

Long-distance Travel in Canada: Multimodal Modeling with a National Network

by

Moloud Hajimoradi

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Moloud Hajimoradi 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Long-distance (LD) travel comprises a disproportionately large share of total passenger-kilometers despite representing a small fraction of trip counts. Yet LD travel remains underexamined in Canada's vast geographic context. This thesis develops and applies a comprehensive modeling framework to analyze LD trip generation and mode choice for Canadian residents, leveraging data from Statistics Canada's National Travel Survey (NTS) (January 2018–February 2020) and a new national multimodal transportation network construct for this thesis. The network integrates geospatial centroids for Census Subdivisions with travel-time estimates for automobile, air, intercity rail, and bus modes.

Trip generation was examined through both disaggregate (person-level hurdle and zero-inflated count models) and aggregate (origin-destination zone-pair hurdle models) approaches, incorporating socioeconomic variables (age, income, gender), trip attributes (distance, season), and accessibility measures. Results indicate that accessibility, rather than traditional demographics, may be an important variable in predicting whether a LD trip occurs: with lower local accessibility and greater distance to airports increasing the likelihood of at least one trip in the given month. However, once the trip "hurdle" is crossed, trip counts are less sensitive to accessibility, underscoring behavioral impacts. Even with the very large dataset, models are very weak suggesting that travel surveys are a weak method for understanding LD travel.

Mode choice was analyzed using a Multinomial Logit (MNL) model alongside Machine Learning (ML) classifiers (Decision Trees, Random Forests, Support Vector Machines, Neural Networks). While MNL yields interpretable elasticities, with intercepts confirming preference for the driving mode and positive income effects for air travel, ML methods achieve superior predictive power. Feature importance from Random Forests highlights travel time (especially driving) as the dominant determinant, followed by accessibility, with sociodemographic and seasonal factors playing secondary roles. Mode choice models with alternative specific travel times are viable with publicly available data and these results support the need to seriously consider use of ML in LD mode choice even though understanding the influence of individual behavioral factors becomes more limited.

Long-distance passenger travel demands models are not typically available in Canada despite their utility for infrastructure, service and environmental planning. This thesis research demonstrates models are viable with existing publicly available data.

Acknowledgements

I am deeply grateful to everyone who supported me during this journey.

First, my sincere thanks go to my supervisor, Dr. Lisa Aultman-Hall, for her expert guidance, thoughtful feedback, and contributions to several sections of the writing. Her insights and dedication have been invaluable in shaping this work and helping me grow as both a researcher and a person.

I also thank my readers, Dr. Clarence Woudsma and Dr. Shi Cao, for their careful review and constructive suggestions, which improved this thesis.

To my husband, thank you for your constant love and support, you have been my rock throughout this process.

I would also like to thank Sharon Lee, Jessica Sparry, Leila Mirzaei and Cassidy Berube for their assistance in tabulating and validating databases.

Finally, I dedicate this thesis to my father. His memory and the values he taught me will always inspire my work and heart.

Table of Contents

Author's Declaration.....	ii
Abstract.....	iii
Acknowledgements.....	iv
List of Figures.....	viii
List of Tables	ix
Chapter 1 Introduction	1
Chapter 2 Literature Review	4
2.1 Long-distance Travel Studies.....	4
2.2 Trip Generation.....	4
2.2.1 Long-Distance Trip Generation Methods	4
2.2.2 Predictors of Trip Generation	5
2.3 Mode Choice.....	6
2.3.1 Discrete Choice Models of Long-distance Mode.....	6
2.3.2 Machine Learning Models of Long-distance Mode Choice.....	7
2.3.3 Predictors of Long-distance Mode Choice.....	8
2.4 Implications from Prior Research	9
Chapter 3 Data	11
3.1 National Travel Survey	11
3.1.1 Description of the NTS	12
3.1.2 National Zone System.....	17
3.1.3 Determination of Zone Centroids	19
3.1.4 Accessibility.....	22
3.1.5 NTS Data Summary	23
3.2 Development of a National Transportation Network.....	25

3.2.1 Road Network Travel Times	26
3.2.2 Air Network Travel Times	27
3.2.3 Rail Network Travel Times.....	31
3.2.4 Intercity Bus Network Travel Times.....	32
3.2.5 Validation of Travel Times	33
Chapter 4 Methods.....	34
4.1 Trip Generation Modeling	34
4.1.1 Person-based (Disaggregate) Trip Generation	35
4.1.2 OD-based (Aggregate) Trip Generation.....	38
4.2 Mode Choice Modeling	38
4.2.1 Multinomial Logit (MNL)	40
4.2.2 Machine Learning (ML).....	41
Chapter 5 Results	44
5.1 Trip Generation Results	44
5.1.1 Disaggregate Trip Generation: Hurdle Model Results.....	44
5.1.2 Disaggregate Trip Generation: ZI Model Results	47
5.1.3 Aggregate Trip Generation: Hurdle Model Results	49
5.2 Mode Choice Results	51
5.2.1 Multinomial Logit.....	51
5.2.2 Machine Learning	53
Chapter 6 Conclusions	56
6.1 Key Findings.....	56
6.2 Broader Implications.....	57
6.3 Study Limitations.....	58
6.4 Recommendations for Future Research	59

References.....	60
Appendix A.....	66
Appendix B.....	67

List of Figures

Figure 1 Percentage Air vs Automobile By One-way Distance	15
Figure 2 Kernel Density Estimate of Distance vs Duration.....	17
Figure 3 Boundary of Different Census Geographic Units.....	18
Figure 4 Spatial Distribution of Population Centers (red) vs. Non-Populated Areas (green).....	19
Figure 5 Comparison of Unweighted (green) vs. Population-Weighted (red) Geographic Centroids.....	20
Figure 6 Donut-shaped CSD Zone Example.....	20
Figure 7 Accessibility Distance Bands	23
Figure 8 Road Network in Canada.....	27
Figure 9 Geographic Distribution of Selected Airports in Canada-US Region	29
Figure 10 Total Air Travel Time Between Origin and Destination Pairs	30
Figure 11 Rail network in Canada	32
Figure 12 Feature Importance in Random Forest Mode Choice Model	55

List of Tables

Table 1 ML studies for Travel Mode Choice.....	8
Table 2 Main variables in NTS Datasets	13
Table 3 Summary Statistics of Travelers and Non-travelers	16
Table 4 Number of Zones by Type	21
Table 5 Summary of Dataset Version	24
Table 6 Sample Size by Mode	25
Table 7 Road's Network Size	26
Table 8 Categories of Airport Activities 2019.....	30
Table 9 Potential Variable for Disaggregate Trip Generation	35
Table 10 Potential Variable for Aggregate Trip Generation.....	38
Table 11 Potential Variables for Mode Choice Models.....	39
Table 12 Estimation Results for Disaggregate Trip Generation (Hurdle)	45
Table 13 Estimation Results for Disaggregate Trip Generation (ZI).....	48
Table 14 Estimation Results for Aggregate Trip Generation (Hurdle).....	50
Table 15 Mode Choice Estimation Results.....	52
Table 16 Models Performance for Mode Choice Prediction with Oversampling Strategy.....	54
Table 17 Activity Type in Visit File of the National Travel Survey	66
Table 18 List of Canadian Airports Included in The Analysis	67

Chapter 1 Introduction

Long-distance travel demand represents a significant portion of overall passenger travel as evidenced by research in some European countries, with trips over 100 km representing just 0.5% of all trips but contributing to approximately 20% of the total kilometers traveled [1]. This ratio is also found in the United Kingdom (UK) where trips over 50 miles were shown to only make up 2% of all journeys, but account for 30% of all vehicle traveled [2]. In the United States (US), long-distance travel comprises a large and increasing percentage (approximately 33%) of total miles traveled [3]. The findings in Germany show that 46% of passenger kilometers come from trips with a one-way distance of 100 kilometers or more [4].

Complicating the quantification of LD travel is the reality that the definition of long-distance travel varies across the literature, with no universal standard. While some studies define long-distance trips based on specific distance thresholds [5], [6], [7], others prioritize criteria such as overnight stays to avoid relying on arbitrary distance limits [3], [8], [9]. The choice of definition typically aligns with the study's objectives and available data, as no single criterion suits all purposes. Understanding LD travel patterns in Canada is complicated by the very large range of domestic one-way distances. The distance from the most distance provincial capitals (St. John's NFLD and Victoria BC) is 5062 km. Iqaluit, the capital of the Arctic territory of Nunavut, is 2088 km from Ottawa, the nation's capital. In the Greater Toronto area some commuter trips are routinely 60 km [10], [11]. Defining long-distance trips is a complication for modeling national travel in Canada. Defining what constitutes "long-distance" travel can vary by context and purpose: for example, a household in a small remote town may make necessary LD trips for medical appointments, whereas wealthier households in a metropolitan area may utilize hub airports to access very distant destinations, which are completely discretionary. LD travel includes highly diverse trip types, reflecting purposes that extend from optional recreational journeys to obligatory business travel.

Historically, transportation planners and demand modelers have primarily focused their research and analysis on shorter distance intraurban trips due to the large number of these trips when compared to LD intercity trips [3]. Furthermore, in many locations planning agencies are responsible for single metropolitan areas whereas long-distance trips occur between multiple regions. However, long-distance travel represents a significant portion of the distances traveled in the overall transportation system, making it critical for environment, congestion, and equity concerns. As Canada's total population has grown, the patterns and demand for long-distance (LD) travel evolve in complexity. The ongoing awareness of environmental issues, economic development, service quality enhancement, and emerging technology underscore the importance of robust LD travel demand models for future infrastructure and service planning. These

models are relevant not only for long-distance intercity travel, but also for non-routine LD trips in rural, remote and northern regions of Canada (overall leisure, holiday, and recreational trips are only approximately one third of the trips in the NTS used in this thesis). LD demand models are used to understand the factors that influence long-distance trip generation and the choice of transportation modes, while also enabling a comprehensive approach to rural accessibility and policy analysis, such as carbon emissions reduction. LD travel for this thesis is defined by the Statistics Canada data source: a long-standing survey program that annually collects a large dataset of all non-commute trips over 40-km. The main goal of this research is to utilize this comprehensive database to address gaps in modeling overall national LD travel in Canada, thereby advancing not only understanding of travel in Canada but also appropriate methods to use for LD demand analysis. The contributions of this thesis are as follows:

- 1. Establishing an Appropriate Multimodal National Network Model from Existing Data:** Modeling LD travel including mode choice behavior in Canada is challenged by vast geographical expanse and data limitations. This thesis develops a national model consisting of origin and destination zones together with a multimodal transportation network including travel times between zones for four modes of LD travel: air, automobile, bus, and intercity rail.
- 2. Analysis Of LD Trip Generation as a Function of Accessibility:** Accessibility, defined as the ease of reaching destinations, is hypothesized to influence trip generation rates in long-distance travel. While existing models often focus on the impact of socioeconomic factors on LD travel, few incorporate accessibility and even those focus only on general urban versus rural comparisons. By addressing this gap, this work advances insights to guide consideration of equity in LD transportation considerations.
- 3. Comparison of Aggregate and Disaggregate Trip Generation Methods:** The survey database used here is much larger than most others and we meaningfully compare aggregate flow-based models and disaggregate individual-based models to assess their relative strengths, data requirements, and forecasting accuracy for LD travel.
- 4. Comparison of Mode Choice Models:** Discrete choice models and machine learning (ML) techniques are compared to consider the most appropriate methods for forecasting and analyzing LD travel mode choice behavior in the Canadian landscape and data context.

In **Section 1** of the thesis, existing literature is reviewed to inform modeling choices and contextualize our findings. **Section 2** details the primary data sources used in this study, points to limitations, and outlines how our approach enhances the data, with a particular focus on constructing a model of Canada's national long-distance travel network. **Section 3** describes both aggregate and disaggregate travel demand methods

for trip generation and mode choice as adapted for the national Canadian context. After presenting our methods, results and comparison of findings are in **Section 4**, while **Section 5** summarizes conclusions, outlines the limitations of the study, and suggests alternatives for future research.

Chapter 2 Literature Review

2.1 Long-distance Travel Studies

It is important for this thesis to consider the research strategies for both trip generation and travel mode choice. Although some activity-based and household simulation models for long-distance travel demand models have been published [12], most analysis and modeling has remained embedded in the four-step travel demand model structure originally developed for the urban context: trip generation, destination choice, mode choice, and traffic assignment. This thesis focuses on trip generation and mode choice. Destination choice is also considered in the sense that generation will also be considered in terms of origin and destination zone pairs. The following sections provide a discussion of prior work on LD trip generation (Section 2.2) and mode choice (Section 2.3). Each section is structured to first review the methods and key predictor variables used in previous studies, followed by a justification of the chosen approach and predictor variables for this study.

2.2 Trip Generation

The trip generation phase of modeling serves as the starting point in the conventional four-step travel forecasting process, where the anticipated number of trips originating from or destined for a specific zone of the study area is estimated. The purpose of trip generation modeling is to create mathematical models that accurately represent levels of travel, whether at the disaggregate (individual or household) or aggregate (zonal) level. A person's ability to participate in LD trips may be defined by the non-discretionary *need* to participate in activities at the destination such as medical appointments, work, or family responsibilities or a discretionary *desire* for leisure or experiences away-from-home. In this way, some LD travel is a benefit to the traveler while other LD travel may be a burden for the traveler.

2.2.1 Long-Distance Trip Generation Methods

While there are various methods available for modeling trip generation, regression and categorical analysis are among the most common techniques [13]. Other approaches include the ordered logit model, the Tobit model, the Poisson model, multiple classification analyses, structural equation models (SEM), machine learning models and more. Frandberg and Vilhelmson [14] used multiple regression models to analyze the number of international trips taken in the previous year. LaMondia et al. [15] employed ordered probit models to assess travel frequency across four categories, considering different purposes (work and leisure/personal) and modes (driving, air, intercity rail, and intercity bus). Aguiléra and Proulhac [16] applied Poisson regression to model the frequency of long-distance business travel. In another study,

LaMondia et al. [9] investigated inter-trip time intervals using data from a longitudinal survey on overnight travel, utilizing Negative Binomial (NB) regression. Using the same dataset, Aultman-Hall et al. [17] used NB regression to model annual trip tour generation for different definitions of long-distance travel. Similarly, Berliner et al. [5] applied NB regression models to analyze trip counts for various purposes, such as leisure and business. Czepkiewicz et al. [6] also used NB regression models to examine domestic ground trips, international leisure trips, and non-work air travel.

Machine learning has seen less application than regression in the LD arena, but Alizadeh et al. [18] used different ML techniques to predict whether an individual undertakes a LD trip or not. All the aforementioned studies, despite employing a variety of approaches for trip generation modeling, predominantly focus on the individual traveler (disaggregate) level which has been established as the state of the practice for intra-urban daily trip modeling where the data to support disaggregate modeling is often available.

2.2.2 Predictors of Trip Generation

Research has identified several factors that significantly affect long-distance trip generation, including travelers' socioeconomic attributes such as gender, age, income, household composition, vehicle ownership. Age [5], [17], income [5], [6], [17], education level [18], and being male [5], [6] show a positive correlation with the frequency of long-distance trips. Geographic factors, including land use and built environment measures, also play a meaningful role [19], [20] but the direction of influence varies.

Accessibility may have impact on long-distance travel behavior, although it has been addressed in relatively few studies and direction of causality remains unclear. Accessibility measures together the quality and quantity of destinations with a measure of the ease with which individuals can reach those destinations [21]. The direct influence on local daily trip-making behavior has been established [22], [23], [24]. Accessibility can be quantified in several ways, including the number of destinations reachable within a given time or distance threshold, travel time to closest destination, distance to key destinations, and the availability of transportation options. By integrating accessibility metrics such as the distance to major airport [25], [17], [26], or the count of the number of opportunities (e.g., jobs, services, population) reachable within a certain travel time or distance [27], [28], we can evaluate how accessibility impacts LD trip generation. Aultman-Hall et al. [27] have previously hypothesized that access to long-distance destinations may be related to the burden of LD travel (does one need to travel or does one want to travel). The casual mechanism is uncertain. For some people accessibility may decrease the number of trips per period but alternatively frequent travelers may choose to live in urban areas with airport access to more easily accomplish LD travel.

However, Grimal [29] and Moeckel [30] found that people living in major cities tend to make fewer long-distance trips. This paradox may be partly due to the lack of readily available data on relevant airports for a given individual's trip and the associated travel distance or time. Additionally, as Kim and Ryerson [31] highlight, airport choice involves a complex interplay of factors such as pricing and connectivity, further complicating this relationship. The challenge in establishing causal relationships may simply be the overall lack of robust long-distance travel data. In this study, with an unparalleled sample size, we build and use more disaggregate measures of accessibility including both travel distances, airport access and size of destination in a physically larger study area than typically modeled.

2.3 Mode Choice

Mode choice models predict a traveler's selection of one option such as driving, bus, train, or air over another for a particular trip from an origin to a destination. By analyzing these decisions, planners and policymakers can design more efficient, user-centered transportation systems. In this review, we first examine the two main methodological paradigms: discrete-choice and machine-learning approaches, highlighting how each has been adapted for long-distance travel contexts and where gaps remain. We then review the key predictor variables and factors that have been found to impact long-distance mode choice in prior studies.

2.3.1 Discrete Choice Models of Long-distance Mode

Researchers and practitioners most often use the discrete choice random utility framework to model people's choice of their travel modes. This approach treats different modes of travel as distinct options, often nested, that cover choices considered by the decision maker [32]. To date the most widely used discrete choice model is the multinomial logit (MNL) model [33]. This modeling relies on the assumption that the traveler is maximizing their own utility with their choice and is designed with a mathematical structure that includes a random unobserved component of utility that simplifies the process of estimating parameters [34]. A limitation of MNL models is that they assume the probabilities of choosing between any two alternatives are not influenced by the presence or attributes of other alternatives [33]. To overcome these constraints, researchers, particularly those focused on daily intraurban trips, have introduced significant advancements beyond the MNL model, including the nested logit model [35], probit model [36], and the mixed logit model [37]. Unlike the MNL model, the probit model does not assume independence among alternatives, yet due to the need for high-dimensional integration, its parameter estimation process, even with today's powerful computers, is computationally intensive [38]. In recent years, the mixed logit model has received considerable use [39]. Unlike the MNL model, the mixed logit does not rely on the IIA

assumption, allowing for the accommodation of preference diversity, making it better suited to represent consumer choices [40].

While MNL models have been extensively used for mode choice of daily trips within urban areas, few models have been developed to address long-distance mode choice. Lu et al. [41], Baik et al. [42], Moeckel et al. [43], Bradley et al. [44] and Wu et al. [45] created a nested structure for mode choice while Juschten and Hossinger [46], Miskeen [47] and Llorca et al. [48] estimated the MNL to model long-distance mode choice. Predictors of LD mode choice are described below in 2.3.3.

2.3.2 Machine Learning Models of Long-distance Mode Choice

Mode choice modeling can also be viewed as a classification problem. As an alternative to logit models which rely on strict statistical assumptions about data distributions, machine learning (ML) models can capture complex relationships in a flexible, data-driven manner. Machine learning techniques present a promising alternative to traditional statistical methods for modeling travel mode choice particularly when there is a large dataset such as a National Travel Survey used here. ML studies related to mode choice are presented in Table 1 and illustrate the diversity of ML methods that have been applied. Most studies to date that have employed ML techniques are for daily and intra-city travel and has demonstrated the utility of machine learning for capturing the complexity of mode choice behavior [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59]. However, the application of ML to mode choice modeling in the context of intercity LD travel remains limited and its suitability for long-distance travel is not self-evident. Therefore, it was important to examine its applicability within this study. To the best of the authors' knowledge, only the study shaded in Table 1 [60] has used ML for LD mode choice, focusing specifically on intercity trips originating from and destined for Xi'an, China. The studies in Table 1 all show that ML methods outperform traditional logit models in accurately predicting travel mode choices in the urban daily trip context. For instance, Tang et al. [50] demonstrated the superiority of the Decision Tree (DT) over the logit model for predicting the travel mode chosen by individuals in Luxembourg. Hagenauer and Helbich [52] and Cheng et al. [55] showed Random Forest (RF) performs significantly better than MNL and other machine learning models. Wang and Ross [53] found the Extreme Gradient Boosting (XGB) model has overall higher prediction accuracy than the MNL model especially when the dataset is not extremely unbalanced. The number of successful mode choice studies using ML resulted in the decision to apply ML to the Canadian intercity LD data in this thesis.

Table 1 ML studies for Travel Mode Choice

Authors	Data	Methodology
Omran, 2015	Daily mobility of Luxembourg 2003	Artificial neural net-MLP**, artificial neural net-RBF, multinomial logistic regression, and SVM
Tang et al., 2015	two-day travel diary from the 2007/2008 Washington, DC region and adjacent areas	DT** and logit model
Hagenauer & Helbich, 2017	Six-day trips from 2010 to 2012 Netherland	ANN-based, Naive Bayes (NB), SVM, ANN, Gradient boosting, Bagging, RF**
Wang & Ross, 2018	Delaware Valley 2012 regional household travel survey data	XGB** vs MNL model
Chapleau et al., 2019	Montreal household travel surveys (2008 & 2013)	Random Forest
Cheng et al., 2019	One day travel diary data from Nanjing, China in 2013	RF**, SVM, adaptive boosting (AdaBoost), and MNL
Zhao et al., 2020	Home-to-work stated-preference survey of University of Michigan faculty, staff, and students	Naive Bayes (NB), classification and regression trees (CART), boosting trees (BOOST), bagging trees (BAG), RF **, SVM, NN, mixed logit, MNL
E.-J. Kim, 2021	daily trip, Seoul 2016	XGB**, RF, ANN
Naseri et al., 2022	2018 Montreal OD; Montreal census-tract; Montreal proximity; Walk Score; national deprivation; non-car infrastructure accessibility	Logistic Regression, Light Gradient Boosting Machine** (LightGBM), XGB, RF, AdaBoost, Linear Discriminant Analysis, and CatBoost
Tamim Kashifi et al., 2022	Six-day trips from 2010 to 2012 Netherland	Logistic Regression, RF, Multilayer Perceptron, LightGBM, DT**
Li et al., 2024	surveys in Xi'an, China, entry–exit passenger survey March 2018	XGB
Note: ** model shows the highest reported accuracy among all compared models.		

2.3.3 Predictors of Long-distance Mode Choice

While different methods, such as multinomial logit (MNL) models and machine learning approaches, have been extensively applied to mode choice modeling, the factors evaluated for influence on mode behavior have been relatively constant. While almost all studies measure or control for trip travel time, cost or impedance, most empirical work still concentrates on socio-demographic variables; relatively few compare socioeconomic, trip-specific, and origin–destination factors within a single framework.

Trip-specific attributes including travel time, travel distance, cost, and trip purpose significantly affect predicting LD travel mode choice [42], [43], [48], [60], [62]. However, in the LD context travel time and

cost data are not often collected and it is particularly challenging to measure accurately, especially for unchosen modes. This is one focus of the national network developed in this thesis for Canada.

In terms of socio-demographic factors, several authors demonstrate that age, income, gender, and household composition strongly influence LD mode choice [63], [46], [48], [41], [47], [45], [60]. Hess et al. [64] extends this by using a hybrid choice model, combining socioeconomic characteristics with individual attitudes, to show that, beyond basic demographics, individual preferences further refine mode decisions among train, automobile, air, and bus users between major U.S. cities. Monchambert [65] explored the willingness to carpool for long-distance trips in France, finding that individuals prefer to travel alone rather than share a ride. Moreover, those who choose to carpool exhibit a higher average value of time than train or bus users.

Origin–destination factors such as accessibility to transportation hubs including airports, rural-urban locations, and weather conditions have also been shown to influence people’s choice of mode [43], [66]. Van der Waerden et al. [67] focused on medium- to long-distance travel decisions between train and automobile, particularly highlighting the attributes of access modes. Their research indicated that travel time and cost are the most critical factors, while additional features of access modes to the rail station had a relatively limited impact in their context.

Long-distance travel can be considered the sale of a service in discretionary situations. There are service-specific case studies in the literature that examine travelers’ responses to particular service offerings. Lannoo et al. [68] and Van Acker et al. [69] investigated the interest levels of Belgian business travelers in intercity coach services compared to other transportation options. Bergantino and Madio [70] analyzed potential shifts between transportation modes with the introduction of planned high-speed rail services in Italy. Their study found that such shifts are more likely to occur from air travel and conventional rail services than from bus, carpooling, or private automobile usage.

Despite this diverse body of literature, very few studies conduct a unified analysis that directly compares socioeconomic, trip-specific, and origin–destination variables within the same model for long-distance travel across a study area as large as the country of Canada, leaving a clear gap for research that integrates all three factor categories.

2.4 Implications from Prior Research

Prior studies of LD travel are particularly constrained by the available datasets. In this study aimed to model LD travel in Canada a data constraint is also present. We must adopt the National Travel Survey’s definition

of long-distance travel (see Section 3.1) and rely on the predefined zone system. Only those predictors that appear in the survey data or can be derived from external sources are viable. Despite these constraints, this thesis advances the field in several ways. In terms of methods, regression-based methods are adopted for trip generation as used by others as they are well-suited for exploring the factors affecting LD travel in Canada using the survey data. The National Travel Survey is much larger ($N = 337,928$ respondents used in this NTS extraction) than those in prior studies and allows for evaluation of the sample size truly needed to capture all the variability in LD travel. However, given the large zones and very large area of Canada, a zonal pair-based, aggregate approach to trip generation is also evaluated. This method, which allows for examination of travel patterns at a more regional level, represents a perspective not recently used in similar LD studies and no longer considered appropriate in urban daily models. For mode choice modeling, this research employs the Multinomial Logit (MNL) model due to its effectiveness. But the results will be compared with those from machine learning (ML) models to assess the strengths and limitations of both techniques, in this nation modeling context. The thesis compares alternative methods for both trip generation and mode choice.

This thesis also builds on the predictors of LD trip generation and mode choice. The typical factors used by others are incorporated into the analysis including **traveler attributes** such as age, income, and gender; **household characteristics** such as size and the presence of children. Typical **trip-specific attributes** are also considered, including the great circle distance between zone pairs, distances from origins to airports and rail stations, and trip purpose. More refined trip-specific measures are available in the NTS compared to other surveys including length of stay, travel parties, accommodation type, and activity types during the trip and traveler spending. Considerable effort was undertaken to advance the measures of zone accessibility beyond those used by others for LD travel. Most importantly, travel surveys do not typically include mode-specific travel time between zones. These critical factors are calculated via development of a national transportation network model. In summary, this study uses a larger survey dataset with more predictors than prior studies as well as generating more disaggregate travel times and accessibility measures for comparison of typical and alternative modeling methods.

Chapter 3 Data

In this study, the main travel data is extracted from the National Travel Survey (NTS) collected by Statistics Canada as described in Section 3.1. However, the national transportation network including zone structure, and travel times by four modes was generated for this thesis using publicly available data from the census (e.g., population, zone boundaries), geographic coordinates, and detailed information on transportation infrastructure such as road networks, rail networks, airports, as described in Section 3.2

3.1 National Travel Survey

The National Travel Survey (NTS) has been conducted by Statistics Canada since 2018, to measure domestic and international tourism demand by Canadian residents. This survey replaced the prior similar Travel Survey of Residents of Canada (TSRC) also conducted by Statistics Canada (since 2005), which focused exclusively on domestic trips. The survey is conducted by electronic questionnaire and captures all overnight trips with either a domestic or international destination, as well as all non-commute same-day trips exceeding a one-way distance of 40 km. The retrospective data is collected continuously for a one-month period for each respondent. In addition to reporting the total number of trips ending in the prior month, the survey collects more detailed information on up to three trips for each respondent including mode choice, activities at the destination, stops (legs) during the trip, and trip expenditures.

The NTS lacks trip-specific details, such as travel time, route and exact travel distance. Moreover, trip details are only collected for a subset of all trips. To determine for which trips detailed data will be collected, Statistics Canada applies a three-stage sampling design. The primary sampling unit (for the first stage) is the household. The sampling unit for the second stage is the person in the dwelling chosen from the first stage. Finally, the sampling unit for the third stage is up to three trips from a given dwelling and a selected person. For the first stage, the household, the sample is first allocated among the provinces using the cubic root of the number of dwellings in each of 10 provinces. For the second stage, one adult per selected dwelling is randomly chosen using a selection method based on the age of household members. For the third stage, the electronic questionnaire asks the respondent for a short description of all trips ending in the reference month. If the respondent reports three trips or less, then all the trips are chosen for the remainder of the questionnaire. Otherwise, an algorithm in the electronic questionnaire selects three trips according to a Poisson sequential sampling from all reported trips in the reference month [71].

Although the survey has been conducted continuously since 2018 (except for March, April, May, and June of 2020 due to the COVID-19 pandemic), this thesis specifically focuses on data from **January 2018 to**

February 2020 to avoid pandemic-related disruptions in travel patterns. Utilizing this dataset allows us to analyze long-distance travel patterns, model travel mode choices, and explore other relevant aspects for Canadian residents. Notably, the survey does not collect commuting trips like routine travel which is a regular part of one's job, nor regular household or grocery shopping, nor commuting to school. It also does not collect travel data on visitors to Canada or non-residents. Participants are selected through a sample-based methodology derived from the Labour Force Survey (LFS) sample [72]. This approach is designed to ensure a representative cross-section of Canadian households. The overall weighted response rate for the NTS varies by quarter. For instance, in the second quarter of 2018, the response rate was 32.7% [73], while in the fourth quarter of 2020, it was 24% [74]. These rates can differ across provinces.

The NTS selects approximately 39,000 households monthly. This sample size may be adjusted based on partnerships or specific survey requirements. It is important to note that the actual number of respondents can be influenced by the response rate and other factors affecting survey participation. The NTS employs weighting adjustments to account for non-response and ensure representativeness. These adjustments help mitigate potential biases and improve the accuracy of the survey estimates [75].

3.1.1 Description of the NTS

This section provides a summary of the NTS dataset used in this study, which includes 337,928 respondents surveyed between Jan 2018 and Feb 2020. Of all respondents, 39% took at least one LD trip (totaling 132,480 travelers), while 61% did not travel long distances in their one-month survey period (205,448 non-travelers). Table 2 provides the main variables that are available in this survey. Many variables are typical in travel surveys, but the activity, spending and accommodation details are relatively unique.

Table 2 Main variables in NTS Datasets

Categorical Variables	Description	% (weighted)
Person gender	Male	49.2
	Female	50.8
Household income	Less than \$50,000	29.9
	\$50,000 to less than \$70,000	18.8
	\$70,000 to less than \$100,000	16
	>\$100,000	33
	Missing	2.3
Main trip purpose (participant defined)	Holiday, leisure or recreation	34
	Visit friends or relatives	39.2
	Non- routine Business (non-commute work trips)	12.3
	Other (shopping, medical, religious, other personal)	14.5
Trip main mode	Automobile	84.2
	Air	9.4
	Bus	1.9
	Train	1.3
	Other	3.1
	Missing	0.1
Overnight Accommodation	Same-day trip (no overnight accommodation)	60.3
	Home of friends and relatives	17.3
	Hotel	11.9
	Motel	1.0
	Campground	2.0
	Other unpaid or commercial	7.0
	Missing	0.5
Trip End date	Year/Month/Day	–
Non-exclusive activities (Appendix A)	34 activities such as camping, boating, fishing, skiing (multiple answers are typical)	–
Categorical Variables		Min, max, mean
Person age	18-99 years	18, 99, 48.2
Household size	Number of adults/children	1, 20, 2.9
Number of Trips	Total domestic and international trips per month	0, 40, 0.8
Travel party	Number of adults/children (in household) that accompanied respondent	1, 11, 1.7
Great circle distance	Distance one-way in km	40, 5543, 263.5
Duration	Number nights at destination	0, 234, 1.6
Spending	Total Spending (\$CAN) for each of accommodation, food, recreational and entertainment activities, clothing, and other	0, 127, 444
Origin/Destination/Stop	Province/CMA*/CD**/CSD***, if domestic	–
	State, if United States	
	Country, if other than Canada and US	

* Census Metropolitan Area, ** Census Division, *** Census Subdivision

A total of 305,941 trips were documented in detail (Unweighted). Among unweighted travelers, 45% made a single trip, whereas 55% made multiple trips. As expected, automobile travel dominates the mode choice, followed by air travel. In contrast, bus and rail account for smaller portions of trips. Additionally, some trips are categorized under other modes, including options like cruises, ferries, private planes, private boats, and snowmobiles. Mode percentages in this paragraph are weighted for population. Air travel is most prominent in British Columbia, accounting for 13.8% of all weighted trips originating from this province. In contrast, New Brunswick has the smallest portion of air trips, at 4.2%, indicating flights are less commonly used for trips from this province. Train and bus travel are more common in Ontario and Quebec compared to other provinces. Seasonal travel patterns indicate that long-distance travel peaked in June and July, with a participation rate of 10.3%, and was lowest in October, at 6.4%. Air travel is more common in April, accounting for 11.8% of all trips made during this month, while August and July see the lowest percentage of trips by flight, at 6.4%, and 7.1%. Automobile travel is most common in August and July, accounting for 87% of trips, and least common in February and March, at 81.7%. This suggests that travelers may favor driving in summer and flights in winter which is consistent with a cultural understanding of Canada as a winter country where driving is challenged by winter weather and where those who can take vacations in the winter to warmer more distance places often do.

Business trips peak in September and October, while they are less common in June and July. This trend suggests that higher air travel in winter may be influenced by business travel, as workers are more likely to use flights for these trips. Conversely, the majority of trips in summer are for vacation, which could explain the increased driving and the decline in air travel during these months in part due to larger travel party size. Additionally, 45% of weighted respondents traveled with companions, while 55% traveled alone, resulting in an average party size of 1.7 household members per trip.

On average, 64.8% of weighted trips ended on weekends compared to 35.2% on weekdays. Driving is higher on weekends, accounting for 87.1% of trips, compared to 82% on weekdays. Air travel is more common on weekdays, making up 11.3% of trips, while it drops to 6.8% on weekends. This may reflect a higher proportion of business travel during the week, with workers using flights for work-related trips.

Figure 1 illustrates how mode share varies with one-way trip distance. For trips up to 750 km, automobile travel is the predominant mode, accounting for almost all trips within the 0-500 km range. This indicates that for shorter distances, people prefer driving over flying. As distance increases, the proportion of trips made by air rises. Starting around the 750 km mark, air travel begins to increase, gradually overtaking driving as the distance increases.

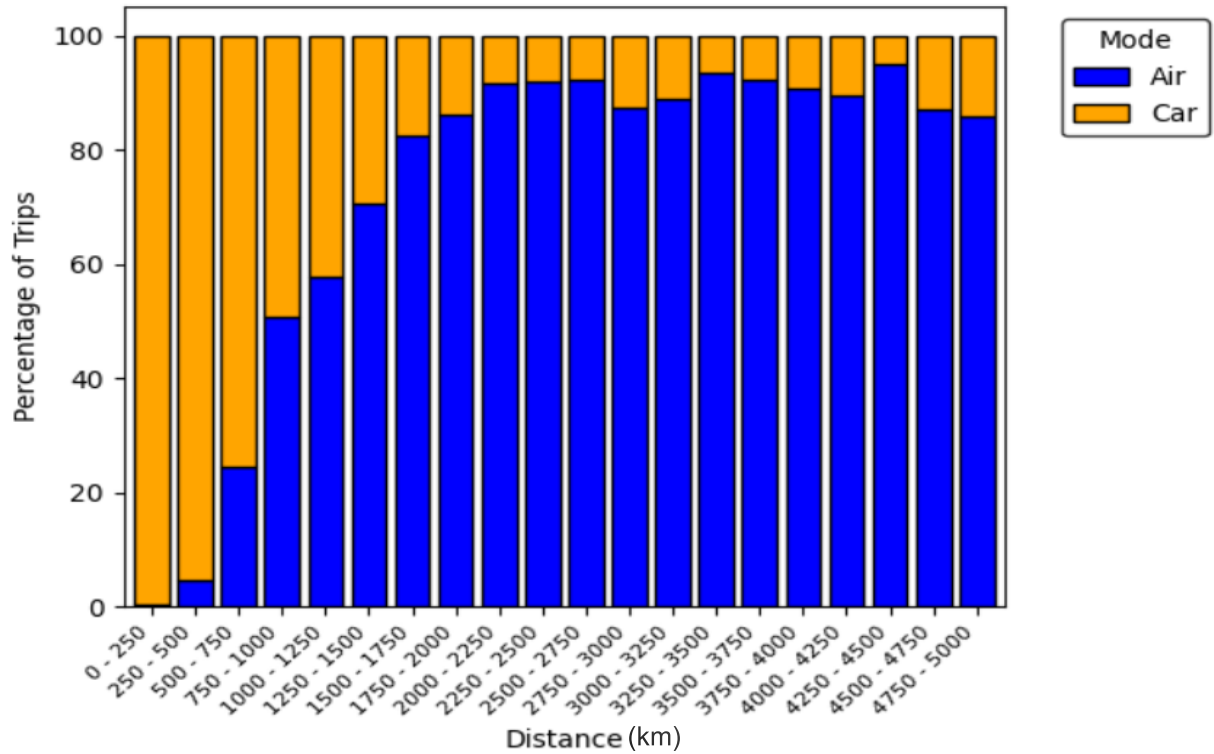


Figure 1 Percentage Air vs Automobile By One-way Distance

Table 3 presents the weighted demographic statistics of people and their households, comparing travelers (those who made a trip in the survey month) and non-travelers in the NTS dataset. Travelers are predominantly from higher-income households, consistent with prior research indicating that higher-income individuals undertake more LD trips [76]. In contrast, gender splits are nearly identical (travelers: 48.9 % male, 51.1 % female; non-travelers: 49.4 % male, 50.6 % female), and household size categories show only marginal variation (e.g., two-person households make up 36.3 % of travelers and 33.3 % of non-travelers). Likewise, the presence of children differs by less than one percentage point (30.8 % of travelers have children versus 29.5 % of non-travelers). However, **among travelers**, 69.2 % are from childless households compared to 30.8 % with children, indicating that households without children may be more likely to make long-distance trips. In other words, aside from income, travelers and non-travelers exhibit very similar demographic profiles which is consistent with prior studies.

Table 3 Summary Statistics of Travelers and Non-travelers

Variable	Category	Travelers % (Weighted)	Nontravelers % (Weighted)
Gender	Male	48.9	49.4
	Female	51.1	50.6
Household Income	Less than \$ 50,000	21.3	34.8
	\$50,000 to less than \$70,000	17.7	19.5
	\$70,000 to less than \$100,000	16.9	15.6
	>\$100,000	41.9	27.8
	unknown	2.2	2.3
Household Size	1	13.1	15.5
	2	36.3	33.3
	3	19.7	20.1
	4	19	18.1
	+5	11.9	13
Children present in Household	No children	69.2	70.5
	Have children	30.8	29.5

Figure 2 is a 2D Kernel Density Estimate (KDE) plot illustrating how trips are distributed by distance and duration. Most trips are relatively short in terms of both distance and duration, with a gradual decline in frequency as trips become longer. This suggests that most long-distance trips remain close to home, around 250 km or less, and typically last fewer than 3 days. Longer journeys are less common. The colored contours indicate regions of higher or lower density: hotter colors (yellows and greens) show where most trips cluster, while cooler colors (blues and purples) indicate fewer trips. The white areas generally represent regions of very low or zero density.

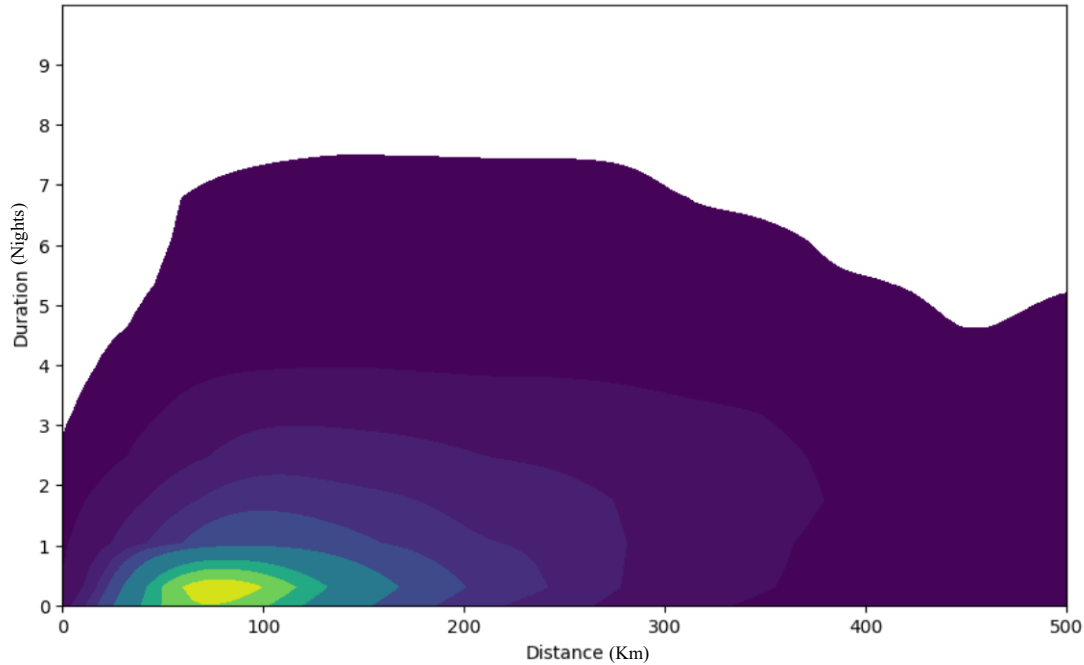


Figure 2 Kernel Density Estimate of Distance vs Duration

3.1.2 National Zone System

The start locations of each trip in the NTS are coded geographically by country, province, tourism region (TR), Census metropolitan area (CMA)¹, Census division (CD)², and the smallest unit, Census subdivision (CSD)³. In Canada, there are 13 provinces and territories, 152 CMAs, 293 CDs, and 5,253 CSDs. Figure 3 illustrates the boundaries of CMA, CD, and CSD for a portion of the province of Ontario. Since the CSD is the smallest resolution used in the dataset they are used as the origin/destination zone in the modeling in this thesis. This choice enhances the precision of our analysis, ensuring calculation of mode-specific details, like travel time. But the variability of CSD zone size also represents a significant challenge for accurate modeling. For example, Toronto (2021 population: 2,794,356) is one CSD, as is Waterloo, Ontario (2021 population: 121,436), despite Toronto being considerably larger (631 square km versus 64 sq km).

¹ Large, densely populated centers made up of adjacent municipalities that are economically and socially integrated. CMAs only cover dense urban areas.

² A group of neighboring municipalities joined together for the purposes of regional planning and managing common services (such as police or ambulance services).

³ Area that is a municipality or an area that is deemed to be equivalent to a municipality for statistical reporting purposes.

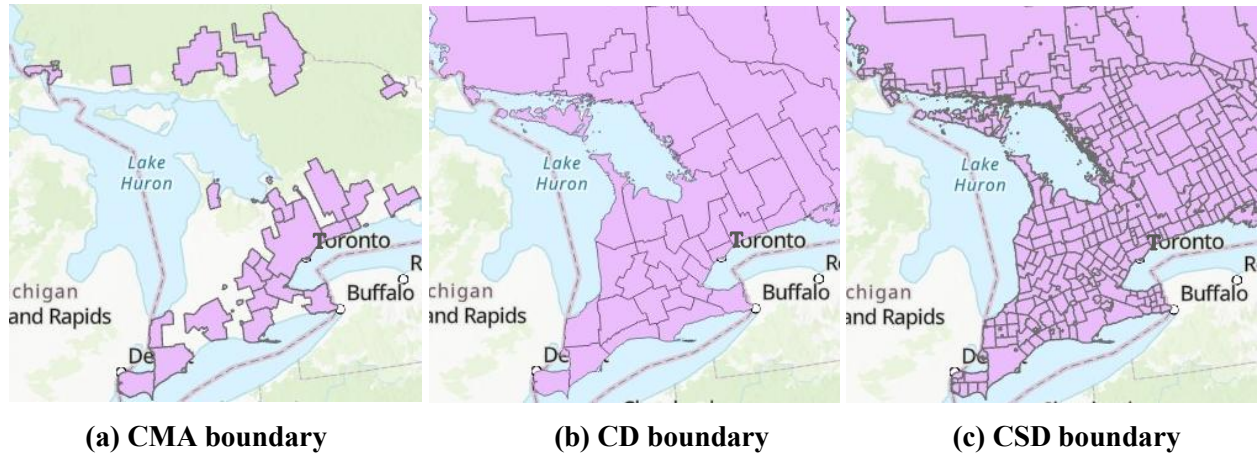
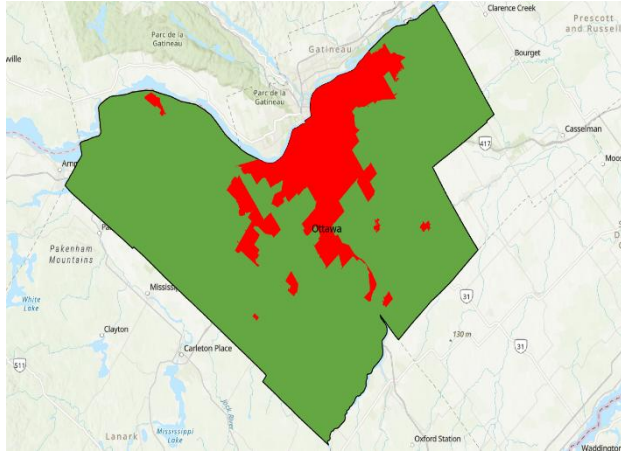


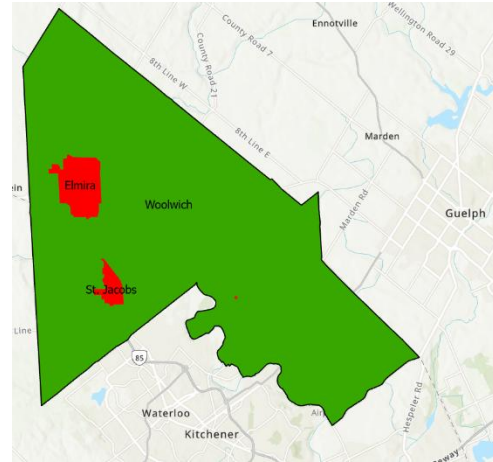
Figure 3 Boundary of Different Census Geographic Units, 1:11,000,000

Additionally, Canada includes numerous rural and sparsely populated CSDs. Each CSD (such as Ottawa or Woolwich in Figure 4) might comprise both populated areas (represented in red as population centers) and unpopulated or sparsely populated areas (depicted in green, often comprising agricultural, forest, or wetland areas). Both the red and green regions together form a single CSD, highlighting the variability in land use within these geographic units. These limitations of the original data zones make establishing centroids and calculating meaningful travel times very challenging, particularly when a single centroid may not adequately represent large or irregularly shaped CSDs. It can affect the accuracy of calculating distances, travel times and network access for origin and destination points.

Before establishing the final set of zones, we remove a set of CSDs that met any of the following criteria: a) zero population, b) categorized as First Nations, or c) located in the Northern Territories. The first criterion assumes that the travelers modeled in this study are more likely to make trips between populated areas. The second and third criteria are based on constraints imposed by the NTS, which excludes persons living on First Nation lands and persons living in the Northern Territories. Therefore, 1208 CSDs were excluded from our modeling leaving 4045 CSD zones in the study area covering 43% of Canada's total area of 9,984,670 square kilometers.



(a) Ottawa



(b) Woolwich

**Figure 4 Spatial Distribution of Population Centers (red) vs. Non-Populated Areas (green),
1:450,000**

3.1.3 Determination of Zone Centroids

Within zone areas, the selection of a representative point for the start and end of trips is needed as input for the estimation of travel distances and times between zones. The chosen point should aptly represent the entirety of its respective region zone. Typically, when population is uniformly distributed, representative points for the origin and destination zones are located at the centroids of zones. However, population is rarely uniform, and many areas of Canada have a low population density. To improve the accuracy of centroid location for the origins and destinations in Canada, we incorporated additional data from the Census dissemination areas (DA) within the CSDs. These areas consist of one or more neighboring dissemination blocks⁴ with an average population of 400 to 700 persons. We calculated the population-weighted mean center as nodes for each remaining CSD zone (4045 zones) by considering the distributed population of all DAs within each CSD zone. Therefore, the representative centroid point for each CSD zone is positioned in areas with higher DA populations, effectively placing it near the population center. For example, the difference between Toronto's mean center (green circle) and Toronto's population-weighted mean center (red circle) is depicted in Figure 5(a). It is evident that the population-weighted mean center is biased toward a densely populated area of Toronto (downtown). In cities with evenly distributed populations, such as Toronto, there is no major distinction between mean center and weighted mean center. However, for many CSDs, population density is not uniformly distributed, and there is more distance

⁴ An area bounded on all sides by roads and/or boundaries of standard geographic areas. It is the smallest geographic area for which population and dwelling counts are disseminated.

between the alternative centroid points, as observed in the Thunder Bay district (Figure 5b) and Pickering (Figure 5c). Baik et al. [42] employed a similar approach, utilizing a weighted average based on Census tract population.

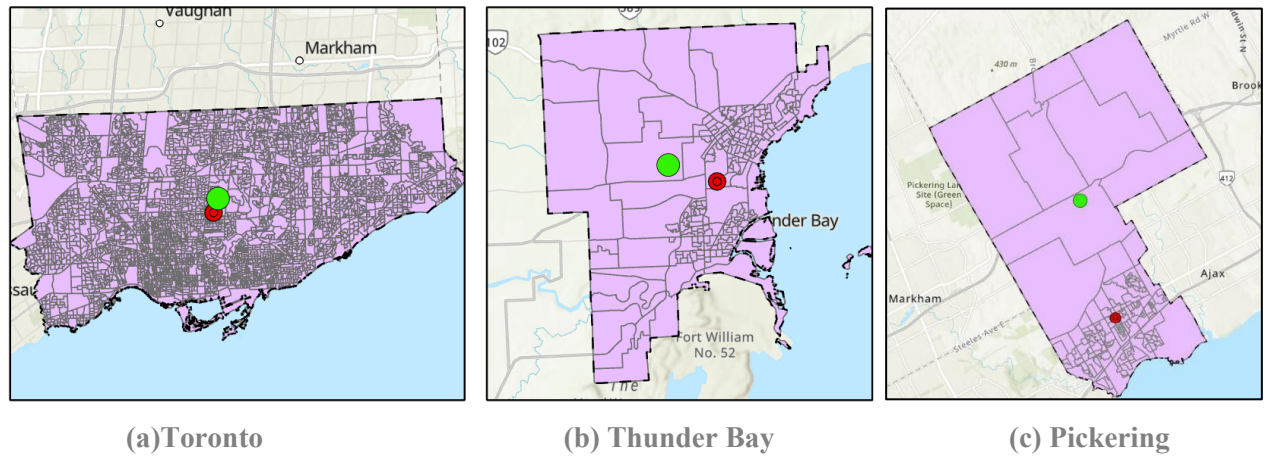


Figure 5 Comparison of Unweighted (green) vs. Population-Weighted (red) Geographic Centroids, 1:450,000

There are several challenges related to the census zones: a) sizes vary degrading the distance calculations using centroids; b) shapes vary causing centroids to be located in some cases outside the zone; c) donut zones of rural areas surrounding more populated villages have their centroid inside the built-up area. Dealing with the latter two problems (Figure 6) can be challenging when using the weighted mean center approach. In such cases, the weighted mean center (red x) may end up located outside the boundaries of the CSD zone (green). To address this issue, we reassign the weighted mean center to the most densely populated dissemination area (DA) within the CSD. This adjustment is made based on the assumption that it is more likely for trips to originate from or be destined to areas with higher population density.

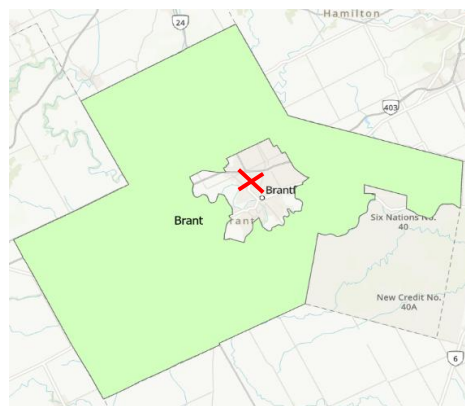


Figure 6 Donut-shaped CSD Zone Example

For international trips to the United States, the smallest geographic unit recorded in the survey is the state with 51 regions identified as destinations. For international trips to countries other than the USA, destination data is limited to the country level, including 248 countries as destinations. To determine the zone centroids for these countries as an international destination for Canadian residents, the centroid of the most populated city within each country was selected, consistent with the methodology adopted for the USA, where the centroid of the most populated city in each state serves as the representative point, where transportation activities are concentrated and accessible. This decision is grounded in the assumption that destinations are more likely to be in major cities. The adoption of the most populated city's centroid serves to establish a representative point for calculating both driving travel time/distance and flight time.

Upon request, Statistics Canada also provided more disaggregation spatial locations for this thesis: the postal forward sortation area (FSA) also known as the three-digit postal code. This geography was provided for the home of respondents and is therefore available for trip origins, but not destinations). However, in remote or rural regions, a three-digit postal code area can exceed the size of the corresponding CSD. To address this, CSDs were classified into two groups: (1) CSDs for which the three-digit Forward Sortation Area (FSA) is smaller than the CSD boundary, and (2) CSDs for which the three-digit Forward Sortation Area (FSA) is larger or equal to the CSD boundary. A CSD is assigned to **Group 1** if all of its DAs lie within a population center and its total population exceeds 30 000⁵, under these conditions, the three-digit postal code provides finer spatial resolution, 145 CSDs meet these criteria. Conversely, **Group 2** is if any DA lies outside a population center (i.e., in a rural area) or if the CSD's population is below 30 000, then the CSD boundary is equal to or smaller than the three digit postal code area; in those cases, the CSD is retained as the primary geographic unit. Table 4 summarize the zone system used in this study.

Table 4 Number of Zones by Type

Country	Zone Type	# Origin Zones	# Destination Zones
Canada	FSA	984 (within 145 CSDs)	-
	CSD	3900	4045
US	States	-	51
Other countries	Country	-	248

⁵ Medium population centers are defined as having at least 30 000 residents [77]

3.1.4 Accessibility

Prior research has shown that travel varies between rural areas, between rural and non-rural areas, and between major cities [78], [79]. In this thesis, we sought to utilize a more refined measure of zone type in terms of accessibility to assess its impact on trip-making. Accessibility is defined in transportation planning research as the ease of reaching valued destinations [80]. Statistics Canada's pre-defines an "Index of Remoteness" [81] which indicates a community's relative isolation by combining population size and travel distance to service centers. However, the earliest available remoteness index corresponds to the 2016 Standard Geographical Classification (SGC), while the NTS dataset relies on 2011 SGC definitions. As a result, several 2011 CSDs lack a remoteness value. To avoid excluding these CSDs accessibility was re-modeled, based on definitions in the literature at the CSD zone level. In the accessibility literature, "opportunities" are commonly proxied by the sum of employment within a band; however, because employment and population are highly correlated, we use DA population as a pragmatic proxy for valued destinations. This choice treats population as an indicator of the depth of nearby services and activities (e.g., retail, health, education), consistent with the idea that larger nearby populations generally support more opportunities. To evaluate accessibility as a measure of travel burden for this study, we categorize accessibility into three incremental one-way distance bands based on distance to the zone (See Figure 7):

1. 0–40 km for local and regional accessibility (reachable within about one-half hour),
2. 40–80 km for nearby regional accessibility (reachable within approximately one hour),
3. 80–160 km for extra-regional accessibility (approximately less than a half-day drive).

Within each of these bands, we calculate the population of Dissemination Areas (DAs) that fall within the specified distance from a given CSD (or 3-digit postal code), shown in Figure 7. It is important to note that the three distance bands (0–40 km, 40–80 km, and 80–160 km) are not accessibility categories in themselves. Rather, accessibility is calculated as a continuous measure, based on the total population of Dissemination Areas (DAs) within each band around an origin zone (CSD or 3-digit postal code). The accessibility value for a given band can range from zero (if no DA falls within that range) to several million (in large metropolitan areas). For example, in the Toronto CMA, the 0-40 km band includes a very large number of densely populated DAs, resulting in an accessibility value exceeding five million.

This method ensures that the spatial distribution of opportunities is effectively captured, allowing us to identify how proximate each CSD is to destinations in surrounding regions. By integrating these metrics,

we aim to better understand how accessibility impacts long-distance travel behavior and the interplay between local and regional connectivity.

After computing accessibility it was compared with the Index of Remoteness. The Pearson's correlation coefficient was -0.82 ($p < 0.01$), indicated a strong, statistically significant inverse relationship. A basic linear regression of accessibility on the Index of Remoteness yielded $R^2 = 0.77$, showing that 77 % of the variation in remoteness can be explained by accessibility. Therefore, our own calculated accessibility was used instead of the Index of Remoteness which suffered missing data.

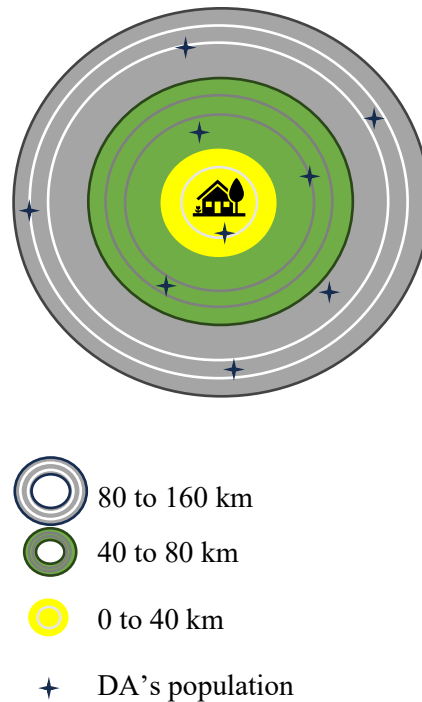


Figure 7 Accessibility Distance Bands

3.1.5 NTS Data Summary

To address different research questions in this study, we employ three distinct extractions of the NTS dataset. Each version is structured and weighted according to the specific requirements of the analyses, particularly for aggregate and disaggregate Trip Generation (TG) and Mode Choice (MC) modeling. Table 5 summarizes the key characteristics of these datasets, including sample size, weighting strategy, and geographic scope.

Table 5 Summary of Dataset Version

Model	Unit of Analysis	Geographic Scope	Sample Size (Unweighted)	Weighting Method
TG-Disaggregate	Individual Person	CSD, FSA, States, Countries	234,835	None
TG-Aggregate	OD Zone Pair	CSD	260,703	NTS Trip Weight
MC-Trips	Trip	CSD, FSA, States	153,713	None

1. **The Person-based Trip Generation (TG-Disaggregate)** dataset is used to model total trips per month based on person or household characteristics and accessibility. Because three-digit postal code data are unavailable for the first seven months of 2018, only data from August 2018 through February 2020 are considered.
2. **The Zone Pair Aggregate Trip Generation (TG-Aggregate)** dataset supports alternative modeling for interzonal trips longer than 40km, that may be useful for policy analyses and transportation planning at the regional or national level, where aggregated trip flows between zones are a useful focus. We only consider interzonal trips within Canada due to lack of interzonal travel time. Therefore, we removed trips involving the US and other countries as well as trips with unspecified exact CSDs as destinations.
3. **The Mode Choice (MC-Trips)** dataset allows us to examine mode choice decisions in detail, capturing the factors that lead travelers to select car, air, rail, bus for specific trips. Only up to three sets of trip details were collected per person. Only out and back trips with one destination stop are included in the mode choice modeling. These are direct journeys from the origin to the final destination without any intermediate stops. Similar to TG-Disaggregate dataset, only data from August 2018 through February 2020 are used. International destinations beyond the U.S. and Canada were removed, since flying is the only realistic mode for those routes. Trips lacking a valid CSD or State destination and overnight trips under 40 km were also excluded. Trips were excluded if the respondent did not report a mode of transportation or selected “other” as their mode. Additionally, some inconsistencies in the dataset led to further exclusions in the MC-Trips dataset. Trips reported as being made by car over a distance of more than 1000 km with a duration of 0 (same-day), or trips that start and end in the same zone (e.g., within Toronto) but are reported as using a plane were excluded. Trips to certain islands, such as Hawaii, where car, bus, and train options are unavailable, and airports are not included in the built flight networks—resulting in no air option—were also excluded. The multistop trips that were excluded lacked a specified visit

sequence between stops with respondents reporting only the main mode and purpose for the entire trip rather than for each leg (N= 7775 trips or 5%). After these exclusions, a total of 153713 trips were considered for analysis. Table 6 summarizes the amount of data available for the mode choice model estimation, trips from 2018 to 2020.

Table 6 Sample Size by Mode

Trip type	Number of surveyed trip records (Unweighted)				Total
	Driving (%)	Air (%)	Bus (%)	Train (%)	
Domestic	136372(88.7)	5756 (3.75)	1474 (0.96)	1718 (1.12)	145320
to U.S.	3760 (2.45)	4464(2.9)	151 (0.1)	18 (0.01)	8393
Total	140132	10220	1625	1736	

3.2 Development of a National Transportation Network

In order to develop measures of impedance and accessibility to consider mode-specific attributes affecting mode choice, the travel time for both chosen and unchosen modes between trip origins and destinations are desired. Travel time is not available in the survey even for the chosen mode as routes are not known. This is typical in travel surveys including the NTS. Although travel times could theoretically be obtained from existing sources such as Google Maps or other travel websites, there are significant limitations to relying on these services for large-scale origin–destination (OD) matrices. With over 4,000 zones in this study, computing a complete OD travel time matrix requires evaluating more than 16 million OD pairs ($4,000 \times 4,000$). Using commercial APIs for this volume of requests would be both cost-prohibitive (running into thousands of dollars) and time-consuming, as these services typically impose strict rate limits on queries. In addition, such APIs are subject to changes in pricing, usage policies, and access restrictions, making them an unreliable long-term solution. Therefore, the dataset is expanded by building a multi-modal national transportation network that includes estimating alternative specific travel times for different modes of transportation between all feasible origin and destination pairs. The network includes automobile, flying, inter-city rail, and inter-city bus where these options are available

An important part of building the national network model is determining the feasible links by mode and associated travel times between all zones. ArcGIS Pro was used to build the network dataset, which links approximately 4000 origins and destinations (based on Section 3.1.3) with travel time. With these link-level attributes in place, ArcGIS Pro’s network analysis tool employs Dijkstra’s algorithm [82] to compute the

shortest path according to the mode specific procedures described in each section below. Validation and quality assurance checks are also described below.

3.2.1 Road Network Travel Times

Driving travel times are available online publicly from various sources such as ESRI or Google. However, the cost of repeatedly accessing the APIs is cost prohibitive and may prevent researchers from sharing them with others. Moreover, part of the value of the online data is its real time or time-stamped specificity which is not of value here as we do not know the exact times or route of travel. As an alternative, creating our own network dataset that represents generalized average travel times by driving was necessary.

To create the road network between all CSD (or FSA) and state zones in Canada and US states, we used actual major roadways, including freeways, primary highways, secondary highways, municipal collectors, and local streets from the Bureau of Transportation Statistics (BTS) [83]. The geodatabase provides segment length and speed limits for roads in Canada and the United States. Table 7 listed the included roads representing the driving network across the Canada-US Region. Travel distance is computed based on the available length of roads. Travel times were estimated based on distances and speed limits, which vary between 20 and 130 km per hour depending on the type of road. ArcGIS Pro network analysis enables us to calculate the shortest driving distances and times between all zone pairs. Note there are zones in the study area that lacked any road connections. A total of 104 CSDs (classified as rural CSD) did not have road access within 20km of the centroid and have been excluded from the road network. This is not unexpected in the rural northern areas of Canada. Figure 8 depicts the extensive road network across Canada, highlighting the region's remote location within the country.

Table 7 Road's Network Size

Country	# of Segments	Total Length (km)
Canada	21,975	124,038
United States	670,066	713,083

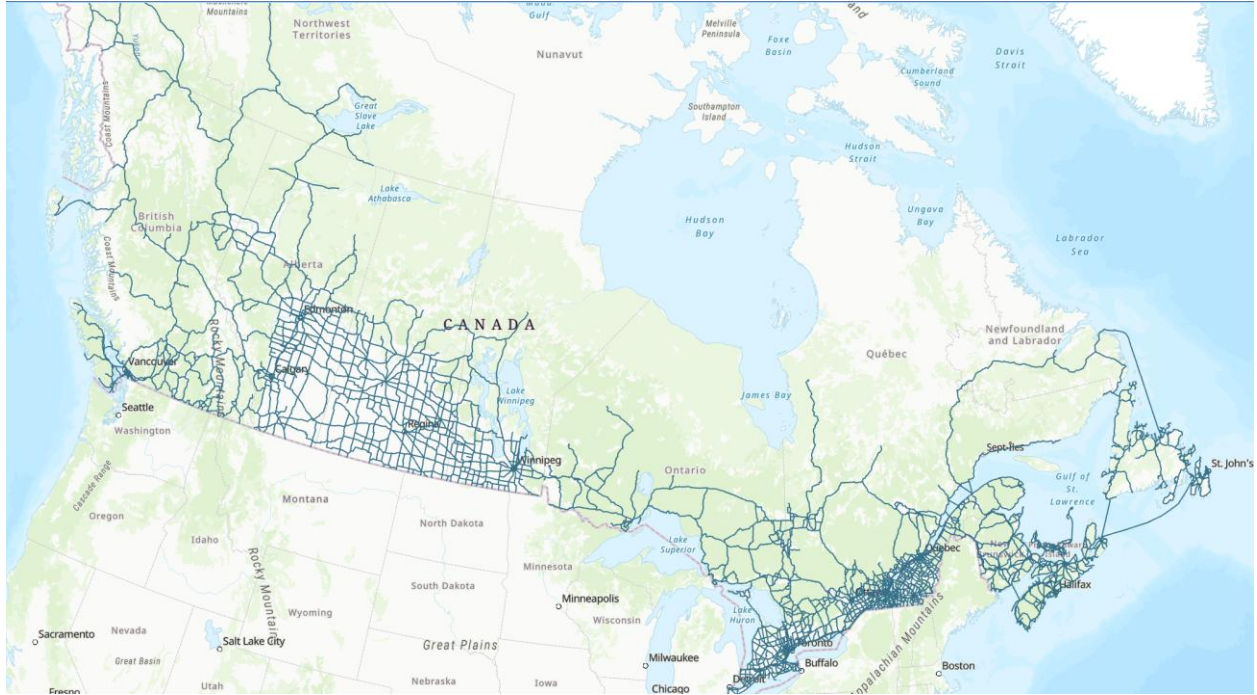


Figure 8 Road Network in Canada, 1:15,000,000

3.2.2 Air Network Travel Times

The multimodal travel network also requires air connections, including realistic travel time, by air with connections between Canadian zones and from Canadian origins to destination zones in the USA. However, not all zone pairs have commercial flight options, and no readily available dataset exists for air travel similar to the road network. Therefore, we first needed to identify the commercial airports that would effectively cover the Canada–U.S. region and confirm the availability of frequent direct flights between them. Once these nodes and their connections were established, they could be imported into ArcGIS for network analysis and travel-time calculations.

To ensure an accurate depiction of the aviation landscape without pandemic-related disruptions, we used flight data from 2019, which reflects a non-COVID-affected period and therefore the general accessibility by air under non-pandemic conditions. Although, Canada has over 300 airports with scheduled commercial flights, a significant portion of them are small, have infrequent flights, operate seasonally, and lack navigation services. Consequently, our analysis focused exclusively on larger commercial airports (excluding private airports) based on the following criteria: (1) the 23 National Airports⁶ [84] owned by

⁶ The National Airports are based on passenger traffic and cover most of the traffic in Canada.

Transport Canada and leased to Canadian airport authorities, (2) the five busiest airports in 2023, and thirteen key airports distributed across various provinces to ensure broad regional coverage. In total, these criteria yielded 41 airports (see Appendix B). We then obtained comprehensive data on direct flight availability and flight times for each of these selected airports in Canada from the FlightStats [85] and FlightAware [86], respectively. Through these sources, we identified frequent direct flights and calculated average flight times between the Canadian airports in our dataset.

For the U.S. component of the air network, we used a subset of 140 commercial airports across the continental United States, identified from 2019 FAA airport data [87]. The airport dataset includes hub size classification, annual enplanements, and annual flights. All medium and large hub airports are included. Small hub airports are included only if they are more than 65 miles (straight-line distance) away from a medium or large hub airport. The non-hub airports are included in the 140-airport set if the airport has greater than 90,000 annual enplanements [87].

Flight data for these 140 selected airports was sourced from the Federal Aviation Administration (FAA) T-100 Segment All Carriers dataset [88], which provides information for every nonstop segment, including cross-border segments between Canadian and U.S. airports. Any entry with zero reported air travel time or zero passengers was excluded. Further, only direct flights repeated at least 365 times per year were retained in the dataset to ensure viable service was represented. Additionally, any airport pair lacking full round-trip service (i.e., if $A \rightarrow B$ exists but $B \rightarrow A$ does not) was removed from the dataset. The distribution of the 41 selected airports in Canada and 140 airports in the U.S. is illustrated in Figure 9. A maximum of 3,712 air segments with service as defined here was found in the air network.

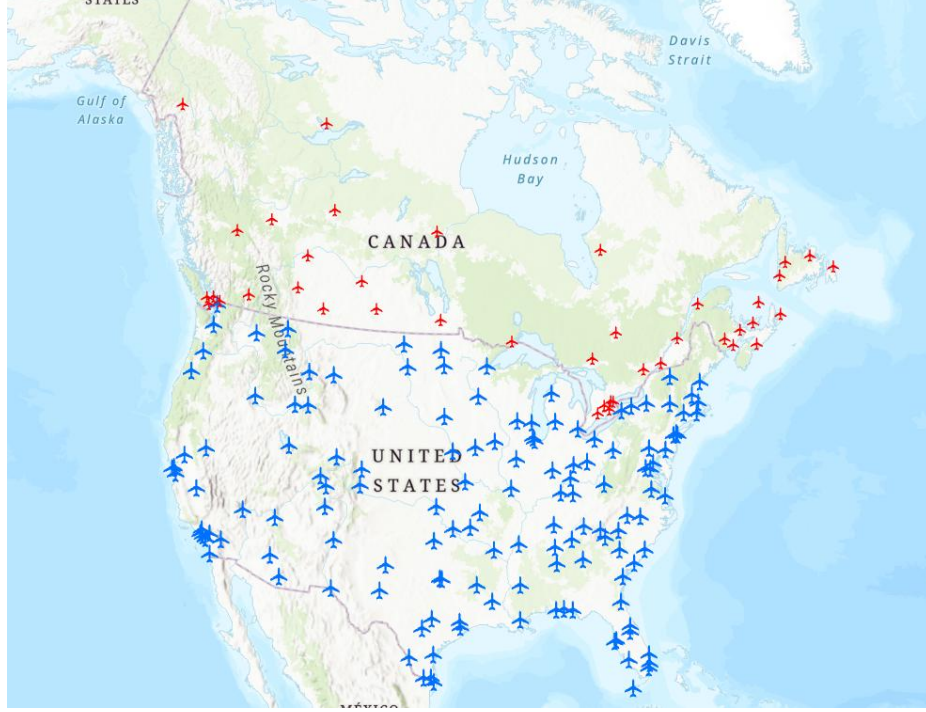


Figure 9 Geographic Distribution of Selected Airports in Canada-US Region

Total air travel time is not simply flying time and consists of the two components calculated with equations 1 and 2. As illustrated in Figure 10, by combining origin airport to destination airport flying time with highway-based access/egress times, we estimated the total travel time by air (equation 1). We break down origin airport to destination airport travel time (equation 2) into several key components: flying time, taxi-in/out time, and waiting time at each airport.

$$\text{Total air travel time} = \text{Access time} + \text{Airport-to-airport travel time} + \text{Egress time} \quad (1)$$

$$\text{Airport – airport travel time} = \sum_{i=1}^N \text{Waiting}_i + \text{Taxi_out}_i + \text{Flying time}_i + \text{Taxi_in}_i \quad (2)$$

Where N represents the number of flight segments, including both direct (N=1) and connecting flights (N > 1) and Waiting_i , Taxi_out_i , Flying time_i , and Taxi_in_i refer to the time associated with the i-th flight segment. We use ArcGIS Pro to identify the shortest flight route between a departure airport and an arrival airport, accounting for all available connections, whether direct or involving multiple segments.

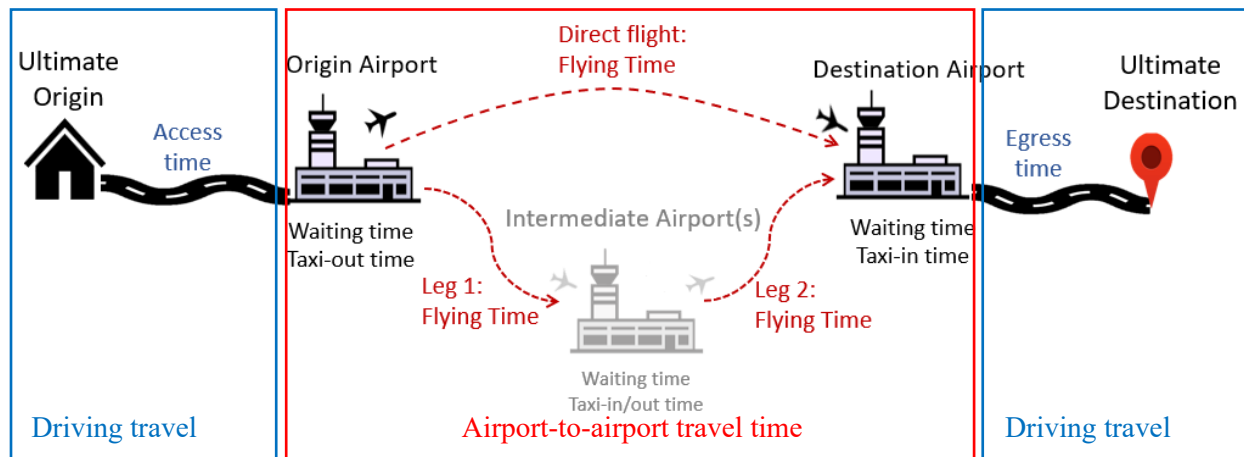


Figure 10 Total Air Travel Time Between Origin and Destination Pairs

Flying time is calculated as the average flight time between the origin and destination airports, obtained from FlightAware, and T-100 data as described above.

Waiting time at the origin airport includes check-in, security screening, and boarding. The duration of these activities varies by airport size, as detailed in Table 8. At the destination airport, waiting time also includes baggage retrieval, with a standard exit time of 30 minutes.

Taxi-in/out time in the United States is sourced directly from the FAA's Airline On-Time Performance Data [89]. For Canadian airports, we classify them as large, medium, or small based on enplanement data from 2019 (Appendix A) and assign average taxi times from the corresponding U.S. airport categories, ensuring consistency in estimating taxi-in/out times, as presented in Table 8.

Table 8 Categories of Airport Activities 2019

Size of airport	Waiting at Origin (min)	Waiting at Destination (min)	Average Taxi-out (min)	Average Taxi-in (min)
Large	120	30	19.1	9.1
Medium	90		13.5	5.6
Small	75		14.9	5.3
Non-hub	60		15.5	4.7

Access and egress times to and from the airport are equivalent to the driving times on the road network from the zone centroid to the airports as outlined in Section 3.2.1. Note that for each origin-destination (OD) zone pair, we evaluate multiple airport pair options. Instead of including all airports for every OD pair, the 3 closest airports to origin centroid and 3 closet airports to destination centroid were identified [87]. This approach results in nine potential airport pairs being evaluated to determine the minimum air travel time. The shortest travel time among these nine pairs is then chosen to represent total air travel time between the OD zone pair.

3.2.3 Rail Network Travel Times

One of the available travel modes between some zones in Canada is intercity rail or train service, primarily operated by VIA Rail but also some regional rail authorities. Canada has numerous rail stations, some of which offer daily service, while others operate as infrequently as twice a week. Some stations in Canada serve only passenger trains, others are dedicated to freight, while some accommodate both. This study considers VIA Rail stations serving passengers across Canada and GO Transit stations in the Greater Toronto Area that provide intercity service. GO operates daily service in a significant portion of Southern Ontario. Note that although rail service exists in other metropolitan areas it was not considered to serve intercity origin and destination pairs. Local rail services, such as the Expo Line in British Columbia and EXO in Quebec, were excluded as they do not serve LD trips, allowing for a simplified rail network model. Additionally, six Amtrak stations in the United States, which serve as common destination points for travel to the U.S. from Canada, are included.

The main rail links for this research are shown in Figure 11. The rail network was developed using ArcGIS Pro, focusing on zones located within 50 km of these rail stations, encompassing a total of 1,686 out of 4,783 zones (3796 CSD + 981 FSA + 6 States). To calculate the total travel time by train, we combined station-to-station travel time with highway-based access and egress times to the nearest station, determined by our driving network. This relationship is shown in Equation (3). The **station-to-station travel times** are obtained from VIA Rail, GO Transit, and Amtrak schedule tables. For international trips, travel times include border control delays of an estimated 30 minutes. Additionally, for routes without a direct connection where transfer at intermediate stations is required average waiting times for transfers are also considered in the overall travel time.

$$\textbf{Total rail travel time} = \text{Access time} + \text{Station_to_station travel time} + \text{Egress time} \quad (3)$$



Figure 11 Rail network in Canada

3.2.4 Intercity Bus Network Travel Times

Canada is a vast country with an unevenly distributed population, resulting in limited intercity bus services in many areas. Estimating the bus travel time between all 4000 zone pairs was not feasible. The NTS trip dataset indicates that the mean great circle distance for bus travel falls between 186.9 km and 212.2 km with 99% confidence, suggesting considering zone pairs within a 600 km threshold would be a reasonable subset to consider for bus trips. For longer distances, bus travel is not considered a practical option due to the extended time and lack of convenience compared to other modes of transportation. Therefore, to construct intercity bus travel times, this study focuses on pairs of zones within 600 km of each other. Even with this limitation there were too many zone pairs to exhaustively enumerate.

From the full set of zone pairs, we filtered out any pairs exceeding 600 km in great-circle distance and then randomly selected 2,500 pairs for manual assessment. We obtained bus travel times from two sources: Rome2Rio and the Google API databases. For trips to the US, we added a penalty of +30 minutes to account for border crossing delays. This dataset formed the basis for training and testing various machine learning models to estimate bus travel times between zone pairs. The selected predictors included origin/destination, great circle distance between two zones (km), driving travel time (hours), and whether the origin is in a metropolitan area. We began with linear regression as a baseline model for comparison and subsequently applied random forest regression to accommodate potential non-linearities and interactions among predictors. Random forests, comprising multiple decision trees trained on bootstrap samples [90], have also been successful in prior travel time studies [91], [92].

To ensure robust model evaluation, we split the dataset into training (80%) and test (20%) subsets, maintaining a representative distribution of predictor variables in each. Using 5-fold cross-validation on the training set, we tuned key hyperparameters (e.g., number of trees, $n_estimators$, maximum depth), evaluating each configuration via Mean Squared Error (MSE). We then selected the parameter settings that yielded the lowest average MSE.

3.2.5 Validation of Travel Times

To validate travel time by different modes obtained from own national network, 100 random samples were selected. The time differences (in minutes) between the travel times estimated by our own national network and those obtained from Google Maps or Rome2Rio were calculated. While most time differences cluster around zero, a few differences could be attributed to real-world conditions (e.g., construction, accidents), or varying assumptions (e.g., different ferry speeds). We also verified that the assigned speed limits in the BTS dataset aligned with known provincial, state, or regional regulations. The mean absolute error in our travel-time estimates is approximately 4% of the average trip duration across all modes. These findings confirm that the underlying reasons for outliers are reasonable and do not undermine the overall alignment between the national network estimates and external sources.

Chapter 4 Methods

This study focuses on understanding long-distance travel in Canada and identifying appropriate methods for demand analysis. Two types of modeling for each of trip generation and mode choice are used. In Section 4.1 two different approaches, one aggregate and one disaggregate, for trip generation are discussed. In Section 4.2, two different approaches to mode choice are presented: one the traditional discrete choice; and the other a more recent machine learning approach that while applied to daily intracity trips has not seen wide application in long-distance analysis. We retained predictors with strong theoretical support, low collinearity, solid univariate signal, measurable importance, and available data. Interaction terms were introduced only when supported by theory or when residual diagnostics indicated meaningful effect modification between variables.

All estimation and analyses were conducted in Python 3.10. Pandas and NumPy were used for data ingestion, cleaning, and numerical operations, while Matplotlib and Seaborn were employed for plotting and graphical summaries. Additional key packages are described in the relevant sections below.

4.1 Trip Generation Modeling

Trip generation stands as the foundational phase in forecasting travel demands and understanding travel behavior. Trip generation can be modeled using both aggregate and disaggregate approaches. **Disaggregate models** analyze trip generation at an individual person or household level. Greater precision in predicting individual person or household travel behavior is expected as this allows for incorporation of personal and behavioral factors which might be sensitive to policy changes [32]. However, **aggregate models** analyze trip generation at a zonal or group level, treating the population or households in each spatial zone in aggregate. It is simpler to implement and computationally efficient and well-suited for large-scale analysis such as infrastructure planning and where detailed individual-level data is unavailable [93]. In Section 5, this thesis compares results from both models and evaluate the strengths and weaknesses of each.

When modeling trip generation, whether using an aggregate or disaggregate approach, the number of trips is typically treated as count data. Traditional regression methods, such as Poisson or Negative Binomial (NB) regression, are commonly employed. However, we observe that many respondents in the National Travel Survey (NTS) may not have experienced or reported trips during the study period, resulting in observed excess of zero counts in our dataset. Therefore, Hurdle [94], [95] or Zero-Inflated (ZI) models [96] are more appropriate, as will be explained in detail in the next section. Model comparison measures, such as Akaike information criterion (AIC) [97], Macfadden R-square are used to compare the goodness of

fit of these two types of models. The Statsmodels package [98] was used for statistical hypothesis testing and count-data regressions, including Poisson, Logit, Negative Binomial, and Zero-Inflated models.

4.1.1 Person-based (Disaggregate) Trip Generation

Given the type of information available, several statistical models have been applied to model the number of LD trips and presented in Section 5. Table 9 shows the potential outcome and predictor variables available from the NTS for person-based trip generation

Table 9 Potential Variable for Disaggregate Trip Generation

Variable	Description	Range
Travel Outcome Measure Options	Made a trip or not	0,1
	Number of monthly individual trips	0,1, 2, ...
Predictor variable	Age	18 to 99 years
	Household Income Category	\$-50k to \$100k+ per year
	Gender Identity	Male/Female
	Have children	0,1
	Household size	1,2,3,4,5+
	Quarter of year	Winter, Spring, Summer, Fall
	Local accessibility at origin	Sum of population within 40 km radius
	Distance to nearest airport in national network	km
	Distance to nearest rail station	km

Different model specifications were estimated using different combinations of predictor variables. We included interaction terms only when they had strong theoretical support and statistically significant p-values (like income $\times$ HH size). The interaction term between household income groups and household size was the only tested interaction that proved significant, capturing the nuanced relationship between income and household composition. Building on the available data and insights from past research, we consider two common approaches for handling count data that are described below: Two-Part (Hurdle) Model [94], [95], and Zero-Inflated Model (ZIM) [96].

Two-Part (Hurdle) Model

The Hurdle Model consists of two distinct components:

a) **A binary model** (e.g., logistic or probit regression):

The first component models the probability of observing at least one long-distance (LD) trip (i.e., crossing the "hurdle"). It is implemented using logistic regression:

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (4)$$

- Where $P(Y = 1)$ is probability of observing at least one LD trip.
- $\beta_0, \beta_1, \dots, \beta_p$ are coefficients for the predictor variables $X_1, \dots, X_p$.

b) **A truncated count model** (Poisson or Negative Binomial):

The second component models the distribution of trip counts **given that at least one trip was made**. It is implemented using a truncated count distribution:

$$\log(E[Y|Y > 0]) = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_q X_q \quad (5)$$

- $E[Y | Y > 0]$ is expected number of trips for individuals who made at least one trip.
- $\gamma_0, \gamma_1, \dots, \gamma_q$ are coefficients for the predictor variables $X_1, \dots, X_q$.

The overall probability mass function (PMF) for the Hurdle Model is:

$$P(Y = y) = \begin{cases} \pi & \text{if } y = 0 \\ (1 - \pi)f(y; \theta) & \text{if } y > 0 \end{cases} \quad (6)$$

- where π is the probability of being in the "always zero" group (i.e., making no trips) from the binary model.
- $f(y; \theta)$ is the probability mass function of the truncated distribution (Poisson or Negative Binomial).

Zero-Inflated Model

The Zero-Inflated Model, such as the Zero-Inflated Poisson (ZIP) or Zero-Inflated Negative Binomial (ZINB) can be used when the data contains two types of zeros: structural zeros (always zero) those that never travel, and sampling zeros (zeros that could have been positive but were observed as zero in the surveyed period). This model consists of two components:

a) **A binary model** (logistic regression)

This component predicts whether an observation belongs to an "always zero" group:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha_0 + \alpha_1 X_1 + \dots + \alpha_r X_r \quad (7)$$

- where π is the probability of belonging to the "always zero" group.
- $\alpha_0, \alpha_1, \dots, \alpha_r$ are coefficients for the predictor variables $X_1, \dots, X_r$.

b) **A standard count model** (Poisson or Negative Binomial)

This component models the count of trips for individuals who are not in the "always zero" group:

$$\log(E[Y | Y \text{ not always zero}]) = \delta_0 + \delta_1 X_1 + \dots + \delta_s X_s \quad (8)$$

- $E[Y | Y \text{ not always zero}]$ is the expected count for observations in the "not always zero" group.
- $\delta_0, \delta_1, \dots, \delta_s$ are coefficients for the predictor variables $X_1, \dots, X_s$.

The overall probability mass function for the Zero-Inflated is:

$$P(Y = y) = \begin{cases} \pi + (1 - \pi)f(0; \theta) & \text{if } y = 0 \\ (1 - \pi)f(y; \theta) & \text{if } y > 0 \end{cases} \quad (9)$$

- where π is the probability of belonging to the "always zero" group.
- $f(y; \theta)$ is the probability mass function of the count distribution (e.g., Poisson or Negative Binomial).

The distinction between hurdle and Zero-Inflated models is subtle but important. For the Hurdle model, zeros only come from the binary model. Once an individual crosses the hurdle ($Y > 0$), they cannot produce zeros in the count model. The count model only predicts strictly positive values ($Y > 0$). In a Zero-Inflated Model, zeros can come from two sources: (1) the binary model (structural zeros) or (2) the count model (sampling zeros). Individuals in the "not always zero" group can still produce zeros through the count model. Structural zeros are individuals who never make a trip (e.g., due to physical constraints or lack of interest), while sampling zeros are individuals who might not make a trip (or did not report) in the observed period but could make one in another period.

All Zero-Inflated models were fit in Python (3.10) with Statsmodels package, employing default maximum likelihood estimation. Specifications were compared via AIC and examined residuals for over-dispersion. Additionally, while a Zero-Inflated Negative Binomial (ZINB) model might also be appropriate, its

application on our extensive dataset (approximately 16 million records) proved to be highly time-consuming and often led to system crashes.

4.1.2 OD-based (Aggregate) Trip Generation

In addition to individual-level trip generation modeling, travel demand can be aggregated by origin–destination (OD) pairs to estimate the total number of trips between zones. The aggregate trip generation model provides distinct insights into the dynamics of inter-zonal travel flows. This approach relies on zonal characteristics and travel impedances as key predictors of long-distance travel volumes (see Table 10). Given that approximately 99% of OD pairs report zero trips in the NTS dataset, both Zero-Inflated Models (ZIM) and Hurdle models are suitable candidates, as discussed in the previous section. However, due to the large size of the dataset (approximately 16 million records), applying a Zero-Inflated Model proved computationally intensive and frequently resulted in system crashes. Therefore, only the Hurdle model was implemented.

Table 10 Potential Variable for Aggregate Trip Generation

Variable	Description	Range
Travel Outcome Measure Options	Made a trip or not	0,1
	Number of weighted OD pair trips per month	0,1, 2, ...
Predictor variable	Great circle distance (GCD)	in km
	Driving travel time	in min
	Distance to nearest airport	in km
	Local Accessibility	Sum of population within 40 km radius
	Population	Sum of population within the CSD zones

4.2 Mode Choice Modeling

The essence of mode choice modeling lies in its ability to capture the influence of predictors on the probability of transportation modes being selected by individuals or groups for their trips. The development of these models involves evaluating various factors that influence travelers' decisions. Based on the literature review conducted in Chapter 2 and the data available in our survey, key variables that commonly influence mode choice include those listed in Table 11. These factors may play a role in individual preferences and choices between different LD transportation modes in Canada.

Table 11 Potential Variables for Mode Choice Models

Variable	Description	Values/Units
Travel Outcome Measure	Mode	Air, Drive, Bus, Train
Predictor variable	Great circle distance	km
	Travel time by specific mode	Hour
	Age	18 to 99 years
	Household Income Dummy Category	\$-50k to \$100k+ / year
	Gender	Male/Female
	Number of travel party	1,2,3,4,5+
	Quarter (season) of year	1,2,3,4
	Main trip purpose	Visit family/Vacation/non-routine Business ⁷ /Other
	Local Accessibility	Sum of population within 40 km radius

Based on a thorough review of the literature, both Multinomial Logit (MNL) models and machine learning (ML) algorithms were applied to predict mode choice in the NTS. These models utilize socio-demographic variables, and travel impedance as predictors. The objective is twofold: to compare the relative predictive accuracy of the methods and to analyze the influence of various factors on mode selection. To ensure a valid comparison, the MNL and ML models were trained, evaluated, and tested on the same dataset. Specifically, the classifiers are trained on 80% of the data, with the remaining 20% reserved for evaluation. In practice, the NTS dataset is imbalanced, with approximately 84% of respondents choosing automobile as their mode. Various techniques, such as oversampling, Synthetic Minority Over-sampling Technique (SMOTE), and undersampling, can be used to balance the dataset. SMOTE generates synthetic data through interpolation, which may introduce unrealistic observations, while undersampling risks losing important information. To address this, oversampling techniques were applied to the training data, increasing the frequency of minority choices. This approach improves the model's ability to predict rare outcomes while preserving overall performance and interpretability [99]. The Imblearn package [100] was used to address class imbalance in discrete-choice and classification tasks, employing techniques such as SMOTE, RandomOverSampler, and RandomUnderSampler.

⁷ Non-commute work trip

Importantly, the test dataset is kept unchanged to ensure it remains representative of real-world data. The goal is to enhance the signal for rare choices without “fooling” the model into learning an unrealistic baseline share. Moreover, evaluating overall accuracy alone can be misleading because it may be dominated by the majority class. For example, if 84% of travelers choose automobile, a model that always predicts “driving” would achieve 84% accuracy, despite being ineffective. Therefore, besides accuracy, we also assess the influence of each variable. Model performance is assessed for accuracy using F1-score [101]. Scikit-learn package [102] was used to support machine learning workflows, including train/test splitting, model fitting, cross-validation, and evaluation using metrics accuracy, and F1 score. The F1-score is especially useful when the classes are imbalanced, as it considers both false positives and false negatives. A higher F1-score indicates better model performance. For the ML models, specifically for Random Forest models we employ tools feature importance rankings [90] to identify which variables most significantly impact predictions. The feature importance is based on the Mean Decrease in the Gini coefficient, which measures how much each feature contributes to reducing impurity across all trees in the forest. Specifically, it quantifies the average decrease in Gini impurity resulting from splits on a given feature. The Gini index (also known as Gini impurity) for a node t in a classification tree is defined from [90] in Equation 10.

$$G(t) = 1 - \sum_{i=1}^k p_{it}^2 \quad (10)$$

Where k is the number of classes (modes) and p_{it} is the proportion of observations of class i in node t . If all observations at node t belong to a single class, then one of the $p_{it} = 1$ and the rest are 0, so $G(t) = 0$ (perfect purity). If observations are evenly split among classes, $G(t)$ reaches its maximum (greatest impurity). A larger Mean Decrease in Gini for a feature therefore indicates that using that feature to split nodes produces purer child nodes and thus contributes more strongly to correct classification.

4.2.1 Multinomial Logit (MNL)

MNL models have been widely used in transportation studies because of their interpretability and strong theoretical foundation in utility theory, allowing estimation of the probability of choosing a specific mode based on predictor variables. In the multinomial logit (MNL) model, each individual n faces a choice among J alternatives. The utility U_{nj} that person n associates with alternative j is composed of a deterministic component V_{nj} (a function of observed characteristics) plus a random error ε_{nj} [103]:

$$U_{nj} = V_{nj} + \varepsilon_{nj} \quad (11)$$

The ε_{nj} is assumed to be independent and identically distributed (i.i.d.) drawn from a Type-I extreme-value (Gumbel) distribution. Under this assumption, the choice probabilities take the closed-form [104]:

$$P_{nj} = \Pr(U_{nj} \geq U_{nk} \forall k) = \frac{\exp(V_{nj})}{\sum_{l=1}^J \exp(V_{nl})} \quad (12)$$

Because of the Gumbel distributed error terms, the difference in utilities maps directly to log-odds [33]:

$$\ln \frac{P_{nj}}{P_{nk}} = V_{nj} - V_{nk} \quad (13)$$

That is, a one-unit increase in the deterministic utility of j (holding all else constant) increases the log-odds of choosing j over k by exactly one.

In the specification of an MNL model, explanatory variables can be categorized into generic variables and alternative-specific variables. Generic variables are those that take the same value across all alternatives for a given decision-maker (e.g., an individual's age, income, or trip purpose) and thus enter the deterministic utility function V_{nj} with alternative-specific coefficients. In contrast, alternative-specific variables differ across alternatives and capture attributes unique to each mode, such as travel time, travel cost, or service frequency. In the present study, the only alternative-specific variable incorporated is travel time. To facilitate comparison across modes within the MNL framework, travel time was expressed as Delta travel time, representing the positive or negative difference between the travel time of each non-driving mode and that of driving. This formulation allows the model to capture the marginal impact of relative travel time differences on mode choice probability, while controlling generic socio-demographic and trip-related characteristics. In practice V_{nj} is specified as a linear index [32]:

$$V_{nj} = \beta_{0j} + \sum_m \beta_{mj} X_{njm} \quad (14)$$

where the X_{njm} is the value of variable m for individual n and alternative j , and the β coefficients are estimated by maximum likelihood. BIOGEME [105] was used for the estimation of the MNL model.

4.2.2 Machine Learning (ML)

Recent advancements in ML have opened up new possibilities for predictive accuracy, leading to exploration of models such as Support Vector Machine (SVM) [106], Neural Network (NN) [107], Decision

Trees (DT) [108] and Random Forests (RF) [109] as complementary approaches in this thesis. The recommended tenfold cross-validation is used for hyperparameter tuning on the training set [110].

Support Vector Machine (SVM): SVM is a supervised machine learning algorithm designed to find a hyperplane that optimally separates data into two categories or classes (e.g., ensuring all observations from one class are positioned above the hyperplane and those from the other class below it). The algorithm aims to maximize the margin between the two classes, ensuring the hyperplane is as far as possible from the closest data points of each class. To enhance flexibility, SVM employs kernel functions that transform input vectors into a higher-dimensional space, enabling effective classification of non-linearly separable data [106].

In this study, the C-Support Vector Machine (C-SVM) variant was implemented, as it is particularly well suited for classification tasks such as predicting travel mode choices. Different kernel types, including linear, polynomial, radial basis function (RBF), and sigmoid, were tested to determine the most accurate model. To evaluate performance, 10-fold cross-validation ($cv=10$) was applied, ensuring robust assessment across multiple training and validation splits. Among the tested kernels, the RBF kernel outperformed the others, indicating its superior ability to capture complex, non-linear relationships in the data. The accuracy and F1 score of the SVM model with the RBF kernel were computed using the test dataset.

Neural Network (NN): Neural networks are a class of machine learning algorithms inspired by the structure and function of the human brain. These models learn complex patterns and relationships in data through iterative training and by comprising interconnected layers of nodes (neurons). Each neuron receives inputs, applies a weighted sum, and passes the result through an activation function to produce an output. This structure allows neural networks to learn and model complex patterns and relationships within large sets of data. Neural networks are capable of representing highly non-linear relationships and can automatically learn relevant features from raw data. One common architecture is the Multi-Layer Perceptron (MLP), a type of feedforward neural network that consists of at least three layers: an input layer, one or more hidden layers, and an output layer [107]. In this study, an MLP classifier is utilized to predict travel mode choices and is initialized with a single hidden layer of 20 neurons and trained. The training process involves up to 1000 iterations, during which the model adjusts its internal parameters (weights and biases) based on the data it is exposed to. This iterative adjustment is aimed at minimizing a predefined loss function. The training method employed is a standard approach where the 'fit' function is used to feed the training data into the model, allowing the neural network to learn from the inputs and corresponding labels. After training, the performance of the MLP is evaluated using accuracy and F1-score metrics on a separate test dataset.

Decision Tree (DT): Decision trees are a fundamental machine learning algorithm used for both classification and regression tasks. They recursively partition the input space into regions by using a series of binary decisions based on feature values. Each internal node represents a decision point, while each leaf node corresponds to a predicted outcome. One of the significant advantages of decision trees is their ability to handle both numerical and categorical data, making them versatile for various applications. Moreover, their structure allows them to be relatively unaffected by outliers and irrelevant features, which can distort the performance of more sensitive algorithms. The code employs cross-validation to determine the optimal hyperparameters for the decision tree classifier. Specifically, the algorithm explores a range of hyperparameters, such as the maximum depth of the tree, to enhance model generalization and prevent overfitting [108]. The Decision Tree Classifier is initialized with a random state parameter to ensure reproducibility, meaning the same results can be achieved every time the model is run with the same parameters and training data. After training the classifier on the training set, predictions are made on the test set to evaluate model performance using accuracy and F1-score metrics.

Random Forest (RF): Random Forest is a supervised learning algorithm widely used for classification tasks. In multi-class classification problems, each decision tree independently votes for a class, and the class with the highest number of votes is selected. To enhance model robustness and prevent overfitting, Random Forest introduces randomness through techniques such as feature subsampling, bootstrap sampling, and random splits during tree construction. Hyperparameter tuning is essential for optimizing Random Forest, particularly in complex tasks like predicting travel mode choices [109]. In the mode choices here, the search was conducted over a specified range: the number of trees varied between 50 and 500, and the maximum tree depth ranged from 1 to 20. The search utilized 10-fold cross-validation ($cv=10$) to ensure robust model evaluation.

Chapter 5 Results

Model estimation results for both aggregate and disaggregate are presented for trip generation in Section 5.1 and mode choice in Section 5.2. Tabulated model results with goodness of fit measures are shown in each section.

5.1 Trip Generation Results

5.1.1 Disaggregate Trip Generation: Hurdle Model Results

The disaggregate Hurdle trip generation was estimated two ways. For the second stage of the Hurdle models, we considered both Poisson and Negative Binomial (NB) regression specifications. However, since the variance of the trip count data exceeds its mean (a condition known as overdispersion), the Poisson model is less appropriate. The NB model, which effectively handles overdispersion, provides a superior fit, as evidenced by its lower Akaike Information Criterion (AIC) and higher log-likelihood compared to the Poisson specification. Consequently, only the NB results are presented. For all predictor variables, Pearson correlation analyses and Variance Inflation Factor (VIF) tests were conducted to assess multicollinearity, ensuring the independence of explanatory variables in the model. The results indicate that the inter-variable correlations are weak, and the VIF values are less than two. This suggests that multicollinearity is not a concern, ensuring that the model estimates are stable and reliable. We estimate a Negative Binomial specification to account for overdispersion in the trip-count data. The dispersion parameter $\alpha = 0.19$ ($p < 0.001$) indicates that the variance exceeds the mean, violating the Poisson model's equal-mean-and-variance assumption and justifying the NB model.

The Hurdle model (Table 12) is split into two components: a binary logit model determining the likelihood of making any trip and a Negative Binomial (NB) model capturing the number of trips conditional on trip-making, shown in Table 12. The NB component, which estimates the number of trips given that a trip is made, the number of significant predictors is limited, and their coefficient magnitude are generally lower. This suggests that while the factors do influence whether a person is a LD traveler, the number of trips is more random. This might imply that once the decision to travel is made, other unobserved factors or situational variables come into play in determining trip frequency. Overall, the Hurdle model has an AIC of 614723 and a McFadden R^2 of 0.04, indicating a very large portion of variability in trip frequency driven by unobserved or situational factors not captured by the current covariates even with very large sample sizes. Note this weak goodness of fit is surprising given the large and representative sample size.

Table 12 Estimation Results for Disaggregate Trip Generation, N = 234,835 (Hurdle)

	1: Binary Logit		2: Negative Binomial	
Variable	Coeff.	P-value	Coeff.	P-value
Constant	0.37	0.000	1.25	0.000
Household Characteristics				
<i>Household Income</i> (Base: \$0 to \$49,999)				
G2: \$50,000 to \$69,999	0.65	0.000	—	—
G3: \$70,000 to \$99,999	0.84	0.000	—	—
G4: \$100,000 or more	1.12	0.000	—	—
G9: Not stated	0.47	0.000	—	—
Respondent Gender (Base: Male)	0.02	0.065	-0.07	0.000
Respondent Age	-0.1	0.000	-0.01	0.000
Household Size	-0.04	0.000	—	—
<i>Interaction term</i> (Base: Income G1)				
Income G2 * Household Size	-0.1	0.000	—	—
Income G3 * Household Size	-0.09	0.000	—	—
Income G4 * Household Size	-0.05	0.000	—	—
Income G9 * Household Size	0.06	0.023	—	—
Zone Characteristics				
Log local accessibility	-0.12	0.000	-0.04	0.000
Log distance to airport	0.08	0.000	0.05	0.000
Log distance to rail station	-0.05	0.000	-0.01	0.000
Trip Characteristics				
<i>Quarter</i> (Base: Winter)				
Q2: Spring	0.19	0.000	0.04	0.000
Q3: Summer	0.56	0.000	0.08	0.000
Q4: Fall	0.13	0.000	0.04	0.000
Alpha		0.19		0.000
McFadden R²		0.02		
Final log-likelihood, LL_F		-307333.5		
AIC		614723.02		
Note: Only variables and interactions significant at 0.05 in either model were included. The bolded numbers indicate coefficients that are statistically significant at the 0.01 level.				
— = not included.				

Not surprisingly, in the binary logit component, household income emerges as a strong predictor: respondents in higher income brackets (\$50,000–\$69,999, \$70,000–\$99,999, and \$100,000 or more) have

significantly higher odds of generating trips compared to the base group (\$0–\$49,999), with the highest income (G4) having approximately three ($e^{1.12} = 3.06$) times the probability of the baseline. In practical terms, higher-income households might have more opportunities or resources to travel. The negative interaction coefficients (income*HH size) in the binary component indicate that the positive effect of higher income on the likelihood of trip-making diminishes as household size increases. In other words, the combined effect of a larger household and higher income does not lead to a proportionally higher likelihood of traveling, suggesting that additional household members might offset some of the mobility advantages typically associated with higher income. Once a trip is made, household size has no impact on how many trips follow (because the count coefficient is not significant). This suggests that larger households may coordinate activities in a way that reduces the likelihood of initiating a trip, but once they do, household size no longer constrains the number of trips. Additionally, demographic factors play a role; for example, older respondents and those from larger households exhibit lower probabilities of initiating any trip. Seasonal effects are controlled for in the model, with summer showing the strongest positive impact on trip-making relative to winter: good weather months encourage more activity/travel.

After accounting for typical predictors, the results for accessibility and airport/rail access were of focus. There may be interaction between different levels of local and regional accessibility. The results suggest that a lack of local access may increase the need to travel longer distances. The negative coefficient on local accessibility (within 40km) suggests that as access to local needs increases, residents are less likely to undertake long-distance trips. One plausible explanation is that higher local accessibility implies that individuals can fulfill most of their needs nearby, reducing the necessity for LD travel. Similarly, the positive coefficient on log-transformed distance to an airport indicates that lower airport access increases the likelihood of making at least one LD trip. This may be interpreted as lower access to an urban area (with an airport) results in a need to make LD trips to access goods, services, or friends and family. This may indicate the need for non-discretionary LD travel is more important than the ease of travel increasing discretionary trip making. Taken together, both findings are directionally consistent: higher accessibility, locally or to regional hubs, corresponds to fewer LD trips, while lower accessibility corresponds to more LD travel. Note that being farther from rail lowers the probability of making a trip, possibly reflecting the opposite and requiring more investigation as very few areas in Canada had rail access.

5.1.2 Disaggregate Trip Generation: ZI Model Results

The alternative disaggregate Zero-Inflated (ZI) model also splits the trip generation model into two parts (Table 13). The inflated (binary logit) component, which models the probability of an observation being an “excess” zero, and the NB count component for trip frequency. Notably, the ZI model yields a lower AIC (566,312) and final log-likelihood compared to the Hurdle model, suggesting a marginally better fit, while its McFadden R^2 is 0.02. The likelihood ratio tests (LRT) indicate that each model performs better than a null model (p -value = 0.000), confirming small predictive improvement over an intercept-only specification.

In the inflated component, the negative coefficients associated with higher income groups indicate that such households are less likely to be classified as excess zeros, in other words, they are more prone to generating trips. This is consistent with the Hurdle model’s findings on income. Older respondents show a small positive inflated coefficient (0.10), implying they are slightly more likely to produce excess zeros which aligns with the Hurdle’s lower binary-logit probability for trip-making among older ages. Female respondents have a negative inflated coefficient (-0.08), indicating they are less likely to be excess zeros (i.e. slightly more likely to travel at least once). In the count component, both older age (-0.02) and female gender (-0.10) reduce expected trip counts once a trip is initiated, mirroring the Hurdle’s NB stage where these demographics had small but significant negative effects.

In the inflated component, local accessibility (0.10), indicating that areas with better local access tend to generate more excess zeros (that is, fewer long-distance trips) just as in the Hurdle model. The inflated coefficient on distance to airport (-0.05) implies that households farther from airports are less likely to be excess zeros (i.e. more likely to travel), consistent with the Hurdle’s positive effect in the binary stage. This dual influence underscores how accessibility issues, whether through local amenities or the ease of reaching long-distance travel hubs, can shape travel behavior.

Table 13 Estimation Results for Disaggregate Trip Generation N = 234,835 (ZI)

	1: Binary Logit (Inflated component)		2: Negative Binomial (Count component)	
Variable	Coeff.	P-value	Coeff.	P-value
Constant	0.41	0.000	1.3	0.000
Household Characteristics				
<i>Household Income</i> (Base: \$0 to \$49,999)				
G2: \$50,000 to \$69,999	-0.7	0.000	—	—
G3: \$70,000 to \$99,999	-0.92	0.000	—	—
G4: \$100,000 or more	-1.25	0.000	—	—
G9: Not stated	-0.5	0.027	—	—
Respondent Gender (Base: Male)	-0.08	0.000	-0.1	0.000
Respondent Age	0.1	0.000	-0.02	0.000
Household Size	0.04	0.000	-0.0005	0.796
<i>Interaction term</i> (Base: Income G1)				
Income G2 * Household Size	0.11	0.000	—	—
Income G3 * Household Size	0.1	0.000	—	—
Income G4 * Household Size	0.06	0.000	—	—
Income G9 * Household Size	-0.06	0.000	—	—
Zone Characteristics				
Log local accessibility	0.1	0.000	0.06	0.000
Log distance to airport	-0.05	0.000	-0.07	0.000
Log distance to rail station	0.05	0.000	-0.01	0.000
Trip Characteristics				
<i>Quarter</i> (Base: Winter)				
Q2: Spring	-0.3	0.000	0.05	0.000
Q3: Summer	-0.8	0.000	0.12	0.000
Q4: Fall	-0.16	0.000	0.07	0.000
Alpha			1.44	0.000
McFadden R²		0.0208		
Final log-likelihood, LL_F		-283127.4		
AIC		566312.87		
Note: The bolded numbers indicate coefficients that are statistically significant at the 0.01 level. = not significant.				

Both the Hurdle and ZI models yield statistically significant results for predictors consistent with prior research. However, the low McFadden R² values (around 2%) across models suggest that a substantial portion of the variability in disaggregate trip generation remains unexplained. This is particularly noteworthy given the very large and representative sample available. Taken together, while the ZI model outperforms the Hurdle model based on AIC and log-likelihood, both models consistently highlight the

general importance of demographic, zonal, and seasonal factors in influencing trip generation and exhibit broadly similar patterns. The modest explanatory power, however, points to the need for richer data to more accurately capture the complexity of individual travel behavior.

5.1.3 Aggregate Trip Generation: Hurdle Model Results

While the overall movement in travel demand modeling is towards more disaggregate models, the variability present in the above regression models despite very large sample size suggests the possibility of reconsidering aggregate models which lack the individual behavior information but still embody the spatial, regional and corridor level information needed for infrastructure and system service planning. Table 14 presents the results for the aggregate (OD-based) trip generation model using a Hurdle framework. The study and data source facilitate a large sample size of approximately 16 million zone-to-zone OD pairs for modeling (43,411 non-zero). Notably, the aggregate model achieves a McFadden R^2 of 0.31, substantially higher than that of the disaggregate trip generation models, indicating better explanatory power, a value (0.31) that falls within what others have deemed an excellent fit (Domencich & McFadden, 1975; McFadden, 1977). Despite a moderate correlation (approximately 0.5) between population, and accessibility variables, the VIF values are approximately 1, with a lower AIC and improved R^2 , confirming that multicollinearity is not a concern. Overdispersion is well captured by the significant alpha (1.65).

The Binary Logit component, which models the probability of trip occurrence between OD pairs, shows that log population at both origin and destination zones has a positive and statistically significant effect on the likelihood of trip generation as would be expected as more populated zones are more likely to generate more trips. In contrast, log local accessibility at both origin and destination exhibits a negative effect, implying that when residents have better access to local services and amenities, they are less likely to travel outside their zones, thereby reducing inter-zonal trips. Similarly, greater distance to the nearest airport at both the origin and destination is associated with a lower probability of trip generation, likely reflecting reduced regional connectivity. Moreover, the log of great circle distance between OD pairs has a strong and negative effect, reaffirming that spatial separation significantly discourages trip formation. These factors all suggest traditional gravity modeling paradigms might still be appropriate for LD modeling. In the binary (trip-occurrence) stage, destination population (1.13) is slightly more influential than origin population (0.98), suggesting that large urban attractors drive whether any long-distance trips occur.

It is worth noting that travel time by automobile and air was initially both considered as predictors; however, due to their high correlation with great circle distance, they were removed from the model and distance was used. In the Negative Binomial component, which models the trip count for zones that generate trips, similar

patterns are observed: the log-transformed population at the destination remains a positive and significant determinant of trip count. Local accessibility at both the origin and destination continues to negatively influence trip counts, while great circle distance is no longer significant in determining the number of trips once they occur. This highlights a key distinction: distance and accessibility are more influential in determining whether a trip occurs, rather than how many. Distance to the airport at origin reduces the likelihood of any trip (−0.37) but increases trip counts for those who do travel (+0.07). This flip suggests a “barrier then bundling” effect: people far from airports travel less often, but when they do, they tend to bundle multiple trips into a single outing (perhaps to justify the overhead of reaching destinations). Both the binary and count components show more negative elasticities for destination accessibility (−0.65 and −0.14) than for origin (−0.36 and −0.08). This indicates that long-distance trips are motivated by needs and desires as would be expected.

Table 14 Estimation Results for Aggregate Trip Generation (Hurdle)

	1: Binary Logit		2: Negative Binomial	
Variable	Coeff.	P-value	Coeff.	P-value
Constant	2.31	0.000	2.47	0.000
Zone Characteristics				
<i>Log population at Origin</i>	0.98	0.000	0.46	0.000
<i>Log distance to airport at Origin</i>	-0.37	0.000	0.07	0.000
<i>Log local accessibility at Origin</i>	-0.36	0.000	-0.08	0.000
<i>Log population at Destination</i>	1.13	0.000	0.41	0.000
<i>Log distance to airport at Destination</i>	-0.26	0.000	-0.03	0.000
<i>Log local accessibility at Destination</i>	-0.65	0.000	-0.14	0.000
Trip Characteristics				
<i>Log Great Circle Distance</i>	-1.92	0.000	—	—
Alpha			1.65	0.000
McFadden R²		0.31		
Final log-likelihood, LL_F		-552019.9		
AIC		1104069.7		
Note: The bolded numbers indicate coefficients that are statistically significant at the 0.01 level.				
— = not significant.				

5.2 Mode Choice Results

The estimation results for mode choice using MNL are summarized in Section 5.2.1, while Section 5.2.2 summarizes the ML results. Comparison between the MNL and machine learning (ML) models using the same test dataset is important as few LD comparisons have yet to be presented in the literature. Insights into how key explanatory variables influence mode choice are important for model applications in transportation and Random Forest approaches can achieve this relative to black-box ML models. This comparative additional analysis helps to highlight differences in variable importance and interpretation between the two modeling techniques.

5.2.1 Multinomial Logit

The MNL results (Table 15) provide insights into how individual, and trip characteristics influence the utility (and, hence, the probability) of choosing each travel mode, with “driving” as the base alternative. Note in particular that most LD mode choice models do not use travel time as an alternative specific variable but here travel time has been estimated and included for each mode. All coefficients in Table 15 represent the change in utility for mode j relative to driving when the corresponding variable increases by one unit, holding all else constant. For each coefficient, the corresponding odds ratio (OR) is also presented to facilitate interpretation. Every discussion of “increase” or “decrease” is therefore about the relative attractiveness of that mode compared to driving. The model achieves a McFadden R^2 of 0.45, which indicates very strong explanatory power for a discrete-choice specification.

All alternative-specific constants are negative, which means the base utility for choosing air, bus, train is lower than driving affirming the relatively higher inherent preference for driving travel. Among individual characteristics, **age** exerts a small but statistically significant deterrent effect on every non-auto mode: each additional year reduces the odds of choosing air by 4% ($\beta = -0.04$, OR = 0.96), bus by 5% ($\beta = -0.05$, OR = 0.95) and train by 9% ($\beta = -0.09$, OR = 0.91), suggesting older travelers increasingly favor driving. **Gender** has no significant impact on air but strongly shapes bus and train choices: women are 54% more likely than men to take the bus ($\beta = 0.43$, OR = 1.54) and 34% more likely to take the train ($\beta = 0.29$, OR = 1.34) relative to driving. This indicates a gendered preference in travel behavior for these particular modes. **Household Income** reveals a clear stratification: those earning \$100 k + see a 31% increase in flying odds ($\beta = 0.27$, OR = 1.31) but a 73% drop in bus use ($\beta = -1.30$, OR = 0.27), reflecting elevated flight propensity and diminished bus use at higher incomes. This divergence underscores income-driven mode stratification suggesting wealth enables flight while discouraging lower-cost bus travel. Intermediate income groups (G2, G3) follow the same pattern to lesser degrees.

Table 15 Mode Choice Estimation Results N = 153,713

Variable	Air	Bus	Train
Intercept	-1.2	-1.3	-0.03
Individual Characteristics			
Age	-0.04	-0.05	-0.09
Respondent Gender (Base: Male)	—	0.43	0.29
Household Income (Base: \$0 to \$49,999)			
G2: \$50,000 to \$69,999	-0.13	-0.54	-0.14
G3: \$70,000 to \$99,999	-0.16	-0.96	-0.04
G4: \$100,000 or more	0.27	-1.3	-0.005
G9: Not stated	-0.13	-1.12	0.2
Trip Characteristics			
Delta Travel Time relative to driving	-0.95	-0.04	-0.02
Quarter (Base: Winter)			
Q2: Spring	0.1	-0.29	-0.39
Q3: Summer	-0.41	-0.51	-0.68
Q4: Fall	0.05	0.07	-0.16
Trip Purpose (Base: Vacation)			
P2: Visit family	0.04	-0.6	-0.14
P3: Non-routine Business	2.01	0.69	1.36
P4: Other	-0.12	-0.45	-0.48
Zone Characteristics			
Log local accessibility at Origin	0.4	0.42	0.1
McFadden R ²	0.45		
Final log-likelihood, LL _F	-119053.6		
AIC	238195.2		
Note: The bolded numbers indicate coefficients that are statistically significant at the 0.01 level.			

Travel Time enters as an alternative-specific variable in the form of Delta Travel Time (ΔTT), defined as the difference between the travel time of each non-auto mode and that of driving (e.g., $\Delta TT \text{ Air} = TT \text{ air} - TT \text{ driving}$): the sign and magnitude of ΔTT determine how strongly each mode's utility is adjusted. When ΔTT is positive (the non-auto mode is slower than driving), the product of the negative coefficient and ΔTT is negative, thereby penalizing the utility of that mode and making driving more likely. Conversely, when ΔTT is negative (the non-auto mode is faster than driving), multiplying the negative coefficient by a negative ΔTT yields a positive contribution to utility, increasing the attractiveness of that mode relative to

driving. The larger the absolute coefficient, the stronger this adjustment: for example, air travel ($\beta = -0.95$) experiences a steep utility gain when it is faster than driving, and a steep penalty when it is slower, while bus ($\beta = -0.04$) and train ($\beta = -0.02$) show much weaker sensitivity to relative time differences. **Seasonality** also matters: in Spring, bus and train odds fall by 25% ($\beta = -0.29$, OR = 0.75) and 32% ($\beta = -0.39$, OR = 0.68,) relative to Winter, with no significant effect on Air; Summer reduces non-auto mode odds by 34% for Air ($\beta = -0.41$, OR = 0.66), 40% for Bus ($\beta = -0.51$, OR = 0.60) and 49% for Train ($\beta = -0.68$, OR = 0.51), perhaps due to driving conditions or recreational purposes. Fall effects are small and not significant. These patterns signal operator and traveler sensitivities to time of year and weather variations.

Trip purpose significantly influence travel mode decisions. Non-routine business trip multiplies air odds by 7.5 ($\beta = 2.01$, OR = 7.46), bus by 2.0 ($\beta = 0.69$, OR = 1.99) and train by 3.9 ($\beta = 1.36$, OR = 3.90) compared to vacation trips, underscoring the heightened utility and preference for scheduled, reliable modes in business contexts where expenses are likely reimbursed. Conversely, visits to family slightly increase air utility ($\beta = 0.04$) but significantly reduces bus odds by 45% ($\beta = -0.60$, OR = 0.55). Trips categorized as 'Other' experience considerable reductions in utility across all modes, reflecting lower propensity to select non-auto modes for miscellaneous travel purposes. Finally, zone characteristics matter: a one-unit increase in log local accessibility at the trip origin raises the odds of choosing air by 49% ($\beta = 0.40$, OR = 1.49) and Bus by 52% ($\beta = 0.42$, OR = 1.52), with a smaller increase for train ($\beta = 0.10$). This suggests that well-connected, urban and transit-friendly origins bolster the probability of travelers opting for alternative modes to driving. This demonstrates the importance of the overall infrastructure systems in promoting different travel behavior. Overall, these patterns confirm as expected that Automobile is the strong baseline mode, older travelers decreased preference for non-auto modes, women are more likely to use bus and train, higher incomes favors flying but discourage bus use. Ultimately, relative to driving, long-distance mode choice in Canada is most strongly driven by travel-time sensitivities (especially for air), trip purpose (business travel), demographic heterogeneity (age and gender), income stratification, seasonality effects, and local transit accessibility.

5.2.2 Machine Learning

Table 16 compares the performance of different ML modeling techniques for mode choice prediction using an oversampling strategy to balance the four classes (driving, air, bus, train). While Multinomial Logit (MNL) is typically evaluated using log-likelihood and McFadden R^2 , in order to enable a direct comparison with machine learning classifiers, the predicted mode was defined as the alternative with the highest estimated utility (i.e., highest predicted probability) for each observation. MNL model serves as a reference. All Machine Learning (ML) techniques outperform MNL in overall accuracy, with Random Forest (RF)

and Decision Tree (DT) showing the highest accuracy (the proportion of correctly classified trips) and F1-score while Neural Networks (NN) provide a more balanced performance across all modes, which might be beneficial if estimating all four modes are a priority. All machine learning models perform well on predicting air mode. The machine learning models, particularly RF, and DT achieve high accuracy (around 0.98 for RF and 0.96 for DT) and handle non-linear interactions, whereas MNL shows moderate performance for the automobile mode. Bus mode is the most challenging mode, most of the wrong predictions are driving or train. The low performance for the bus mode across most models suggests that bus trips may be underrepresented or inherently more difficult to predict possibly due to limited sample size and the challenges estimating accurate travel times. Notably, the Support Vector Machine (SVM) model failed to accurately predict test instances for bus mode choice. One plausible explanation is the sensitivity of SVM to feature scaling, as it typically requires standardized or normalized input variables to achieve optimal performance. However, due to the consistent use of the same dataset across all models without specific preprocessing adjustments, this step was intentionally omitted, which may have impacted SVM's predictive capability. Air trips are most accurately predicted with every ML model exceeding 90 % accuracy for this class. This high performance may reflect interactions amongst distinctive characteristics of air trips (longer travel times, higher traveler incomes, and frequent association with business travel), which ML is better able to capture compared to MNL.

Table 16 Models Performance for Mode Choice Prediction with Oversampling Strategy

Technique	Total accuracy (F1_average)	Accuracy of Air	Accuracy of Driving	Accuracy of Bus	Accuracy of Train
Multinomial Logit	0.69(0.75)	0.52	0.71	0.28	0.41
SVM	0.87 (0.9)	0.89	0.89	0.02	0.48
Random Forest	0.95 (0.94)	0.89	0.98	0.11	0.23
Neural Network	0.78 (0.86)	0.9	0.79	0.36	0.65
DT	0.93 (0.93)	0.8	0.96	0.24	0.21
Note: The bold numbers indicate better accuracy. For reference, a random guesser following the observed distribution would achieve accuracies of 69.7% (Car), 30.2% (Air), 0.09% (Bus), and 0.03% (Train).					

It is well known that the interpretability of an MNL model often contrasts with better predictive power of ML. To consider the machine learning models' interpretability, results of feature importance ranking by random forest are shown in Figure 12. The X-axis represents the Mean Decrease in Gini, quantifying each

feature's contribution to reducing impurity across the trees within the Random Forest model. Higher values indicate greater importance in determining the choice of travel mode. In other words, a larger Mean Decrease in Gini value indicates that a feature more effectively separates trips into the correct mode classes. The Random Forest model identifies travel time (especially driving) as the primary determinant of long-distance mode choice, more important than sociodemographic and seasonal variables, aligning with the MNL results. This suggests that individuals prioritize convenience and duration when choosing a mode, and interventions that reduce travel time may have the most impact on shifting behavior. Log of the distance to airport also holds moderate importance, highlighting the influence of transit accessibility at origin points. Age has a visible effect, more than other demographic features, indicating systematic mode preferences among different age groups.

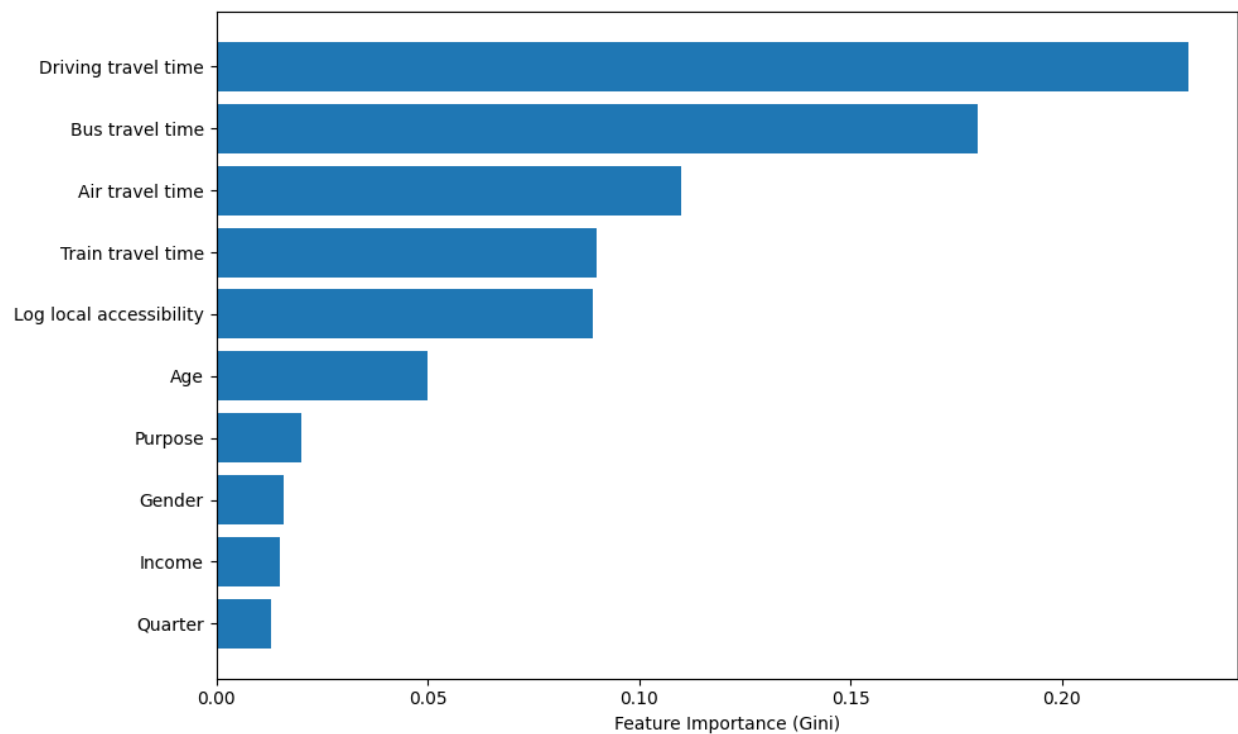


Figure 12 Feature Importance in Random Forest Mode Choice Model

Chapter 6 Conclusions

Long-distance (LD) travel remains understudied compared to daily urban travel, yet it represents a significant portion of overall travel distances and raises distinct issues related to sustainability and social equity. This thesis examined LD travel behavior across Canada using comprehensive public survey data collected between 2018 and 2020. Two primary types of modeling were explored with this very large national dataset: trip generation (TG) and mode choice, both evaluated with alternative methodological approaches. To facilitate these analyses, an accurate multimodal national transportation network, encompassing travel times by airplane, driving, bus, and intercity rail, was established. In addition to providing critical insights into LD travel behavior in Canada, this research addresses methodological gaps and informs future LD research. The national multimodal network and comparative methodological analyses demonstrate that more national LD modeling is viable to evaluate effective transportation policies aimed at infrastructure, sustainability and equity planning.

6.1 Key Findings

Multimodal National Network Development: A significant challenge in modeling LD travel stems from inherent large geography and sparse data availability. This research successfully developed a detailed national transportation model incorporating origin-destination zones and multimodal networks. The resulting model effectively allows computation of travel times across Canada for airplane, driving, bus, and intercity rail, demonstrating sufficient accuracy to support comprehensive LD travel analysis. Importantly, we demonstrate that national-scale multimodal models can be built efficiently, paving the way for scenario testing or what-if analyses. Efforts to estimate mode specific travel time as undertaken here with publicly available data are recommended.

Influence of Accessibility on Trip Generation: While socioeconomic variables commonly underpin LD trip generation models, the explicit role of accessibility has rarely been explored beyond basic urban-rural contrasts. This thesis filled this gap by evaluating accessibility's impact on LD trips. Results revealed that lower local accessibility (greater difficulty accessing nearby amenities) increases the propensity to make a LD trips, while poor access to an airport (a possible proxy measure for being in an urban area) was also associated with a higher number of LD trip. Both results underscore the importance of considering accessibility when addressing travel behavior and also transportation equity. LD maybe a burden to some and a benefit to others.

Comparative Evaluation of Aggregate and Disaggregate Trip Generation Models: Two methodological frameworks: aggregate OD zone-based models and disaggregate individual-based models were compared. Findings showed that both models consistently emphasize demographic, zonal, and seasonal factors. However, the low McFadden R^2 values for disaggregate models highlight considerable unexplained variance, reflecting the inherent complexity of individual LD travel behavior. Often the need for better and larger surveys to capture variation in LD trip making is recommended in survey-based studies. But the NTS survey was very large ($N=234,835$ in just this extract), and the model fit with a large number of typical predictor variables was dismal. This points to the need to seek different methods and different data for LD modeling. The aggregate gravity style modeling here performed better. Richer, more granular but still aggregate data sources in future research might include passive cellular or GPS data because even with very large survey sample sizes such as the NTS variability in trip making cannot be captured with the current covariates. Moreover, qualitative studies to better understand why some travel and some do not could potentially provide deeper insights into disaggregate travel patterns and improve future data collection and ultimately model performance.

Mode Choice Methodologies: Two distinct modeling strategies were examined for mode choice: traditional discrete choice Multinomial logit (MNL) and machine learning (ML) methods. Predictors had the expected significance and direction. However, ML approaches (particularly Decision Trees and Random Forrest) delivered superior predictive. These findings are consistent with daily travel model results and illustrate the potential advantages of ML in forecasting LD mode choice behaviors and highlight a crucial insight: characteristics of travel modes themselves (e.g., travel time) exert greater influence on mode choice than demographic factors alone (see Figure 12).

By combining MNL's behavioral interpretability with ML's predictive power, these results suggest a dual-track framework:

- **Use ML** (e.g., RF, NN) for higher accuracy prediction in network-wide simulations and scenario analysis for comparison between air and driving.
- **Use MNL** when the focus is on policy insights and behavioral interpretation, especially for evaluating elasticities or marginal effects.

6.2 Broader Implications

In the travel demand research community, the relative lack of LD research despite its importance is shocking. LD travel in several countries has been studied often with dated or limited data. This thesis adds

a comprehensive analysis from Canada, a large country with significant domestic LD travel and a long-standing LD travel survey. The implications for research and planning are numerous:

- I. National LD modeling is viable and could be used to support infrastructure and service planning.
- II. Lower-income households travel significantly less, signaling a potential equity issue. Travel surveys by their nature do not include unrealized travel demand. Targeted planning or enhanced services could mitigate an LD mobility gap which is mostly unmeasured.
- III. Findings from aggregate trip generation show that distance and accessibility are significantly more influential in determining whether trips occur at all, but not how many trips are taken.
- IV. Local accessibility strongly influences LD trip initiation: residents with nearby opportunities are less inclined to undertake LD travel, also signaling a potential equity issue. Needs may be unmet due to lack of access.
- V. Ironically, increased distance to airports and rail, also increase LD trips, supporting the need to better evaluate how LD modes interact.
- VI. Despite statistically significant findings, substantial unexplained variance remains in disaggregate trip generation analyses, suggesting the need for new types of data beyond surveys and new frameworks of LD modeling to support policy and planning questions.
- VII. Travel times matter most for mode choice. If aiming to influence mode choice (e.g., shift from driving to other modes), reducing travel time or perceived effort is likely the most effective lever.

6.3 Study Limitations

This study has several limitations that should be acknowledged. First, the National Travel Survey (NTS) used for analysis documents detailed information for a maximum of three trips per respondent. While this reduces respondent burden, it limits the completeness of observed travel behavior and may omit some important long-distance trips. Second, the dataset lacks key explanatory variables, including car ownership (an important predictor highlighted in the literature). Third, LD encompasses a wide range of trip types, from optional leisure travel to mandatory medical visits and non-routine business trips. Each of these categories is influenced by different sets of explanatory variables, yet in this study they are all treated within a single survey questions and therefore modeling framework. The definition of a LD trip was defined by the survey program and 40km is not necessarily an appropriate threshold for all applications. This simplification may reduce the model's ability to capture trip-type-specific behavioral drivers. Fourth, for the aggregate (zone-based) trip-generation models, we relied on population-weighted centroids to represent all trip origins and destinations. Because zone areas vary substantially, this approach can introduce spatial

inaccuracies in great-circle distance, travel time and accessibility measures. Fifth, the mode share in the dataset is highly imbalanced, particularly for train and bus trips, which are underrepresented. To address this, oversampling techniques were applied; however, the small sample size for these modes may still limit the model's ability to accurately capture their choice patterns. Lastly, all models assume static travel preferences, whereas in reality, individuals' decisions may shift over time due to changing contexts, personal habits, or recent travel experiences.

6.4 Recommendations for Future Research

Future studies could benefit from incorporating post-COVID travel data to capture the pandemic's lasting impacts on long-distance behavior. More advanced discrete choice modeling methods, such as Mixed Logit models, might better capture individual-level heterogeneity in travel behavior and decision-making processes including relative travel times. Assessing specific high-demand corridors, such as Toronto–Montreal, using this data could provide valuable insights into corridor-level dynamics and policy implications. Additionally, future research should consider integrating a broader range of behavioral and built-environment variables such as vehicle ownership, detailed land-use characteristics, interaction terms, and ratios of modal travel times to enhance model performance and explanatory power. In particular, incorporating data on business locations from Desktop Mapping Technologies Inc (DMTI) Spatial's "Points of Interest" dataset could improve the calculation of accessibility measures and enable more nuanced evaluation of their role in long-distance travel behavior.

Incorporating emerging passive data sources into LD travel analyses, such as mobile phone or GPS datasets, can further improve the accuracy and spatial-temporal granularity of trip generation and mode choice analyses. To enhance the interpretability of machine learning models in mode choice studies, future work should consider using techniques like SHAP (SHapley Additive exPlanations) values [111], which can help clarify and quantify the contribution of each predictor.

References

- [1] K. W. Axhausen, "Methodological research for a European survey of long-distance travel," in *Personal Travel: the Long and Short of It. Conference Proceedings June 28–July 1, 1999 Washington, DC*, 2001, pp. 321–342.
- [2] J. Dargay and S. Clark, "The determinants of long distance travel: an analysis based on the British national travel survey," in *12th World Congress on Transport Research*, 2010.
- [3] L. Aultman-Hall, "Incorporating long-distance travel into transportation planning in the United States: A White Paper from the National Center for Sustainable Transportation," 2018.
- [4] A. Schulz, M. Magdolen, B. Chlond, T. Kuhnimhof, and K. Köhler, "Identifying and understanding long-distance travel demand by combining official transport statistics and survey data," *Transp. Res. Procedia*, vol. 76, pp. 458–469, 2024.
- [5] R. M. Berliner, L. Aultman-Hall, and G. Circella, "Exploring the Self-Reported Long-Distance Travel Frequency of Millennials and Generation X in California," *Transp. Res. Rec.*, vol. 2672, no. 47, pp. 208–218, Dec. 2018, doi: 10.1177/0361198118798478.
- [6] M. Czepkiewicz, J. Heinonen, P. Næss, and H. Stefansdóttir, "Who travels more, and why? A mixed-method study of urban dwellers' leisure travel," *Travel Behav. Soc.*, vol. 19, pp. 67–81, 2020, doi: <https://doi.org/10.1016/j.tbs.2019.12.001>.
- [7] P. S. Hu, T. Reuscher, and others, "Summary of travel trends: 2001 national household travel survey," Prepared for US Department of Transportation Federal Highway Administration, Washington, DC, 2004.
- [8] J. Dowds, L. Aultman-Hall, and J. J. LaMondia, "Comparing alternative methods of collecting self-assessed overnight long-distance travel frequencies," *Travel Behav. Soc.*, vol. 19, pp. 124–136, 2020.
- [9] J. J. LaMondia, M. Moore, and L. Aultman-Hall, "Modeling intertrip time intervals between individuals' overnight long-distance trips," *Transp. Res. Rec.*, vol. 2495, no. 1, pp. 23–31, 2015.
- [10] S. C. Government of Canada, "Results from the 2016 Census: Long commutes to work by car." Accessed: Jul. 03, 2025. [Online]. Available: <https://www150.statcan.gc.ca/n1/pub/75-006-x/2019001/article/00002-eng.htm>
- [11] M. Mendonca, J. Schabas, K. Kellam, and P. Forestell, "Smart Commute Workplace Program Business Case Review,". Presented at 95th Annual Meeting of the Transportation Research Board, Washington, D.C., pp. 1–18, 2016.
- [12] M. Outwater, M. Bradley, N. Ferdous, S. Trevino, and H. Lin, "Foundational Knowledge to Support a Long-Distance Passenger Travel Demand Modeling Framework: Implementation Report," FWH Exploratory Advanced Research Program NO. DTFH61-10-R-00036, 2015.
- [13] J. S. Chang, D. Jung, J. Kim, and T. Kang, "Comparative analysis of trip generation models: results using home-based work trips in the Seoul metropolitan area," *Transp. Lett.*, vol. 6, no. 2, pp. 78–88, 2014.
- [14] L. Frändberg and B. Vilhelmson, "Personal mobility: a corporeal dimension of transnationalisation. The case of long-distance travel from Sweden," *Environ. Plan. A*, vol. 35, no. 10, pp. 1751–1768, 2003.
- [15] J. J. LaMondia, L. Aultman-Hall, and E. Greene, "Long-Distance Work and Leisure Travel Frequencies: Ordered Probit Analysis Across Non-Distance-Based Definitions," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2413, no. 1, pp. 1–12, Jan. 2014, doi: 10.3141/2413-01.
- [16] A. Aguiléra and L. Proulhac, "Socio-occupational and geographical determinants of the frequency of long-distance business travel in France," *J. Transp. Geogr.*, vol. 43, pp. 28–35, 2015.
- [17] L. Aultman-Hall, C. Harvey, J. Sullivan, and J. J. LaMondia, "The implications of long-distance tour attributes for national travel data collection in the United States," *Transportation*, vol. 45, pp. 875–903, 2018.

- [18] H. Alizadeh, C. Morency, and M. TRÉPANIÉ, “Long-Distance Travel Demand Modeling Through Rare Event Modeling Approach,” *Available SSRN 4009461*, 2022.
- [19] W. A. O’Neill and E. Brown, “Long-distance trip generation modeling using ATS,” in *conference Personal Travel: The Long and Short of It, Transportation Research Board, June*, 1999.
- [20] C. Llorca, J. Ji, J. Molloy, and R. Moeckel, “The usage of location based big data and trip planning services for the estimation of a long-distance travel demand model. Predicting the impacts of a new high speed rail corridor,” *Res. Transp. Econ.*, vol. 72, pp. 27–36, 2018.
- [21] A. Páez, D. M. Scott, and C. Morency, “Measuring accessibility: positive and normative implementations of various accessibility indicators,” *J. Transp. Geogr.*, vol. 25, pp. 141–153, 2012.
- [22] K. Maria Kockelman, “Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area,” *Transp. Res. Rec.*, vol. 1607, no. 1, pp. 116–125, 1997.
- [23] B. Van Wee, K. Geurs, and C. Chorus, “Information, communication, travel behavior and accessibility,” *J. Transp. Land Use*, vol. 6, no. 3, pp. 1–16, 2013.
- [24] A. M. El-Geneidy and D. M. Levinson, “Access to destinations: Development of accessibility measures,” Technical Report, Minnesota Department of Transportation Research Services Section, 2006.
- [25] H. B.ENZLER, “Air travel for private purposes. An analysis of airport access, income and environmental concern in Switzerland,” *J. Transp. Geogr.*, vol. 61, pp. 1–8, 2017, doi: <https://doi.org/10.1016/j.jtrangeo.2017.03.014>.
- [26] S. H. Kim and P. L. Mokhtarian, “Who (never) makes overnight leisure trips? Disentangling structurally zero trips from usual trip generation processes,” *Travel Behav. Soc.*, vol. 25, pp. 78–91, Oct. 2021, doi: 10.1016/j.tbs.2021.05.011.
- [27] L. Aultman-Hall, “Do regional and extra-regional accessibility affect long-distance intercity travel? Insights from California,” 2024.
- [28] K. T. Geurs and B. Van Wee, “Accessibility evaluation of land-use and transport strategies: review and research directions,” *J. Transp. Geogr.*, vol. 12, no. 2, pp. 127–140, 2004.
- [29] R. GRIMAL, “Spatial analysis of long-distance mobility,” in *Proceedings of the 12th World Conference on Transport Research. Lisbon, Portugal*, 2010.
- [30] R. Moeckel, R. Donnelly, and P. Brinckerhoff, “Nationwide Estimate of Long-Distance Travel (NELDT),” in *Third International Conference on Innovations in Travel Modeling (ITM) of the Transportation Research Board, Tempe, AZ*, 2010.
- [31] A. M. Kim and M. S. Ryerson, “A long drive: Interregional airport passenger ‘leakage’ in the US,” *Tour. Manag.*, vol. 65, pp. 237–244, 2018.
- [32] M. E. Ben-Akiva and S. R. Lerman, *Discrete choice analysis: theory and application to travel demand*, vol. 9. MIT press, 1985.
- [33] D. McFadden, “Conditional logit analysis of qualitative choice behavior,” Zarembka P. (Ed.), *Frontiers in Econometrics*, Academic Press (1974), pp. 105-142 1972.
- [34] F. S. Koppelman and C.-H. Wen, “Alternative nested logit models: structure, properties and estimation,” *Transp. Res. Part B Methodol.*, vol. 32, no. 5, pp. 289–298, 1998.
- [35] H. C. Williams, “On the formation of travel demand models and economic evaluation measures of user benefit,” *Environ. Plan. A*, vol. 9, no. 3, pp. 285–344, 1977.
- [36] C. Daganzo, *Multinomial probit: the theory and its application to demand forecasting*. Elsevier, 2014.
- [37] D. McFadden and K. Train, “Mixed MNL models for discrete response,” *J. Appl. Econom.*, vol. 15, no. 5, pp. 447–470, 2000.
- [38] J. K. Dow and J. W. Endersby, “Multinomial probit and multinomial logit: a comparison of choice models for voting research,” *Elect. Stud.*, vol. 23, no. 1, pp. 107–122, 2004.
- [39] C. Román *et al.*, “Valuation of travel time savings for intercity travel: The Madrid-Barcelona corridor,” *Transp. Policy*, vol. 36, pp. 105–117, 2014.

- [40] D. A. Hensher and W. H. Greene, "The mixed logit model: the state of practice," *Transportation*, vol. 30, pp. 133–176, 2003.
- [41] M. Lu, H. Zhu, X. Luo, and L. Lei, "Intercity travel demand analysis model," *Adv. Mech. Eng.*, vol. 6, p. 108180, 2014.
- [42] H. Baik, A. A. Trani, N. Hinze, H. Swingle, S. Ashiabor, and A. Seshadri, "Forecasting model for air taxi, commercial airline, and automobile demand in the United States," *Transp. Res. Rec.*, vol. 2052, no. 1, pp. 9–20, 2008.
- [43] R. Moeckel, R. Fussell, and R. Donnelly, "Mode choice modeling for long-distance travel," *Transp. Lett.*, vol. 7, no. 1, pp. 35–46, Feb. 2015, doi: 10.1179/1942787514Y.0000000031.
- [44] M. Bradley, N. Ferdous, M. L. Outwater, S. Hess, A. Daly, C.R. Bhat, R.M. Pendyala, and J. J. LaMondia, "A Tour-Based National Model System to Forecast Long Distance Passenger Travel," 2014.
- [45] L. Wu, J. Zhang, and A. Fujiwara, "Dynamic analysis of Japanese tourists' three stage choices: tourism participation, destination choice, and travel mode choice," *Transp. Res. Rec.*, vol. 2322, no. 1, pp. 91–101, 2012.
- [46] M. Juschten and R. Hössinger, "Out of the city—but how and where? A mode-destination choice model for urban–rural tourism trips in Austria," *Curr. Issues Tour.*, vol. 24, no. 10, pp. 1465–1481, 2021.
- [47] M. A. A. B. Miskeen, A. M. Alhodairi, and R. A. A. B. O. Rahmat, "Modeling a multinomial logit model of intercity travel mode choice behavior for all trips in Libya," *Int. J. Civ. Environ. Eng.*, vol. 7, no. 9, pp. 636–645, 2013.
- [48] C. Llorca, J. Molloy, J. Ji, and R. Moeckel, "Estimation of a Long-Distance Travel Demand Model using Trip Surveys, Location-Based Big Data, and Trip Planning Services," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2672, no. 47, pp. 103–113, Dec. 2018, doi: 10.1177/0361198118777064.
- [49] H. Omrani, "Predicting Travel Mode of Individuals by Machine Learning," *Transp. Res. Procedia*, vol. 10, pp. 840–849, 2015, doi: 10.1016/j.trpro.2015.09.037.
- [50] L. Tang, C. Xiong, and L. Zhang, "Decision tree method for modeling travel mode switching in a dynamic behavioral process," *Transp. Plan. Technol.*, vol. 38, no. 8, pp. 833–850, Nov. 2015, doi: 10.1080/03081060.2015.1079385.
- [51] X. (Michael) Chen, M. Zahiri, and S. Zhang, "Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach," *Transp. Res. Part C Emerg. Technol.*, vol. 76, pp. 51–70, Mar. 2017, doi: 10.1016/j.trc.2016.12.018.
- [52] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Syst. Appl.*, vol. 78, pp. 273–282, Jul. 2017, doi: 10.1016/j.eswa.2017.01.057.
- [53] F. Wang and C. L. Ross, "Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2672, no. 47, pp. 35–45, Dec. 2018, doi: 10.1177/0361198118773556.
- [54] X. Chang, J. Wu, H. Liu, X. Yan, H. Sun, and Y. Qu, "Travel mode choice: a data fusion model using machine learning methods and evidence from travel diary survey data," *Transp. Transp. Sci.*, vol. 15, no. 2, pp. 1587–1612, Nov. 2019, doi: 10.1080/23249935.2019.1620380.
- [55] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behav. Soc.*, vol. 14, pp. 1–10, Jan. 2019, doi: 10.1016/j.tbs.2018.09.002.
- [56] R. Chapleau, P. Gaudette, and T. Spurr, "Application of Machine Learning to Two Large-Sample Household Travel Surveys: A Characterization of Travel Modes," *Transp. Res. Rec.*, vol. 2673, no. 4, pp. 173–183, Apr. 2019, doi: 10.1177/0361198119839339.
- [57] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models," *Travel Behav. Soc.*, vol. 20, pp. 22–35, Jul. 2020, doi: 10.1016/j.tbs.2020.02.003.

- [58] E.-J. Kim, “Analysis of Travel Mode Choice in Seoul Using an Interpretable Machine Learning Approach,” *J. Adv. Transp.*, vol. 2021, pp. 1–13, Mar. 2021, doi: 10.1155/2021/6685004.
- [59] H. Naseri, E. O. D. Waygood, Z. Patterson, M. Alousi-Jones, and B. Wang, “Travel Mode Choice Prediction: Developing New Techniques to Prioritize Determinant and Interpret Black-Box Machine Learning Techniques,” *SSRN Electron. J.*, 2022, doi: 10.2139/ssrn.4035093.
- [60] X. Li *et al.*, “Exploring interactive and nonlinear effects of key factors on intercity travel mode choice using XGBoost,” *Appl. Geogr.*, vol. 166, p. 103264, May 2024, doi: 10.1016/j.apgeog.2024.103264.
- [61] M. Tamim Kashifi, A. Jamal, M. Samim Kashefi, M. Almoshaogeh, and S. Masiur Rahman, “Predicting the travel mode choice with interpretable machine learning techniques: A comparative study,” *Travel Behav. Soc.*, vol. 29, pp. 279–296, Oct. 2022, doi: 10.1016/j.tbs.2022.07.003.
- [62] F. S. Koppelman and V. Sethi, “Incorporating variance and covariance heterogeneity in the Generalized Nested Logit model: an application to modeling long distance travel choice behavior,” *Transp. Res. Part B Methodol.*, vol. 39, no. 9, pp. 825–853, Nov. 2005, doi: 10.1016/j.trb.2004.10.003.
- [63] L. Aultman-Hall, C. Harvey, J. Sullivan, and J. LaMondia, “The impact of long-distance tour generation and travel attributes on data collection in the United States,” *Transportation*, vol. 43, no. 6, 2016.
- [64] S. Hess, G. Spitz, M. Bradley, and M. Coogan, “Analysis of mode choice for intercity travel: Application of a hybrid choice model to two distinct US corridors,” *Transp. Res. Part Policy Pract.*, vol. 116, pp. 547–567, 2018.
- [65] G. Monchambert, “Why do (or don’t) people carpool for long distance trips? A discrete choice experiment in France,” *Transp. Res. Part Policy Pract.*, vol. 132, pp. 911–931, 2020.
- [66] N. Limtanakool, M. Dijst, and T. Schwanen, “The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips,” *J. Transp. Geogr.*, vol. 14, no. 5, pp. 327–341, Sep. 2006, doi: 10.1016/j.jtrangeo.2005.06.004.
- [67] P. van der Waerden and J. Van der Waerden, “The relation between train access mode attributes and travelers’ transport mode-choice decisions in the context of medium-and long-distance trips in the Netherlands,” *Transp. Res. Rec.*, vol. 2672, no. 8, pp. 719–730, 2018.
- [68] S. Lannoo, V. Van Acker, R. Kessels, D. P. Cuervo, and F. Witlox, “Getting business people on the coach: A stated preference experiment for intercity long distance coach travel,” *Transp. Res. Rec.*, vol. 2672, no. 8, pp. 165–174, 2018.
- [69] V. Van Acker, R. Kessels, D. P. Cuervo, S. Lannoo, and F. Witlox, “Preferences for long-distance coach transport: Evidence from a discrete choice experiment,” *Transp. Res. Part Policy Pract.*, vol. 132, pp. 759–779, 2020.
- [70] A. S. Bergantino and L. Madio, “Intermodal competition and substitution. HSR versus air transport: Understanding the socio-economic determinants of modal choice,” *Res. Transp. Econ.*, vol. 79, p. 100823, 2020.
- [71] S. C. Government of Canada, “National Travel Survey (NTS).” Accessed: Jan. 08, 2025. [Online]. Available: <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5232>
- [72] S. C. Government of Canada, “National Travel Survey.” Accessed: Dec. 03, 2024. [Online]. Available: https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=getInstrumentList&Item_Id=438439&UL=1V
- [73] S. C. Government of Canada, “National Travel Survey: Response Rate at the estimation stage - Q2 2018.” Accessed: Jun. 03, 2025. [Online]. Available: https://www.statcan.gc.ca/en/statistical-programs/document/5232_D1_V2
- [74] S. C. Government of Canada, “National Travel Survey: Response Rate at the estimation stage – Q4 2020.” Accessed: Jun. 03, 2025. [Online]. Available: https://www.statcan.gc.ca/en/statistical-programs/document/5232_D1_V11

- [75] S. C. Government of Canada, “National Travel Survey (NTS).” Accessed: Jan. 03, 2025. [Online]. Available: <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1504469>
- [76] C. R. Bhat and R. Gossen, “A mixed multinomial logit model analysis of weekend recreational episode type choice,” *Transp. Res. Part B Methodol.*, vol. 38, no. 9, pp. 767–787, Nov. 2004, doi: 10.1016/j.trb.2003.10.003.
- [77] S. C. Government of Canada, “Illustrated Glossary - Population centre (POPCTR).” Accessed: Jun. 03, 2025. [Online]. Available: <https://www150.statcan.gc.ca/n1/pub/92-195-x/2021001/geo/pop/pop-eng.htm>
- [78] H. Millward and J. Spinney, “Time use, travel behavior, and the rural–urban continuum: results from the Halifax STAR project,” *J. Transp. Geogr.*, vol. 19, no. 1, pp. 51–58, Jan. 2011, doi: 10.1016/j.jtrangeo.2009.12.005.
- [79] J. Pucher and J. L. Renne, “Rural mobility and mode choice: Evidence from the 2001 National Household Travel Survey,” *Transportation*, vol. 32, no. 2, pp. 165–186, Mar. 2005, doi: 10.1007/s11116-004-5508-3.
- [80] H. Wu and D. Levinson, “Unifying access,” *Transp. Res. Part Transp. Environ.*, vol. 83, p. 102355, Jun. 2020, doi: 10.1016/j.trd.2020.102355.
- [81] S. C. Government of Canada, “Measuring remoteness and accessibility: A set of indices for Canadian communities.” Accessed: Jun. 04, 2025. [Online]. Available: <https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2017002-eng.htm>
- [82] “Algorithms used by Network Analyst—ArcGIS Pro | Documentation.” Accessed: Jul. 12, 2025. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/help/analysis/networks/algorithms-used-by-network-analyst.htm?utm_source=chatgpt.com
- [83] “North American Roads.” Accessed: Oct. 31, 2024. [Online]. Available: <https://geodata.bts.gov/datasets/usdot:north-american-roads/about>
- [84] T. Transport Canada, “List of airports owned by Transport Canada,” AHPP 12832219. Accessed: Oct. 31, 2024. [Online]. Available: <https://tc.canada.ca/en/aviation/operating-airports-aerodromes/list-airports-owned-transport-canada>
- [85] “FlightStats - Global Flight Status & Tracker, Airport Weather and Delays,” FlightStats. Accessed: Oct. 31, 2024. [Online]. Available: <https://www.flightstats.com/v2/>
- [86] “FlightAware - Flight Tracker / Flight Status,” FlightAware. Accessed: Oct. 31, 2024. [Online]. Available: <http://www.flightaware.com/>
- [87] L. Aultman-Hall, A. Onayev, J. Dowds, and J. J. LaMondia, “Understanding the relationship between mode and destination choices for personal out-of-town travel,” *Travel Behav. Soc.*, vol. 31, pp. 399–409, 2023.
- [88] Bureau of Transportation Statistics, “OST_R | BTS | Transtats Fields.” Accessed: Jan. 07, 2025. [Online]. Available: https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FMG
- [89] Bureau of Transportation Statistics, “OST_R | BTS | Transtats Data Tables.” Accessed: Oct. 31, 2024. [Online]. Available: https://www.transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r0%FDb0-gvzr0%FDcr4s14zn0pr0%FDQn6n&QO_fu146_anzr=b0-gvzr
- [90] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [91] W. Song and Y. Zhou, “Road travel time prediction method based on random forest model,” in *Smart Trends in Computing and Communications: Proceedings of SmartCom 2019*, Springer, 2020, pp. 155–163.
- [92] H. Taghipour, A. B. Parsa, and A. K. Mohammadian, “A dynamic approach to predict travel time in real time using data driven techniques and comprehensive data sources,” *Transp. Eng.*, vol. 2, p. 100025, 2020.
- [93] J. de Dios Ortúzar and L. G. Willumsen, *Modelling transport*. John Wiley & sons, 2024.
- [94] D. C. Heilbron, “Zero-altered and other regression models for count data with added zeros,” *Biom. J.*, vol. 36, no. 5, pp. 531–547, 1994.

- [95] J. Mullahy, "Specification and testing of some modified count data models," *J. Econom.*, vol. 33, no. 3, pp. 341–365, 1986.
- [96] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.
- [97] H. Akaike, "Akaike's information criterion," *Int. Encycl. Stat. Sci.*, pp. 25–25, 2011.
- [98] S. Seabold, J. Perktold, and others, "Statsmodels: econometric and statistical modeling with python.," *SciPy*, vol. 7, no. 1, pp. 92–96, 2010.
- [99] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.
- [100] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.
- [101] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.
- [102] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [103] C. F. Manski, "The structure of random utility models," *Theory Decis.*, vol. 8, no. 3, p. 229, 1977.
- [104] K. E. Train, *Discrete choice methods with simulation*. Cambridge University press, 2009.
- [105] M. Bierlaire, "BIOGEME: A free package for the estimation of discrete choice models," in *Swiss transport research conference*, 2003.
- [106] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
- [107] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [108] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.
- [109] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [110] R. Kohavi, *Wrappers for performance enhancement and oblivious decision graphs*. stanford University, 1996.
- [111] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

Appendix A

Table 17 Activity Type in Visit File of the National Travel Survey

Item	Activity Type
1	Visit friends or family
2	Shopping
3	Sightseeing
4	Museum or art gallery
5	Historic site
6	Zoo or aquarium
7	Sports event as a specter
8	Festival or fair
9	Performance such as a play and concert
10	Casino
11	Theme or amusement park
12	Aboriginal event
13	Go to the movies
14	Dine out, restaurant, bur, club
15	Go wildlife viewing or bird watching
16	National, provincial or nature park
17	Go to medical or other health treatment
18	Business meeting, conference, seminar
19	All terrain vehicle
20	Boating
21	Canoeing or kayaking
22	Camping
23	Hiking or backpacking
24	Fishing
25	Beach
26	Hunting
27	Golfing
28	Cycling
29	Snowmobiling
30	Downhill skiing or snowboarding
31	Cross-country skiing or snowboarding
32	Play individual or team sport
33	Other activity
34	Did not do any activities

Appendix B

Table 18 List of Canadian Airports Included in The Analysis

Airport	Location (Province/Territory)	IATA Code	Category	Enplanements	Hub
Nanaimo	British Columbia	YCD	Other	121,341	M
Deer Lake	Newfoundland and Labrador	YDF	Other	163,012	M
Edmonton Intl	Alberta	YEG	National	3,927,952	M
Fredericton International	New Brunswick	YFC	National	204,807	M
La Grande Riviere	Quebec	YGL	Other	-	S
Iles-de-la-Madeleine	Quebec	YGR	Other	31,348	S
Hamilton	Ontario	YHM	Busiest	473,372	M
Halifax Robert L. Stanfield International	Nova Scotia	YHZ	National	2,062,298	M
Stephenville Dymond International	Newfoundland and Labrador	YJT	Other	-	S
Kitchener/Waterloo	Ontario	YKF	Busiest	445,312	M
Kelowna	British Columbia	YLW	National	988,745	M
Fort McMurray	Alberta	YMM	Busiest	300,758	M
Ottawa/Macdonald-Cartier International	Ontario	YOW	National	2,485,901	M
Québec City Jean Lesage International Airport	Quebec	YQB	National	863,217	M
Moncton/Greater Moncton International	New Brunswick	YQM	National	329,067	M
Regina International	Saskatchewan	YQR	National	582,257	M
Thunder Bay	Ontario	YQT	National	387,307	M
Grande Prairie	Alberta	YQU	Other	-	S
Gander International	Newfoundland and Labrador	YQX	National	79,842	S
JA Douglas McCurdy Sydney	Nova Scotia	YQY	Other	87,834	S
Sudbury	Ontario	YSB	Other	122,580	M
Saint John	New Brunswick	YSJ	National	-	S

Airport	Location (Province/Territory)	IATA Code	Category	Enplanements	Hub
Thompson	Manitoba	YTH	Other	84,696	S
Toronto City Centre	Ontario	YTZ	Busiest	-	M
Montréal-Pierre Elliott Trudeau International	Quebec	YUL	National	9,741,252	L
Val-d'Or	Quebec	YVO	Other	56,284	S
Vancouver International	British Columbia	YVR	National	12,867,297	L
Winnipeg James Armstrong Richardson International	Manitoba	YWG	National	2,195,576	M
Saskatoon John G. Diefenbaker International	Saskatchewan	YXE	National	739,949	M
Medicine Hat	Alberta	YXH	Other	38,625	S
Prince George	British Columbia	YXS	National	237,239	M
London	Ontario	YXU	National	340,514	M
Abbotsford	British Columbia	YXX	Busiest	503,232	M
Erik Nielsen Whitehorse	Yukon	YXY	National	183,092	M
Calgary International	Alberta	YYC	National	8,580,388	L
Charlottetown	Prince Edward Island	YYG	National	169,473	M
Victoria International	British Columbia	YYJ	National	927,534	M
St John's Intl	Newfoundland and Labrador	YYT	National	720,737	M
Mont-Joli	Quebec	YYY	Other	37,990	S
Toronto Pearson International	Ontario	YYZ	National	24,429,919	L
Yellowknife	Northwest Territories	YZF	Other	247,736	M