Technische Universität München - Modellierung Räumlicher Mobilität

Univ.-Prof. Dr.-Ing. Rolf Moeckel

Augustenstraße 44 80333 München, www.msm.bgu.tum.de

MASTER'S THESIS

Long-Distance Mode Choice Modeling of Ontario Province

Author:

Joanna Yuhang Ji

Supervision:

Prof. Dr. -Ing. Rolf Moeckel

Date of Submission: 2017-06-02

Abstract

Transportation planning has always been limited by data availability. Data aggregation and collection can be time and resource intensive. Luckily, today, we are in an era of hitherto unknown abundance, with large amounts of data and data aggregators being readily available online. Though there are many mode data sources for short-distance travel, such as the General Transit Feed Specification (GTFS), long-distance travel data are still quite scarce and disaggregated. It is not yet clear how best to harness the potential of these data for travel demand modeling.

The modeling area of this thesis, Ontario, Canada, presents a special challenge in this respect, as it is geographically large and has an extremely unevenly distributed population density. This paper describes the development process of a long-distance mode-choice model for Ontario, using a novel approach of open data from an online trip planner combined with traditional survey data.

Rome2rio is an international door-to-door multi-modal trip building travel search platform. The website builds complete trips using available long- and short-distance, public and private modes. By utilizing its search API, it is possible to quickly collect vast amounts of precise, origin-to-destination mode data. To estimate this mode choice model, mode attributes of auto, air, train and bus travel were gathered between all zone pairs using the Rome2rio API. Traditionally, the scope of this project would have involved significant efforts to manually collect or estimate such data from disparate sources.

The Rome2rio data was then combined with the Travel Survey of the Residents of Canada to estimate multinomial logit mode choice models for the business, leisure and visit trip purposes. The models include attributes specific to the mode, person and trip. The resulting estimated models match the overall modal share trends and are sensitive to level-of-service changes. Results show the viability of applying aggregated open source online travel data in long-distance mode choice modeling.

Acknowledgements

I would like to acknowledge my advisor, Prof. Dr. Rolf Moeckel, for his knowledge, guidance, his encouragement to pursue new ideas, and for his trust in my abilities. Furthermore, I want to thank Dr. Carlos Llorca, for his help and advice all along the way. I would like to acknowledge the WSP team as well for their background support with necessary data and suggestions. I also want to say thank you to Dr. Ana Moreno and my other lovely colleagues who are always kind, helpful and good company.

To my family, thank you for providing me with the opportunity to do what I want to do. To my friends near and far, I send my love and gratitude. Finally, I want to thank Christian Mynster Andersen, without whose unwavering patience and support I would have been adrift.

Table of Contents

1. Introduction

In an era of increasing global mobility, we must address the additional strain on our intercity travel systems, from congested highways to overcrowded airports. This, coupled with the current zeitgeist of environmental impact awareness and economic sensitivity, makes it imperative to have reliable intercity mode choice models that can assess proposed intercity transportation improvements.

According to the 2009 National Household Travel Survey (NHTS), though long-distance trips account for less than one-percent of all vehicle trips, they make up 15.5 percent of all vehicle miles traveled. (Schiffer, 2012). Long-distance travel has an outsized impact on transport systems and travel-related emissions. This highlights the importance of accurately modeling long-distance travel, but also hints at its inherent difficulty: the small actual number of long-distance journeys conducted means long-distance travel modeling suffers from a lack of data.

The Ministry of Transportation of Ontario (MTO) is building a provincial transport model. An integral part of the model is the long-distance travel model, of which mode choice is a component. Making a mode choice model for Ontario presents some unique challenges. The region to be modeled is quite large, as Ontario boasts an area of more than 1 million km^2 ("Ontario Fact Sheet", 2017). The population concentration is imbalanced and congregate in the southwestern part of the province ("Ontario Fact Sheet", 2017).

I was provided with the Travel Survey of Residents of Canada (TSRC), which gives revealed preference data of Canadian residents' long-distance travel behavior and presents a chance to derive an intercity mode choice model econometrically. However, it was not a targeted survey for mode-choice modeling and is therefore missing relevant level-of-service attributes of the mode used. In addition, to compare the utility across modes, it is necessary to know the attributes of all modes in the choice set, chosen and unchosen. The time and resources spent on gathering this data depends on the zonal resolution and availability. There must be a balanced compromise between spatial resolution and data availability, especially for such a large model area.

With the advent of the internet, more data has been made available than ever before. Many open sourced data are already being integrated into travel demand modeling (Toole et al., 2014). However, this has been happening relatively slowly in long-distance modeling. More recently, several international intermodal, multimodal trip planning platforms have arisen.

These platforms have centralized short- and long-distance travel data for all available modes. Rome2rio in particular has door-to-door global coverage and supplies an applicable programming interface (API) for research purposes. By using the trip planner as a data resource, I can consistently and quickly gather modal data for all origin-destination (OD) pairs. The model is then estimated with data sourced from Rome2rio and the TSRC.

1.1. Outline

This thesis goes over the state-of-the-art and relevant background literature in Chapter 2. Chapter 3 discusses the data sources and presents the data collection methods used to estimate the model. Chapter 4 details the model estimation process of the multinomial logit and nested logit models and shows the model results and performance. Chapter 5 focuses on the discussion and results of the thesis, as well as limitations and suggestions for future work.

2. Literature review of long-distance mode choice modeling

Transportation projects and policies are often costly. To evaluate their impact, planners turn to travel demand forecasting and modeling as one of primary tools to predict travel demand. The current most accepted paradigm of travel demand forecasting is the seminal four-step model (FSM), so named because it breaks down the travel demand estimation into four steps. The four-step model was originally pioneered in the 1950s and 60s in Detroit, Chicago (Miller, 2001). In the time since, FSM has seen many developments and weathered much criticism, but it remains the most popular and practical approach to this day. The basic four steps of the process are as follows:

1. Trip Generation
2. Trip Distribution
3. Mode Choice
4. Trip Assignment

This thesis focuses on the third step of the FSM, mode choice. This step splits the trips output by the Trip Distribution step into different modes. The output of this step are person trips by mode. Mode choice models are a crucial component in a travel demand model. Koppelman and Bhat state that “mode choice is arguably the single most important determinant of the number of vehicles on roadways” and “the most easily influenced travel decision” (2006, p. 3). In the context of intercity travel demand models, Eric Miller refers to them as the “‘heart’ of most intercity travel demand modeling efforts.” (2004, p. 97).

2.1. Discrete mode choice models

In the beginning, travel demand forecasting relied on aggregate analysis. The models aggregated the collected travel data into travel analysis zones (TAZs) and then applied simple zonal averages or distributions of characteristics (Weiner, 1999, p. 90).

However, with the idea from econometrics and psychometrics that travel choices were discrete, the field began to shift to a disaggregate approach (Weiner, 1999, p. 91). Currently, mode choice models are most often disaggregate, discrete choice models. These models analyze and predict an individual's choice of one alternative from a set of finite alternatives (Koppelman & Bhat, 2006). They have distinct advantage over aggregate models in that they can explain a mode choice based on the traveler's individual characteristics rather than statistical associations based on a larger group; are more applicable to different time and space contexts because they are causal and less tied to the estimation data; and are more data efficient, i.e. able to include a range of relevant variables versus the loss in variation of aggregate models (Koppelman & Bhat, 2006).

Discrete choice analysis is based on the principles of utility maximization, i.e. the individual selects the alternative that has the highest utility in the set of choices, with the utility being how much they value each option (Schiffer, 2012, p. 32). This means that individuals with the same characteristics will always select the same alternative, which is not true to life, since similar individuals can still select different choices. To account for this, a certain random component is added to the utility function, making it a random utility model. (Koppelman and Bhat, 2006).

Equation 1

$$U_{it} = V_{it} + \varepsilon_{it}$$ (Koppelman & Bhat, 2006, p. 18)

Here, U_{it} is the utility of taking alternative *i* to trip-maker *t,*, V_{it} is the observable portion of the utility and ε_{it} is the error portion of the utility, which is assumed to be Gumbel-distributed in the multinomial logit (MNL) estimation (Koppelman & Bhat, p. 19, 2006).

The observable portion of the utility, or V, is calculated as follows:

Equation 2

$$V_{it} = V(S_t) + V(X_i) + V(S_{t,}X_i)$$ (Koppelman & Bhat, 2006, p. 19)

where V_{it} is the observation portion of utility of alternative *i* for individual t, $V(S_t)$ is the part of utility associated with characteristics of individual *t*, $V(X_i)$ is the utility from the attributes of

alternative *i,* and $V(S_t, X_i)$ is the portion of the utility which results from interactions between the attributes of alternative *i* and the characteristics of individual *t* (Koppelman & Bhat, 2006, p. 19).

The parameters for each attribute are estimated using a maximum likelihood function. The error component of the utility is represented by a probability distribution. Disaggregate discrete mode choice models rely on an S-curve distribution to represent the error components and thus determine the probability of a choice being made (Weiner, 1999, p 92). The curve is usually a probit or logit function. Arguably the most common model used in travel forecasting is the logit model, and the most common logit models being the multinomial logit model (MNL) (Schiffer 2012, p. 31).

2.1.1. Multinomial logit model

The multinomial logit model is a class of logit model that addresses more than two alternatives. It is based on the assumptions that the error terms follow a Gumbel distribution and are independently distributed across alternatives and individuals (Koppelman & Bhat, 2006, p. 26).

The equation of the probability of choosing an alternative is as follows:

Equation 3

$$\Pr(i) = \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}}$$ (Koppelman & Bhat, 2006, p. 26)

where $\Pr(i)$ is the probability of choosing alternative i, and V_j is the observable component of utility of alternative j (Koppelman & Bhat, 2006, p. 26).

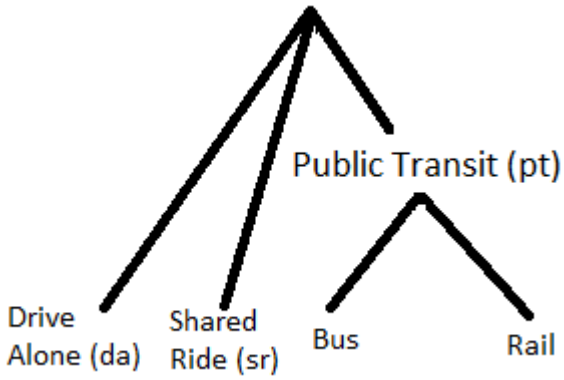
A troublesome characteristic of MNL models is the independence of irrelevant alternatives (IIA). This means that the ratio between any two alternatives are not affected by the presence of a third alternative (Koppelman & Bhat, 2006, p. 39). The implication of this is that the parameters are not affected by adding or removing an alternative from the choice set. However, in real life, sometimes alternatives are in fact dependent on and affect each other, as illustrated by the famous red bus/blue bus thought example (Koppelman & Bhat, 2006, p. 40): When a model has an equal modal split between car and red bus, and later on introduces another bus alternative, only painted blue, one would expect more current red bus takers to switch to the new bus service, but in a simple MNL model, an equal number of car and bus takers would switch to the new bus service (Koppelman & Bhat, 2006, p. 41).

To address the IIA problem, many modelers employ another popular mode choice model, the nested logit model, which will be discussed in the next section.

2.1.2. Nested logit model

The nested logit model (NL) addresses the IIA problem by grouping together alternatives that are similar and making the choice as a multi-step decision (Schiffer, 2012, p 40). Take as an example the Figure 1:

Figure 1.Sample nesting structure (adapted from Koppelman & Bhat, 2006)



The nested logit model assumes that the random error terms are shared between some alternatives (Koppelman & Bhat, 2009, p. 160). This makes the utility equation of the alternative bus:

Equation 4

$$U_{bus} = V_{pt} + V_{bus} + \varepsilon_{pt} + \varepsilon_{bus}$$ (Koppelman & Bhat, 2006, p. 161)

with ε_{pt} being the common random component and V_{pt} being the common observed component (Koppelman & Bhat, 2006, p. 161).

The error components are still assumed to follow a Gumbel distribution, but with a scale factor μ_{pt}, or commonly $\theta_{pt} = \frac{1}{\mu_{pt}}$ (Koppelman & Bhat, 2006, p. 161).

The probability of choosing a nested alternative is based on the conditional probability of choosing the nested alternative times the marginal probability of choosing the nest, as shown in the following equations:

Equation 5

$$Pr_{bus} = Pr_{\frac{bus}{pt}} * Pr_{pt}$$ (Koppelman & Bhat, 2006, 161)

where $Pr_{\frac{bus}{pt}}$ is

Equation 6

$$Pr_{\frac{bus}{pt}} = \frac{e^{\frac{V_{bus}}{\theta_{pt}}}}{e^{\frac{V_{bus}}{\theta_{pt}}} + e^{\frac{V_{rail}}{\theta_{pt}}}}$$ (Koppelman & Bhat, 2006, 162)

and Pr_{pt} is

Equation 7

$$Pr_{pt} = \frac{e^{(V_{pt}+\theta_{pt}\tau_{pt})}}{e^{V_{da}} + e^{V_{sr}} + e^{(V_{pt}+\theta_{pt}\tau_{pt})}}$$ (Koppelman & Bhat, 2006, 162)

τ_{pt} is the log of sum of exponents of the nested utilities:

Equation 8

$$\tau_{pt} = \log[e^{\frac{V_{bus}}{\theta_{pt}}} + e^{\frac{V_{rail}}{\theta_{pt}}}]$$ (Koppelman & Bhat, 2006, 162)

The logsum parameter, or nesting coefficient, corresponds to how similar alternatives are within a nest. It should be between zero and one (Koppelman & Bhat, 2006, p. 163). When logsum is one, it implies that there are no correlation between mode pairs in the nest, and the model is equivalent to an MNL model (Koppelman & Bhat, 2006, p. 163). When the logsum is zero, there is perfect correlation between the mode pairs in the nest, and the model becomes deterministic (Koppelman & Bhat, 2006, p. 163).

The selection of an appropriate nest structure for a model is a blend of reasonable judgment and statistical evidence. Potential nesting structures are narrowed down based on conventional wisdom, and the proposed nests are tested against each other and the MNL model to see which is the more representative model.

2.2. Intercity mode choice models

Intercity travel demand modeling, or long-distance travel demand modeling, dates almost as far back as urban mode choice models (Miller, 2004). They are often applied to a well-defined travel corridor that has a small number of origin and destination cities (Miller, 2004). There is a dearth of models that cover a larger, overall region (Moeckel, Fussell, & Donnelly, 2015).

The models are usually segmented by trip purpose, distance, party size, region (Koppelman & Bhat, 2006; Koppelman & Wen, 1998; Bhat, 1997) The modes typically studied are auto, rail,

bus, and air (Koppelman & Bhat, 2006; Koppelman & Wen, 1998; Bhat, 1997). The usual explanatory variables considered are level-of-service attributes of the mode, characteristics of the trip maker, and characteristics of the trip (Zhang et al., 2015).

Large regional models of Canada have been attempted before (Wilson et al., 1990. Abdelwahab, 1991). The models were based on the Canadian Travel Survey (CTS), the precursor to the TSRC used in this thesis. Both models were MNL intercity mode choice models split for eastern and western Canada at Thunder Bay (Wilson et al., 1990. Abdelwahab, 1991). Due to data limitations, both models only used trips to and from Census Metropolitan Areas (Wilson et al., 1990. Abdelwahab, 1991). The models were segmented by trip purpose, and in Abdelwahab's case, also by distance (Wilson et al., 1990; Abdelwahab, 1991). Abdelwahab concluded that there is very low transferability of estimated coefficients for different model regions, which reinforces the importance of estimating a model based on local data (1991). In recent decades, there has been some interest in a high-speed rail in the Windsor – Quebec corridor. This has inspired various intercity travel demand models and studies of the region. These models rely on the data assembled by VIA rail in 1989 and vary from MNL, NL, to the heteroscedastic model (Koppelman and Wen 2000, 1998; Bhat 1995, 1997). The data was limited to business travelers only (Koppelman and Wen 2000). More recently, Wong and Habib derived an NL intercity mode choice model for the Windsor-Quebec corridor. They concluded that access and egress was more important to travelers than in-vehicle travel time (2015).

Intercity modeling has a number of inherent difficulties that do not apply to its short-distance counterpart. Eric Miller succinctly points out these major challenges in his 2004 paper, "The Trouble with Intercity Travel Demand Models":

- The modes are overly aggregated.
- The effect of access and egress are not adequately accounted for in line-haul modes.
- Current explanatory variables are limited due to data limitations and aggregation.
- New modes can only be modeled based on Stated Preference surveys.

Moeckel et al. gave a thorough overview of the state of the art of long-distance mode choice models (2013). They concluded that the nested logit model has been proven preferable to the simple multinomial logit model and that the transferability of models from one region to another is not recommended (Moeckel et al., 2010).

Though model parameters can be asserted (Moeckel et al., 2010; Alliance Transportation Group, 2015), there is value in deriving a model, given the relative scarcity of intercity mode

choice models. The challenge is that long-distance travel is somewhat rare and thus suffers from a lack of data. The next section addresses this issue.

2.3. Data for long-distance mode choice modeling

Although there is an extensive literature on mode choice modeling, comparatively few studies focus on long-distance mode choice modeling. Those that do are often focused on the statistical methods of the model. However, the input data to the model itself has serious impact on the accuracy of the model. In fact, according to Zhang et al., the entire model framework largely hinges on the quality of the data available (2015).

The primary input data source for long-distance mode choice modeling are often surveys, both stated preference and revealed preference. Most models in the U.S. are based upon the National Household Travel Survey (Cho, 2009; Schiffer, 2012). Per Zhang et al.'s review of data sources used in long-distance models, the most commonly used are household and person travel surveys from public agencies, followed by revealed or stated preference surveys conducted for the express purpose of the project, operational data from mode providers, and data purchased from private sources (2015). There can often be more than one data source per model (Zhang et al. 2015). Zhang et al. conclude that there is a general lack of data for building detailed, complete OD matrices, especially for bus, rail and auto modes in the U.S. (2015). Like Miller, they also point out the need for data on access and egress (2014; Zhang et al., 2015).

Since the mode choice model is designed to select one out of several available modes, the level-of-service (LOS) variables of alternative modes are also needed. These tend to be, as Zhang et al. note, "monetary costs, travel times, and frequency of service" (2015, p. 417). Traditionally, for short-distance mode choice models, these could be derived as skims from an existing travel demand model. However, for intercity models, such networks may not be readily available, as was the case for this thesis.

For long-distance travel, there are several methods for acquiring this data. One common method, as demonstrated by Cho, is to manually search actual LOS data from randomly selected OD pairs in the air, bus and train networks and then extrapolating linearly within certain distance groups (2013). Wilson et. al acquired mode data from the Strategic Planning Division of Transport Canada, and, because of the relatively manageable number of OD pairs, they could fill in the rest manually from published timetables and other such sources. (1990). In his 1997 model of the Toronto – Montreal corridor, Bhat relied on data provided by the major

Canadian rail operator, VIA (1997). Often the survey would ask for the LOS associated with the chosen mode of a journey. However, the TSRC, which was designed to gauge the state of domestic tourism, lacked the level-of-service variables associated with the chosen mode, not to mention other available modes.

Miller pointed out that multimodal network data are hard to get since they are often run by the private sector (2011). Bus and rail data are proprietary, at least in the U.S., and hard to acquire in sufficient detail (Zhang et al. 2015). In the U.S., there are 10% samples of ticket survey data available for air, bus and train travel (Cho, 2013). Even if datasets exist, they are usually from disparate sources and must be combined and synthesized (Zhang et al. 2015). With an often-limited budget for data collection (Zhang et al. 2015), it is in this area that aggregated web-based data sources can shine.

The use of new data sources in travel demand modeling, particularly the usage of big data, has become a trending topic. There have been various studies on using crowdsourced geo-spatial data, mobile data, etc. in travel demand modeling (Toole et al., 2014). The use of various GTFS data in modeling has also been explored, but GTFS data only pertains to short distance public transport options and does not cover long-distance travel (Antrim & Barbeau, 2013). Online trip planners, however, have not been explored as a data source. The platform Rome2rio, which is used in this thesis, has been the subject of occasional research, but only in its capacity as a trip planner (Antrim & Barbeau, 2013; Klock, Owens, & Schwartz, 2012). As far as I am aware, there has not been a documented case of using data from trip planners such as Rome2rio in travel demand modeling.

2.4. Thesis implications

As the Transportation Research Board NCHRP 735 report notes, “even for applications with similar circumstances, unless models have identical specifications, the values for specific coefficients may differ significantly between models” (Schiffer, 2012, p. 63). This points to how easily the coefficients of the model are affected by the definition of variables, and the value of an econometric estimation of a mode choice model using locally applicable data.

This thesis aims to contribute to the transportation modeling field by developing a new long distance mode choice model for Ontario. An econometrically derived model can reflect the unique transportation behavior of Ontario, Canada and serve as reference. As a model that will be part of work done for the Ministry of Transportation of Ontario, it is publicly owned and

will hopefully enlarge the body of knowledge that future researchers can draw on for comparison.

This model also goes towards addressing some of the concerns pointed out by Eric Miller in his 2004 paper, namely the lack of openly documented intercity models and the lack of mode data from privately owned organizations.

The common modes of travel in intercity trips are often privately owned, and these actors may be unwilling to share operational data. The thesis attempts to circumvent this by using a relatively new aggregated web-based multimodal data source to derive a long-distance mode choice model.

In addition, although combinations of GTFS and Web 2.0 data have been employed in transportation modeling, the specific application of Rome2rio in mode-choice modeling, especially intercity mode-choice modeling, has not yet been done. Comprehensive global trip planners such as Rome2rio have only recently emerged. Their application to solve the age-old data scarcity problem of long distance mode modeling has not been widely explored. By developing a model using mode-specific data from Rome2rio, this thesis aims demonstrate the plausibility and validity of using such data sources in intercity mode choice modeling. With the advent of big data, travel demand modeling is gaining access to many promising emerging data resources, such as social network-based location tracking, and wireless network location services (Schiffer, 2012). Updating methods to utilize novel data resources may thus herald a new way of building travel demand models.

3. Data collection

The model estimation and calibration is reliant on two main sources of data, the Travel Survey of Residents of Canada and data collected from Rome2rio. The former is a survey conducted by Statistics Canada to collect the characteristics of domestic travel. This data gives the various trips and socioeconomic characteristics of the trip taker. It provides the origin, destination and mode taken for the trip. However, it lacks mode-specific details, such as travel time and travel costs. To this end, additional data was collected from Rome2rio, a trip-building web platform with comprehensive multimodal travel information aggregated from multiple sources. It can build a door-to-door trip itinerary of all possible modal connections between an origin and a destination, providing details such as travel time and costs. This thesis employs Rome2rio's free API, which offers limited requests to access its database. The mode choice parameters were thus derived by combining these two sources of data.

3.1. Travel Survey of Residents of Canada

The Travel Survey of Residents of Canada is conducted periodically to assess the status of Canada's tourism industry. It focuses on domestic travel and has information on the volume and characteristics of the trips and trip makers. The survey is a voluntary supplement of the compulsory household survey Labour Force Survey (LFS). The LFS has a sample size of 54,000 households and a response rate of 90% (Schiffer, 2012). The TSRC provides Visit, Trip, and Person data files. The Person data file has the characteristics of individuals who answered the survey. The Visit file contains information about the places visited in each trip. However, the only relevant information for the estimation of this model was the Trips data.

3.1.1. Trips data characteristics

The TSRC Trips data for the years 2010 – 2013 were provided. Originally, the data was in a microfile format, which was processed for the destination choice model. It counts all non-routine same-day trips with destinations more than 40 km away, and overnight trips with at least one night spent in Canada as long-distance trips. Each trip record contains information on the trip purpose, origin, destination, distance, mode, socioeconomic factors, activities done on trip and money spent. The relevant data and data categories are listed in the Table 1.

Table 1.Relevant TSRC Trips data categories

Variables	Categories
Reference month	January - December
Origin	Census division, census metropolitan area
Destination	Census division, census metropolitan area
Trip purpose	Leisure; Visit; Business; Other
Mode	Car or truck; Air; Camper RV; Bus; Train; Ship/ferry; Boat; Other
Weights	Trip weight
Age	18 - 24; 25 - 34; 35 - 44; 45 - 54; 55-64; >65
Sex	Male; Female
Education	<High school; High school; Post-secondary; University
Employment	Employed; Unemployed
Income	<$50,000; $50,000 - $70,000; $70,000 - $100,000; >$100,000
Travel party size	0 - 95
Household members on trip	0 - 6+
Household adults on trip	0 - 5+
Household children on trip	0 - 4+
Self reported trip distance	Kilometers
Type of trip	Overnight - Canadian; Sameday; Overnight - International

In this estimation, the ‘Other’ trip purpose is grouped together with leisure, forming three trip purposes, business, leisure and visit. The model is only concerned with overland modes; therefore, ‘Ship/ferry’, ‘Boat’ and ‘Other’ modes are excluded. Of the modes modeled, ‘Car or truck’ and ‘Camper RV’ are included in the auto mode. ‘Overnight-International’, meaning a trip with at least one night spent outside of Canada, was not considered a domestic long-distance trip and excluded.

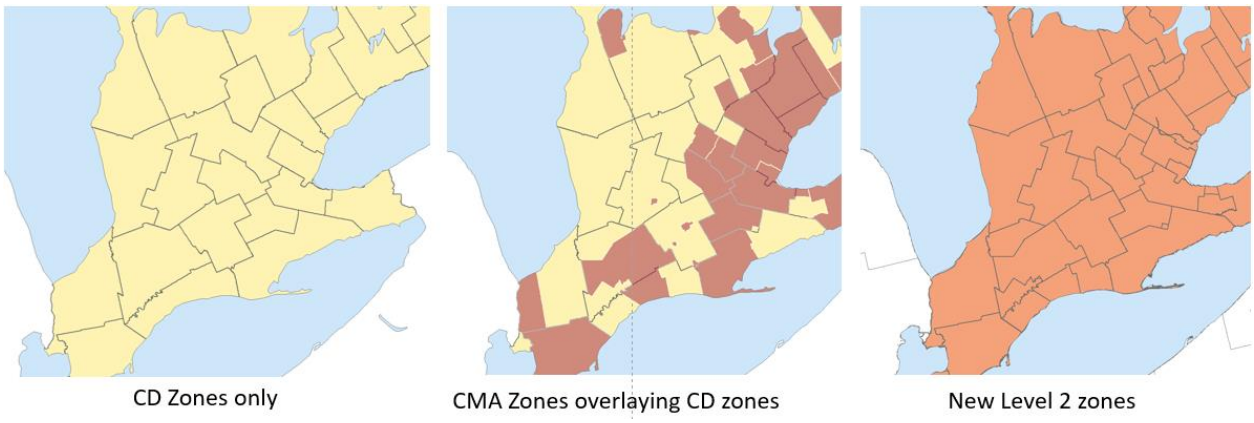
3.1.2. Zone system

The TSRC records trips at the resolution of Census Divisions and Census Metropolitan Areas. Though more detailed traffic analysis zones were given for the project, the model is estimated using the broader zone system given by the TSRC.

Since the model will eventually be disaggregated into TAZs, which are much finer in resolution than TSRC zones, it was worthwhile to pursue as much resolution as possible from the TSRC data. Therefore, CDs and CMAs were combined to form a new zonal system.

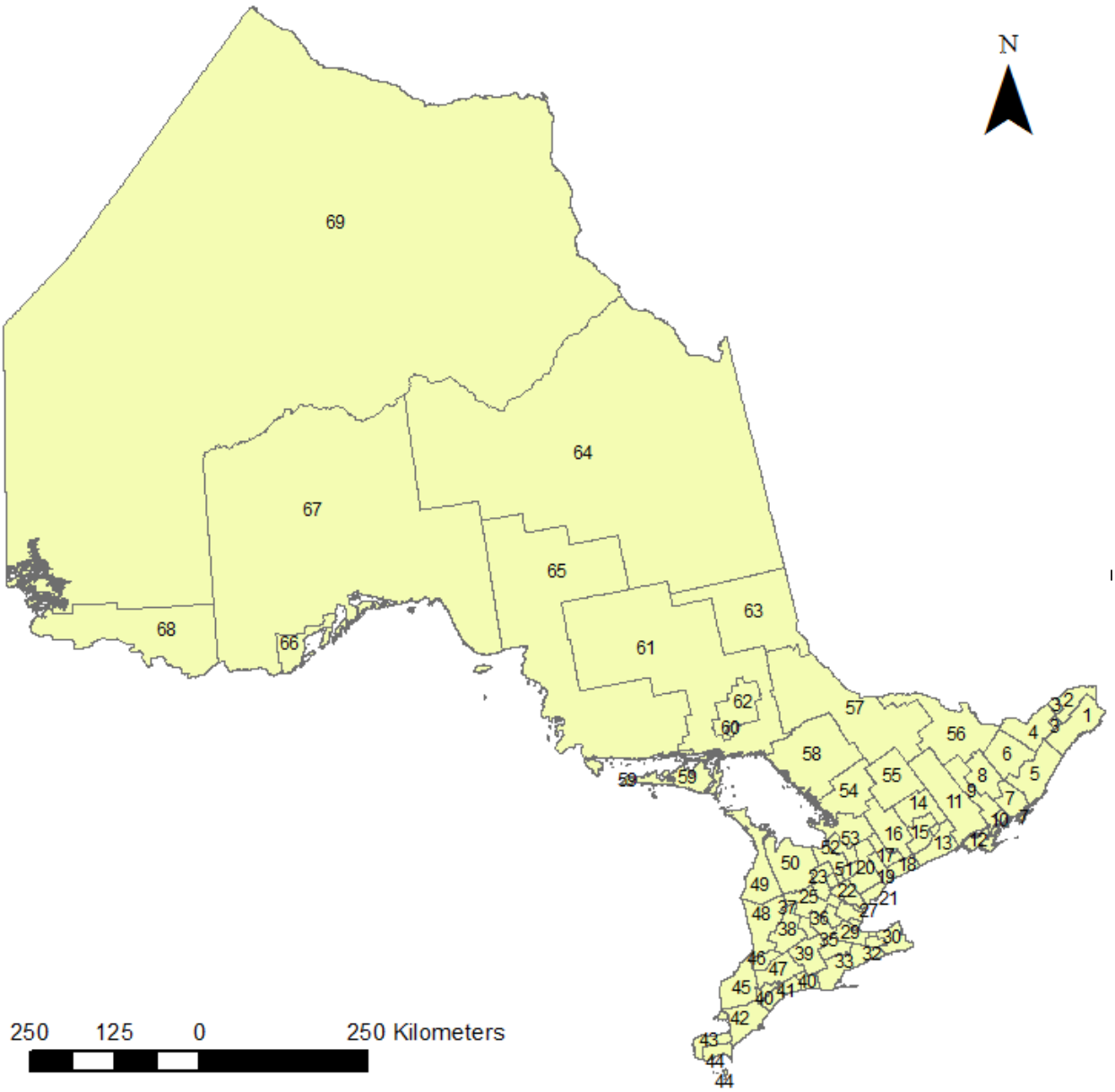
There are 49 CDs and 15 CMAs in Ontario. By intersecting the CDs and CMAs and taking the intersected areas as new zones, a total of 69 zones were derived from the TSRC data.

Figure 2.Level 2 zone creation process



This was done by Joe Molloy for the destination choice model. Since the mode-choice model is based on the same TSRC trips data, it is also estimated using the same zone system, leaving the disaggregation into TAZs to a later step.

Figure 3.Ontario level 2 combined zones



3.1.3. Relevant trip records

All four years of TSRC trips data amounted to a total of 219,997 records. As is the norm in surveys, not all data fields were filled out for all records. This renders some trip records unsuitable for consideration in the mode choice model estimation. For example, since income level is considered as a parameter in the estimation, when it is not reported for a trip record, that record is filtered not and not considered. Therefore, as a first step, the following trip records were filtered out:

- Records that did not state income level

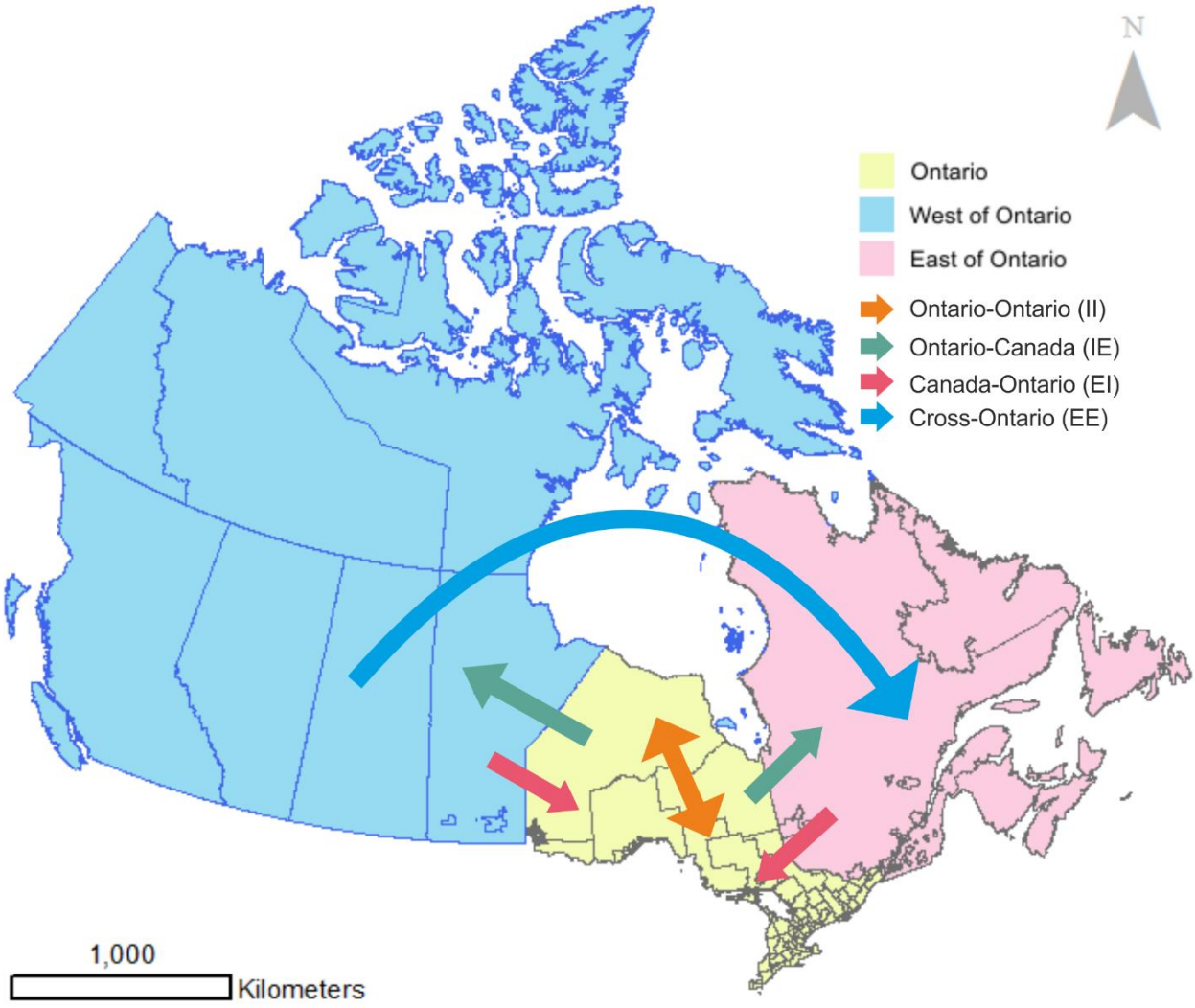
- Records with number of travelers in travel group greater than 8, since it is assumed that an ordinary private vehicle could only carry up to 8 passengers
- Records that do not have the main travel mode as either auto, rail, bus, air, or camping RV
- Records of overnight trips that are under 40 km in distance, since this model defines long-distance travel as trips to destinations more than 40 km away, and the TSRC includes overnight trips that may be less than 40 km between origin and destination. Though these trips may be defined as long-distance by the TSRC, they are not representative of the long-distance travel behavior this model is trying to capture.

Secondly, trips were geographically filtered out. Since the model is being built for Ontario, the only trips that were considered were those with at least one trip endpoint in Ontario or those going across Ontario. Trips considered were:

- Trips with both origin and destination within Ontario
- Trips with an origin in Ontario but a destination outside of Ontario
- Trips with an origin outside of Ontario but with a destination inside Ontario
- Trips with an origin and destination outside of Ontario, but which, due to geography, must go across Ontario

The map in Figure 4 depicts Canada as Ontario and external zones.

Figure 4. Types of trips modeled



Since Ontario province bisects Canada's geography, it is reasonable to assume that all trips originating east of Ontario and ending west of Ontario and vice versa must travel across Ontario. Therefore zones east of Ontario and west of Ontario were identified, and external trips that cross Ontario were retained.

There were also two zones identified in Quebec province that could potentially result in cross-Ontario trips. These were zones 85 and 103, or Ottawa-Gatineau and Montreal. Trips to and from these zones and from the rest of the Quebec province CMA zones were also retained.

Figure 5. Trips also crossing Ontario from zones 85 and 103

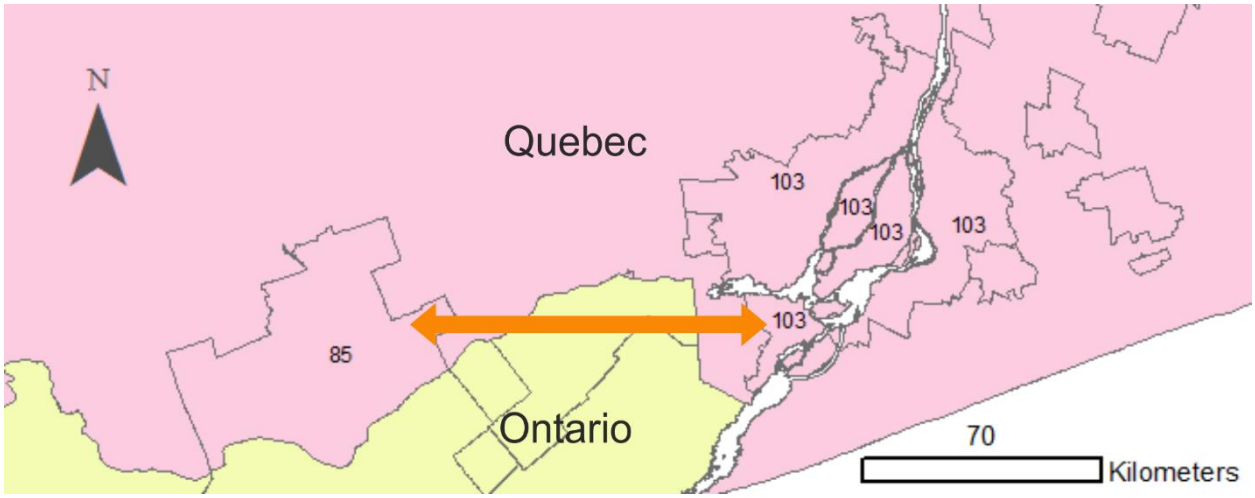


Table 2. Trips retained per trip purpose

Trip Purpose	Trips Retained
Business	6,028
Leisure	25,922
Visit	31,744

3.1.4. Weights

There are two relevant weights in the TSRC Trips data: the person-trip weight (WTTP) and the trip weight (WTEP). The person-trip weight (WTTP), was calculated from the person-weight for the LFS survey adjusted with factors that approximate how many identical trips were taken. The trip weight, or (WTEP), was calculated by dividing the person-trip weight by number of adults from household on the trip. As the survey user guideline suggests, the person-trip weight (WTTP) was the appropriate weight to use for all socioeconomic characteristics of all-travelers, same-day or overnight.

3.2. Data from Rome2rio

A mode choice model relies heavily on level-of-service factors, especially travel time and travel cost (Koppelman & Bhat, 2006). That is to say, an individual often chooses a travel mode by comparing what each mode offers in terms of travel time, travel cost, service frequency, etc. Having the level-of-service data for all available mode choices is therefore crucial to the derivation of a mode choice model. The TSRC did not include travel time. Though it did ask for travel costs, the travel costs were only recorded for the mode taken, so it is not possible to

form a comparison between travel costs of all available modes and between an origin and destination pair. Therefore, alternative sources of data were pursued.

At first, I considered the option of manually searching for each OD pair in the appropriate online trip planner for each mode, but this was not feasible as we had over 20,000 relevant OD pairs. Another option we explored was to extract data from available resources. For example, the travel time for rail mode was calculated by manually entering rail network location and time table data from VIA into the transport modeling software EMME and extracting the calculated skims. However, this would require enormous effort and disparate data sources that might or might not be readily available, such as the location of all long-distance bus stops in North America, or long-distance bus timetables from each bus provider. Furthermore, this method could not be applied to air transportation, which does not rely on traditional time tables and networks.

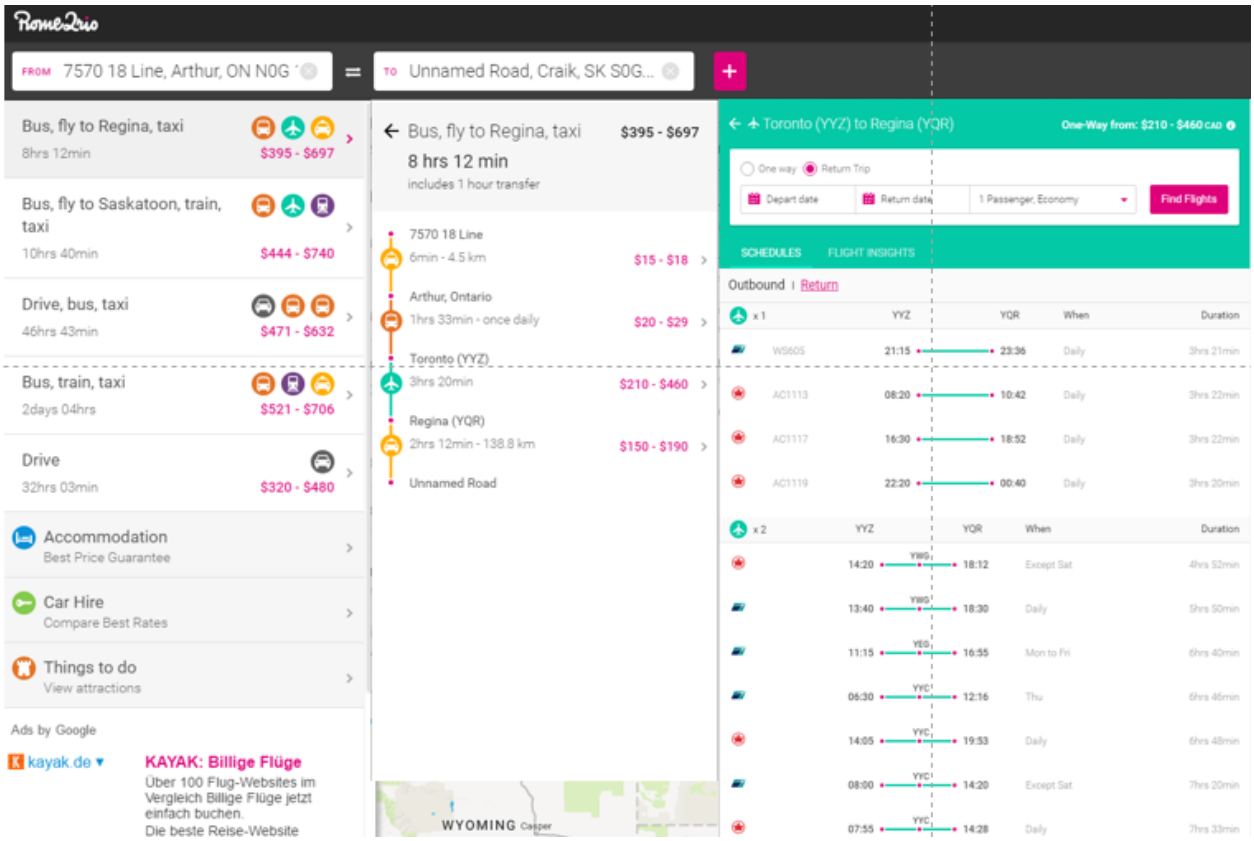
In the end, I turned to web-based open resources and discovered Rome2rio. Rome2rio is an online travel metasearch platform that provides door-to-door journey planning. The platform is unique in that it has global multimodal and intermodal capabilities, meaning that it can provide long-distance trip planning options across multiple modes, including flight, train, bus, ferry and driving. Its strength lies in its ability to build trips from door-to-door. That is, it provides the first and last leg journey information, such as the access to and egress from the airport. This is achieved by drawing on multiple data sources such as API feeds from other travel search websites, GTFS data, etc., which aggregate into a comprehensive multimodal route information database of both long- and short-distance travel. Rome2rio boasts a global coverage: it contains information from 670 airlines worldwide, and, in North America, it covers the rail providers Amtrak and ViaRail and 170 bus transport providers (“Transport Coverage Overview,” 2017). Driving and walking directions are supplied using OpenStreetMaps (“Transport Coverage Overview,” 2017).

3.2.1. Rome2rio API

Rome2rio offers a free API key with a limit of 100,000 searches per month and 300 requests per hour. This API allows the user to specify an origin, a destination, and gives the Rome2rio search result back in XML or JSON formats. The API query returns all information contained in a normal search request on Rome2rio. However, it does not provide live pricing data as that is done through a third-party website. Instead, it returns the Rome2rio general price estimate based on its historical data.

To illustrate the details available in the data, below is a sample query using the normal Rome2rio interface.

Figure 6. Example request and route suggestions from Rome2rio (rome2rio.com, 2016)



As can be seen above, a query returns several possible trip route options combining different available modes. The exact algorithm used by Rome2rio for trip-building is not public, but I assume it is optimized to give all possible, reasonable travel options. For each route option, the travel duration, transfer time, travel distance, and estimated price range are given. Each route option is composed of one or more travel segments, separated by mode or transit provider. In the Figure 6, the route option “Bus, fly to Regina, Taxi” is composed of four segments, taxi, bus, plane, then taxi again. For each travel segment, in addition to travel duration, distance and price, Rome2rio gives details such departure schedules, service frequency and transit provider. The Rome2rio API query returns all of the above information and more.

3.2.2. Methodology

1. Request a free API key from Rome2rio for research purposes

2. Build query URLs usingzone centroid geographical coordinates. Below is an example query URL:

*http://free.rome2rio.com/api/1.4/xml/Search?key=&**oPos**= &**dPos**= &**currencyCode**=CAD*

where *oPos* is the origin latitude and longitude, and *dPos* is the destination latitude and longitude, and *currencyCode* is the international currency code, in our case, Canadian dollars (CAD).

A query was performed

   a. from each Canadian zone to all other zones, meaning from each Canadian zone to each other Canadian zone, and
   b. from each Canadian zone to each zone outside of Canada.

   This resulted in a total of 21,878 OD pairs.
3. A python script was composed to automate the API querying process to adhere to the request limit of 300 per hour and 100,000 per month.
4. The data was collected on November 11th – 13th, 2016. The Rome2rio data extracted is in JSON format. Due to the nature of the trip building search platform, it does not clearly distinguish between drive, air, bus and train modes, nor does it distinguish access, egress and main modes. Therefore definitions and assumptions were made to categorize and process the data into a useful format for the task at hand.
5. For more information on the API search request and response variables, please refer to the Rome2rio API documentation page in the bibliography.

3.2.3. Data processing and parsing

The JSON files returned from the API request were parsed using Python to extract useful information. The following presents a detailed list of how each variable was extracted and calculated.

Although Rome2rio does give total travel time and price for the entire route, they were calculated using segments because Rome2rio does not distinguish between access, egress and main parts of the journey. Furthermore, Rome2rio may give alternative segments that may be faster than the original suggested segment, therefore making it necessary to calculate the total route characteristics using segments instead of taking the given value.

Main mode

Rome2rio may build a travel route from several different modes. For this thesis, a main mode must be determined to be used in mode choice analysis. A mode hierarchy was used to

determine the main mode of a route option. The hierarchy used here is air, rail, bus, and auto. This meant that if any segment of a travel route listed flying as a mode, then flying was taken as main mode of the route. If there was no flying segment, then rail took precedence and so on and so forth. The consecutive segments of a route with the main mode were then considered together as the main trip. In the example above, the main mode of the suggested route “Bus, Fly to Regina, Taxi” would be air. The main mode travel time, main mode price, main mode transfer time, and main mode distance were summed using these segments of the route. The main mode frequency was taken as the minimum service frequency of all main mode segments instead of an average, as the segment of a trip with the lowest service frequency would be the limiting factor of the journey. The number of transfers of a journey was taken as the number of main mode segments minus one.

In some cases, the main mode segments are broken up by other modes.

Example route:

1. Taxi (10km) – 2. Bus (50 km) – 3. Taxi (20km) – 4. Bus (100 km) – 5. Taxi (10 km)

For such cases, the segment with the longest distance travelled – here, bus – was taken as the main mode, and everything from 2. Bus (50 km) to 4. Bus (100 km) segments were taken as the main mode.

In another special case, such as

1. Train (10 km) – 2. Bus (20km) – 3. Train (100 km) – 4. Taxi (10 km)

the main mode was train, and the journey would be taken as segments 1 – 3, which left no access time and taxi as egress. In this case the first segment of the journey, train (10km), was taken as access mode.

Access and egress

Access and egress are any segments of the trip that comes before and after the main mode segments respectively. Access and egress time and distance were taken to be the travel times and distances of all travel segments before and after the main mode segments, respectively. In the first example above, the first taxi and bus segments would be counted as access, and the last segment by taxi would be counted as egress.

Total trip

Total travel time, total transit time, and distance were calculated using all segments of a route. Total travel time is the travel time of all segments combined. Total transit time is the total in-vehicle travel time of a route, i.e. omitting transfer time. The average price took the price of access and egress modes into account, but the model estimation only used the price of the main mode.

Alternative segments

Rome2rio might give alternative suggestions for some minor segments of a route. In that case, if the alternative segments were faster, they were substituted in place of the original segments, and all calculations were done with the substituted alternative segments.

Flight

Flight segment data was formatted differently from other segments. Due to the large amount of different flight options that can be available between an OD pair, one single flight segment can contain many different potential flight routes and their respective details. Therefore, when flying was the main mode, the main mode frequency was the sum of the frequencies of all flight options.

Transfer time

Transfer time in Rome2rio is given as the wait time between two travel segments. Access and egress transfer times were not counted in access and egress time or in total transfer time.

Travel cost

The travel cost for transit modes were taken as given by Rome2rio. However, Rome2rio assumes all auto travel segments are done by taxi or other car services, and calculates costs accordingly. Therefore driving costs were not taken from Rome2rio but rather calculated using distance and fuel price.

Train fare

Rome2rio gives fares differently for VIA and Amtrak trains: VIA rail is shown with escape, economy and economy plus fares, while Amtrak is given Coach, Business and Room fares, which are not equivalent and ended up tilting price averages so that Amtrak train fares are much more expensive than VIA rail train fares for similar distances travelled. Therefore train fares were taken as Economy Plus train fares for VIA rail, as Coach train fares for Amtrak-

operated trains, and as Business Seat fares for some Amtrak trains when the only fares available are Business Seat and First Class Seat. However, Amtrak trains are only relevant for international travel into the U.S., which is not considered in this portion of the model.

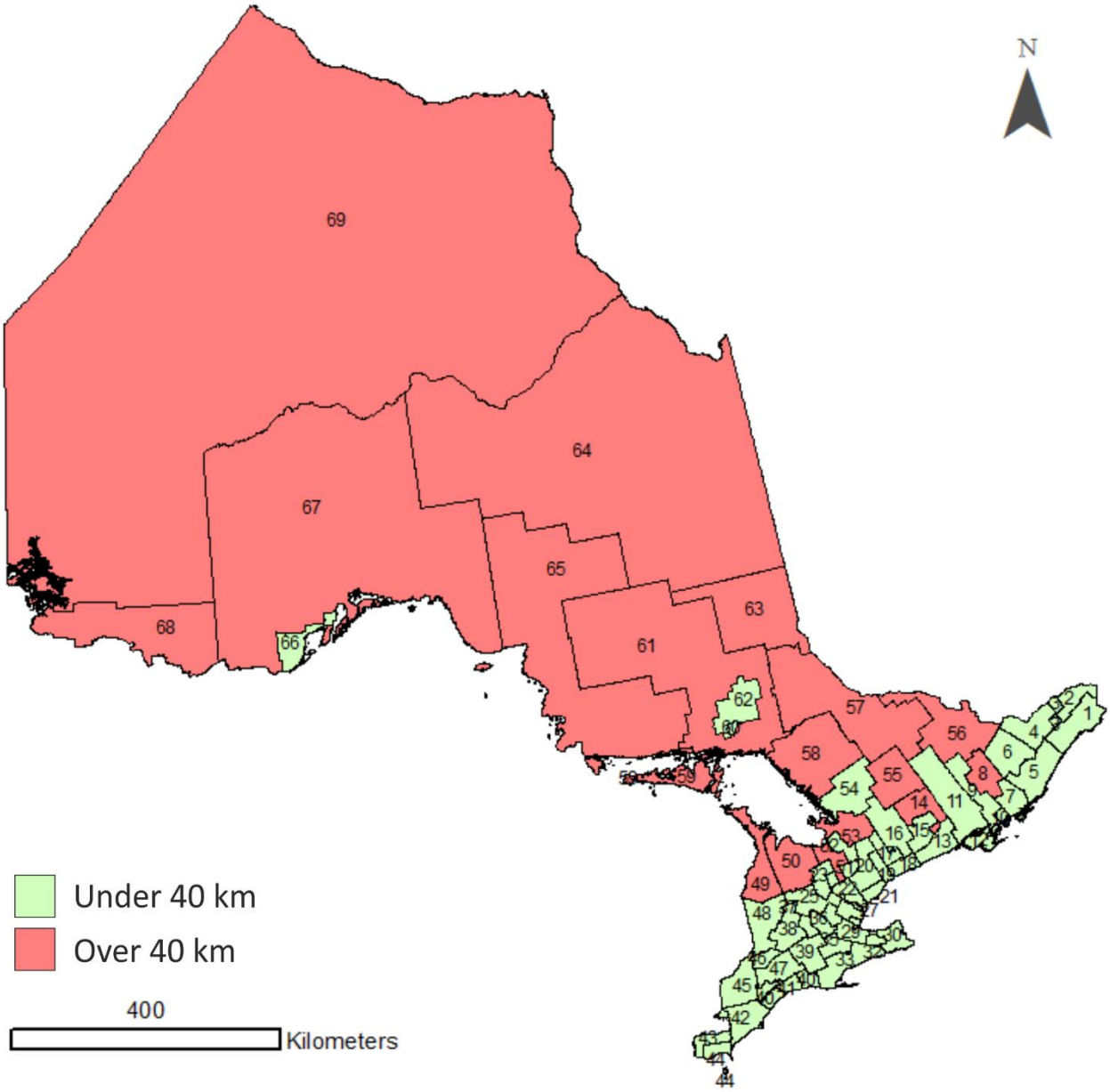
Frequency

Frequency was taken as the number of times service is offered per week. When there was more than one segment to a route, the minimum frequency is taken.

3.2.4. Intrazonal mode data

The prepared URLs that use centroids of Level 2 zones neglected long distance trips made within the same zone. Therefore, a separate URL list was prepared to acquire intrazonal travel data.

Figure 7. Zones with intrazonal distances longer than 40 km



For zones within Ontario with intrazonal distances of under 40 km, I assume that the only intrazonal mode available is driving, with intrazonal travel time assumed to be 35 minutes. For zones outside of Ontario, I assume the CMA zones only have driving as the long-distance mode. Therefore, zones with an intrazonal driving distance of over 40 km in Ontario and non-CMA zones outside of Ontario but within Canada are considered to have significant long-distance mode competition. To estimate the intrazonal travel times by each mode, I pick the two municipalities or metropolitan areas within that zone with the largest populations as the origin and destination, with the assumption that most intrazonal trips are made between them.

3.3. Combine results with TSRC survey data

The data was spotchecked for reliability, and then combined with the TSRC trips data in R, using the Level 2 zone origin and destination to match. Since there can be multiple routes for the same mode per O-D pair, only the route with the fastest overall travel time for each mode was used.

Rome2rio was not able to find all corresponding mode options for all reported trips. This could be due to reasons such as error in survey reporting, or the coarseness in our assumption of level 2 zone geographical centroids as origin and destination points. Out of all trip records in the TSRC data, Table 3 was the percentage of trips for each relevant mode that was not found by Rome2rio.

Table 3. Percentage of trip records by mode in TSRC not matched by Rome2rio

Mode	% of Trip Records not Found
Air	1.3%
Bus	1.1%
Train	3.6%

3.3.1. Auto assumptions

The auto price from Rome2rio was not used in the estimation since it assumes taxi or other commercial car rental services as part of auto travel costs. Instead it was calculated using average fuel efficiency and average fuel prices. According to the Canadian Company Average Fuel Consumption, the estimated average fleet fuel consumption for 2010 passenger cars was 6.8L/100km (“Canada Light-Duty Fuel Consumption and GHG”, 2016). Per the Ontario Ministry of Energy website, the latest Ontario average unleaded gasoline price for the year 2016 is 105 cents per liter (“Fuel Price”, 2017). Multiplying the two and an estimate of roughly CAD $0.072/km was used to estimate auto travel price. Estimated auto access and egress times were assumed to be 1 minute if in a non-metropolitan area and 5 minutes in a metropolitan origin or destination.

3.4. Data summary statistics

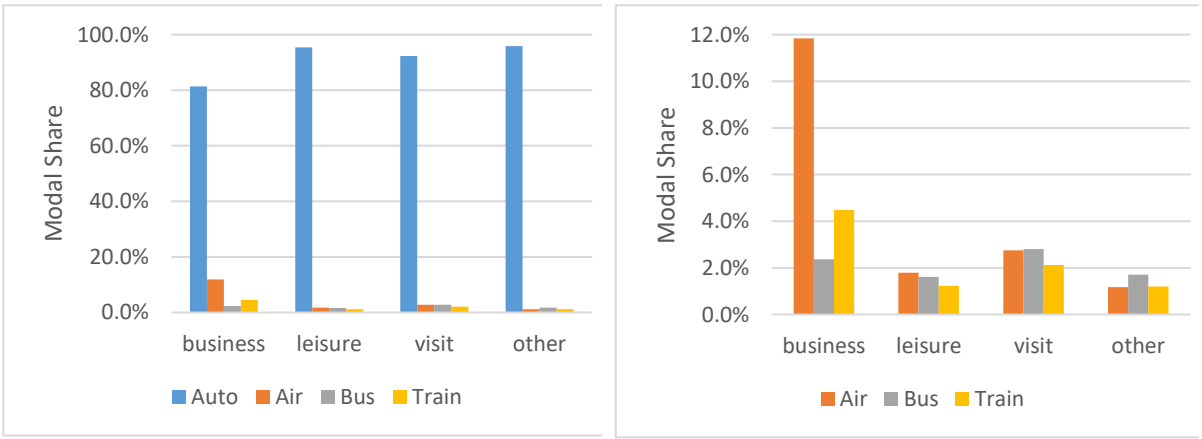
The data was analyzed statistically to help determine the relevant explanatory variables and other specifications of the model. The relationship between variables and modal choice were explored.

Since the auto mode is dominant in this dataset, often another graph is made showing just transit modes for clarity. If there is no auto displayed in a graph, please consult Appendix A for graphs that include auto.

3.4.1. Socioeconomic characteristics

Figure 8. Trip purpose vs. modal share (a) with and (b) without auto mode shows that the trip purposes business and visit have unique modal share patterns, with business trips exhibiting the highest transit modal share, followed by visit. Business also has higher air modal share than any other trip purposes. Other and leisure purpose modal shares are very similar to each other. As such, trips with 'other' purpose are counted as leisure trips and used together to estimate the leisure model. Figure 9 shows that higher income seems to correlate with lower bus and train use and more trips by flying. Bus is particularly favored by those making under $50,000. Figure 10 shows that, as education level goes up, so too does the modal share of all transit modes. The general pattern in Figure 11 seems to be that the transit modal share declines past age bracket 45 – 54. The youngest age group has lower air modal share than bus and train modes. Figure 12 show that gender does affect modal share - females are more likely to take all forms transit modes than males and are less likely to drive. This echoes the findings of Bhat (1997), using the Canadian Travel Survey.

Figure 8. Trip purpose vs. modal share (a) with and (b) without auto mode



(a) (b)

Figure 9. Income bracket vs. modal share without auto mode

Figure 10. Education level vs. modal share

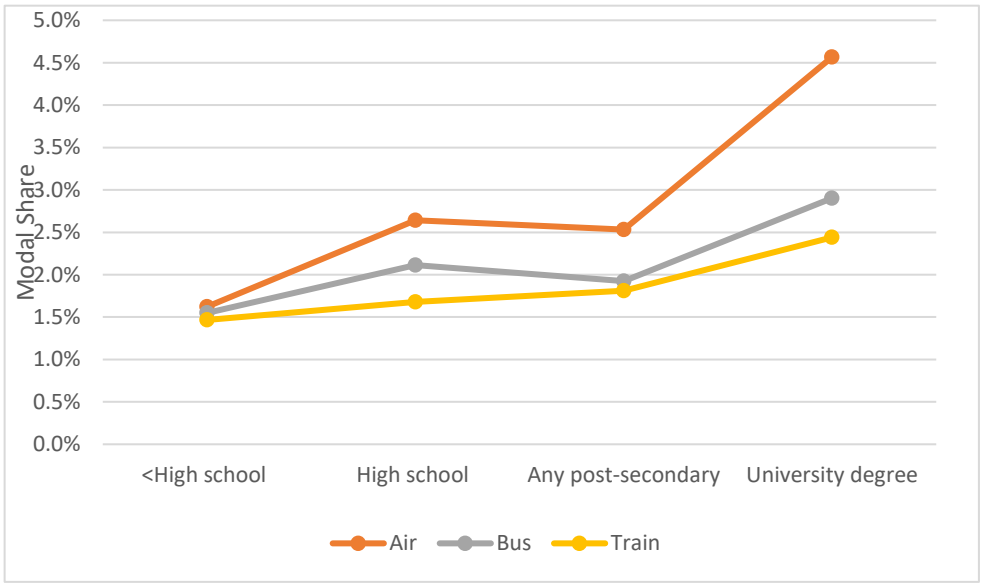


Figure 11. Age bracket vs. modal share

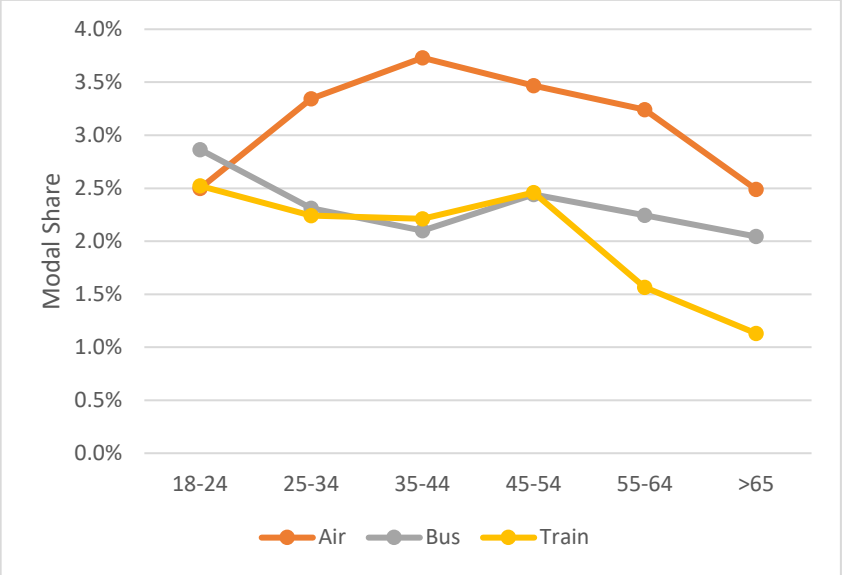
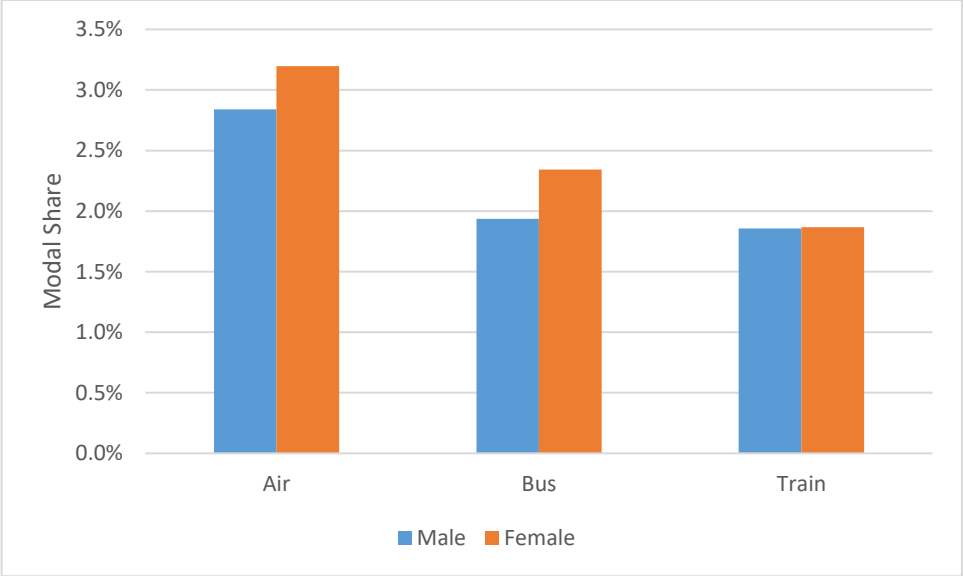


Figure 12. Gender vs. modal share



3.4.2. Trip characteristics

In Figure 13, as travel party size goes up, modal share of auto goes up while transit shares go down. This peaks at a travel party size of five, and then auto share goes down slightly. This could be due to most passenger autos holding five occupants. A season segmentation was also tested, with the months November until March labeled as winter and the rest as summer. It was found to be of no significance to modal share, as seen in Figure 14, and was thus not considered further. As demonstrated in Figure 15, trips that start and end in non-metropolitan areas have the least transit modal share usage, followed by trips with at least one trip-end in

a metropolitan area. Trips that start and end in metropolitan areas have the highest transit modal shares and the lowest auto share. Therefore whether a trip is ‘intermetro’ or ‘interrural’ is used as a variable in the estimation. In Figure 16, when a trip has at least one night spent under way, the proportion of trips made by transit and especially by air is much higher. Therefore, whether a trip is overnight or same-day is used as a potential explanatory variable.

Figure 13. Travel party size vs. modal share

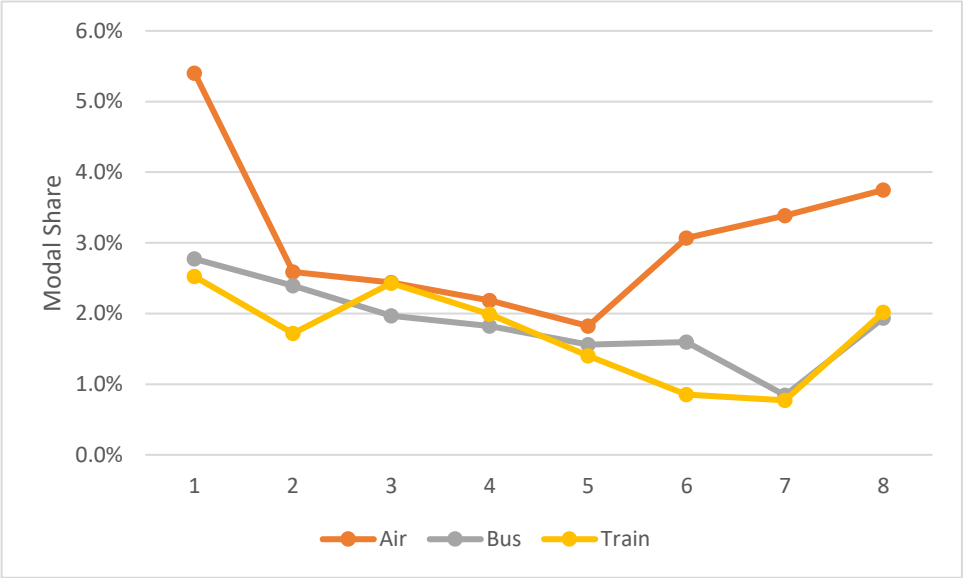


Figure 14. Modal share by season

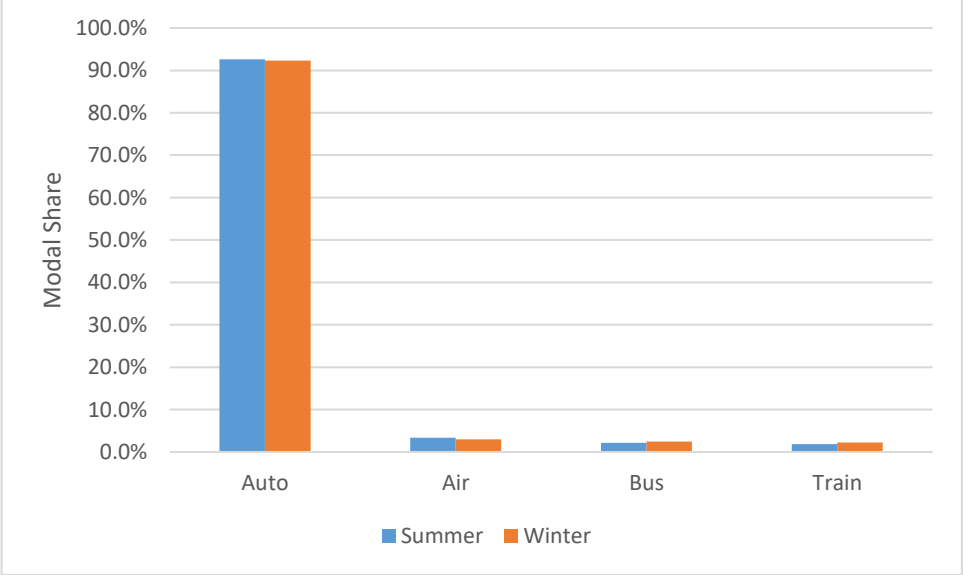
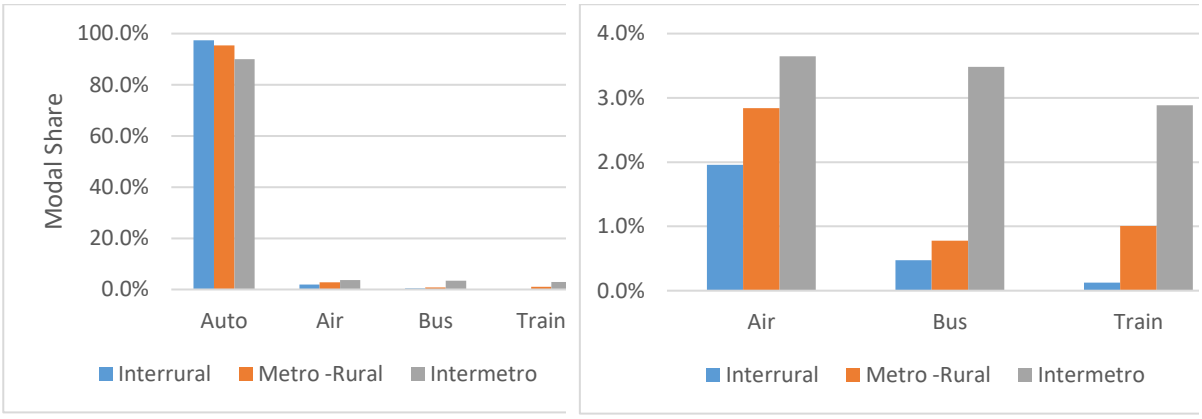
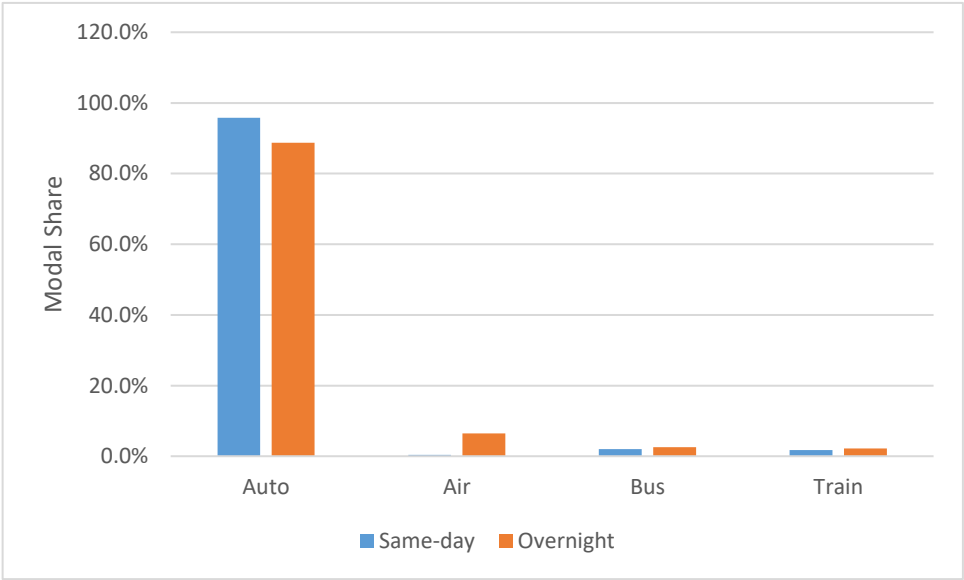


Figure 15. Origin and destination in metro or rural areas vs. modal share (a) with auto and (b) without auto



(a) (b)

Figure 16. Same-day or overnight trip vs mode share



3.4.3. Mode by distance

It is clear that modal distribution is strongly influenced by trip distance. As trip distance increases, so too does modal share by air, with air gaining majority at around 1,300 km. The almost mirrored trends between auto and air reflect how as one mode goes up, the other goes down, and signifies the predominance of auto and air modes, as compared to bus and rail. For perspective, the total number of trips shows that the majority of trips made are under 200 km, with 45% of trips between 40 and 100 km, 30% between 100 and 200 km, and 96% of these trips being made with auto.

Figure 17. Modal distribution by trip distance bands

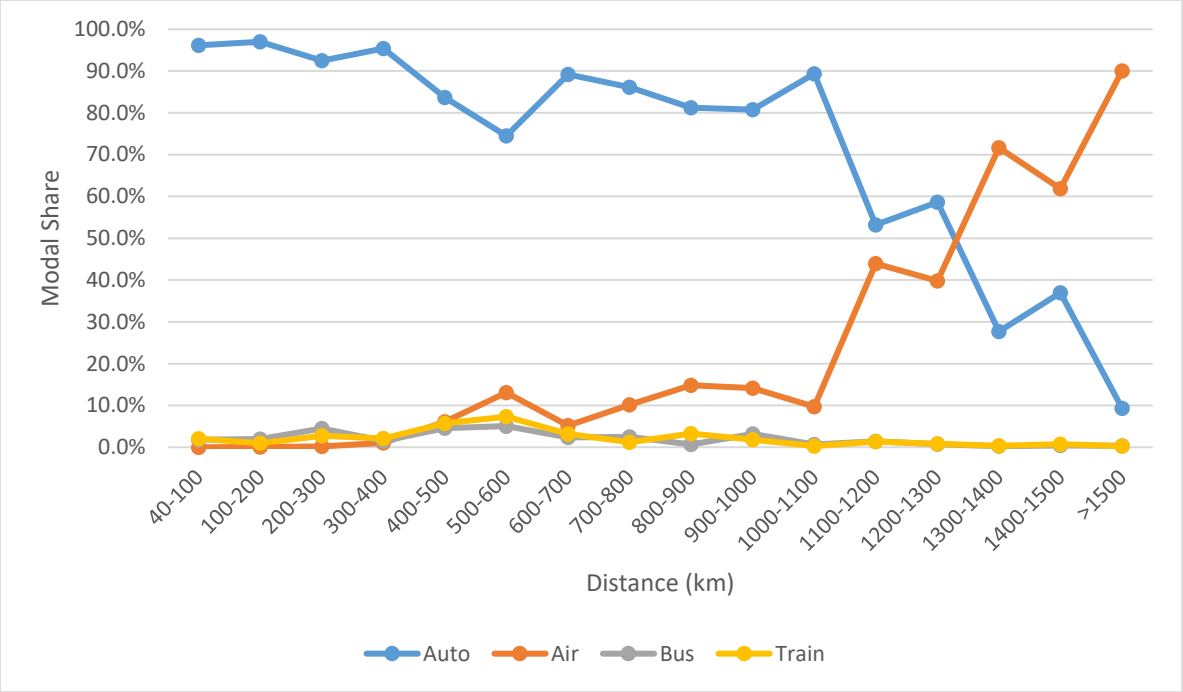
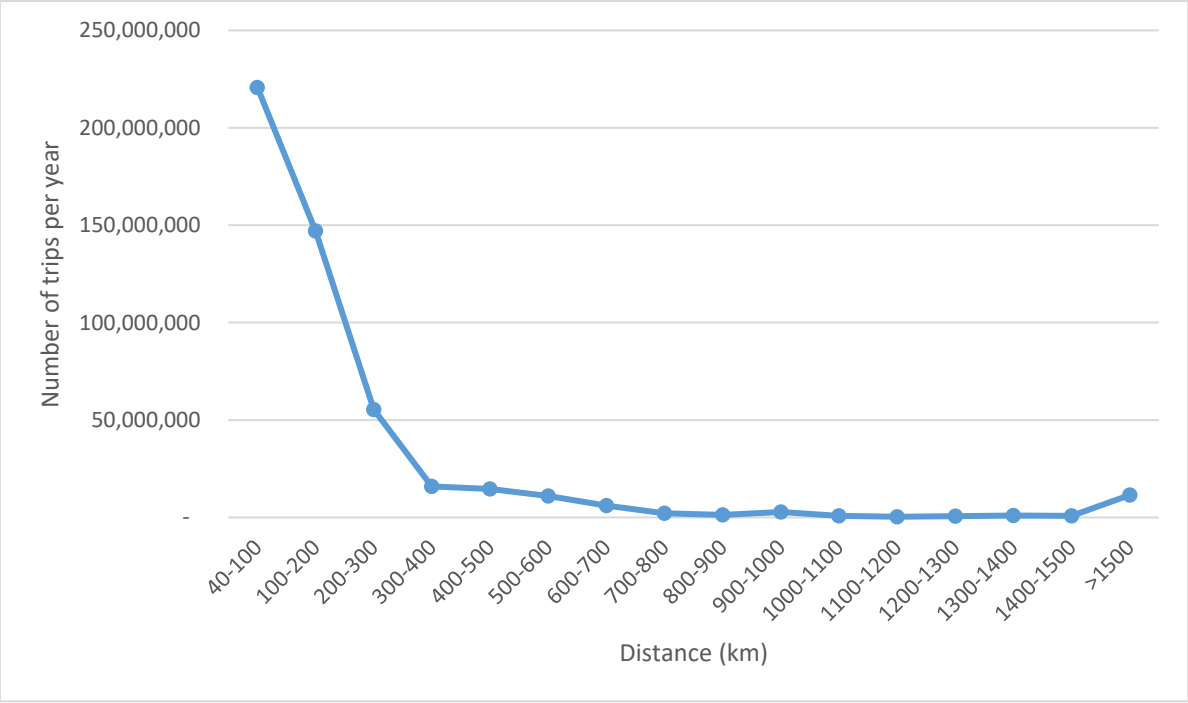


Figure 18. Total trips per year by trip distance bands (km)



3.4.4. Variable correlation

To check for possible correlation and interaction between variables, a correlation plot was created, as seen on the next page. There is some correlation between the socioeconomic

variables. Age is correlated with employment status – the older someone is, the less likely they are to be employed. This is so because the age group in TSRC starts at the working adult age 18 – 24 and continues past the retirement age of 65. Employment status is corrected with income – being employed means having higher income. There is also some negative correlation between number of household members on a trip and travel party size, which is to be expected. During estimation, only one of these correlated variables were tested at a time. In the mode-specific characteristics, correlations are prevalent between travel time and price for all modes. Travel cost for air is less correlated to travel time, which is reflective of the volatile and opaque pricing structure of air travel. Travel times and costs are also highly correlated to distance and between modes. These correlations have implications for model estimation, as will be discussed in the next section.

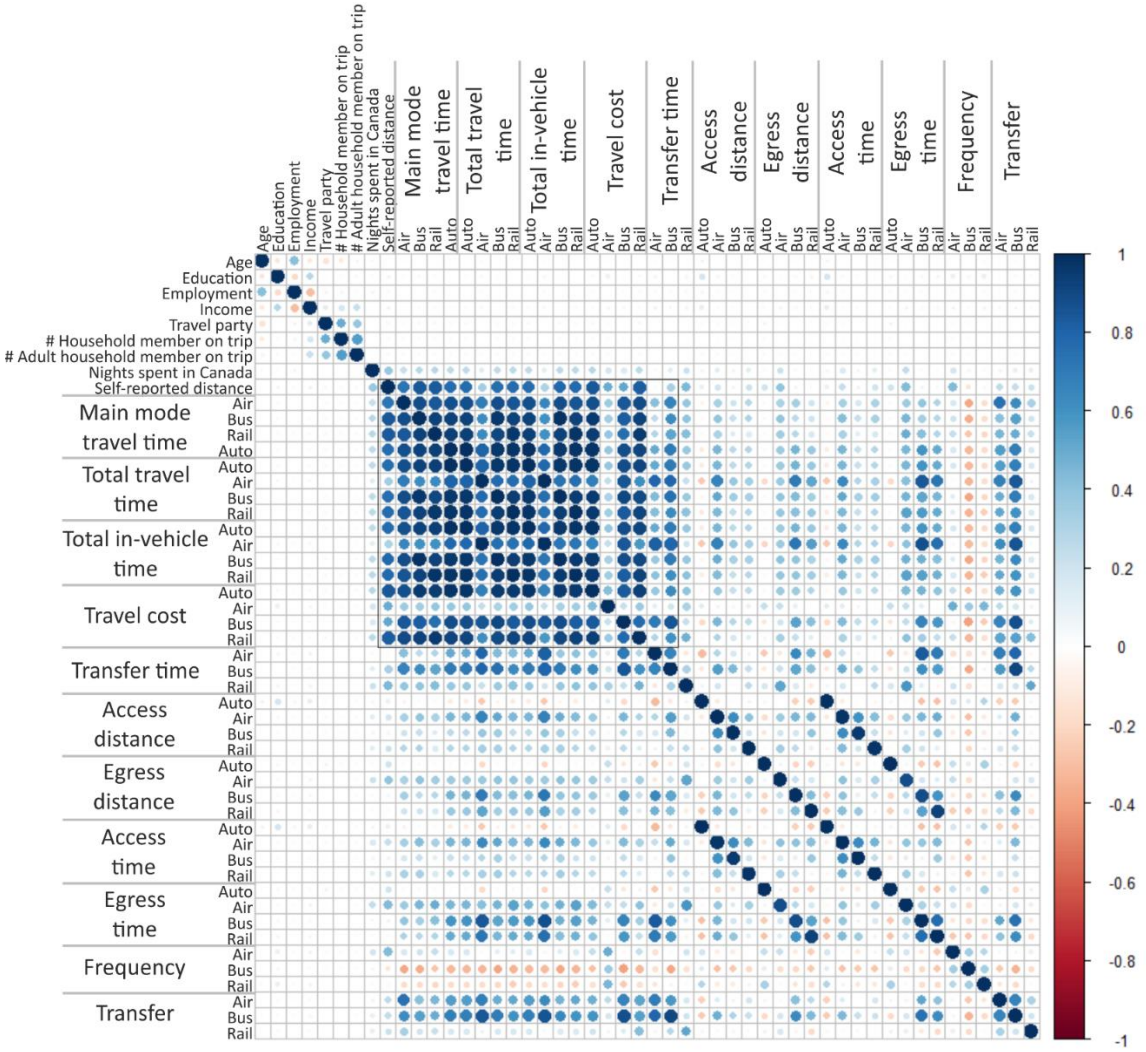


Figure 19. Correlation matrix

4. Model estimation

4.1. Model specification

The model is specified before the estimation process. This involves specifying the alternative choice set, model segmentation, explanatory variables and model structure. In the previous chapter, the choice set was determined to be auto, air, bus and rail, and the segmentation to be by trip purpose into business, leisure and visit.

4.1.1. Variables considered

The variables considered were heavily dependent on the mode choice modeling state-of-the-art and on the characteristics of the data, as shown in the previous chapter. I considered both variables that have often been shown to influence mode choice and variables that seem to have significant correlations with mode share in the data. Variables highly correlated to each other were discarded or only considered one at a time. The variables are broken down into three groups: mode characteristics, trip characteristics, and individual trip-maker characteristics. Mode characteristics are those that represent the LOS of the mode, such as travel time and frequency of service. Trip characteristics describe characteristics of the trip such as the time of day a trip is made and the travel party size of the trip. Individual specific variables are the characteristics of the trip maker, such as income or education level, or the trip maker's household size, such as the number of children.

Table 4 summarizes the variables considered for estimation. Note that some variables represent categorical data and are used in the model as so-called dummy variables. This means that they are coded as either 1 or 0, 1 meaning the characteristic is present, and 0 meaning it is not. For example, a dummy variable for income could be a 1 representing those in the lowest income bracket, and 0 representing those who are in all other income brackets.

Table 4. Variables considered in the estimation

Variables considered		Type of variable
Individual characteristics		
Age		Dummy
Gender		Dummy
Education		Dummy
Employment status		Dummy
Income		Dummy
Number of household members on trip		Continuous
Number of adult household members on trip		Continuous
Trip characteristics		
Travel party size		Continuous
Overnight or same-day		Dummy
Intermetro/interrural		Dummy
Mode characteristics (per each mode)		
Frequency (excl. auto)		Continuous
Number of transfers (excl. auto)		Continuous
Travel cost		Continuous
Travel time	Access time	Continuous
	Egress time	Continuous
	Transfer time (excl. auto)	Continuous
	Main mode travel time	Continuous
	Total travel time	Continuous
Travel distance	Access distance	Continuous
	Egress distance	Continuous
	Main mode travel distance	Continuous
	Total travel distance	Continuous

4.1.2. Utility equation

The utilities are estimated against a base alternative. In this model, the alternative auto was taken as the base alternative, with its constant set to zero.

Equation 9

$$U_{it}^{k} = V_{it}^{k} + \varepsilon_{it}^{k}$$

where U_{it}^{k} is the utility of taking alternative *i* (*i* = auto, air, bus and rail) to trip-maker *t* for purpose *k*, V_{it}^{k} is the observable portion of the utility and ε_{it}^{k} is the error portion of the utility, which is assumed to be Gumbel-distributed in the MNL estimation (Koppelman and Bhat, p. 19, 2006).

The observable part of the utility, V_{it}, is composed of portions related to the mode characteristics, the trip-maker's characteristics and the trip's characteristics.

Equation 10

$$V_{it}^{k} = V_{mode}^{k} + V_{trip-maker}^{k} + V_{trip}^{k}$$

Each of these portions of utility is a linear addition of estimated parameters multiplied by the attribute.

Equation 11

$$V_{mode}^{k} = \beta_{im}^{k} * A_{im}^{k} \dots \dots$$

where V_{mode}^{k} is the portion of the utility related to the characteristics of the mode, β_{im}^{k} is the *m*th parameter for mode *i* for purpose *k*, and A_{im}^{k} is the *m*th attribute of mode *i* for purpose *k.*

Equation 12

$$V_{trip-maker}^{k} = \beta_{im}^{k} * A_{mt}^{k} \dots \dots$$

where $V_{trip-maker}^{k}$ is the portion of the utility related to the characteristics of the trip-maker, β_{im}^{k} is the *m*th parameter for mode *i* for purpose *k*, and A_{mt}^{k} is the *m*th attribute of traveler *t* for purpose *k.*

Equation 13

$$V_{trip}^{k} = \beta_{im}^{k} * A_{mp}^{k} \dots \dots$$

where V_{trip}^{k} is the portion of the utility related to the characteristics of the trip, β_{im}^{k} is the *m*th parameter for mode *i* for purpose *k*, and A_{mp}^{k} is the *m*th attribute of trip *p* for purpose *k.*

This makes the utility for a mode *i,* individual *t* for purpose *k*:

Equation 14

$$U_{it}^{k} = \beta_{0i}^{k} + V_{mode}^{k} + V_{trip-maker}^{k} + V_{trip}^{k}$$

where β_{0i} is the alternative specific constant, or mode constant for mode *i* for purpose *k*, which represents the portion of utility that is not estimated by the variables.

4.2. Estimation using R package mlogit

The model is estimated using the package for multinomial logit models, *mlogit*, developed by Yves Croissant. The package has the capability to estimate the basic MNL model and other popular logit class models such as NL (Croissant, 2011).

4.2.1. Mlogit data format

The *mlogit* function accepts data in wide and long format. The wide format has one row per each choice while the long format has one row per each alternative. In this case, there would be four rows per trip record, one per each mode. The long data format is used here to account for the mode specific characteristics.

4.2.2. Mlogit function

The *mlogit* function accepts a formula and a dataset. The *mlogit* formula consists of three types of variables:

- Alternative specific variables with a generic coefficient across all alternatives
    - Ex. Generic travel time coefficient
- Individual and trip specific variables
    - Ex. Income, age, number of travelers
- Alternative specific variables with different coefficients for each alternative (Croissant, 2011)
    - Ex. Travel time for auto, air, bus, and rail separately

The function then outputs estimated coefficients, statistical measures of each parameter and of the overall model, e.g. the t-statistics and the log-likelihood, etc.

For nested logit models, the *mlogit* function offers the same specifications in terms of variables, with an added specification of nesting structures. The function then outputs the estimated coefficients as well as the nesting coefficient.

4.3. Multinomial logit model estimation results

The process of model specification and estimation is not strictly linear. Model estimation informed the specification of the model, i.e. the results from model estimation showed whether variables worked reasonably together, and variables were then modified, added or dropped to form a coherent estimation. This process continues until one reaches a final specification of

the utility equation with good statistical performance and theoretical soundness. This sort of feed-back loop process is common in model estimation.

Estimations were performed for different combinations of explanatory variables. Variables were added gradually, in the order of alternative-specific variables, trip-specific variables, and individual-specific variables. Variables that were theoretically consistent and had at least a 95% significance were retained.

As this model should be policy sensitive for MTO, it was important to have both travel time and travel cost variables. This was problematic due to the high correlation between travel time and travel cost in long distance travel, which resulted in a positive price coefficient in some estimations. In the end, the most consistent model results came from having one aggregated coefficient for travel price. Travel time worked well as an aggregated coefficient across all modes for business and visit trips, and was disaggregated for the leisure purpose. Number of transfers was discarded when the coefficient became positive and illogical, but frequency of service was kept.

The models and coefficients of this model, which will be called Model 1, are shown Table 5, although the variables are only explained in detail later in the chapter since this is not the final version of the model.

Table 5. Coefficients of Model 1 with travel cost and travel time as separate variables (Significant codes: *** 99.9% significance level, ** 99%, * 95%)

Mode	Parameter	Coefficent	Significant	Coef.	Sig.	Coef.	Sig.
Air	Intercept	-1.293378	***	-5.055	***	-6.45	***
Bus	Intercept	-3.54818	***	-4.124	***	-3.354	***
Rail	Intercept	-4.380824	***	-3.892	***	-3.078	***
	Frequency	0.0029796	***	0.0031	***	0.0027	***
	Travel cost	-0.00598	***	-0.002	***	-0.001	**
	Travel time	-0.004449	***			-0.005	***
Auto	Travel Time			-0.003	***		
Air	Travel Time			-8E-04	*		
Bus	Travel Time			-0.001	***		
Rail	Travel Time	-0.0049	***	-0.002	***		
Air	Intermetro	0.3828137	**				
Bus	intermetro	0.4362147	.			1.7715	***
Rail	Intermetro	1.6676012	***			0.6605	***
Air	Interrural			-1.056	***		
Bus	Interrural			-1.144	***		
Rail	Interrural			-3.748	**		
Air	Overnight	1.1448042	***	1.7124	***	3.565	***
Bus	Overnight	1.0552277	***	0.4485	***	1.5068	***
Rail	Overnight	0.8852667	***	0.9029	***	0.8427	***
Air	Group size	-0.27165	***	-0.174	***	-0.532	***
Bus	Group size	-0.428316	**	-0.404	***	-1.17	***
Rail	Group size			-0.578	***	-0.926	***
Air	Young (<25)	-1.777155	***				
Bus	Young (<25)	1.1630496	***			1.5996	***
Rail	Young (<25)					1.5898	***
Air	Male	-0.541929	***	1.217	***	-0.673	***
Bus	Male	-0.582315	**	1.4509	***	-0.465	***
Rail	Male	-0.888204	***				
Air	Highly educated	0.6087562	***				
Bus	Highly educated	0.8073368	***				
Rail	Highly educated	0.8343339	***				
Air	High income					0.6112	***
Bus	High income					-0.89	***
Rail	High income					-0.908	***
Air	Low income	-0.876095	***				
Bus	Low income	0.5372872	***	1.3243	***		
Rail	Low income			0.5619	***		
	Log-likelihood	-2074.8		-3923.7		-6090.8	
	McFadden's R^2	0.47		0.35		0.46	

The following chart shows the log-likelihood and Mcfadden's R^2 for a mode-specific constants-only model, a variables-only model, and the full model. This is done to assess the contribution of the variables. In other words, this shows how much the variables improve the model, and how much of the model is determined by the constants alone. In general, smaller constants are preferred, as this means the variables are capturing most of the behavior.

Table 6. Full model comparison with constants-only model and model without constants

Purpose	Constants Only	No Constants	Full Model
Log-likelihood			
Business	-3910.2	-2315.4	-2074.8
Leisure	-6038.5	-4630	-3923.7
Visit	-11261	-6504.1	-6090.8
McFadden's R^2			
Business		0.41	0.47
Leisure		0.23	0.35
Visit		0.42	0.46

The comparison indicates that the variables do go some way towards explaining mode choice behavior. The log-likelihood and R^2 values of the full model improve over the constants-only model by almost 50%. The model with variables but no constants is still a signification improvement compared to constants-only and does not increase in log-likelihood substantially with constants added (full model). This could mean that the constants do not play as strong a role when compared to the variables, which is encouraging.

4.3.1. Confusion matrix

A confusion matrix, or a misclassification matrix, was also created to see the effectiveness of the model. The sum of each row represents the actual number of observations while the sum of each column represents the predicted number. The diagonal cells are the matched observations while the non-diagonal cells show how much each mode is misclassified as another mode choice. The individual match rate is the percentage of correctly predicted observations over the number of actual observations, which shows the accuracy of predictions. The aggregate match rate is the total number of predicted observations, correctly and incorrectly predicted, over the actual number of observations per mode.

Table 7. Confusion matrix for the (a) business, (b) leisure and (c) visit mode choice model including time and price

Business Confusion Matrix								
		Predicted						
		Auto	Air	Bus	Rail	Actual total	Individual match rate%	Aggregate match rate%
Actual	Auto	22,612	858	645	1,149	25,264	89.50%	99.86%
	Air	929	2,986	28	111	4,054	73.66%	99.33%
	Bus	574	46	39	89	748	5.18%	102.64%
	Rail	1,115	137	56	203	1,510	13.46%	102.78%
	Predicted total	25,229	4,027	768	1,552	31,575		

(a)

Leisure Confusion Matrix								
		Predicted						
		Auto	Air	Bus	Rail	Actual total	Individual match rate%	Aggregate match rate%
Actual	Auto	121,864	754	2,040	1,408	126,066	96.67%	100.00%
	Air	787	1,432	38	17	2,274	62.97%	98.99%
	Bus	1,969	32	109	94	2,204	4.92%	102.54%
	Rail	1,451	33	74	109	1,666	6.51%	97.69%
	Predicted total	126,071	2,251	2,260	1,628	132,209		

(b)

Visit Confusion Matrix								
		Predicted						
		Auto	Air	Bus	Rail	Actual total	Individual match rate%	Aggregate match rate%
Actual	Auto	131,646	865	3,252	2,546	138,309	95.18%	99.99%
	Air	924	3,549	27	29	4,528	78.37%	100.04%
	Bus	3,186	50	753	422	4,411	17.07%	101.85%
	Rail	2,541	66	461	371	3,438	10.78%	97.96%
	Predicted total	138,296	4,530	4,493	3,368	150,686		

(c)

It is reassuring to see that the model has a very good aggregate match rate. There is a high misclassification rate for the modes bus and rail. It seems most bus and rail trips are misclassified as auto trips, which may indicate these modes share many similar characteristics. This could also be due to heavy dominance of the auto mode and the low number of observations for the modes bus and rail in the dataset.

However, this was not the final iteration of the model, as the implied values of time were not theoretically consistent. This issue is discussed in full in the next section.

4.3.2. Value of time

In mode choice modeling, when the coefficient for time and the coefficient of cost are known, it is possible to calculate the hidden value of time (VOT), i.e. how much the traveler values a unit of their time.

Equation 15

$$Utility_{mode} = \beta_1 \cdot travel\ time + \beta_2 \cdot travel\ cost + \varepsilon$$

Equation 16

$$Utility_{mode} = \frac{\beta_1}{\beta_2} \cdot travel\ time + travel\ cost + \varepsilon$$

Equation 17

$$Value\ of\ Time\ (VOT) = \frac{\beta_1}{\beta_2}$$

Though both travel time and travel cost coefficients were negative across all trip purposes, the calculated hidden value-of-time implied by them were not consistent with expectations.

Table 8. The implied value of time for each trip purpose (Note that the leisure model has disaggregated travel time coefficients and therefore has a value of time per mode while the other two models do not)

	Business	Leisure	Visit
Auto		$ 96.89	
Air		$ 25.74	
Bus		$ 33.23	
Rail		$ 61.32	
General	$ 44.64		$ 235.09

The VOT for trip purpose business is smaller than non-business, which goes against the common understanding that business travelers tend to be more sensitive to time and should thus have a higher VOT. In the disaggregated VOTs calculated for the leisure model, air travelers had the lowest VOT, followed by bus, rail and auto. This ranking is illogical, as flying is usually the fastest and most expensive option, and normally we would expect air travelers to have higher VOTs than other modes. Bus travelers, for example, accept a longer journey time in return for a cheaper fare.

To investigate further, models were estimated with a travel cost coefficient per each income category. The Table 9 shows the implied value of time per each income category for each trip purpose.

Table 9. The implied Value-of-Time per income category for each trip purpose

Income Category	Business	Leisure	Visit
< $50,000	$ 69.36	$ 134.36	$ 116.35
$50,000 - $70,000	$ 45.34	$ 21.44	$ 86.76
$70,000 - $100,000	$ 34.65	$ 24.79	$ 97.57
>$100,000	$ 46.00	$ 64.44	$ (353.58)

Table 9 shows no clear trend of value-of-time across income categories. The lowest income bracket has the highest implied value-of-time across all purposes. The highest value-of-time belong to the lowest income bracket leisure travelers. There was even a negative VOT for high income bracket visit travel. So, this segmentation of travel cost by income bracket was discarded.

The next possibility was to fix the value of time. This means a VOT is asserted rather than derived from the estimated model. The asserted VOT is used to convert either time into cost or vice versa, forming one general impedance factor that incorporates both time and cost. However, there is no consensus in the modeling community on the right VOTs, and VOTs can vary depending on demographics and trip characteristics. To mitigate this uncertainty, the Ministry of Transportation of Ontario proposed VOTs based on past and current relevant travel demand models and rail forecasting models in the region.

Table 10. Proposed value of time in Canadian dollars per hour

Mode		Business	Leisure	Visiting
Option 1				
Air		$ 100	$ 50	$ 50
Bus		$ 35	$ 20	$ 20
Auto		$ 60	$ 30	$ 30
Rail		$ 60	$ 30	$ 30
Option 2				
		$ 65	$ 32	$ 32
Option 3				
IVTT	Air	$ 75	$ 45	$ 45
	Bus	$ 20	$ 15	$ 15
	Auto	$ 50	$ 30	$ 30
	Rail	$ 30	$ 20	$ 20
	average	$ 45	$ 25	$ 25
Access/Egress	Air	$ 110	$ 65	$ 65
	Bus	$ 35	$ 20	$ 20
	Auto	$ 50	$ 30	$ 30
	Rail	$ 70	$ 40	$ 40
	average	$ 65	$ 40	$ 40

The VOTs are separated by trip purpose. Option 1 differentiates VOT by mode. Option 2 has one generic VOT for all modes. Option 3 has VOTs by mode and by in vehicle travel time (IVTT) and access/egress time (OVTT). Option 3 was not considered because the model does not distinguish access/egress time from IVTT. Options 1 and 2 were tested by converting travel cost into travel time using VOT and adding the two together to form a generalized time variable (GT).

Equation 18

$$GT\ (general\ time) = \frac{travel\ cost}{VOT} + travel\ time$$

Equation 19

$$u_{mode} = \beta_1 \cdot GT + \varepsilon$$

Using VOT Options 1 and 2, three different models were derived and compared, making a total of four models including the already estimated model with separate travel time and travel cost coefficients

Model 1. Existing model estimation with separate coefficients for travel time and travel cost

Model 2. One general VOT (Option 2) used to calculate GT and one coefficient derived for GT

Model 3. One general VOT (Option 2) used to calculate GC and four coefficients derived for GT

Model 4. Four VOT (Option 1) to calculate GT and one coefficient derived for GT

Table 11. Log-likelihood and Mcfadden's R^2 values across all four models

Model	1	2	3	4
	Time & Price	1VOT 1 COEF	1 VOT 4 COEF	4 VOT 1 COEF
Log-likelihood				
Business	-2074.8	-2077.3	-2034.4	-2096.2
Leisure	-3923.7	-3923.9	-3826.3	-3923.7
Visit	-6090.8	-6092.2	-6097.7	-6077
McFadden's R^2				
Business	0.47	0.47	0.48	0.46
Leisure	0.35	0.35	0.37	0.35
Visit	0.46	0.46	0.46	0.46

As can be seen Table 11, the likelihoods and R^2 values were similar. No one model performed consistently better. Further analysis was conducted to differentiate the model performances.

4.3.3. Scenario analysis

As a continuation of the VOT discussion, a sensitivity analysis was also conducted with the help of Dr. Carlos Llorca. The city pair Toronto – Montreal was selected. To see the sensitivities of the models to travel time and travel cost, we analyzed a few scenarios: doubled air travel time, doubled air fare, and halved rail travel time. The analysis was run 1000 times to account for stochastic effects.

Air Scenarios

Below are the results for the trip purpose business:

Figure 20. Comparison of business modal changes for Toronto – Montreal with doubled air travel time

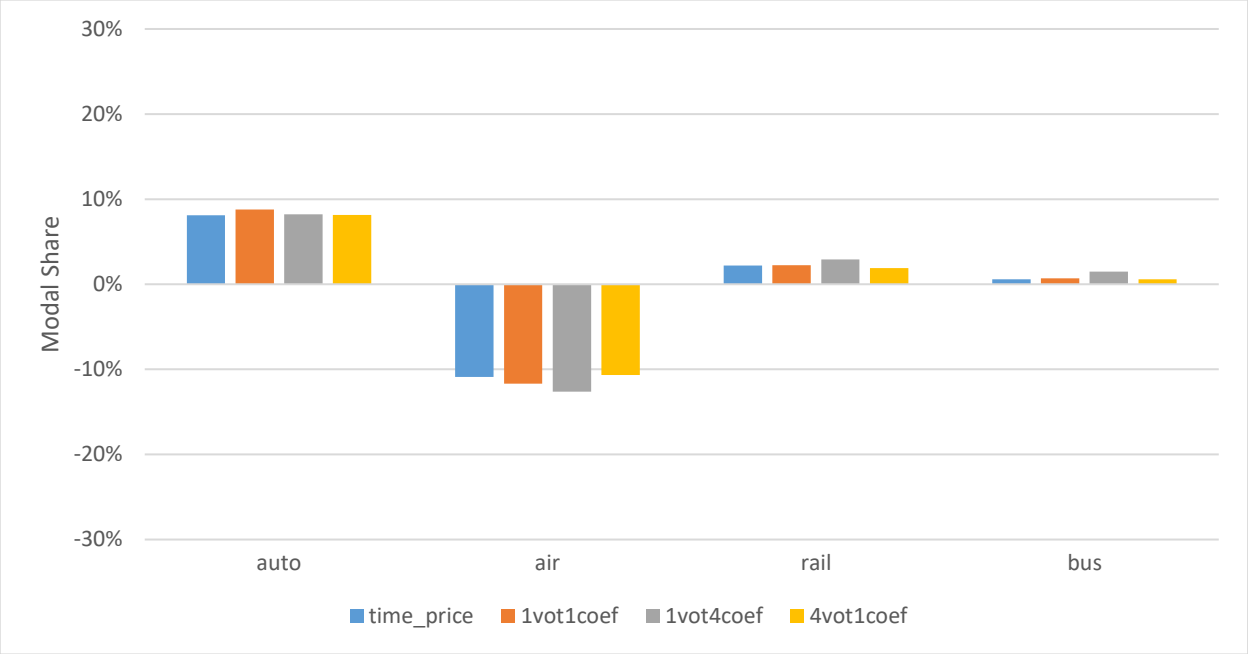
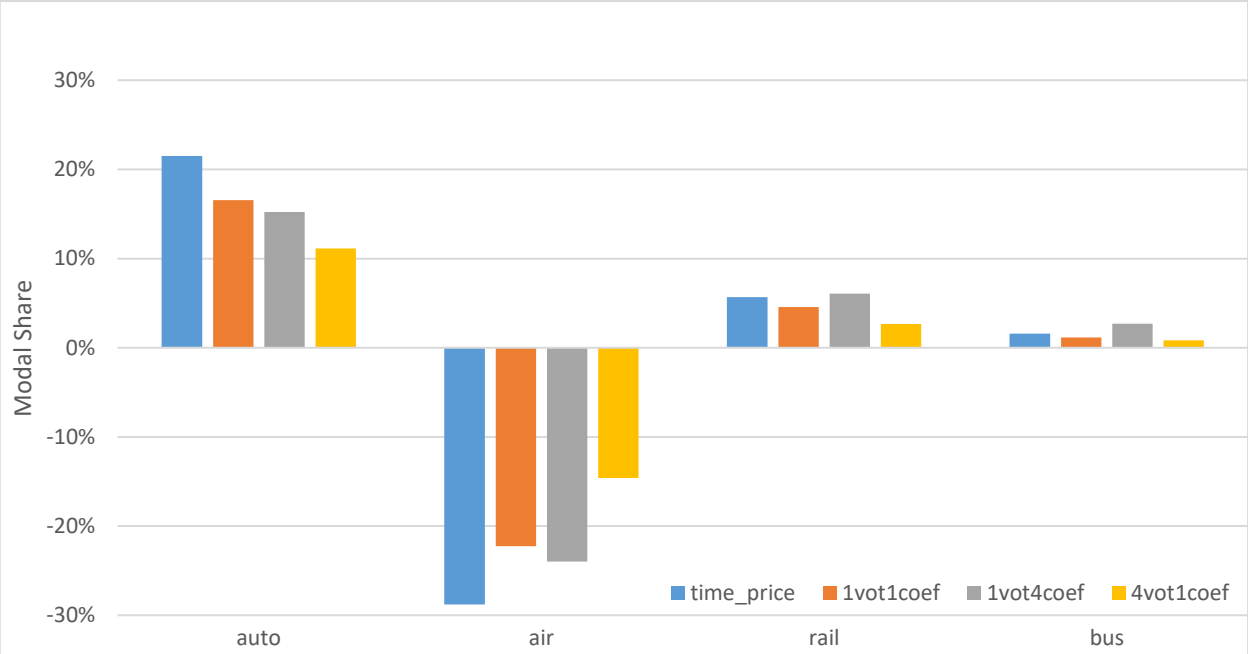


Figure 21. Comparison of business modal changes for Toronto – Montreal with doubled air fare



As expected, when airfare or air travel time doubled, the attractiveness of air mode went down, resulting in less air modal share and more of the other three modes. Auto had the biggest modal share increase, followed by rail and bus, which is also in accordance with their current modal share. Since all models show the correct sign of sensitivities, and we do not have

information on the correct, expected sensitivity, it was hard to use this as a measurement of model performance.

High Speed Rail Scenario

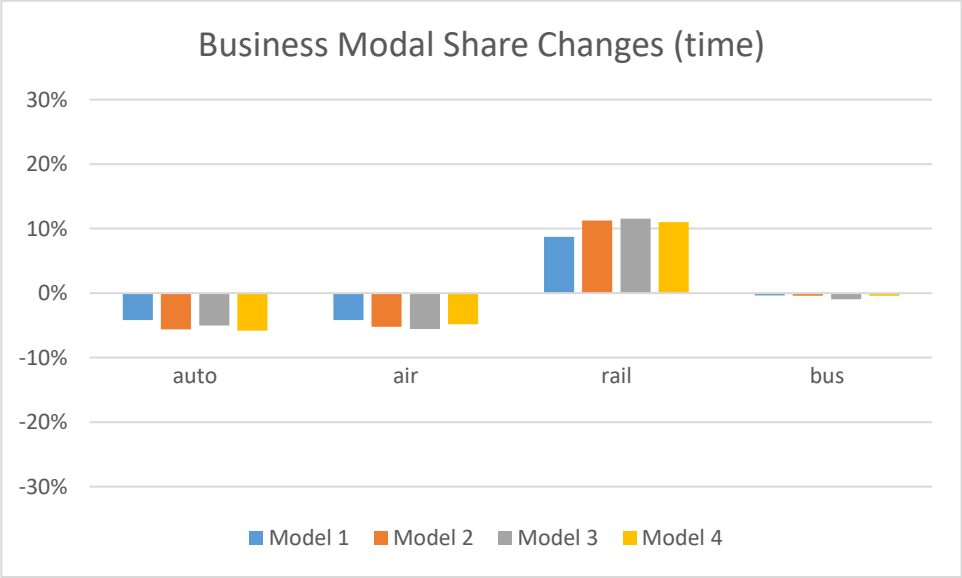
For the past few decades there has been continued interest in the potential of high speed rail in Canada, particularly for the southern Ontario region. The Quebec City – Windsor corridor region is the most densely populated corridor in Canada and is already well-serviced by air, bus and existing rail. Thus, at the behest of MTO, a scenario analysis with rail was also conducted for Toronto – Montreal, two of the biggest metropolitan areas in this corridor and in Canada.

Currently, the travel speed by train between the OD pair is five hours. Taking the straight driving distance between the pair to be 550 km, the average train speed is currently about 100km/hr. Assuming a high-speed rail speed of 200 km/h, the new travel time would be 2.5 hours, approximately halving the current travel time by rail.

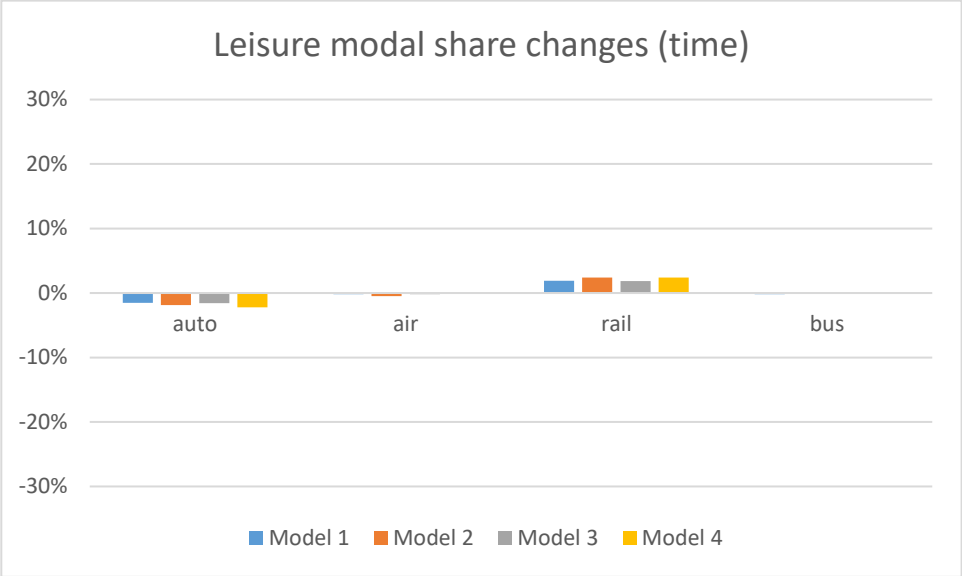
This scenario is rather simplified and is only concerned with Toronto to Montreal one-way trips.

As is the case with the air mode sensitivity analysis, there is no clearly superior model. All four models show the correct signs – rail modal share goes up while auto, air and bus modal shares go down.

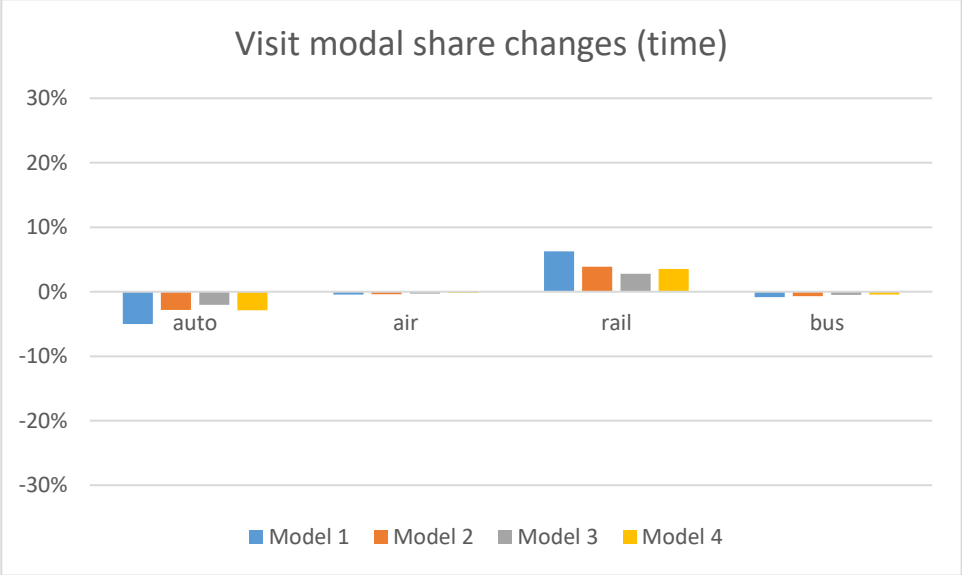
Figure 22. Comparison of modal share changes for Toronto – Montreal with high speed rail for trip purposes: (a) business, (b) leisure, (c) visit



(a)



(b)



(c)

4.3.4. Comparison of coefficients

The coefficients were also considered. All four models had similar statistically significant variables. It was noted that the coefficients for Models 2 and 4 were similar while Model 3 had markedly different coefficients. Since Models 2 and 4 had similar coefficients and thus validate each other, it was more likely that the disaggregation of the GC coefficient in Model 3 might have affected the behavior of other parameters. Furthermore, the relative ranking of the estimated GT coefficients were also not strictly logical. In the business model, the most negative GT was auto, followed by rail, air, and bus. This implies that, for an increase in travel

time and/or travel cost, auto would be the least attractive, followed by rail, air and finally bus. It is hard to argue that bus is the most attractive mode for business travelers unless we assume these are captive riders. The rankings of Model 3 for the other two purposes follow different patterns (from most to least negative: leisure: auto, rail, bus, air; visit: auto, bus, rail, air) and are similarly hard to justify. Thus we decided to reject Model 3 and pick either Models 2 or 4. Of these two models, Model 2 was favored with the argument that when two models are similar, we should strive for the more parsimonious model, i.e. the model that can explain the observed behavior with as few variables as possible. Therefore the rest of the analysis was conducted with Model 2. Table 12 shows the coefficients of all four models for the trip purpose visit. Please refer to Appendix D for the detailed coefficient comparisons across models for the other trip purposes.

Table 12. Coefficients for models 1 – 4 for trip purpose visit (Significant codes: *** 99.9% significance level, ** 99%, * 95%)

		Model 1		**Model 2**		**Model 3**		**Model 4**	
		Time & Price		**1VOT1Coef**		**1VOT4Coef**		**4VOT1Coef**	
Mode	Variable	Coef	Sig	Coef		Coef	Sig	Coef	Sig
Air	Intercept	-6.45048	***	-4.17917	***	-6.12773	***	-5.19775	***
Bus	Intercept	-3.3544	***	-3.30337	***	-2.10221	***	-3.24767	***
Rail	Intercept	-3.07761	***	-2.90239	***	-2.90149	***	-2.911	***
	Frequency of service	0.002678	***	0.00277	***	0.003782	***	0.00274	***
	Travel cost	-0.00127	**						
	Travel time	-0.00498	***						
	Generalized Time (4 VOT)							-0.00399	***
	Generalized Time (1 VOT)			-0.00425	***				
Auto	Generalized Time (1 VOT)					-0.0044	***		
Air	Generalized Time (1 VOT)					-0.00199	***		
Bus	Generalized Time (1 VOT)					-0.00276	***		
Rail	Generalized Time (1 VOT)					-0.0026	***		
Bus	Intermetro	1.771512	***	1.722853	***			1.70792	***
Rail	Intermetro	0.660516	***	0.691231	***			0.695349	***
Air	Interrural					-0.78429	***		
Bus	Interrural					-2.38203	***		
Rail	Interrural					-0.98862	**		
Air	Male	-0.67324	***	-0.69495	***	-0.78386	***	-0.67762	***
Bus	Male	-0.46476	***	-0.45977	***	-0.46706	***	-0.45681	***
Bus	Young (<25)	1.599585	***	1.601469	***	1.593182	***	1.602354	***
Rail	Young (<25)	1.589757	***	1.581962	***	1.579817	***	1.579972	***
Air	Group size	-0.5325	***	-0.47254	***	-0.59804	***	-0.47815	***
Bus	Group size	-1.16971	***	-1.16337	***	-1.19603	***	-1.15705	***
Rail	Group size	-0.92632	***	-0.92223	***	-0.94602	***	-0.92195	***
Air	High income	0.611178	***	0.526003	***	0.584993	***	0.55181	***
Bus	High income	-0.89031	***	-0.90202	***	-0.87169	***	-0.90533	***
Rail	High income	-0.90797	***	-0.89541	***	-0.91935	***	-0.89352	***
Air	Overnight	3.565031	***	3.509203	***	3.442622	***	3.499783	***
Bus	Overnight	1.506792	***	1.543912	***	1.223359	***	1.597125	***
Rail	Overnight	0.842706	***	0.998013	***	0.669765	***	1.001357	***
	Log-likelihood	-6090.8		-6092.2		-6097.7		-6077	
	McFadden's R^2	0.46		0.46		0.46		0.46	

5. Model Results

5.1. Model 2 coefficients

The final coefficients of Model 2, using 1 VOT for all modes and deriving 1 coefficient for all modes.

Table 13. Final model coefficients by trip purpose (Significant codes: *** 99.9% significance level, ** 99%, * 95%)

		Business		Leisure		Visit	
Mode	**Variable**	**Coefficent**	**Significant**	**Coef.**	**Sig.**	**Coef.**	**Sig.**
Air	Intercept	-1.4863	***	-3.8029	***	-4.1792	***
Bus	Intercept	-3.6936	***	-3.4143	***	-3.3034	***
Rail	Intercept	-4.3905	***	-4.2292	***	-2.9024	***
	Frequency	0.0028	***	0.0023	***	0.0028	***
	Generalized Time (1 VOT)	-0.0049	***	-0.0028	***	-0.0042	***
Air	Intermetro	0.3827	**	0.7141	***		
Bus	intermetro	0.4718	*	0.8660	***	1.7229	***
Rail	Intermetro	1.6727	***	1.3069	***	0.6912	***
Air	Overnight	1.1602	***	1.9594	***	3.5092	***
Bus	Overnight	1.0972	***	0.8945	***	1.5439	***
Rail	Overnight	0.8553	***	1.1964	***	0.9980	***
Air	Group size	-0.2663	***	-0.1693	***	-0.4725	***
Bus	Group size	-0.4433	***	-0.3740	***	-1.1634	***
Rail	Group size			-0.5407	***	-0.9222	***
Air	Young (<25)	-1.9031	***	-0.4283	*		
Bus	Young (<25)	1.0482	***	1.0870	***	1.6015	***
Rail	Young (<25)			1.3471	***	1.5820	***
Air	Male	-0.5222	***			-0.6949	***
Bus	Male	-0.6176	**	-0.3875	***	-0.4598	***
Rail	Male	-0.8944	***	-0.3475	***		
Air	Highly educated	0.6580	***				
Bus	Highly educated	0.8695	***				
Rail	Highly educated	0.8313	***				
Air	High income			0.1952	*	0.5260	***
Bus	High income			-1.3484	***	-0.9020	***
Rail	High income			-0.3724	**	-0.8954	***
Air	Low income	-1.2428	***				
Bus	Low income	0.6940	**				
	Log-likelihood	-2077.3		-3923.9		-6092.2	
	McFadden's R^2	0.47		0.35		0.46	

5.2. Mode specific variables

These are variables that represent the level-of-service of the mode. This model employs gen eric alternative-specific variables that do not vary across different modes.

Generalized Time

Though a model was estimated with travel time and travel cost as separate variables, a combined generalized time coefficient was estimated, in the interest of having a coherent and defensible VOT. The calculation of the generalized time is found in the previous section. The variable signifies how much disutility travel time and travel cost contributes. As expected, the generalized time has a negative sign across all trip purposes, meaning as travel time and/or travel cost of a mode increases, the attractiveness of traveling by that mode decreases. The GT for business purpose is the most negative, followed by visit, then leisure. This would mean that business travelers are most sensitive to travel time and/or cost changes, while leisure travelers are the least sensitive.

Frequency

The frequency represents the service frequency of transit modes. The frequency used was the number of times the service is offered per week. It has a positive coefficient, which is in line with the logic that more service frequency raises the attractiveness of the mode.

5.3. Trip-specific variables

These variables pertain to the characteristics of the trip, such as the number of travelers, time of day the trip is made, etc.

Intermetro

This is a dummy variable for trips that are made between two metropolitan areas. A zone is designated as a metropolitan area if it is tagged as a CMA in the TSRC data. The coefficients are positive for transit modes, which reflects the fact that transit modes offer better connections between metropolitan areas. The variable is strongest for rail for business and leisure trip purposes, indicating that if a trip takes place between two cities, rail becomes much more

attractive. As Canada has a rail network through southern Ontario and Quebec along big metropolitan areas, this is not surprising.

Overnight

This is a dummy variable indicating whether a trip was overnight or completed on the same day. Again the estimated coefficients were positive for transit modes. Trips with overnight stay likely require more planning, and taking transit is part of that planning, whereas trips that can be made in the same day can easily be done more spontaneously with auto. The coefficient is strongest for the air mode. This could be due to overnight trips also being longer in distance. It could also be that ground transit modes may take more time than driving, thus requiring a stay overnight.

Group size

The travel party size coefficients were negative for all transit modes. As travel party goes up in number, transit becomes less attractive compared to auto. Since auto travel cost is perceived as per vehicle, often only considering the fuel cost, it can be seen as more economical to have more passengers per car. The coefficient is least negative for air, indicating that as a large group, taking an airplane is less onerous than taking the bus or train.

5.4. Individual specific variables

These variables describe relevant characteristics of the individual trip maker, such as socio-economic variables, household size, presence of children, etc.

Age

Various combinations of the age brackets were tested, and the young demographic, 18 to 25 years old, was found to be the one with the most significantly different behavior. Being young decreases the attractiveness of flying, but increases the attractiveness of bus and transit. Young people are often more flexible with their time than with their budget and are more accepting of discomfort. They may also have lower auto ownership rates, thus necessitating trips by ground transit.

Gender

Here a dummy variable was used to flag male trip makers. When a trip maker is male, they are more likely to take auto compared to transit. This reflects that males tend to have stronger car habits (Matthies, Kuhn & Kloeckner, 2002).

Education

After testing various combinations of the education brackets, it was found that high education, i.e. having a university degree, was significant for business travel. Those who are highly educated are more likely to travel by transit. Highly educated people may be more socially and environmentally conscious. They may also have better access to information and can better plan trips by transit. Though education could be a proxy for income, the income variable was also estimated and theoretically consistent, meaning that education is significant as a stand-alone variable. Moreover, if education was merely a proxy for income, one would expect the coefficients for bus mode to be negative.

Income

For the business trip purpose, those in the lowest income bracket are less likely to fly and more likely to take the bus, as expected. For the leisure and visit purposes, those in the highest income bracket are more likely to fly than drive, and less likely to take ground transit modes. The coefficient for bus is more negative than for rail, meaning for high income earners, bus is the least attractive mode, followed by rail.

5.5. Model constants

All constants are negative in comparison to auto, which has a 0 constant. This indicates that, without comparing any other characteristics, auto is the most attractive mode. This is expected as auto is by far the dominant mode in the data. For the business model, air has the least negative constant, which is in line with the fact that air travel is the second most frequently chosen mode after auto. The ranking of air, rail and bus constants fluctuates between the other two trip purposes. This could be because the other two purposes are even more overwhelmingly auto dominated and all transit modes only make up a few percentage of the data.

5.6. Model performance

There are various ways to evaluate the performance of a model. During estimation, the log-likelihood and McFadden's R^2 value are relied upon as comparisons between model iterations.

To check the overall fit, the model is used to predict mode choice probabilities, and the prediction is compared against the observed number of trips from the original data. The prediction was done using R on the TSRC dataset. To account for stochasticity, the prediction here is an average of 100 runs.

5.6.1. Model fit by OD pair

The number of trips per mode per OD pair was calculated. The predicted number of trips is plotted against the observed number of trips for the two dominant modes, auto and air. OD pairs of interest such as Toronto-Montreal, Toronto-Ottawa and Toronto-Windsor are highlighted.

Figure 23. Number of observed vs. predicted trips by OD pair for modes (a) auto and (b) air

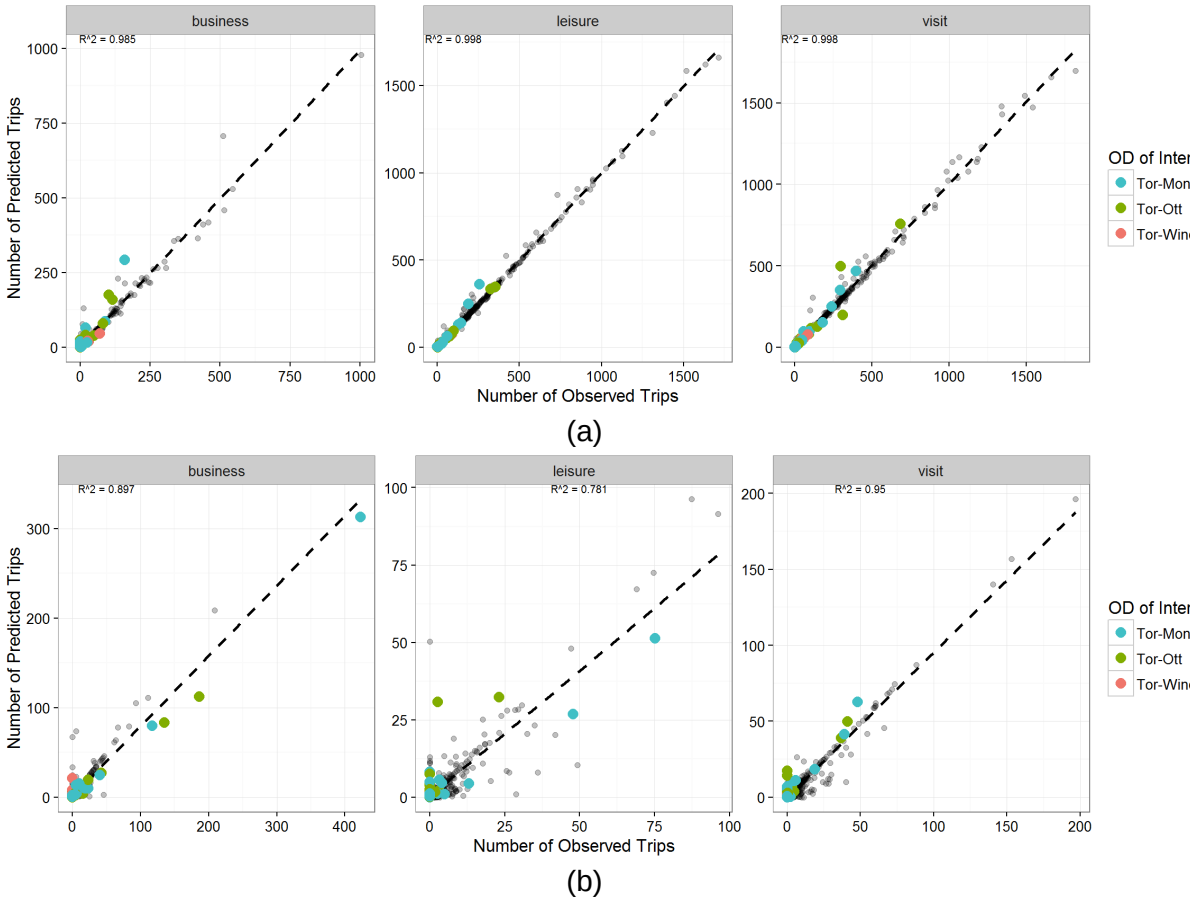
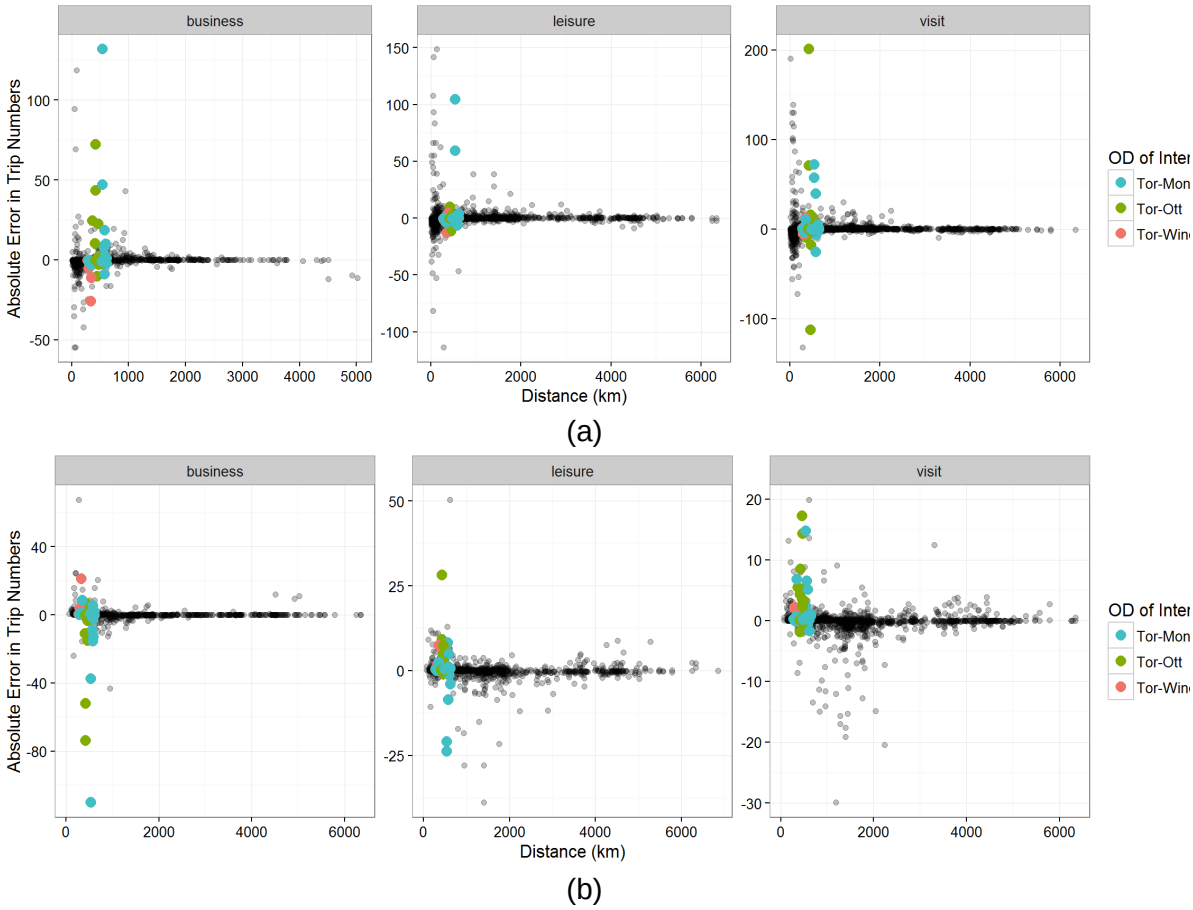


Figure 23 show that the predicted versus observed trips correlate very strongly. The weakest correlation belongs to air mode trip number for leisure trips. However, do note that the leisure purpose has the smallest number of absolute trips by air by OD pair, only going up to 100 trips per day. The important OD pairs are also generally well predicted, with the exception being Toronto – Montreal business trips by air, which is underpredicted by 100 trips.

Figure 24. Absolute error in trip numbers by OD pair vs. trip distance for modes (a) auto and (b) air

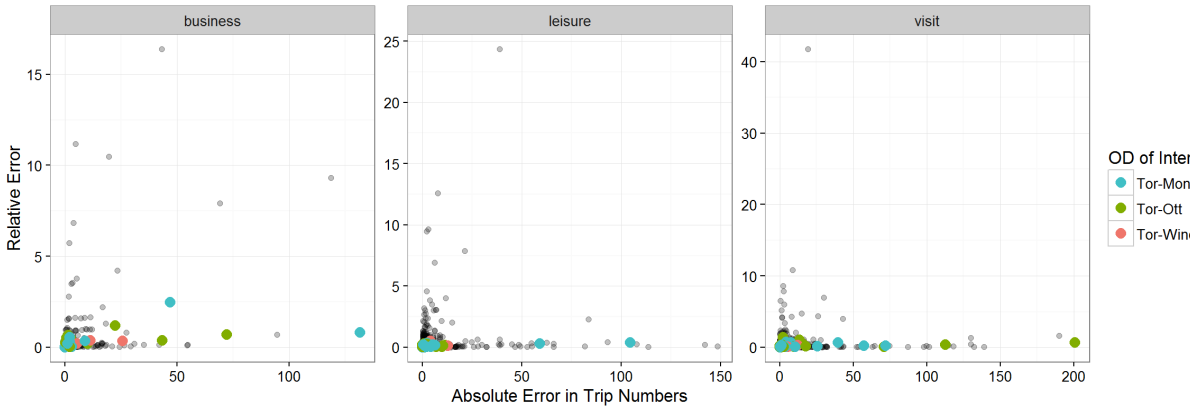


(a)

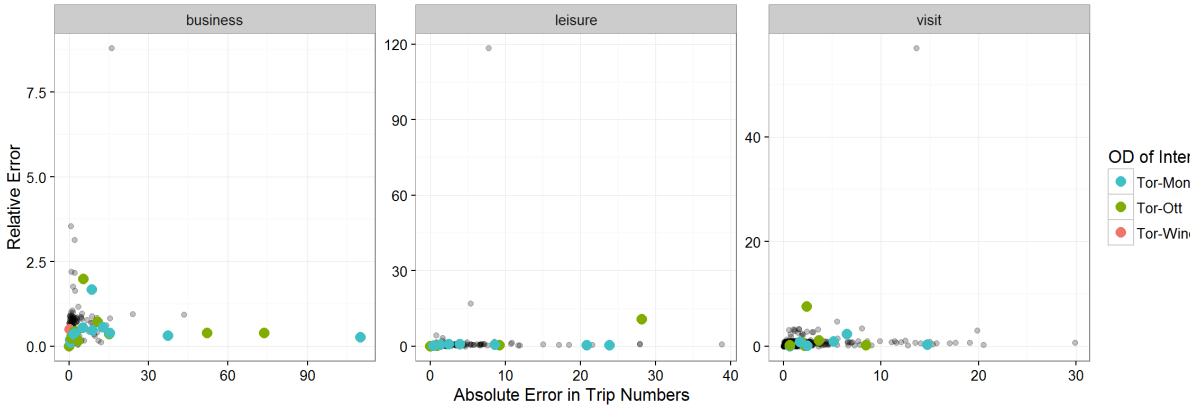
(b)

Figure 24 shows that the error is higher for short distance OD pairs and air trips tend to be underestimated for leisure and visit purposes. However, it is hard to interpret this plot on its own because there is no information on the number of trips per OD pair. The error be quite large in absolute numbers but only account for a small percentage error. To give context to the absolute error, the relative error is calculated and plotted against absolute error.

Figure 25. Relative vs. absolute error in trip numbers by OD pair for modes (a) auto and (b) air



(a)



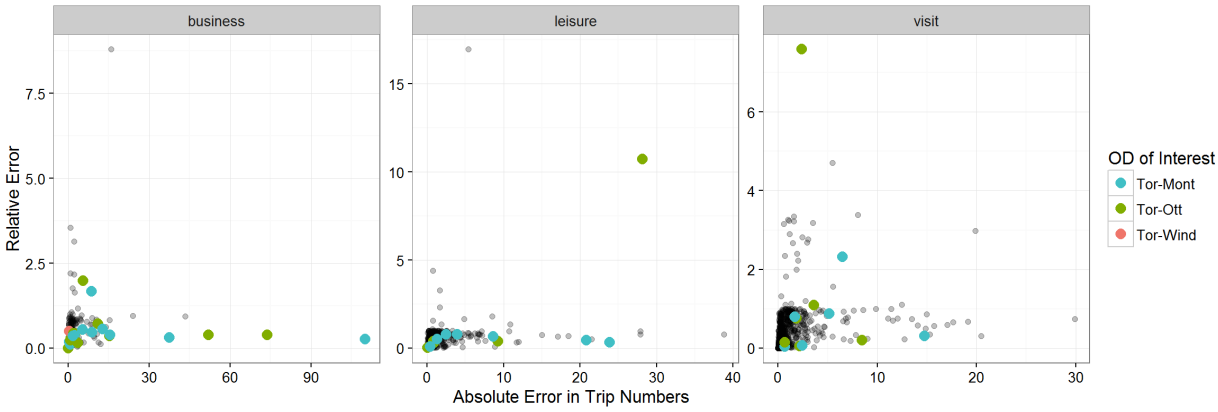
(b)

Ideally both absolute and relative errors would be small. In a relative versus absolute error graph, we hope to see that a large absolute error corresponds to a small relative error and vice versa. This would signify that a large absolute error is due to the OD pair having a large number of trips, and a large relative error could be due to the OD pair having a very small number of trips, thus tolerating a very small margin of error. In other words, the points should lie along the x or y axis.

The graphs here show the clear majority of OD pairs, including the important OD pairs, have either low absolute or low relative error. For auto trips, the business model had the lowest relative error. For the other two trip purposes, the OD pair with the highest relative error is identified as Toronto to Newfoundland and Labrador. For air trips, two outliers were identified: leisure – London to Toronto, visit – Quebec (non-CMA) to Ottawa. Taking away these OD pairs, the rest of the OD pairs had relatively small relative errors or absolute errors with the exception of Toronto to Ottawa for the leisure model, which is overestimated. Nevertheless, keep in mind that, because the model is to fit a large range of trips for the entire province of

Ontario, the number of predicted versus actual trips by OD pair is perhaps the best indication that the model is correctly predicting overall mode choice trends.

Figure 26. Relative vs. absolute error in trip numbers by OD pair for air without outliers



5.6.2. Model fit by distance

The number of trips per mode per distance is aggregated in 100 km distance brackets, with all trips above 1600 km in one bracket. The predicted modal shares are graphed against the observed modal shares.

Figure 27. Distribution of modal share by trip length for trip purpose business

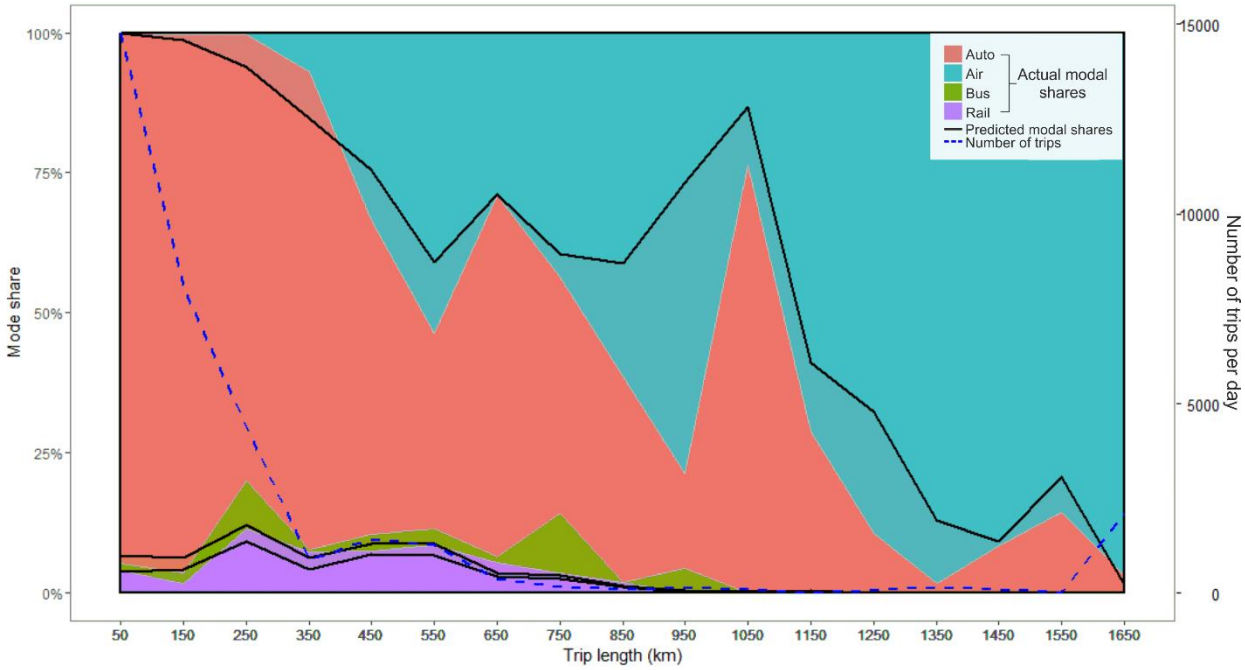


Figure 28. Distribution of modal share by trip length for trip purpose leisure

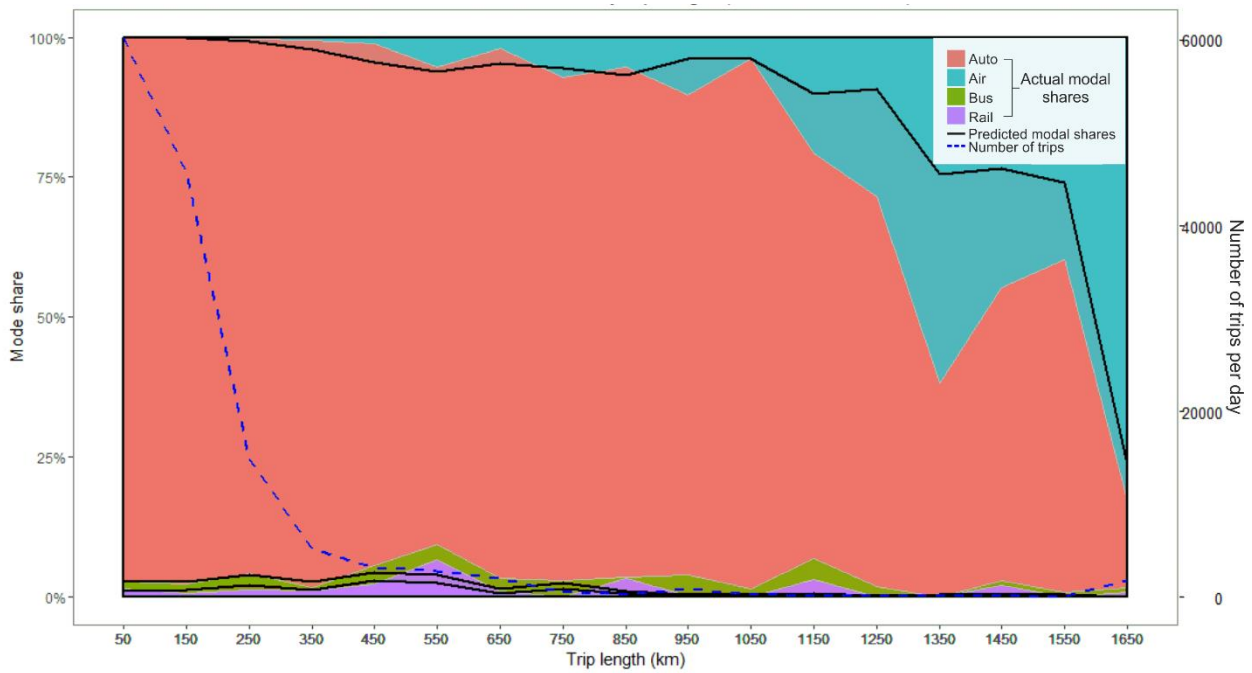
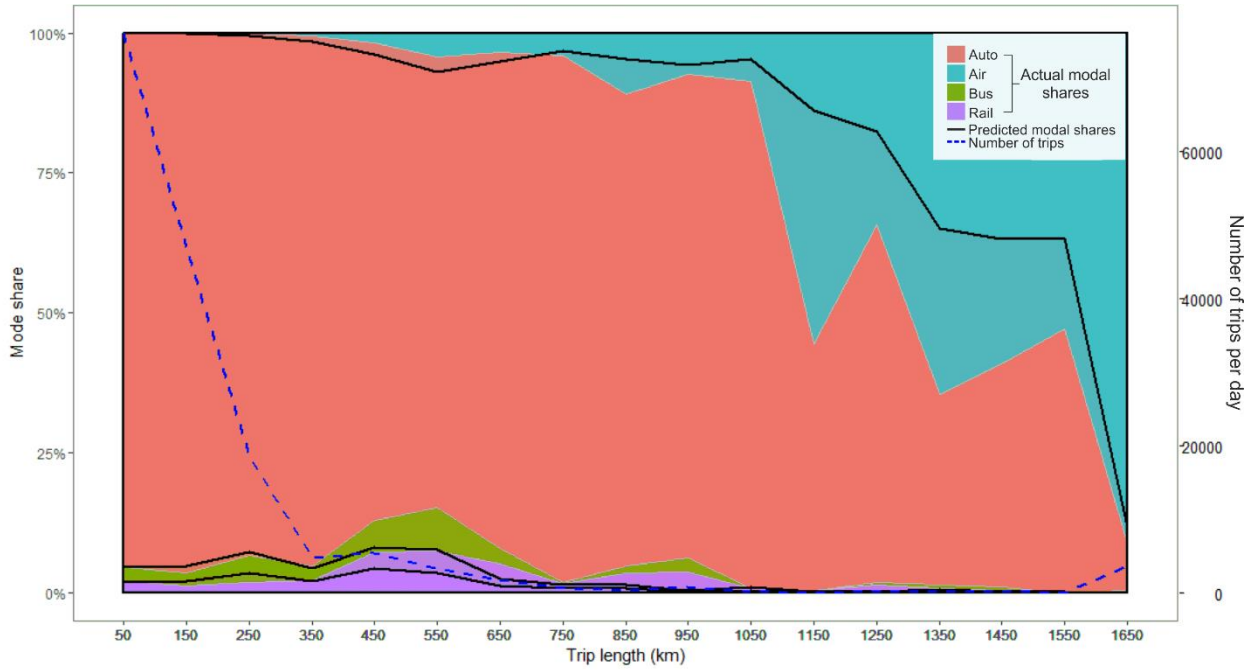


Figure 29. Distribution of modal share by trip length for trip purpose visit



These graphs indicate that the predicted modal shares follow the general pattern of the observed modal shares, especially for shorter distances. The pattern is not so closely matched at longer distances because there are very few actual trips made at such distances. The vast majority of trips are short-distance auto trips, and the model captures that behavior.

Figure 30. Actual vs. predicted number of trips by trip length for trip purpose business

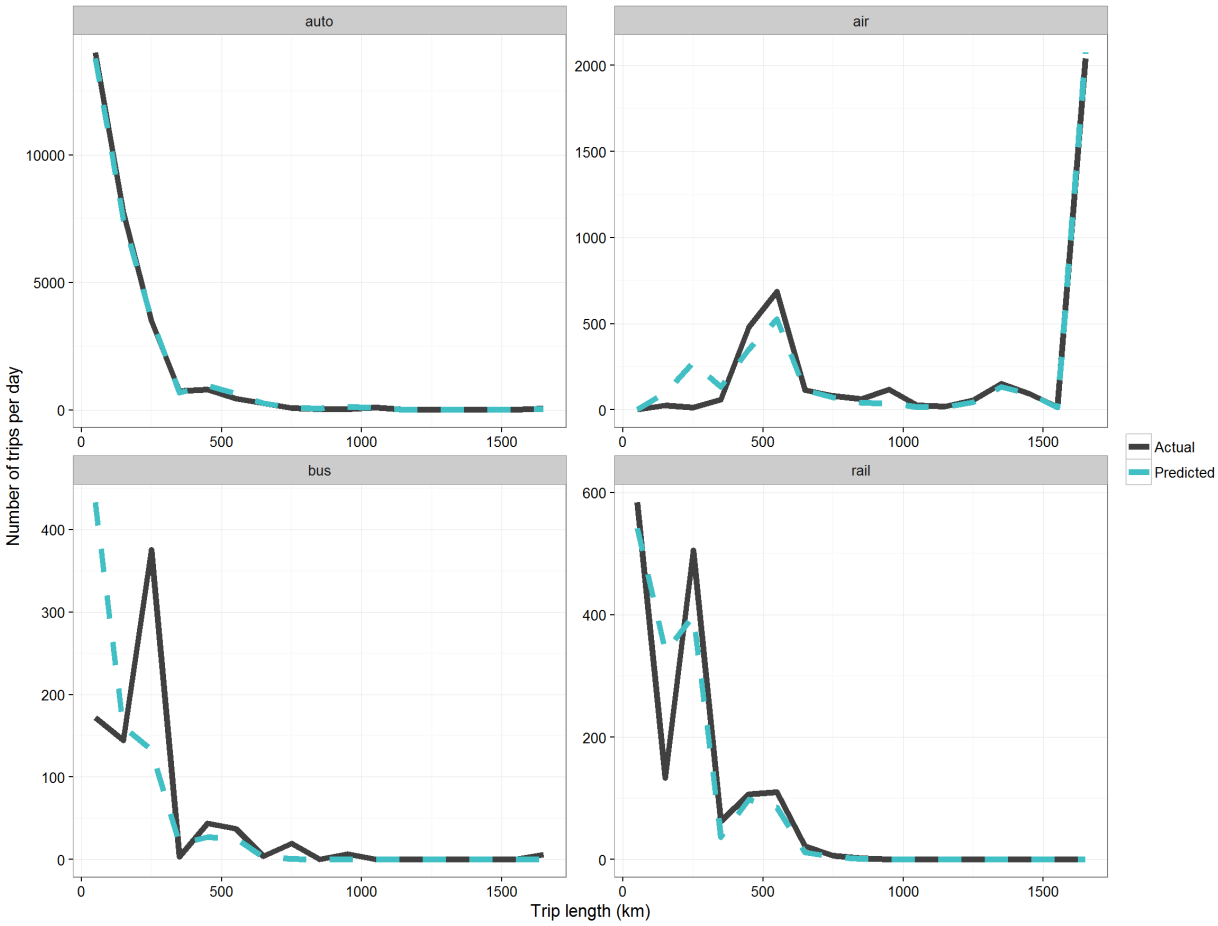


Figure 31. Actual vs. predicted number of trips by trip length for trip purpose leisure

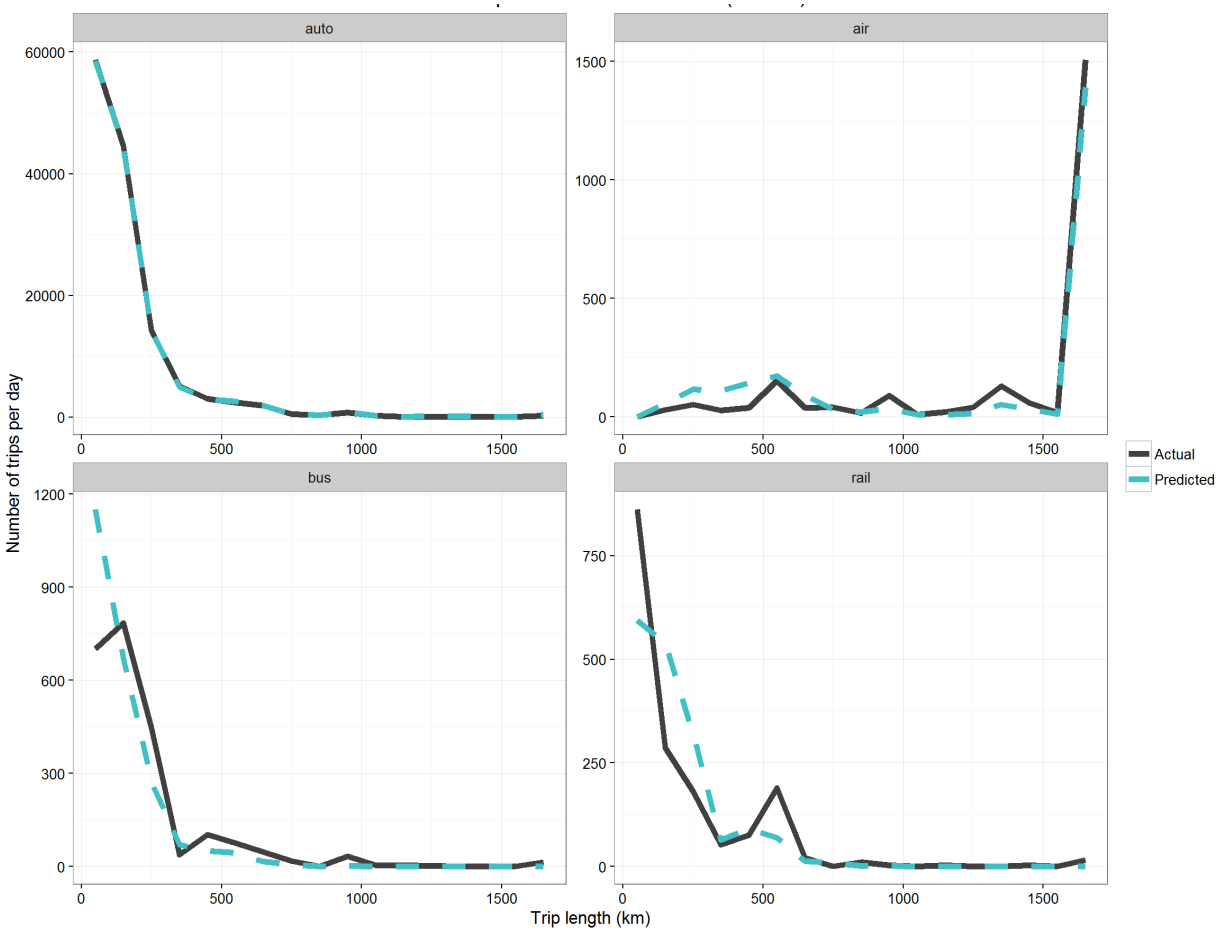


Figure 32. Actual vs. predicted number of trips by trip length for trip purpose visit

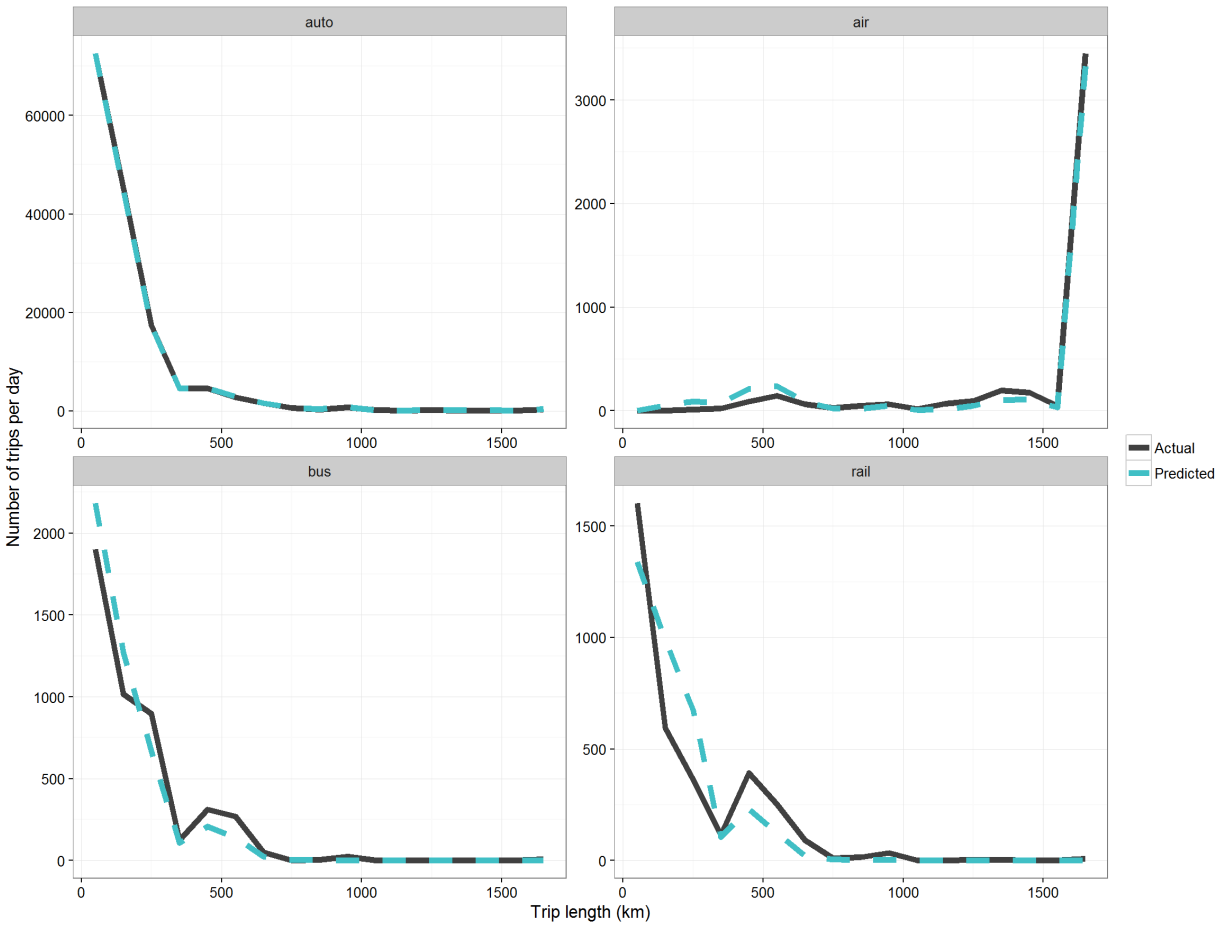
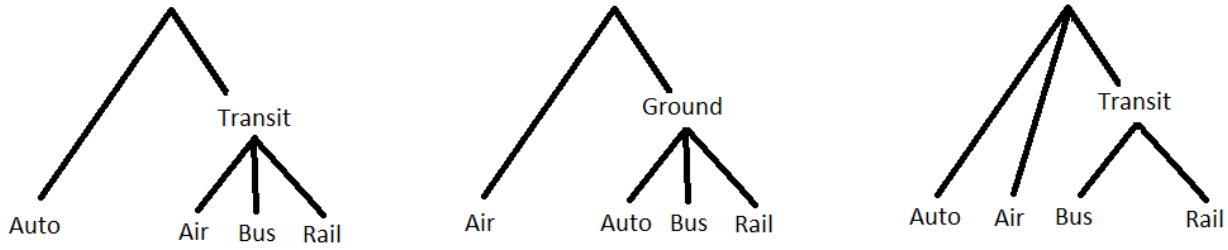


Figure 30, Figure 31, and Figure 32 show the number of actual versus predicted trips using Model 2. Auto trips are very well matched for all three trip purposes. The other three modes are not as exact, but the general trends are picked up by the model. The peak at around 500 km indicates trips between Toronto to Ottawa and Montreal. Some of the transit modal shares of these trips are underestimated. However, it is worth noting that the general bump in trips is shown in the predicted trips.

5.7. Nested logit

For nested logit, the same explanatory variables and choice set was used. Three specific nest combinations were tested:

Figure 33. Nesting structures tested



Each nest was tested for Models 1 – 4 travel time and price combinations. Models with the log of each travel time and travel price combination were also tested. Estimations with nesting coefficients of greater than 1 were ruled out, as that implies the alternatives correlated more with alternatives outside of the nest than within (Koppelman & Bhat, 2006, p. 163).

Many combinations yielded unreasonable nesting coefficients or could not be estimated at all using the *mlogit* package. The only nesting structure and variable combination that yielded reasonable nesting coefficients across all three trip purposes was auto vs. transit nest with the log of travel cost and the log of travel time as variables. This is shown in Table 14.

The variables used were similar to those discussed in the previous section. Though the coefficients here were technically under 1, they were very close to 1, implying an almost flat nesting structure and no great correlation between the alternatives within the nest. In addition, the log-likelihoods for these models were worse for the visit and leisure purposes compared to Model 2, though the business model had a slight improvement. Overall, the MNL model is preferable to the estimated NL.

Table 14. Nested logit model with log of travel time and log of travel cost

		Business		Leisure		Visit	
Mode	Variable	Coefficen	Significan	Coef.	Sig.	Coef.	Sig.
Air	Intercept	-1.69366	***	-4.92306	***	-3.23927	***
Bus	Intercept	-2.87557	***	-4.45811	***	-1.97862	***
Rail	Intercept	-3.18362	***	-4.56428	***	-1.16263	***
	Frequency	0.002438	***	0.002964	***	0.002718	***
	log(travel time)	-1.99561	***	-1.26407	***	-1.52436	***
	log(travel cost)	-1.33964	***	-1.02672	***	-1.22479	***
Air	Intermetro			0.512005	**		
Bus	intermetro	1.165862	***	1.142403	***	1.90861	***
Rail	Intermetro	2.230063	***	1.778067	***	1.033737	***
Air	Overnight	1.216991	***	2.19534	***	2.812266	***
Bus	Overnight	0.382717	**	0.430963	***	1.034165	***
Rail	Overnight	0.208304	*	0.499401	***	0.433203	***
Air	Group size			-0.24873	***	-0.35254	***
Bus	Group size	-0.26815	***	-0.53389	***	-1.25256	***
Rail	Group size	-0.53332	***	-0.69437	***	-1.04904	***
Air	Young (<25)	-1.46499	*				
Bus	Young (<25)	0.938794	***	1.380773	***	1.692177	***
Rail	Young (<25)			1.647043	***	1.639282	***
Air	Male	-0.73333	***	-0.33965	.	-0.44542	***
Bus	Male	-0.93988	***	-0.30204	***	-0.44157	***
Rail	Male	-1.27379	***				
Air	High income	-0.54501	.			0.267279	*
Bus	High income					-1.06275	***
Rail	High income					-1.08642	***
Air	Low income	0.666573	***	1.374713	***		
Bus	Low income			0.696847	***		
	Nesting Coef.	0.9005	***	0.868582	***	0.972912	***
	Log-likelihood	-2013.4		-3987.3		-6092.2	
	McFadden's R^2	0.49		0.34		0.46	

Unfortunately R's *mlogit* package does not allow any direct setting of the nesting coefficient. There is other software, such as Biogeme, that allows greater control over the estimation of nesting coefficients, but due to time constraints, this was not attempted.

6. Remaining work

The model must now be calibrated using the TSRC data. This would involve applying the model to destination choice model outputs and then adjusting the mode-specific constants to better match the TSRC mode choice pattern.

The scope of the project is building a long-distance travel demand model of Ontario. This model assigns modes to domestic trips within Canada. Now remains the trips between Ontario and the rest of the world. Since Canada only shares a land border with the U.S., the majority of long-distance trips into and out of Ontario are from and to America. The other international trips can safely be assumed to be by plane. This part of the model can be derived from the Canadian International Travel Survey and the modal data from Rome2rio. A similar methodology for estimation could be applied since the travel characteristics were gathered for all zones in North America excluding Mexico.

Finally, after the modes have been assigned to each trip, the trips must be disaggregated into TAZs and integrated into the short-distance portion of the Ontario province model for route assignment.

7. Discussion and conclusion

This section gives a final summary of the model development process before diving into a discussion of model results. I then discuss the limitations, give suggestions for future research and give final conclusions.

7.1. Summary

This thesis documents the development of a multinomial logit long-distance mode choice model for the province of Ontario, Canada. The model is based primarily on the TSRC trip data and an open web-based aggregate data source.

To achieve this, the spatial resolution was defined using the TSRC data. The model was defined to only account for domestic trips with at least one trip end in Ontario, and trips that geographically cross Ontario. It was segmented by trip purpose into business, leisure and visit. I then analyzed the to explore the relationship between mode choice and possible explanatory variables.

The lack of mode-specific data was addressed by utilizing the API of the multi-modal global trip planner Rome2rio. I defined the origin and destination trip positions as the centroids of zones and queried all possible mode choice data for all OD pairs.

The variables considered were LOS variables, characteristics of the trip-maker and of the trip. The parameters were chosen based on theoretical consistency and statistical relevance (at least 95% interval). I tested many LOS variables such as in-vehicle time, transfer time, total travel time until an optimal configuration of generic coefficients across all modes of total travel time, travel price and frequency was reached. Other non-LOS variables, such as income bracket, gender, age, education level, travel group size, and whether a trip has metropolitan OD, were then added gradually.

To achieve a theoretically consistent value-of-time, a generalized time variable was used. It was calculated by converting travel cost into travel time using a specific value-of-time. The model was re-estimated with this GT variable instead of separate travel time and travel cost variables. Three additional models with different configurations of GT were estimated for each trip purpose. To determine the right configuration of value-of-time, sensitivity tests were conducted on the four models. All four models showed the right signs in their sensitivities, but it was difficult to gauge which model demonstrated the most correct sensitivity. In the end, the

model with one VOT for all four modes was chosen, as it had the least amount of built-in assumptions and was the most parsimonious of the VOT models.

A nested logit model was attempted, but discarded due to the nesting coefficient not being significantly different from 1.

7.2. Discussion of results

The model considers LOS variables of frequency of service, travel time, and travel cost, which were combined into one generalized time variable. The generalized time coefficient is negative for all trip purposes, meaning that the perceived utility of modes go down as travel time and travel cost rises. The positive frequency parameter demonstrates as frequency increases, so does utility.

Explanatory socioeconomic factors in the model are age, income, gender, and for the business model only, education level. The youngest age group tends to travel less by air and more by bus and rail. Men have a propensity for driving. Higher income makes the traveler more likely to fly and less likely to take the bus and train. Having a higher education level makes transit modes more attractive to business travelers.

The relevant trip characteristics are whether the trip is made between two metropolitan areas, whether the trip is completed on the same day or has at least one overnight stay, and the travel group size. The results show that trips made between two metropolitan areas are more likely to be made with transit options. Trips longer than one day also favor transit options, especially air. As group size goes up, the traveler tends to prefer auto to transit.

The model closely matches the overall modal share. Predicted and actual modal share graphs show that the model correctly predicts the mode of the majority of trips. The modal share of relatively short distance trips and relatively long distance trips are very closely matched.

The model underestimates the air modal share between certain mid-distance city pairs such as Toronto -Montreal and Toronto – Ottawa, especially for the business purpose. In his corridor city of Toronto-Montreal in 1997, Bhat utilized so-called large city indicators to identify a trip originating or terminating in a large city, which is similar to the intermetro variable employed in this model (1997). The California Statewide Model for High Speed Rail accounted for this with regional dummy variables for large metropolitan regions and calibrated to match the observed trips (Outwater et al., 2009). Since the scale of this model encompasses all of Ontario and trips to and from Ontario in all of Canada, it can be argued that the model cannot have the fineness

of a corridor study without incorporating specific dummy variables and geographic segmentation, as in the California model.

7.3. Limitations and suggestions for future research

Many studies have attested to the superiority of NL models over ones using MNL (Koppelman and Wen, 2000; Bhat, 1997) However, this study was restricted by the software employed and lack of time to explore more malleable software options such as Biogeme. Future research should explore the NL model more thoroughly, preferably by using software that allows direct manipulation of the nesting coefficient.

Though Rome2rio is capable of detailed door-to-door, mode-specific data for all possible modes from and to any specific geographic location, it is limited here by the spatial resolution of the TSRC data. Since there are only 69 distinct zones in Ontario for the origin and destination of the trips, any additional refinement in the Rome2rio data went unused. It can readily accommodate data with a finer resolution and more specific trip start and end points. Future research with finer resolution, perhaps coupled with crowd-based network data sources, could take full advantage of this capability.

Another shortcoming is the lack of access and egress modes modeled. Miller and the study conducted by Habib and Wong demonstrated the importance of modeling access and egress modes (2004, 2015). Habib and Wong could construct such a model because they had a targeted survey for the express purpose of building a model for that corridor (2015). The mode-choice model here assumes simple single mode trips, since that is the data available from TSRC. Due to lack of information on access and egress, this model did not account for access and egress choice. Rome2rio is capable of building intermodal routes with alternative modal suggestions for some parts of a route. With more detailed data, Rome2rio could presumably be used to model intermodal mobility. Though this thesis attempted to use the provided complete route, including access and egress possibilities, this may be unreasonable due to the coarseness of zones. The access and egress routes chosen by Rome2rio may not be reflective of the real access and egress modes chosen by the trip maker.

Some routes were not found on Rome2rio but reported in the survey. This could be due to the assumption that trips start and end at zone centroids. The centroids, though weighted by population, could still be in very remote locations, leading to incorrect travel times, costs or perhaps no possible route suggestions. Another possibility is updated transport networks – routes that used to operate may have changed since the time of the survey. It is important to

bear in mind that Rome2rio relies on a multitude of sources, both online and offline, for its data. It could have imperfect coverage of certain areas. For example, some less official bus services, such as the so-called Chinatown buses, and chartered planes, etc., may not be accounted for.

Platforms like Rome2rio show promising potential. Since this travel platform and others like it have global coverage and APIs, relevant modal data could be gathered very quickly once the process to query and integrate data is automated. Traditionally, such data were either estimated, manually compiled, or at best partly automated via methods specifically targeted to different websites and data structures. If a similar long-distance model were to be built for another region, the same procedure of building a list of URLs out of all origin and destination points could be used to pull all necessary multi-modal data. This could alleviate the need to rely on an actual multi-modal network model that may not yet exist and take significant resources to build.

The advantage to using Rome2rio is that it compiles all possible modes into one database and offers the ability of a free-to-try API key. Platforms such as Google, Qixxit, waymate, likely have similar capabilities. Future research should further explore the use of such databases in modeling, as it would significantly reduce the time and effort spent on data collection, and potentially increase the accuracy of the model.

7.4. Conclusions

Travel demand models, especially long-distance models, are often plagued by a lack of quality data. This thesis addresses an aspect of the data scarcity by utilizing a multimodal online trip planner platform to gather necessary mode-specific data. The API for Rome2rio allowed me to gather all mode-specific data for all modes within days, saving the effort required to gather data from other disaggregated sources. It avoids the difficulty of acquiring proprietary data from private enterprises, saves the man hours required to manually gather data from different specific information sources, and cuts out the effort and assumptions that would go into building a model of the network and extracting the skims. The data was then combined with the TSRC survey with data on trip and person characteristics to estimate an MNL mode-choice model for Ontario, Canada. The resulting model has a number of useful parameters and, as part of the Ontario provincial model, can help assess policy impacts. This thesis demonstrates the possibility and the viability of combining new data resources with traditional surveys in the estimation of a long-distance mode choice model.

List of References

Aaron, A., & Sean, B. J. (2012). *THE MANY USES OF GTFS DATA – OPENING THE DOOR TO TRANSIT AND MULTIMODAL APPLICATIONS* (Publication). ITS America's 23rd Annual Meeting & Exposition. doi:10.1.1.391.5421

Abdelwahab, W. M. (1991). Transferability of intercity disaggregate mode choice models in Canada. *Canadian Journal of Civil Engineering,18*(1), 20-26. doi:10.1139/l91-003

About Rome2rio. (n.d.). Retrieved June 01, 2017, from https://www.rome2rio.com/about/

Alliance Transportation Group (2015). AHTD National County-Level Long Distance Travel Model: Model Development and Validation Report. *15th Transportation Research Board National Transportation Planning Applications Conference*

Behrens, C., & Pels, E. A. (n.d.). Intermodal Competition in the London-Paris Passenger Market: High-Speed Rail and Air Transport. *SSRN Electronic Journal*. doi:10.2139/ssrn.1416663

Bhat, C. R. (1995). A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B: Methodological,29*(6), 471-483. doi:10.1016/0191-2615(95)00015-6

Bhat, C. R. (1997). Covariance heterogeneity in nested logit models: Econometric structure and application to intercity travel. *Transportation Research Part B: Methodological,31*(1), 11-21. doi:10.1016/s0191-2615(96)00018-5

Canada: Light-duty: Fuel Consumption and GHG. (n.d.). Retrieved December 19, 2016, from http://www.transportpolicy.net/index.php?title=Canada%3A_Light-duty%3A_Fuel_Consumption_and_GHG

Cho, H. D. (2013). *The factors that affect long distance travel mode choice decisions and their implications for transportation policy* (Unpublished master's thesis). UNIVERSITY OF FLORIDA.

Croissant, Y. (n.d.[2011]). *Estimation of multinomial logit models in R: The mlogit Packages*. Retrieved from: https://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf

Fuel Prices. (n.d.). Retrieved February 19, 2017, from http://www.energy.gov.on.ca/en/fuel-prices/

Government of Ontario, Ministry of Finance. (n.d.). Ontario Fact Sheet. Retrieved May 29, 2017, from http://www.fin.gov.on.ca/en/economy/ecupdates/factsheet.html

Hasan, Asad, Wang Zhiyu, and Alireza S Mahani (2014). Fast Estimation of Multinomial Logit Models: R Package mnlogit. *Journal of statistical software* 75.3.

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

"Transport Coverage". Rome2rio coverage. Retrieved June 01, 2017, from https://www.rome2rio.com/coverage

Matthies, E., Kuhn, S., & Klockner, C. A. (2002). Travel Mode Choice of Women: The Result of Limitation, Ecological Norm, or Weak Habit? *Environment and Behavior,34*(2), 163-177. doi:10.1177/0013916502034002001

Microdata User Guide for Travel Survey of Residents of Canada. (2014). Manuscript, Canada, Statistics Canada.

Miller, E. (2001). The Greater Toronto Area Travel Demand Modelling System Version 2.0. Volume I: Model Overview. Toronto: Department of Civil Engineering, University of Toronto, January

Miller, E. (2004). The Trouble with Intercity Travel Demand Models. *Transportation Research Record: Journal of the Transportation Research Board,1895*, 94-101. doi:10.3141/1895-13

Miskeen, M. A., Alhodairi, A. M., & Riza Atiq Abdullah Bin O.k. Rahmat. (2014). Modeling of Intercity Travel Mode Choice Behavior for Non-Business Trips within Libya. *Research Journal of Applied Sciences, Engineering and Technology,7*(3), 442-453. doi:10.19026/rjaset.7.274

Moeckel, R., Fussell, R., & Donnelly, R. (2014). Mode choice modeling for long-distance travel. *Transportation Letters,7*(1), 35-46. doi:10.1179/1942787514y.0000000031

Ortuzar, J. de D. and L.G. Willumsen (1994). *Modelling Transport*, New York: John Wiley and Sons.

Outwater, M., Tierney, K., Bradley, M., Sall, E., Kuppam, A., & Modugula, V. (2010). California Statewide Model for High-Speed Rail. *Journal of Choice Modelling,3*(1), 58-83. doi:10.1016/s1755-5345(13)70029-0

Rich, J., & Mabit, S. L. (n.d.). A long-distance travel demand model for Europe. *European Journal of Transport and Infrastructure Research,12*(1), 1-20. Retrieved from http://orbit.dtu.dk/en/publications/a-longdistance-travel-demand-model-for-europe(0c27739c-1c67-4c3c-a5d9-4c25faf8b78e).html

Ryan, K., David, O., Henry, S., & Robert, P. (2012). *Integrated Intermodal Passenger Transportation System* (Tech.). Hanover, MD: NASA Center for Aerospace Information.

Schiffer, R. G. (2012). *NCHRP Report 735: Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models* (Rep. No. 735). Washington D.C.: Transportation Research Board of the National Academies.

Search API. (n.d.). Retrieved March 01, 2017, from https://www.rome2rio.com/documentation/1-4/search/

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies,58*, 162-177. doi:10.1016/j.trc.2015.04.022

Travel Survey of Residents of Canada 2014 Public Use Microdata File - Trip. (n.d.). Manuscript, Statistics Canada, Canada.

Weiner, E. (1999). *Urban transportation planning in the United States: an historical overview*. Washington, D.C.: U.S. Dept. of Transportation.

Wen, C., & Koppelman, F. S. (1998). The generalized nested logit model. *Transportation Research Part B: Methodological,35*(7), 627-641. doi:10.1016/s0191-2615(00)00045-x

Wilson, F. R., Damodaran, S., & Innes, J. D. (1990). Disaggregate mode choice models for intercity passenger travel in Canada. *Canadian Journal of Civil Engineering,17*(2), 184-191. doi:10.1139/l90-023

Zhang, L., Southworth, F., Xiong, C., & Sonnenberg, A. (2012). Methodological Options and Data Sources for the Development of Long-Distance Passenger Travel Demand Models: A Comprehensive Review. *Transport Reviews,32*(4), 399-433. doi:10.1080/01441647.2012.688174

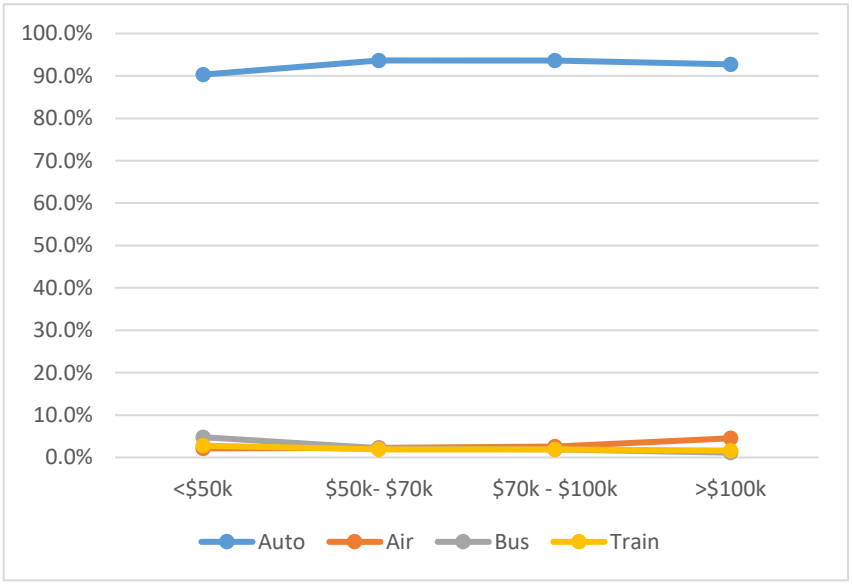
List of Abbreviations

API	Application programming interface
CD	Census division
CMA	Census metroplitan area
FSM	Four-step model
GC	General cost
GT	General time
GTFS	General Transit Feed Specification
IIA	Independence of irrelevant alternatives
IVTT	In-vehicle travel time
LFC	Labour Force Survey
LOS	Level-of-service
MNL	Multinomial logit
MTO	Ministry of Transportation of Ontario
NHTS	National Household Travel Survey
NL	Nested logit
OD	Origin-destination
OVTT	Out-of-vehicle travel time
TAZ	Travel analysis zones
TSRC	Travel Survey of Residents of Canada
VOT	Value-of-time

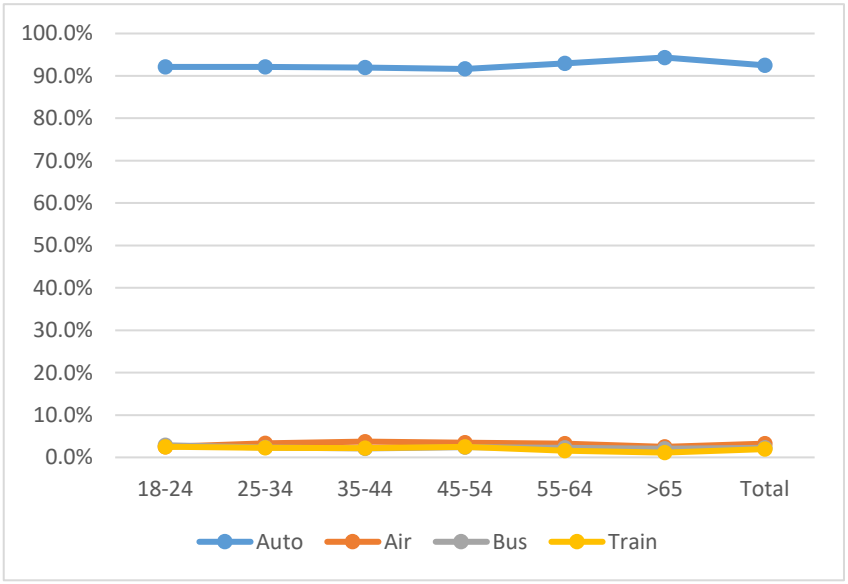
List of Figures

List of Tables

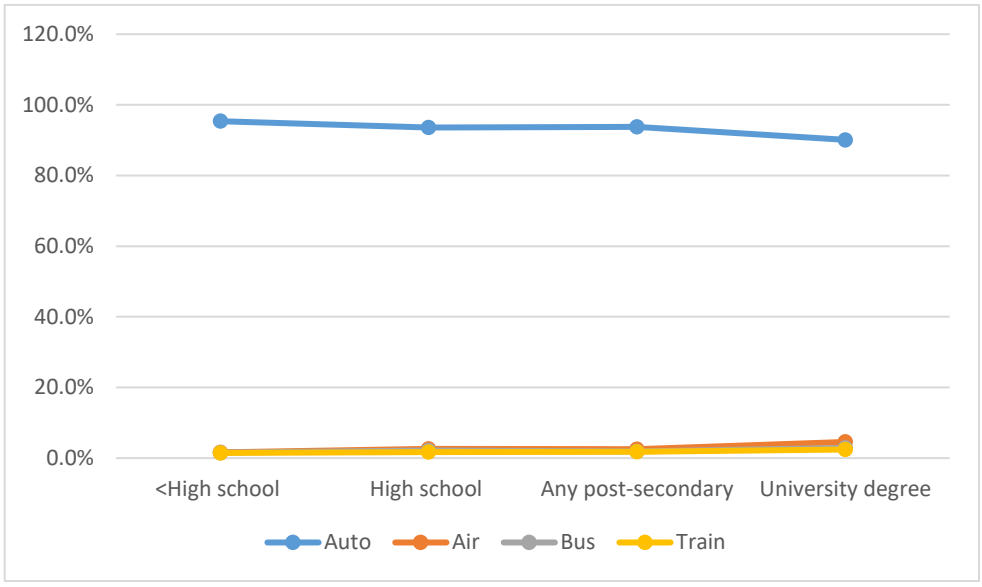
Appendix A: Further data analysis



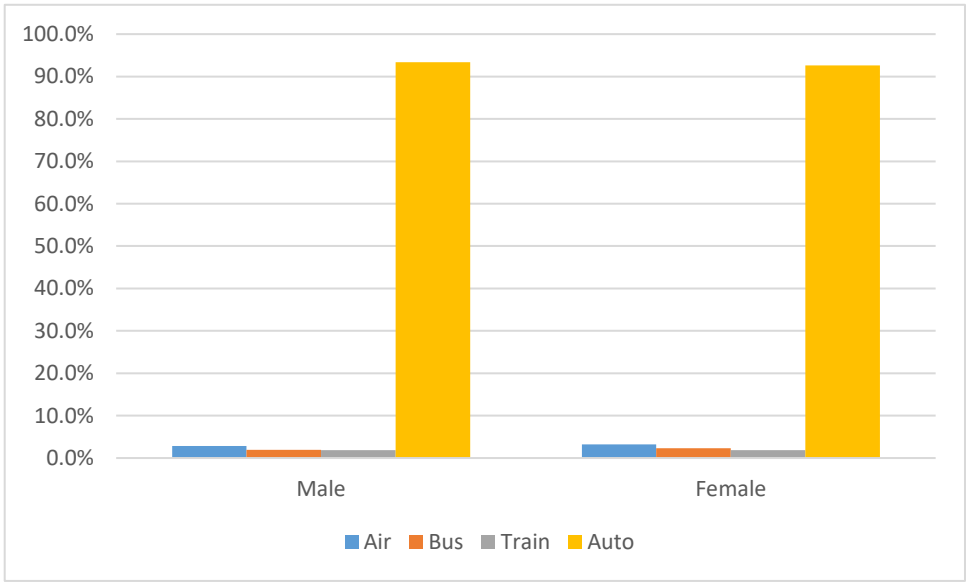
Income bracket vs. mode share



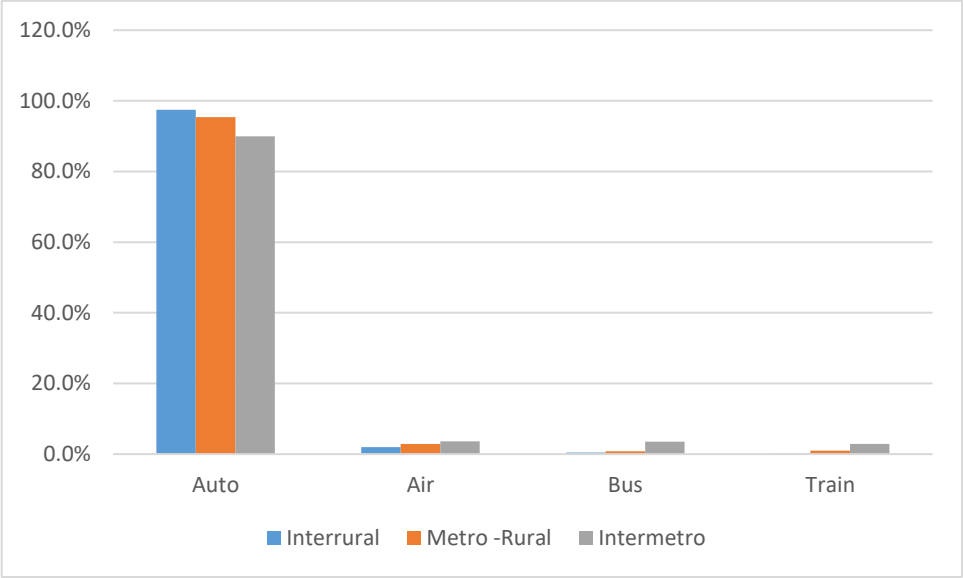
Age bracket vs. mode share



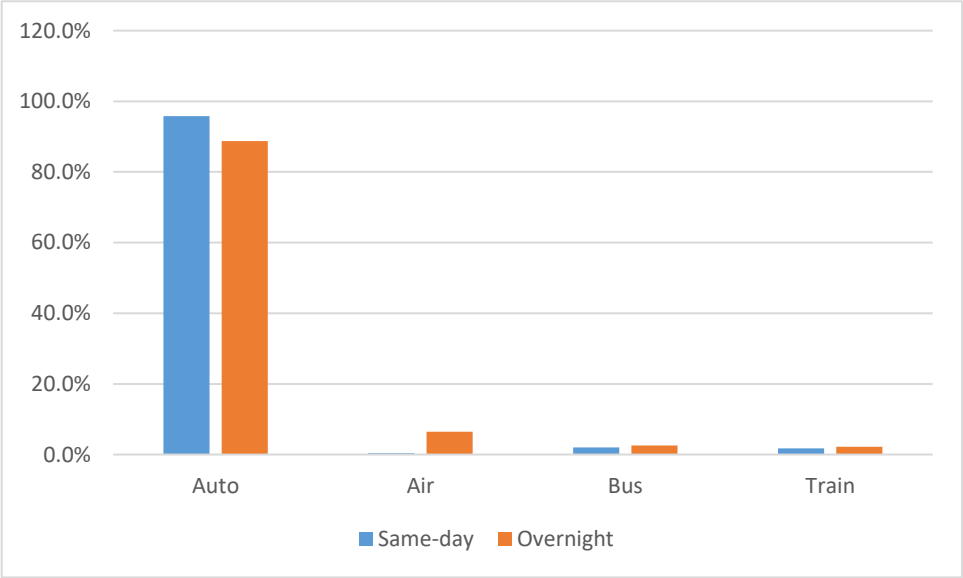
Education level vs. mode share



Gender vs. mode share



Origin and destination in metro or rural area vs. mode share



Whether a trip is made the same-day or has at least one night spent outside vs. mode share

Appendix B: Multinomial logit model call to *mlogit* for Model 2

Business

```
> B <- mlogit(bf,brun, weights = brun$wttp)
> summary(B)

Call:
mlogit(formula = bf, data = brun, weights = brun$wttp, method = "nr",
    print.level = 0)

Frequencies of alternatives:
      1       2       4       5
0.651294 0.280856 0.021068 0.046782

nr method
49 iterations, 0h:0m:2s
g'(-H)^-1g = 9.54E-07
gradient close to zero

Coefficients :
                                          Estimate  Std. Error  t-value  Pr(>|t|)
2:(intercept)                           -1.48627019  0.22965029  -6.4719 9.679e-11 ***
4:(intercept)                           -3.69359226  0.35196454 -10.4942 < 2.2e-16 ***
5:(intercept)                           -4.39053779  0.29592230 -14.8368 < 2.2e-16 ***
I((( == "2") | (alt == "4") | (alt == "5")) * mmfreq)  0.00282164  0.00042771   6.5970 4.195e-11
***
impedence_op2                            -0.00492942  0.00020656 -23.8643 < 2.2e-16 ***
I(( == "2") * (age1))                    -1.90308069  0.51758421  -3.6769 0.0002361 ***
I(( == "4") * (age1))                     1.04823547  0.31658764   3.3110 0.0009295 ***
I(( == "2") * inc1)                      -1.24280895  0.22449778  -5.5360 3.095e-08 ***
I(( == "4") * inc1)                       0.69402520  0.22186846   3.1281 0.0017594 **
I(( == "2") * tp_d01)                    -0.26629578  0.05039677  -5.2840 1.264e-07 ***
I(( == "4") * tp_d01)                    -0.44332138  0.13093985  -3.3857 0.0007100 ***
2:intermetro                              0.38274096  0.12592711   3.0394 0.0023706 **
4:intermetro                              0.47179440  0.23669703   1.9932 0.0462350 *
5:intermetro                              1.67272560  0.25585412   6.5378 6.243e-11 ***
2:triptype                                1.16023208  0.18137890   6.3967 1.587e-10 ***
4:triptype                                1.09719469  0.19812617   5.5379 3.062e-08 ***
5:triptype                                0.85533444  0.14343457   5.9632 2.473e-09 ***
2:sex                                    -0.52218018  0.11471618  -4.5519 5.316e-06 ***
```

```
4:sex                                -0.61757413  0.18899590  -3.2677 0.0010844 **
5:sex                                -0.89436059  0.13261836  -6.7439 1.542e-11 ***
2:edu4                                0.65799863  0.11866946   5.5448 2.943e-08 ***
4:edu4                                0.86954461  0.20987524   4.1432 3.426e-05 ***
5:edu4                                0.83134027  0.14420355   5.7650 8.164e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -2077.3
McFadden R^2:  0.59954
Likelihood ratio test : chisq = 6219.8 (p.value = < 2.22e-16)
```

Visit

```
> V <- mlogit(vf,vrun, weights = vrun$wttp)
> summary(V)

Call:
mlogit(formula = vf, data = vrun, weights = vrun$wttp, method = "nr",
    print.level = 0)

Frequencies of alternatives:
      1       2       4       5
0.867692 0.085780 0.026997 0.019531

nr method
51 iterations, 0h:0m:11s
g'(-H)^-1g = 4.86E-07
gradient close to zero

Coefficients :
                                      Estimate  Std. Error  t-value  Pr(>|t|)
2:(intercept)                        -4.1792e+00  6.2781e-01  -6.6568 2.799e-11 ***
4:(intercept)                        -3.3034e+00  1.6746e-01 -19.7262 < 2.2e-16 ***
5:(intercept)                        -2.9024e+00  1.6052e-01 -18.0817 < 2.2e-16 ***
I((( == "2") | (alt == "4") | (alt == "5")) * mmfreq)  2.7695e-03  2.5197e-04  10.9913 < 2.2e-16
***
impedence_op2                           -4.2473e-03  9.5934e-05 -44.2731 < 2.2e-16 ***
I(( == "2") * (sex))                    -6.9495e-01  9.0618e-02  -7.6690 1.732e-14 ***
I(( == "4") * (sex))                    -4.5977e-01  7.7182e-02  -5.9569 2.570e-09 ***
I(( == "4") * (age1))                    1.6015e+00  8.3399e-02  19.2025 < 2.2e-16 ***
```

```
I(( == "5") * (age1))                    1.5820e+00  9.3722e-02  16.8793 < 2.2e-16 ***
I(( == "4") * (intermetro))              1.7229e+00  1.1879e-01  14.5034 < 2.2e-16 ***
I(( == "5") * (intermetro))              6.9123e-01  1.0007e-01   6.9076 4.928e-12 ***
2:triptype                               3.5092e+00  6.2433e-01   5.6208 1.901e-08 ***
4:triptype                               1.5439e+00  9.5361e-02  16.1902 < 2.2e-16 ***
5:triptype                               9.9801e-01  1.0067e-01   9.9136 < 2.2e-16 ***
2:tp_d01                                -4.7254e-01  4.1858e-02 -11.2890 < 2.2e-16 ***
4:tp_d01                                -1.1634e+00  6.3263e-02 -18.3894 < 2.2e-16 ***
5:tp_d01                                -9.2223e-01  6.3409e-02 -14.5442 < 2.2e-16 ***
2:inc4                                   5.2600e-01  8.9575e-02   5.8722 4.301e-09 ***
4:inc4                                  -9.0202e-01  1.1177e-01  -8.0704 6.661e-16 ***
5:inc4                                  -8.9541e-01  1.1997e-01  -7.4635 8.415e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -6092.2
McFadden R^2:  0.62236
Likelihood ratio test : chisq = 20080 (p.value = < 2.22e-16)
```

Leisure

```
Call:
mlogit(formula = lf, data = lrun, weights = lrun$wttp, method = "nr",
    print.level = 0)

Frequencies of alternatives:
      1       2       4       5
0.911542 0.056284 0.017360 0.014814

nr method
47 iterations, 0h:0m:10s
g'(-H)^-1g = 3.49E-07
gradient close to zero

Coefficients :
                                        Estimate  Std. Error  t-value  Pr(>|t|)
2:(intercept)                           -3.8029e+00  3.3597e-01 -11.3193 < 2.2e-16 ***
4:(intercept)                           -3.4143e+00  1.6485e-01 -20.7114 < 2.2e-16 ***
5:(intercept)                           -4.2292e+00  2.0823e-01 -20.3101 < 2.2e-16 ***
I((( == "2") | (alt == "4") | (alt == "5")) * mmfreq)  2.2758e-03  3.1654e-04   7.1895 6.504e-13
***
```

```
I(( == "2") * (sex))                -3.8746e-01  9.5675e-02  -4.0497 5.128e-05 ***
I(( == "4") * (sex))                -3.4750e-01  1.0361e-01  -3.3538 0.0007971 ***
impedence_op2                          -2.7695e-03  7.5949e-05 -36.4654 < 2.2e-16 ***
2:intermetro                         7.1408e-01  9.5866e-02   7.4487 9.415e-14 ***
4:intermetro                         8.6598e-01  1.1459e-01   7.5575 4.108e-14 ***
5:intermetro                         1.3069e+00  1.4592e-01   8.9563 < 2.2e-16 ***
2:triptype                          1.9594e+00  3.2166e-01   6.0916 1.118e-09 ***
4:triptype                          8.9448e-01  1.0536e-01   8.4895 < 2.2e-16 ***
5:triptype                          1.1964e+00  1.2760e-01   9.3756 < 2.2e-16 ***
2:tp_d01                            -1.6931e-01  3.6838e-02  -4.5960 4.306e-06 ***
4:tp_d01                            -3.7396e-01  4.8715e-02  -7.6765 1.643e-14 ***
5:tp_d01                            -5.4065e-01  5.8756e-02  -9.2016 < 2.2e-16 ***
2:age1                             -4.2832e-01  2.0459e-01  -2.0936 0.0362953 *
4:age1                              1.0870e+00  1.3369e-01   8.1305 4.441e-16 ***
5:age1                              1.3471e+00  1.3792e-01   9.7676 < 2.2e-16 ***
2:inc4                             1.9520e-01  9.5738e-02   2.0389 0.0414647 *
4:inc4                            -1.3484e+00  1.4902e-01  -9.0485 < 2.2e-16 ***
5:inc4                            -3.7241e-01  1.2455e-01  -2.9902 0.0027884 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -3923.9
McFadden R^2:  0.60075
Likelihood ratio test : chisq = 11808 (p.value = < 2.22e-16)
```

Appendix C: Nested logit call to *mlogit*

Business

```
Call:
mlogit(formula = bf, data = brun, weights = brun$wttp, reflevel = "1",
    nests = list(auto = "1", transit = c("2", "4", "5")), unscaled = TRUE,
    constPar = c("iv.auto"))

Frequencies of alternatives:
      1       2       4       5
0.651294 0.280856 0.021068 0.046782

bfgs method
25 iterations, 0h:0m:3s
g'(-H)^-1g = 4.96E-07
gradient close to zero

Coefficients :
                                                    Estimate    Std. Error  t-value Pr(>|t|)
2:(intercept)                                       -1.69366411 0.38760461  -4.3696  1.25E-05 ***
4:(intercept)                                       -2.87556835 0.38225152  -7.5227  5.37E-14 ***
5:(intercept)                                       -3.18362176 0.47995532  -6.6332  3.29E-11 ***
I((( == "2") | (alt == "4") | (alt == "5")) * mmfreq) 0.00243839 0.00030485  7.9986 1.33E-15 ***
I(log(mmprice))                                     -1.33964195 0.06063587 -22.0932< 2.2e-16 ***
I(log(tot_time))                                    -1.99561171 0.10355498 -19.271  < 2.2e-16 ***
I(( == "2") * tp_d01)                               -0.26814861 0.05685865  -4.7161  2.41E-06 ***
I(( == "4") * tp_d01)                               -0.53332322 0.0833556   -6.3982  1.57E-10 ***
I(( == "2") * inc1)                                 -0.54500818 0.29888498  -1.8235  0.068232 .
I(( == "2") * inc4)                                 0.66657311  0.1417679   4.7019 2.58E-06 ***
```

I((== "2") * (age1))	-1.46499394	0.64469323	-2.2724	0.023063	*
I((== "4") * (age1))	0.93879389	0.14965185	6.2732	3.54E-10	***
I((== "2") * (ruralrural))	-1.29969774	0.89690398	-1.4491	0.147311	
I((== "4") * (intermetro))	1.16586168	0.17556687	6.6406	3.13E-11	***
I((== "5") * (intermetro))	2.23006268	0.36545705	6.1021	1.05E-09	***
2:triptype	1.21699056	0.13175448	9.2368	< 2.2e-16	***
4:triptype	0.38271651	0.14218814	2.6916	0.007111	**
5:triptype	0.20830406	0.08845298	2.355	0.018524	*
2:sex	-0.73333458	0.13360087	-5.489	4.04E-08	***
4:sex	-0.93988235	0.12448047	-7.5504	4.33E-14	***
5:sex	-1.27378964	0.0867736	-14.6795	< 2.2e-16	***
iv.transit	0.90049422	0.0172857	52.0947	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -2013.4
McFadden R^2: 0.61185
Likelihood ratio test : chisq = 6347.6 (p.value = < 2.22e-16)

Leisure

```
>
> Lnest <- mlogit(lf,lrun, reflevel="1", weights = lrun$wttp,  nests=list(auto="1",transit=c("2","4","5")), unscaled = TRUE, constPar = c("iv.auto"))

> summary(Lnest)

Call:
mlogit(formula = lf, data = lrun, weights = lrun$wttp, reflevel = "1",
    nests = list(auto = "1", transit = c("2", "4", "5")), unscaled = TRUE,
    constPar = c("iv.auto"))

Frequencies of alternatives:
      1       2       4       5
```

0.911542 0.056284 0.017360 0.014814

bfgs method
25 iterations, 0h:0m:25s
g'(-H)^-1g = 5.46E-07
gradient close to zero

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)	
2:(intercept)	-4.92305648	0.4501625	-10.9362	< 2.2e-16	***
4:(intercept)	-4.45810759	0.34349182	-12.9788	< 2.2e-16	***
5:(intercept)	-4.56427747	0.36682681	-12.4426	< 2.2e-16	***
I(((== "2") | (alt == "4") | (alt == "5")) * mmfreq)	0.00296394	0.00023369	12.683	< 2.2e-16	***
I((== "4") * (age1))	1.38077287	0.0773644	17.8477	< 2.2e-16	***
I((== "5") * (age1))	1.64704327	0.0964155	17.0828	< 2.2e-16	***
I((== "4") * (inc1))	1.37471321	0.06739835	20.3968	< 2.2e-16	***
I((== "5") * (inc1))	0.69684736	0.09015406	7.7295	1.09E-14	***
I((== "2") * (sex))	-0.33964816	0.179033	-1.8971	0.057811	.
I((== "4") * (sex))	-0.30203881	0.0698978	-4.3211	1.55E-05	***
log(mmprice)	-1.02672249	0.04470382	-22.9672	< 2.2e-16	***
log(tot_time)	-1.26407268	0.08589163	-14.7171	< 2.2e-16	***
2:intermetro	0.51200504	0.19492062	2.6267	0.008621	**
4:intermetro	1.14240252	0.10288291	11.1039	< 2.2e-16	***
5:intermetro	1.77806699	0.11811622	15.0535	< 2.2e-16	***
2:tp_d01	-0.24873355	0.05900466	-4.2155	2.49E-05	***
4:tp_d01	-0.53389461	0.03097914	-17.234	< 2.2e-16	***
5:tp_d01	-0.6943679	0.04179013	-16.6156	< 2.2e-16	***
2:triptype	2.19533991	0.24928926	8.8064	< 2.2e-16	***
4:triptype	0.43096289	0.0841882	5.119	3.07E-07	***
5:triptype	0.49940127	0.08008667	6.2358	4.50E-10	***
iv.transit	0.8685816	0.02117763	41.0141	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -3987.3
McFadden R^2: 0.5943
Likelihood ratio test : chisq = 11682 (p.value = < 2.22e-16)

Visit

Call:
mlogit(formula = vf, data = vrun, weights = vrun$wttp, reflevel = "1",
    nests = list(auto = "1", transit = c("2", "4", "5")), unscaled = TRUE,
    constPar = c("iv.auto"))

Frequencies of alternatives:
      1   2   4   5
0.867692 0.085780 0.026997 0.019531

bfgs method
25 iterations, 0h:0m:15s
g'(-H)^-1g = 4.3E-07
gradient close to zero

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)	
2:(intercept)	-3.23926699	0.37545727	-8.6275	< 2.2e-16	***
4:(intercept)	-1.97861545	0.21487784	-9.2081	< 2.2e-16	***
5:(intercept)	-1.16263205	0.1999649	-5.8142	6.09E-09	***
I(((== "2") | (alt == "4") | (alt == "5")) * mmfreq)	0.00271844	0.00013758	19.7586	< 2.2e-16	***
log(mmprice)	-1.22478654	0.02476579	-49.4548	< 2.2e-16	***
log(tot_time)	-1.52436438	0.05541078	-27.5102	< 2.2e-16	***
I((== "2") * (sex))	-0.44541877	0.12140193	-3.669	0.0002435	***
I((== "4") * (sex))	-0.44157138	0.04190919	-10.5364	< 2.2e-16	***
I((== "4") * (intermetro))	1.90860989	0.11343231	16.826	< 2.2e-16	***
I((== "5") * (intermetro))	1.03373735	0.05444647	18.9863	< 2.2e-16	***

I((== "4") * (age1))	1.69217671	0.04460222	37.9393	< 2.2e-16	***
I((== "5") * (age1))	1.63928227	0.04993379	32.8291	< 2.2e-16	***
2:tp_d01	-0.35253985	0.04219986	-8.3541	< 2.2e-16	***
4:tp_d01	-1.25255539	0.03066874	-40.8414	< 2.2e-16	***
5:tp_d01	-1.04903882	0.02593269	-40.4524	< 2.2e-16	***
2:inc4	0.26727929	0.12867235	2.0772	0.0377823	*
4:inc4	-1.06274757	0.04635702	-22.9253	< 2.2e-16	***
5:inc4	-1.08641702	0.06284829	-17.2863	< 2.2e-16	***
2:triptype	2.8122663	0.32910207	8.5453	< 2.2e-16	***
4:triptype	1.03416495	0.04117335	25.1173	< 2.2e-16	***
5:triptype	0.43320331	0.04347866	9.9636	< 2.2e-16	***
iv.transit	0.97291217	0.0120332	80.8523	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -6506.3
McFadden R^2: 0.59669
Likelihood ratio test : chisq = 19252 (p.value = < 2.22e

Appendix D: Model comparison for business and leisure

		Model 1		Model 2		Model 3		Model 4	
		time & price		1VOT1Coef		1VOT4Coef		4VOT1COEF	
Mode	Parameter	Coef	Sig	Coef	Sig	Coef	Sig	Coef	Sig
Air	Intercept	-1.29338	***	-1.48627	***	-1.84871	***	-2.06995	***
Bus	Intercept	-3.54818	***	-3.69359	***	-4.39378	***	-3.62479	***
Rail	Intercept	-4.38082	***	-4.39054	***	-4.63684	***	-4.41808	***
	Frequency	0.00298	***	0.002822	***	0.002878	***	0.002983	***
	Travel cost	-0.00598	***						
	Travel time	-0.00445	***						
	Generalized Time (4 VOT)							-0.00474	***
	Generalized Time (1 VOT)			-0.00493	***				
Auto	Generalized Time (1 VOT)					-0.00616	***		
Air	Generalized Time (1 VOT)					-0.00514	***		
Bus	Generalized Time (1 VOT)					-0.00216	***		
Rail	Generalized Time (1 VOT)					-0.00443	***		
Air	Young (<25)	-1.77715	***	-1.90308	***	-1.75492	***	-1.90554	***
Bus	Young (<25)	1.16305	***	1.048235	***	1.080331	***	1.034022	**
Air	Low income	-0.87609	***	-1.24281	***	-1.21515	***	-1.21781	***
Bus	Low income	0.537287	***	0.694025	**	0.761155	***	0.670685	**
Air	Group size	-0.27165	***	-0.2663	***	-0.30552	***	-0.24641	***
Bus	Group size	-0.42832	**	-0.44332	***	-0.50434	***	-0.43642	***
Air	Intermetro	0.382814	**	0.382741	**	0.589983	***	0.325084	**
Bus	intermetro	0.436215	.	0.471794	*	0.686399	**	0.425184	.
Rail	Intermetro	1.667601	***	1.672726	***	1.710404	***	1.670637	***
Air	Male	-0.54193	***	-0.52218	***	-0.59915	***	-0.50546	***
Bus	Male	-0.58231	**	-0.61757	**	-0.68941	***	-0.61777	**
Rail	Male	-0.8882	***	-0.89436	***	-0.92698	***	-0.89737	***
Air	Highly educated	0.608756	***	0.657999	***	0.73448	***	0.622664	***
Bus	Highly educated	0.807337	***	0.869545	***	0.895217	***	0.865331	***
Rail	Highly educated	0.834334	***	0.83134	***	0.824757	***	0.833025	***
Air	Overnight	1.144804	***	1.160232	***	1.081526	***	1.165753	***
Bus	Overnight	1.055228	***	1.097195	***	0.464027	*	1.170732	***
Rail	Overnight	0.885267	***	0.855334	***	0.64353	***	0.851077	***
	Log-likelihood	-2074.8		-2077.3		-2034.4		-2096.2	
	McFadden's R^2	0.47		0.47		0.48		0.46	

Model comparison for trip purpose business (Significant codes: *** 99.9% significance level, ** 99%, * 95%)

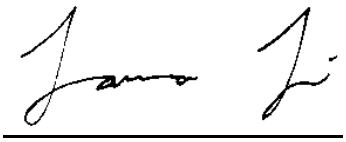
		Model 1		Model 2		Model 3		Model 4	
		Time & Price		1VOT 1Coef		1VOT4Coef		4VOT1Coef	
Mode	parameter	Coef	Sig	Coef	Sig	Coef	Sig	Coef	Sig
Air	2:(intercept)	-5.05459	***	-3.80291	***	-5.73302	***	-4.50578	***
Bus	4:(intercept)	-4.12423	***	-3.41434	***	-3.84903	***	-3.36843	***
Rail	5:(intercept)	-3.89237	***	-4.22924	***	-4.55696	***	-4.24733	***
	Frequency	0.003086	***	0.002276	***	0.002467	***	0.002276	***
	Travel Cost	-0.00197	***						
Auto	Travel Time	-0.00318	***						
Air	Travel Time	-0.00084	*						
Bus	Travel Time	-0.00109	***						
Rail	Travel Time	-0.00201	***						
	Generalized Time (4 VOT)							-0.00255	***
	Generalized Time (1 VOT)			-0.00277	***				
Auto	Generalized Time (1 VOT)					-0.0028	***		
Air	Generalized Time (1 VOT)					-0.00067	***		
Bus	Generalized Time (1 VOT)					-0.00098	***		
Rail	Generalized Time (1 VOT)					-0.00156	***		
Air	Interrural	-1.05603	***						
Bus	Interrural	-1.14432	***						
Rail	Interrural	-3.7477	**						
Air	Intermetro			0.714077	***	1.03073	***	0.670497	***
Bus	Intermetro			0.865981	***	0.981001	***	0.846755	***
Rail	Intermetro			1.306913	***	1.3723	***	1.314138	***
Air	Overnight	1.712379	***	1.959414	***	1.757376	***	1.990293	***
Bus	Overnight	0.448488	***	0.894484	***	0.543838	***	0.933818	***
Rail	Overnight	0.902908	***	1.196368	***	0.966236	***	1.19345	***
Bus	Male	1.216964	***	-0.38746	***	-0.37698	***	-0.3702	***
Rail	Male	1.450933	***	-0.3475	***	-0.32956	***	-0.35181	***
Air	Young (<25)			-0.42832	*			-0.41844	*
Bus	Young (<25)			1.086986	***	1.112197	***	1.083143	***
Rail	Young (<25)			1.347141	***	1.362606	***	1.347479	***
Air	High income			0.195196	*			0.206705	*
Bus	High income			-1.34843	***	-1.35628	***	-1.34684	***
Rail	High income			-0.37241	**	-0.38643	**	-0.37264	**
Bus	Low income	1.324296	***						
Rail	Low income	0.561899	***						
Air	Group size	-0.17414	***	-0.16931	***	-0.17081	***	-0.16806	***
Bus	Group size	-0.40425	***	-0.37396	***	-0.38645	***	-0.37191	***
Rail	Group size	-0.57824	***	-0.54065	***	-0.5561	***	-0.5411	***
	Log-likelihood	-3923.7		-3923.9		-3826.3		-3923.7	
	McFadden's R^2	0.35		0.35		0.37		0.35	

Model comparison for trip purpose leisure (Significant codes: *** 99.9% significance level, ** 99%, * 95%)

Declaration concerning the Master's Thesis

I hereby confirm that the presented thesis work has been done independently and using only the sources and resources as are listed. This thesis has not previously been submitted elsewhere for purposes of assessment.

Munich, June 2nd, 2017



Joanna Yuhang Ji