ASCE)UP.1943-5444.0000776

Circella, G., Alemi, F., Berliner, R., Tiedeman, K., Lee, Y., Fulton, L., Handy, S., Mokhtarian, P. L. (2017). *The multimodal behavior of millennials: exploring differences in travel choices between young adults and Gen Xers in California.* Paper presented at the 96th Transportation Research Board Annual Meeting.

Cirillo, C., & Axhausen, K. W. (2002). Mode choice of complex tours: A panel analysis. *Arbeitsberichte Verkehrs-und Raumplanung, 142*.

Currie, G., & Delbosc, A. (2011). Exploring the trip chaining behaviour of public transport users in Melbourne. *Transport policy, 18*(1), 204-210. doi:https://doi.org/10.1016/j.tranpol.2010.08.003

Delbosc, A., & Nakanishi, H. (2017). A life course perspective on the travel of Australian millennials. *Transportation Research Part A: Policy and Practice, 104*, 319-336.

Ding, C., Cao, X., & Wang, Y. (2018). Synergistic effects of the built environment and commuting programs on commute mode choice. *Transportation Research Part A: Policy and Practice, 118*, 104-118. doi:https://doi.org/10.1016/j.tra.2018.08.041

Duncan, M., & Christensen, R. K. (2013). An analysis of park-and-ride provision at light rail stations across the US. *Transport policy, 25*, 148-157. doi:https://doi.org/10.1016/j.tranpol.2012.11.014

Elias, W., Benjamin, J., & Shiftan, Y. (2015). Gender differences in activity and travel behavior in the Arab world. *Transport policy, 44*, 19-27. doi:https://doi.org/10.1016/j.tranpol.2015.07.001

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

Frank, L., Bradley, M., Kavage, S., Chapman, J., & Lawton, T. K. (2008). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation, 35*(1), 37-54. doi:10.1007/s11116-007-9136-6

Gardner, N., Cui, J., & Coiacetto, E. (2017). Harassment on public transport and its impacts on women's travel behaviour. *Australian Planner, 54*(1), 8-15. doi:10.1080/07293682.2017.1299189

Godefroy, F. (2011). *Méthodologie de caractérisation du vélopartage et d'estimation du marché potentiel du vélo à Montréal.* École Polytechnique de Montréal,

Goulias, K. G., & Kitamura, R. (1991). Recursive model system for trip generation and trip chaining.

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications, 78*, 273-282. doi:https://doi.org/10.1016/j.eswa.2017.01.057

Hasnine, M. S., & Nurul Habib, K. (2021). Tour-based mode choice modelling as the core of an activity-based travel demand modelling framework: a review of state-of-the-art. *Transport Reviews, 41*(1), 5-26.

Hensher, D. A., & Ton, T. T. (2000). A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. *Transportation Research Part E: Logistics and Transportation Review, 36*(3), 155-172.

Hilgert, T., von Behren, S., Eisenmann, C., & Vortisch, P. (2018). Are Activity Patterns Stable or Variable? Analysis of Three-Year Panel Data. *Transportation Research Record, 2672*(47), 46-56. doi:10.1177/0361198118773557

Ho, C. Q., & Mulley, C. (2013). Multiple purposes at single destination: A key to a better understanding of the relationship between tour complexity and mode choice. *Transportation Research Part A: Policy and Practice, 49*, 206-219. doi:https://doi.org/10.1016/j.tra.2013.01.040

Holzapfel, H. (1986). Trip relationships in urban areas.

Horowitz, A. J. (1982). A comparison of socioeconomic and structural determinants of trip tour length. *Papers of the Regional Science Association, 50*(1), 185-195. doi:10.1007/BF01940120

Horowitz, J. L. (1993). Semiparametric estimation of a work-trip mode choice model. *Journal of Econometrics, 58*(1-2), 49-70.

Huang, Y., Gao, L., Ni, A., & Liu, X. (2021). Analysis of travel mode choice and trip chain pattern relationships based on multi-day GPS data: A case study in Shanghai, China. *Journal of Transport Geography, 93*, 103070. doi:https://doi.org/10.1016/j.jtrangeo.2021.103070

Islam, M. (2010). Unraveling the relationship between trip chaining and mode choice using structural equation models.

Islam, M. T., & Habib, K. M. N. (2012). Unraveling the relationship between trip chaining and mode choice: evidence from a multi-week travel diary. *Transportation Planning and Technology, 35*(4), 409-426.

Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies, 19*(3), 387-399.

Keijer, M. J. N., & Rietveld, P. (2000). How do people get to the railway station? The dutch experience. *Transportation Planning and Technology, 23*(3), 215-235. doi:10.1080/03081060008717650

Kim, K., Kwon, K., & Horner, M. W. (2021). Examining the effects of the built environment on travel mode choice across different age groups in Seoul using a random forest method. *Transportation Research Record, 2675*(8), 670-683.

Kim, S., Ulfarsson, G. F., & Todd Hennessy, J. (2007). Analysis of light rail rider travel behavior: Impacts of individual, built environment, and crime characteristics on transit access. *Transportation Research Part A: Policy and Practice, 41*(6), 511-522. doi:https://doi.org/10.1016/j.tra.2006.11.001

Koppelman, F. S., & Bhat, C. (2006). A self instructing course in mode choice modeling: multinomial and nested logit models.

Krizek, K. J. (2003). Neighborhood services, trip purpose, and tour-based travel. *Transportation, 30*(4), 387-410. doi:10.1023/A:1024768007730

Krygsman, S. (2004). *Activity and travel choice (s) in multimodal public transport systems*: Utrecht University.

Krygsman, S., Arentze, T., & Timmermans, H. (2007). Capturing tour mode and activity choice interdependencies: A co-evolutionary logit modelling approach. *Transportation Research Part A: Policy and Practice, 41*(10), 913-933. doi:https://doi.org/10.1016/j.tra.2006.03.006

Lee, J., He, S. Y., & Sohn, D. W. (2017). Potential of converting short car trips to active trips: The role of the built environment in tour-based travel. *Journal of Transport & Health, 7*, 134-148. doi:https://doi.org/10.1016/j.jth.2017.08.008

Li, S. (2021). Research on the Choice Behavior of American Elderly Trip Chain Based on MNL. *International Journal of Social Science and Education Research, 4*(3), 43-52.

Limtanakool, N., Dijst, M., & Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips. *Journal of Transport Geography, 14*(5), 327-341. doi:https://doi.org/10.1016/j.jtrangeo.2005.06.004

Lin, T., Xia, J., Robinson, T. P., Olaru, D., Smith, B., Taplin, J., & Cao, B. (2016). Enhanced Huff model for estimating Park and Ride (PnR) catchment areas in Perth, WA. *Journal of Transport Geography, 54*, 336-348. doi:https://doi.org/10.1016/j.jtrangeo.2016.06.011

Lindsey, M., Schofer, J. L., Durango-Cohen, P., & Gray, K. A. (2010). Relationship between proximity to transit and ridership for journey-to-work trips in Chicago. *Transportation Research, Part A (Policy and Practice), 44*(9), 697-709. doi:10.1016/j.tra.2010.07.003

Litman, T. (2012). Evaluating accessibility for transportation planning. *Victoria: Victoria Transport Policy Institute*.

Liu, C., Sun, Y., Chen, Y., & Susilo, Y. O. (2018). The effect of residential housing policy on car ownership and trip chaining behaviour in Hangzhou, China. *Transportation Research Part D: Transport and Environment, 62*, 125-138. doi:https://doi.org/10.1016/j.trd.2018.02.008

Lu, X., & Pas, E. I. (1999). Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice, 33*(1), 1-18. doi:https://doi.org/10.1016/S0965-8564(98)00020-2

Ma, J., & Goulias, K. (1998). *Forecasting home departure time, daily time budget, activity duration and travel time using panel data.* Paper presented at the Transportation Research Board Annual Meeting, Washington, DC.

Ma, J., Mitchell, G., & Heppenstall, A. (2014). Daily travel behaviour in Beijing, China: An analysis of workers' trip chains, and the role of socio-demographics and urban form. *Habitat International, 43*, 263-273. doi:https://doi.org/10.1016/j.habitatint.2014.04.008

Maat, K., & Timmermans, H. (2006). Influence of Land use on Tour Complexity:A Dutch Case. *Transportation Research Record, 1977*(1), 234-241. doi:10.1177/0361198106197700127

Malichová, E., Cornet, Y., & Hudák, M. (2022). Travellers' use and perception of travel time in long-distance trips in Europe. *Travel Behaviour and Society, 27*, 95-106. doi:https://doi.org/10.1016/j.tbs.2021.12.003

Martel Poliquin, É. (2012). *Mieux comprendre les déterminants du choix modal.* École polytechnique de Montréal,

Martínez, H. S., Moyano, A., Coronado, J. M., & Garmendia, M. (2016). Catchment areas of high-speed rail stations: a model based on spatial analysis using ridership surveys. *European Journal of Transport and Infrastructure Research, 16*(2). doi:10.18757/ejtir.2016.16.2.3143

McFadden, D. (1974). The measurement of urban travel demand. *Journal of public economics, 3*(4), 303-328.

McGuckin, N., & Murakami, E. (1999). Examining Trip-Chaining Behavior: Comparison of Travel by Men and Women. *Transportation Research Record, 1693*(1), 79-85. doi:10.3141/1693-12

Mohd Ali, N. F., Mohd Sadullah, A. F., Abdul Majeed, A. P. P., Mohd Razman, M. A., & Musa, R. M. (2022). The identification of significant features towards travel mode choice and its prediction via optimised random forest classifier: An evaluation for active commuting behavior. *Journal of Transport & Health, 25*, 101362. doi:https://doi.org/10.1016/j.jth.2022.101362

Montréal, C. M. d. (2019). *Un Grand Montréal attractif, compétitif et durable : Plan Métropolitain d'Aménagement et de Développement*. Retrieved from https://cmm.qc.ca/wp-content/uploads/2019/03/pmad_plan_metropolitain_amenagement_developpement.pdf

Niemeier, D. A., & Morita, J. G. (1996). Duration of trip-making activities by men and women. *Transportation, 23*(4), 353-371. doi:10.1007/BF00223061

Noland, R. B., & Thomas, J. V. (2007). Multivariate Analysis of Trip-Chaining Behavior. *Environment and Planning B: Planning and Design, 34*(6), 953-970. doi:10.1068/b32120

Olszewski, P., & Wibowo, S. S. (2005). Using equivalent walking distance to assess pedestrian accessibility to transit stations in Singapore. *Transportation Research Record, 1927*(1), 38-45.

Ortega, J., Hamadneh, J., Esztergár-Kiss, D., & Tóth, J. (2020). Simulation of the daily activity plans of travelers using the park-and-ride system and autonomous vehicles: work and shopping trip purposes. *Applied Sciences, 10*(8), 2912.

Paleti, R., Pendyala, R. M., Bhat, C. R., & Konduri, K. C. (2011). *A joint tour-based model of tour complexity, passenger accompaniment, vehicle type choice, and tour length*. Retrieved from

Płoński, P. (2020). Random forest feature importance computed in 3 ways with python. In.

Primerano, F., Taylor, M. A., Pitaksringkarn, L., & Tisato, P. (2008). Defining and understanding trip chaining behaviour. *Transportation, 35*(1), 55-72.

Rodrigue, J.-P. (2020). *The Geography of Transport Systems (5th ed.)* (5th ed.).

Roorda, M. J., Passmore, D., & Miller, E. J. (2009). Including minor modes of transport in a tour-based mode choice model with household interactions. *Journal of Transportation Engineering, 135*(12), 935-945.

Sabouri, S. (2021). Assessing polycentric development in terms of trip chaining efficiency. *Cities, 117*, 103300. doi:https://doi.org/10.1016/j.cities.2021.103300

Schlossberg, M., Agrawal, A. W., Irvin, K., & Bekkouch, V. L. (2007). How far, by which route, and why? A spatial analysis of pedestrian preference.

Schmöcker, J.-D., Su, F., & Noland, R. B. (2010). An analysis of trip chaining among older London residents. *Transportation, 37*(1), 105-123. doi:10.1007/s11116-009-9222-z

Schneider, F., Ton, D., Zomer, L.-B., Daamen, W., Duives, D., Hoogendoorn-Lanser, S., & Hoogendoorn, S. (2021). Trip chain complexity: a comparison among latent classes of daily mobility patterns. *Transportation, 48*(2), 953-975. doi:10.1007/s11116-020-10084-1

Sekhar, C. R., Minal, & Madhu, E. (2016). Mode Choice Analysis Using Random Forrest Decision Trees. *Transportation Research Procedia, 17*, 644-652. doi:https://doi.org/10.1016/j.trpro.2016.11.119

Shiftan, Y. (1998). Practical Approach to Model Trip Chaining. *Transportation Research Record, 1645*(1), 17-23. doi:10.3141/1645-03

Sicotte, G. (2014). *Modélisation de l'interdépendance entre modes de transport et chaînes de déplacements.* École Polytechnique de Montréal,

Souza, F., Bodmer, M., Zuidgeest, M., Brussel, M., & Amer, S. (2010). *To cycle or not to cycle? Factors influencing the decision to use the bicycle as access mode to public transport.* Paper presented at the 12th World Conference on Transport Research.

Stopher, P. (1969). A multinomial extension of the binary logit model for choice of mode of travel. *Northwestern University, unpublished, 5*.

Taru, S., Kr, V. A., & Rao, N. V. M. (2021). Gender Differences in Influence of Socio-demographic Characteristics on Mode Choice in India. *The Journal of Asian Finance, Economics and Business, 8*(1), 531-542. doi:10.13106/JAFEB.2021.VOL8.NO1.531

Thill, J. C., & Thomas, I. (1987). Toward conceptualizing trip-chaining behavior: A review. *Geographical Analysis, 19*(1), 1-17.

Timmermans, H., van der Waerden, P., Alves, M., Polak, J., Ellis, S., Harvey, A. S., . . . Zandee, R. (2003). Spatial context and the complexity of daily travel patterns: an international comparison. *Journal of Transport Geography, 11*(1), 37-46. doi:https://doi.org/10.1016/S0966-6923(02)00050-9

Valiquette, F. (2010). *Typologie des chaînes de déplacements et modélisation descriptive des systèmes d'activités des personnes.* École Polytechnique de Montréal,

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.

Vleugels, I., Steenbergen, T., Vande Walle, S., & Cornélis, E. (2005). Déterminants des choix modaux dans les chaînes de déplacements, plan d'appui scientifique à une politique de développement durable. In: Suisse.

Wang, X. (2019). Has the relationship between urban and suburban automobile travel changed across generations? Comparing Millennials and Generation Xers in the United States. *Transportation Research Part A: Policy and Practice, 129*, 107-122. doi:https://doi.org/10.1016/j.tra.2019.08.012

Xianyu, J. (2013). An exploration of the interdependencies between trip chaining behavior and travel mode choice. *Procedia-Social and Behavioral Sciences, 96*, 1967-1975.

Ye, N., Gao, L., Juan, Z., & Ni, A. (2018). Are People from Households with Children More Likely to Travel by Car? An Empirical Investigation of Individual Travel Mode Choices in Shanghai, China. *Sustainability, 10*(12), 4573. Retrieved from https://www.mdpi.com/2071-1050/10/12/4573

Ye, X., Pendyala, R. M., & Gottardi, G. (2007). An exploration of the relationship between mode choice and complexity of trip chaining patterns. *Transportation Research Part B: Methodological, 41*(1), 96-113. doi:https://doi.org/10.1016/j.trb.2006.03.004

Yun, M., Chen, Z., & Liu, J. (2014). *Comparison of Mode Choice Behavior for Work Tours and Non-work Tours Considering Trip Chain Complexity*.

Yun, M., Liu, J., & Yang, X. (2011). Modeling on mode choice behavior based on trip chaining: a case study in Zhongshan city. In *ICCTP 2011: Towards Sustainable Transportation Systems* (pp. 825-835).

Zhang, M., & Li, Y. (2022). Generational travel patterns in the United States: New insights from eight national travel surveys. *Transportation Research Part A: Policy and Practice, 156*, 1-13. doi:https://doi.org/10.1016/j.tra.2021.12.002

Zhu, P., & Guo, Y. (2022). Telecommuting and trip chaining: Pre-pandemic patterns and implications for the post-pandemic world. *Transportation Research Part D: Transport and Environment, 113*, 103524. doi:https://doi.org/10.1016/j.trd.2022.103524