Technical University of Munich

Development of a Destination Choice Model for Ontario

by

Joseph Molloy

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the
Chair of Modeling Spatial Mobility
Department of Civil, Geo and Environmental Engineering

December 2016

Technical University of Munich – Professor for Modeling Spatial Mobility

Prof. Dr.-Ing. Rolf Moeckel

Arcisstraße 21, 80333 München, www.msm.bgu.tum.de

MASTER'S THESIS

of Joseph Molloy

Date of Issue: 2016-06-28

Date of Submission: 2016-12-23

<u>Topic:</u> Destination Choice Modelling for Long Distance Travel

A Long Distance Travel Model is being developed for the region of Ontario, Canada. This thesis will focus on the destination choice component of that model.

Some Tasks that should be included in this thesis are:

- Evaluation of the role of destination choice in long distance travel modelling literature.
- Review of the state of the art for such models.
- Analysis of the data availability and limitations for destination choice modelling for the Ontario region.
- Development of the Destination Choice model.
- Scenario analysis using the developed Destination Choice model.

The student will present intermediate results to the mentor Prof. Dr.-Ing. Rolf Moeckel in the fifth, tenth, 15th and 20th week.

The student must hold a 20-minute presentation with a subsequent discussion at the most two months after the submission of the thesis. The presentation will be considered in the final grade in cases where the thesis itself cannot be clearly evaluated.

Prof. Dr.-Ing. Rolf Moeckel

Technical University of Munich

Abstract

Chair of Modeling Spatial Mobility
Department of Civil, Geo and Environmental Engineering

Master of Science

by Joseph Molloy

Standard socioeconomic variables are not always accurate indicators of destination attractiveness in destination choice models. For example, areas such as ski resorts are significant attractors of long distance trips, yet have small populations and little employment. This thesis presents a destination choice model for long distance travel in Ontario that uses data from the location based social network, Foursquare, to address this issue. Points of interest and their historical check-in counts are collected and processed to define measures of destination attractiveness based on common long distance trip activities. The design, estimation and calibration of the multinomial logit model are covered in detail, and the implemented model used to perform a scenario analysis. The results show that big data can be successfully used to represent destination attractiveness, and that such an approach is particularly effective for modeling long distance leisure travel.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Prof. Rolf Moeckel, for his encouragement, knowledge and advice. Secondly, thank you to Ana Moreno and Carlos Llorca who also provided advice throughout the project. Rick Donnelly and the team at WSP Parsons Brinckerhoff must also be recognized for their always timely assistance supplying data on which this thesis depends. Finally, I would like to thank my Mum and Dad, family and friends around the world for their constant support.

Declaration

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

Munich, December 21st, 2016

Joseph Molloy

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information Criterion
API	Application Programming Interface
CD	Census Division
CMA	Census Metropolitan Agglomeration
EE	External Trip
EI	Incoming Trip
GIS	Geographical Information System
IE	Outgoing Trip
II	Internal Trip
ITS	International Travel Survey
JSON	JavaSript Object Notation
LBSN	Location Based Social Network
LL	Loglikelyhood
MNL	Multinomial logit (model)
MTO	Ministry of Transportation, Ontario
NAICS	North American Industry Classification System
NRMSE	Normalized Root Mean Square Error
OD	Origin-Destination
POI	Point of Interest
RMSE	Root Mean Squared Error
TAZ	Traffic Analysis Zone
TSRC	Travel Survey of Residents of Canada

Chapter 1

Introduction

The ability to travel long distances for business or personal reasons depends on large amounts of infrastructure such as highways, airports and railways. Such infrastructure is expensive, and can take decades and billions of dollars from the concept to implementation. It is also hard to alter after it is constructed. Therefore, is important to understand the potential impacts of such projects. To investigate these impacts, transport planners need to be able to observe, visualize, understand and forecast where, why, and how people travel.

These questions have been answered using the 4-step model since the 1960's (Dios Ortúzar, Willumsen, et al. 1994). The steps, in order, are trip generation, trip distribution, mode choice analysis, and route assignment. The classic 4 step model works as an aggregate process. In trip generation, the number of trips leaving each origin is calculated. In the second step, trip distribution, the number of trips between each origin and destination (OD) pair is calculated. In the third step, the mode share (of auto, bus, train) for each OD pair is calculated. Finally, a route assignment is performed to allocate these trips to the transport network. The disaggregate approach instead generates a list of trips for each origin, and using the 4 steps, assigns each a destination, mode and route. In disaggregate models, trip distribution is referred to as 'destination choice'.

The majority of the literature on transport demand modeling focuses on urban mobility and regular travel patterns such as commuting and shopping. It is important that the study of long distance travel behavior is not neglected. Rohr et al. (2010) found that in Great Britain, "Trips over 50 miles in length account for just 2.3% of all trips, but about a third of all distance traveled". Furthermore, with consensus on the existence of climate change in the scientific community (Oreskes 2004), the environmental impact of long distance trips needs to be quantified.

The Ministry of Transportation in Ontario (MTO) is in the process of developing a new provincial transport model, and one important component is the disaggregate model for long distance travel. Canada is made up of ten provinces and three territories, the largest of which is Ontario. Statistics Canada estimate that around 100 million long distance trips are performed every year that interact with Ontario, totally over 45 billion kilometers traveled (Statistics Canada 2014). The choice of destination is a key determinate of the length of a trip, and as such, an accurate destination choice model is vital to a working long distance model.

Big data is a "topic du jour" in the fields of data analysis and transport modeling. There is no clear boundary between normal data and big data. However, it has come to be defined by the "four v's": volume, velocity, variety and veracity (Beyer and Laney 2012). It is these characteristics that make it attractive as a data source for transportation modeling, particularly volume and veracity. Traditional transport models still rely on travel surveys and census data. Even comprehensive travel surveys such as the TSRC in Canada often have a sample size of only around 50,000 records per year. On the other hand, big data sources can track the movements of millions of individuals (volume), and provide unprecedented spatial and temporal accuracy (veracity). In some recent examples, GPS data is increasingly being incorporated into various transport models, particularly destination choice (Schönfelder et al. 2006; Pan et al. 2006) and route assignment (Broach, Dill, and Gliebe 2012; Menghini et al. 2010).

The choice of the destination made by a traveler is not necessarily made based on how many people live and work there. Although common parameters in destination choice models, these measurements are just simplifications for the complex attraction factors of a destination. National parks have no population, and little employment, but are large attractors of leisure trips. Ski areas are another example. Big data, particularly that available from social networks, present the opportunity to better represent the utility of alternatives in destination choice models. This thesis presents the development of a destination choice model for long distance travel in Ontario which incorporates data from the location based social network Foursquare to model destination utility.

The structure of this thesis is as follows: Chapter 2 presents an introduction to the topic and a review of the relevant literature; Chapter 3 details the data sources used in the model, and describes the methodology for the collection of Foursquare data; a gravity model is implemented in Chapter 4 as a baseline to judge the effectiveness of the destination choice model; Chapter 5 covers the estimation, implementation and application of a multinomial logit (MNL) model for destination choice; finally, Chapter 6 provides a discussion of the results and their relevance.

Chapter 2

Literature Review

Long-distance transport models, Also known as intercity models, were first proposed in the 1960s, with two of the earliest being developed in the United States and Canada respectively (Canadian Transport Commission 1971). These models were comparatively basic, with the demand component of the model only incorporating zonal population and income as attraction measures, and trip time, cost and convenience as impedance measures. More recent demand models include attributes such as auto ownership and household size.

Long-distance models are commonly defined to contain trips of certain length or longer, as opposed to the much more common urban model. TRB's NCHRP Report 735 notes that current state-wide models and travel surveys in the United States have used a range of thresholds to define long-distance trip-making, with "either 50, 75, or 100 miles as the minimum threshold for trips to be considered long-distance."(Schiffer 2012).

Per Miller (2004), a distinct class of intercity travel demand models exist, which have unique characteristics when compared to urban models. "An intercity travel demand model is designed to forecast travel demand between two or more urban areas ... rather than travel within a given urban region". He also highlights two main features of such models. Firstly, he argues that an intercity travel demand model should apply to a well-defined travel corridor, containing a small number of major cities. Secondly, he suggests that such models are almost always designed to model the impact of new travel modes such as high speed rail, or other policy initiatives.

Miller also notes that while urban models and methods are well documented in open literature, and applied in published policy analysis, intercity models are often the intellectual property of the consultants involved. The models are infrequently published in the scientific literature, and the travel data on private travel modes is often closely guarded, meaning that the models are hard to replicate, if they are published at all. It follows that "intercity travel demand models tend to be a less attractive/feasible application area for academic researchers than the more data-rich urban field." (Miller 2004).

2.1 Aggregate intercity transport models

Until the 1980s, intercity transport models were exclusively designed as aggregate models, which distributed trips between origin-destination pairs and modes using the gravity model proposed by Casey (1955). However, as early as 1962, the deficiencies in this approach were identified by numerous researchers (Oi 1962; Warner 1962). In 1967, Wilson (1967) first proved the theoretical validity of the gravity model, following two decades of its use in practice. These theoretical foundations encouraged further use of the gravity model in the development of transport demand models.

2.2 Behavioral Models

Travel behavior is based on a series of choices; if the individual travels, where they travel from, where they travel to, and how they make the journey. Each individual takes different factors into account in their decision-making process. These choices depend not just on concrete facts such as the location of the workplace, accessibility and auto ownership, but also the daily weather, family schedule, personal preference for different transport modes and even just their current mood.

Since we cannot hope to replicate the decision-making process of every individual exactly, particularly in the response to future changes in their environment, we use behavioral models to simplify, define and quantify human behavior and its impacts. We can either consider the population as a whole (aggregate modeling), or predict the choice of each individual (disaggregate modeling).

Where the disaggregate modeling of human behavior involves making a choice between alternatives, such as a destinations for a journey, or transport mode, discrete choice models can be used. In such scenarios, only one alternative can be selected, i.e. a person cannot travel to two places at the exact same time. Discrete choice models calculate the probability that an individual chooses a particular alternative. It is then assumed that each individual will choose the alternative that provides the highest utility. The resulting model can only be probabilistic in nature as in practice complete information on the individual and the alternatives is lacking. Although the individual choice may not be exact, when the model is applied over a whole sample population, the behavior of the population can be represented.

2.3 Discrete destination choice models

As an alternative to the gravity model, McFadden (1973) and Ben-Akiva (1974) proposed the use of the logit model as a disaggregate method to model travel demand. McFadden, in his pioneering paper, noted that "When the model of choice behaviour under examination depends on unobserved characteristics in the population, the testable implications of the individual choice model are obscured."(McFadden 1973)

Further research focused on the use of disaggregate models for the trip distribution step of the classic four step model. These came to be known as destination choice models, the focus of this thesis. A thorough investigation of the suitability of discrete choice models as opposed to aggregate methods for transportation modelling was conducted by Spear (1977). Spear noted that the:

- Individual choice models are more data efficient than conventional (i.e. gravity) models.
- They can utilize the variation in socioeconomic data much better, to avoid ecological fallacies.
- The probabilistic nature of the dependent variable allows for the modeling of interdependent choices, such as mode choice and trip chaining decisions.

Since then, the application of disaggregate models in transport demand modeling has been continually refined, with important research done in both modeling destination choice and mode choice in this manner. Daly (1982) focused on representing the attractiveness of a destination in a destination choice model, while further work was done by Ben-Akiva (1974) and Anas (1983) in defining the structure and application of such models. Train (2009) comments that "discrete choice models cannot be calibrated using a simple curve fitting, as since the dependent variable, as a probability cannot be observed". Instead, maximum likelihood estimation is most commonly used. Since the utility of every alternative must be calculated, this technique was prohibitive for large scale problems before the advent of modern computers. This may go some way to explaining the persistent popularity of aggregate models, due to their simplified computational requirements.

In chapter 9 of Discrete Choice Analysis, Ben-Akiva and Lerman (1985) present a comprehensive discussion of destination choice models. They note that “destination choice is characterized by a very large number of alternatives”, and the selection of resolution of the choice set is a very important consideration. They further discuss the challenge of data availability for destination attractiveness. Since the attractiveness of data is not always available at a destination level, “the alternatives in a destination choice model must be based on aggregate alternatives”. Even with the modern GPS and social data available to the modern modeler, this is clearly still an issue. This is an important point that Ben-Akiva and Lerman make: that while destination choice models can model the decisions of individual travelers, they still need to rely on some level of aggregation for modeling the utility of each destination.

Simma, Schlich, and Axhausen (2001) developed a destination choice model for leisure travel in Switzerland that considered many variables of destination attractiveness such as the number of swimming pools, ski area qualities and land use attributes. They found that while the distance terms played the most significant role, measures of destination attractiveness were still important and improved the model.

Feedback loops are also often included to consider the impacts of steps in the transport model on destination choice. In particular, researchers have found that mode choice influences destination choice, and there are two common approaches that can be taken to account for this. The logsum of the mode choice model is included as a parameter in destination choice (Jonnalagadda et al. 2001; Mishra, Ye, et al. 2011) or a combined destination-mode-choice model is calculated (Newman and Bernardin Jr 2010; T. J. Adler and Ben-Akiva 1976; Boyce et al. 1983; M. L. Outwater et al. 2015). (Mishra, Wang, et al. 2013) implemented a gravity model and multinomial logit destination choice model for Maryland and compared the results. They found that the destination choice model performed much better than the gravity model for state-wide travel demand.

2.4 Trip chaining

T. Adler and Ben-Akiva (1979) were some of the first to model the inter-dependencies between links in a trip chain. They defined a theoretical and empirical model of trip chaining behavior to do so, based on utility theory, and accounting for the tradeoffs involved in multi-step chain trips. They, like most researchers in the field, focused on daily travel patterns within urban models. However, they note that “It is important that the determinants of non-work travel patterns that include multiple-sojourn tours be better understood”. To do this, they model the utility to a given household of a particular travel pattern as a function of scheduling convenience, activity duration, income, destination attributes and socioeconomic characteristics of the household. One of the significant advantages of a disaggregate destination model is the ability to model tours. Due to the nature of the data, trip chaining commonly is not included in long distance travel models. Moeckel, Fussell, and Donnelly (2015) considered its inclusion, however the proportion of multi-link trips was found to be too small, and the trip lengths between stops were not recorded in the National Household Travel Survey (NHTS) data they used.

Kitamura (1984) incorporated trip chaining directly into an analysis of destination choice. He used an approach called Prospective Utility that "represents the expected utility of the visit to that zone and also those visits that may be made". In essence, this theory postulates that with two destinations, A and B, of equal utility, opportunity B will be more attractive than A to a trip maker when it is surrounded by destinations supporting other opportunities that the trip maker wishes to pursue.

2.5 Recent long distance transport models

M. Outwater et al. (2010) developed a state-wide model for high speed rail in California. They combined both stated and revealed preference data in their attributes. For destination choice, they looked at destination attraction, employment and household characteristics, the region and area type, trip purpose, distance class, and party size. While not a combined destination-mode choice model, they combined some network data to calculate auto and non-auto accessibility, for peak and off-peak periods respectively. Destination was predicted using a simple multinomial logit model. The authors also note that their modeling shows "that an individual may value different trip characteristics for different distance-categories of travel". They also modeled the area type of a zone, as rural, suburban or urban. Interaction terms were also created between zones, under the assumption that urban to urban trips are much more common.

More recently, models are also being designed on a larger, more ambitious scale. One such example is the new national model of long-distance travel in the United States (M. L. Outwater et al. 2015). This incorporates multiple advancements on previous models, including modeling at an individual household level, a high level of spatial detail for destination choice, and the vertical integration of all 4 steps of the transport model. Unlike activity based models, it uses a temporal resolution of months and weeks, not days, and jointly predicts destination and mode choice together.

2.6 Mnlogit R package

In this thesis, the R package *mnlogit* (Hasan, Zhiyu, and Mahani 2014) is used to estimate a multinomial model of destination choice. While the input formats are the same as the original *mlogit* package, *mnlogit* requires less computer memory usage, allowing for more model parameters. It also performs the maximum likelihood estimation in parallel for significant decrease in estimation runtime.

Data format

The model input must be provided in a long format. The format of the input can be described as follows: Let S be the set of input trips, and R be the total choice set of possible destination alternatives.

For each trip $s \in S$, an arbitrary choice set R_s is required, such that exactly one element in the choice set is marked as selected, and all alternatives in R are included in the choice set for at least one trip. All trips much have the same number of alternatives. The model input is constructed by adding a row for each trip s and alternative $a \in R_s$, giving a total number of rows $\sum_{s \in S} |R_s|$ for the model input. A boolean column is also added to indicate whether a particular choice was chosen for that trip or not.

Formula Specification

The *mnlogit* package accepts model formulas structured using the R formula package. A mlogit formula consists of 4 parts: $choice \sim Y|X|Z$, where X, Y, Z are as followed:

- Choice: the LHS of the equation, the column that indicates if an alternative was chosen or not.
- X: Individual i specific variables with alternative k specific coefficients $\vec{X_i}\vec{\beta_k}$.
- Y: Alternative specific variables with alternative independent coefficients $\vec{Y_i}k\vec{\alpha}$.
- Z: Alternative specific variables with alternative specific coefficients $\vec{Z_i}k\vec{\gamma_k}$.

In the context of destination choice, individual variables are those such as income, gender and education level that pertain to the traveler. A coefficient must be estimated for every destination, as the value of the individual does not vary across the choice set for each trip, only between trips. Alternative specific variables can have coefficients that are independent or dependent of the choice set. The coefficient only needs to be dependent on the alternative Z when the parameter has a different meaning across the choice set. This is more commonly used in mode choice modeling; for example, the calculation of cost might vary between car and train journeys.

Parts of the equation can also be excluded by specifying 0 or -1 in the respective section. Intercepts can also be removed by adding the term $+0$ in the same way. The estimation will return an error if the equation cannot be solved, and the most common reason for this is high multi-colinearity between parameters specified in the model.

2.7 Location based social networks

The ubiquity of mobile GPS transceivers, especially in the smart phone market, has enabled a new category of social networks, called location based social networks (LBSNs), which associate social networking data with a geo-referenced location. Different social networks have taken advantage of this opportunity in different ways. Facebook enables a user to mark themselves as safe during a natural disaster, Flicker can show a map of where your images were taken, and Google maps can provide accurate travel times by identifying areas of congestion.

Most location based social networks, such as Facebook, Tripadvisor and Foursquare enable users to 'check-in' to a point of interest (POI), such as a shop, tourist attraction or airport, and provide tips, ratings and reviews. These POIs are referred to as 'venues'. When these services are used by millions of people around the world, in different countries and cities, an enormous amount of data is collected, which can be used in a multitude of ways to explore mobility patterns.

Lindqvist et al. (2011) looked at how and why people use location sharing services such as Foursquare, and discussed how users manage their privacy when using such services. (Cheng et al. 2011) collected 22 million check-ins across 220,000 users to quantitatively assess human mobility patterns. 53% of their check-ins came from Foursquare, highlighting the dominance of Foursquare in the LBSN space.

Noulas et al. (2012) used Foursquare data to design gravity model based on the theory of intervening opportunities (Stouffer 1940). They found that while no universal law exists between mobility and distance, a universal behavior in all cities when measured with their rank-distance variable exists. Regarding the potential applications of LBSNs in future research, they note that the scale of data collected by Foursquare provides the means to analyze and compare mobility patterns in different parts of the world, surpassing cultural, geographical and political borders. They also warn "there may be a strong demographic bias in the community of Foursquare users", before noting that "it is encouraging that the analysis and models developed in the context of the present work demonstrate strong similarities across multiple urban centers and different countries."

Abdulazim et al. (2015) introduced a framework for inferring activity travel given nearby land use information gathered from LBSNs. Their results suggest that daily activity travel can be automatically inferred from LBSN data, and they present a generic method for acquiring land use data from LBSN services such as Foursquare. The authors also present a case study for the greater Toronto and Hamilton area, Ontario, Canada, a subset of the study area for this thesis. SA et al. (2015) investigated the potential for cell phone and Foursquare data to replace the use of travel surveys in calculating an origin-destination demand matrices. They found that the the number of cell phone and Foursquare records was higher for OD pairs expecting to have higher trip volumes, but that some differences existed. Jin et al. (2014) proposed a doubly constrained gravity model based on LBSNs. They were able to achieve significant reductions in estimation errors caused by sampling bias when compared to a singly constrained model.

2.8 Objectives

Disaggregate models provide clear advantages over aggregate methods in modeling trip distribution. Due to the availability of data and more powerful computers, the modeling of destination choice using logit models is becoming increasingly attractive. Destination choice models provide more flexibility in attribute selection, and more efficient use of data.

It is often difficult or simply not appropriate to take a model that has been designed for another geographical region and apply it in study area of concern. Firstly, the data available for the study area will most likely be different to those available for other study areas. The data may provide more variables that were not available to modelers working in other regions, or be more restricted, forcing the modeler to be creative in designing parameters that can represent the travel behavior. Secondly, it is very difficult to design accurate models that work effectively when transposed to new study areas. This is due, not just to obvious geographical differences, but variations in policy and culture that are difficult to reflect in a destination choice model. If every possible parameter that was possibly relevant was added to the model, not only would it be computationally infeasible, but there would be a high risk of over-fitting in the model. The fact that the destination choice models already presented in the literature are mostly individually unique supports this notion that modeling is both a science and an art, and that there is no "one size fits all" model.

For these reasons, the design alone of a destination choice model for Ontario reflects a new contribution to the field. Transportation modeling as as discipline advances every time a new model is designed, implemented and evaluated. Future researchers can then look at the body of previous models, and use statistical analysis, their experience and intuition to select variables that best suit the requirements and use cases for which their model will be designed.

The second objective of this thesis is the successful application of data from LBSNs to represent zonal utility in the destination choice model. While work has been done on investigating mobility patterns, and the generation of OD matrices using LBSNs, their application to the representation of destination attractiveness in destination choice modeling has not been widely considered. This thesis will explore how Foursquare check-in data can be used in the calculation of destination utility. Check-in data provides an opportunity to model important traits of destination utility, such as the presence of national parks, that are not reflected in standard socioeconomic variables.

Chapter 3

Data Acquisition and Analysis

3.1 Travel Survey of Residents of Canada

3.1.1 Introduction

The Transport Survey of Residents of Canada (TSRC) is a monthly, cross-sectional survey collected by Statistics Canada to measure the volume, characteristics and economic impact of domestic travel . The survey provides a large quarterly sample of performed trips within Canada, along with socioeconomic data and the activities and expenditures performed on each trip. Results are released yearly, with the data available at a monthly temporal resolution.

The TSRC was designed to measure the size and economic impacts of Canada's domestic tourism industry. It was first performed in 2005, and replaces the Canadian Travel Survey. In 2011, the survey was redesigned to bring the questionnaire more in line with the World Tourism Organization guidelines, and align the recorded activities with the International Travel Survey (ITS).

In this thesis, the TSRC acts as the main data source for the estimation and calibration of the destination choice model. This section provides an overview of the aspects of the TSRC and its design that relate to the development of the destination choice model for Ontario. In particular, the methodology behind the survey is detailed, and the relevant variables and weightings available in the resultant microdata are highlighted.

3.1.2 Method

The survey is performed as a voluntary supplement to the compulsory Labour Force Survey (LFS). The LFS is a mandatory household survey of around 54,000 households to measure employment, and has a 90% response rate. The LFS sample consists of the entire civilian, non-institutionalized population over 15 years of age. A sub-sample of these households is selected to answer the TSRC, excluding residents of the Yukon, the Northwest Territories and Nunavut and people living on Native Reserves. A respondent is randomly selected from the household and asked to complete the travel survey. The survey is a computer-assisted telephone interview (CATI) available in both of Canada's official languages, English and French. 15 minutes are allocated for each respondent, with as many trips being collected as possible in that time.

3.1.3 Data

3.1.3.1 Spatial resolution

All spatial data points, namely those for home location, trip origins and destinations and stopovers are provided in the microdata at three resolutions: Province or Territory, Census Division, and Census Metropolitan Agglomeration (CMA). Canada is made up of ten provinces and three territories, the largest of which is Ontario, the focus of this thesis.

Census Divisions are the next largest geographical area in Canada. Census Divisions represent groups of neighboring municipalities combined to aid regional planning and the provision of common services. After the provinces and territories, they are the most stable spatial unit. They were last modified for the 2011 census, and therefore are consistent between each TSRC dataset since the revised version was introduced. In most provinces and territories, these census divisions are defined in legislation, however in Newfoundland and Labrador, Manitoba, Saskatchewan, Alberta, Yukon, Northwest Territories and Nunavut, provincial or territorial law does not provide for these administrative geographic areas. In these cases, the census divisions are allocated by Statistics Canada.

Census sub-divisions are the next smallest geographical area, representing individual municipalities. These are recorded as part of the survey, however are not available in the TSRC microdata. The finest level of aggregation available is that of the Census Metropolitan Areas (CMA) and Census Agglomerations (CA). CMAs and CAs represent certain clustered areas of population around an urban core. More specifically, a CMA is defined as an area having a total population of at least 100,000, with half of those living in a core urban area. CAs, which related to CMAs but require a core population of only 10,000, are not recorded in the TSRC data.

Since CMAs do not topographically cover the whole Canadian study area, but only identify particular dense urban areas, census divisions are the most detailed resolution available for consistent use when working with the TSRC data. Although CMAs are only recorded for 51.5% of trip origins and 48.3% of trip destinations, they can be used to improve the resolution of the zone system for the destination choice model (see Section 3.3.1).

3.1.3.2 Error detection and imputation

The computer-assisted nature of the survey allows for real-time error detection and consistency checking during the interview process. “Dont Know” and “Refused” are also valid options for many questions, to prevent false answers been recorded. Sanity checks against extreme values are also performed, and the coding of geographical areas is mostly performed automatically.

Two forms of imputation are performed for the survey, for trip details and expenditure amounts respectively. Since the survey only allows 15 minutes for the recording of trip details, the details of non-selected trips are imputed from other trips recorded for that resident. This imputation process is multi-staged, and is performed per respondent. A donor pool of trips is selected that are similar to the non-selected trip. A distance function is then used to select the closest donor-trip to the recipient, and the detailed variables (activities, expenditures, etc) are copied over to the recipient trip.

3.1.3.3 Weighting

The weighting of records is particularly important when working with survey data that represents a larger population. Weightings allow a researcher to scale up the results of a sample to build an accurate representation of population, taking into account under- and over- represented groups within the survey. Four weightings are provided for the TSRC, with two relating to trip variables: Full-sample person weights First-month person weights Person-trip weights Trip weights

As the TSRC sample is based on the LFS survey, person weights are applied from the LFS and re-calibrated to reflect sub-sampling, non-response, and known control groups.

After the 2011 redesign, respondents are asked about same-day trips that ended in the previous month, and overnight trips that ended in the previous two months. This means that effectively only half the sample is asked about same-day trips. To account for this, two weights are provided for each person record. A first month weight, that can be used for any person variable, and a second "full sample" weight that can be applied to person characteristics and overnight travel variables.

The person-trip weight, used to estimate trip volume, is then calculated by accommodating for identical trips, declared and reported trips, missing data and non-response. These weights are corrected for outliers and recall bias during the creation of the microdata. In calculating the person-trip weight, the person weight is also multiplied by the number of identical trips that this trip represents. The person-trip weight (WTEP) can be used against all socioeconomic characteristics, as well all trip and visit variables, excluding expenditures. Trip weight (WTTP) is then calculated by dividing the trip-person weight by the number of household members that went on the trip. the WTTP is only used to calculate expenditures, and as such is not relevant to the destination choice model design.

3.1.4 Microfile format

The results from the TSRC are provided as yearly collections, separated into individual files for persons, trips and visits. The survey results are provided as fixed with delimited .dat files. A code book and data dictionaries are provided to decode the values stored in each line. The schema for encoded variables such as province are consistent across files and years (i.e. Ontario is always coded as 35), meaning that once read from the correct position on a line, values do not need to be decoded before being compared with each other.

Each person record is associated with one or more trips. Not all persons recorded in the person microdata necessarily have a trip recorded for a particular time period, as the survey records the travel behavior of both travelers and non-travelers.

Each recorded trip record has at least two associated visit records, and more if intermediate overnight stops were recorded. Visits are classified into two types, origins or destination/airport. Each Trip has one origin visit record, and at least one destination record. Where the main mode of travel for the trip is "Air", two or more airports are specified as visit records, along with the 3-digit airport code for the respective Canadian airport. The survey codebook notes that these airport records may be adjusted to protect respondent privacy.

Trip datafile

Trips included in the TSRC include same-day trips to destinations more than 40km from their place of residence, and overnight trips with at least one night in Canada. Domestic same-day and overnight trips are recorded in full. International trips with no nights in Canada are not recorded in the TSRC. For trips with an overnight destination, but some nights in Canada, only the domestic portion of the trip is recorded, with the point of departure from Canada recorded in the MDxxx variables for trip destination. The TR_D11 variable records the number of times this trip was performed in the reference month, and must be taken into account when estimating trip frequencies.

Socioeconomic variables for the traveler are recorded for each trip record; namely age, gender, education level, employment status and income. The number of household members who participated on the trip is also recorded.

Trip purpose is recorded at two categorical levels. In the first, which is used in this thesis, purposes are split into four options:

- Holidays, leisure or recreation
- Visit friends or relatives
- Business - All business and work related trips, except routine travel which is a regular part of the job
- Other - All trips for other reasons except regular household chores

Visit datafile

The visit data file provides a stops performed on each trip, which can be linked to the relevant trip by the Public Use Microdata File Number (PUMFID) and the Trip Identification Number (TRIPID). Each trip has at least two visits associated with it, an origin and a destination visit, differentiated by the VISRECFL variable. The AIRFLAG variable is used to identify visit records that refer to an airport entry or exit.

If a location is visited twice during a single trip, only one visit is recorded for that location. The visits are not guaranteed to be recorded in the chronological order of visitation, even though the visits are collected in chronological order during the survey process. This lack of order prevents the modeling of trip chaining from the TSRC visit records.

3.1.5 Season of travel

Canada has starkly contrasting seasons which influence travel choices of residents. The TSRC provides the month of travel for each trip, and these are aggregated into two seasons, designed to highlight the impact of winter conditions on long distance travel behavior. For this thesis, the months from November to March are considered winter, with the rest as summer. With this classification, summer covers 7 out of 12 months of year, or 58.4%. Table 3.1 shows how leisure and visit trip counts occur disproportionately in the summer months. The P value indicates the probability that this result is not by chance.

TABLE 3.1: Seasonal split of TSRC trips

	Summer	Winter	Summer %	P
Business	11,750	8,641	57.62%	0.02
Leisure	52,639	19,774	72.69%	1.00
Visit	61,630	39,856	60.73%	1.00

It is self explanatory that destination choice should depend on seasonal factors. The example of winter sports is a prime example. Winter sports are an activity that people willing travel long distances for, but only in winter. In 2014, TSRC respondents reported participating in winter sports in 4% of overnight trips, and 2% of same day trips.

3.2 Filtering of trip records

For the model input, the TSRC trip records from 2011 to 2014 were collated together, giving 220,512 trip records. Not all these trips were relevant to the estimation of the destination choice model. Firstly, records were removed where:

- Either an origin or destination is not stated
- The trip purpose is not leisure, visit or business
- A distance is not recorded
- The mode is recorded as air and the destination and origin airports are identical

The TSRC trip files provide trip records not just for Ontario, but for all of Canada. However, as a model for Ontario, we are only concerned with the following categories of trips that influence travel in Ontario:

- Internal trips within Ontario - Internal (II)
- Trips entering Ontario - Incoming (EI)
- Trips leaving Ontario - Outgoing (IE)
- Trips that cross Ontario - External (EE)

Any trips that did not fit one of these categories are excluded from the trip dataset used to estimate the destination choice model. External trips must be selectively filtered to remove trips that do not cross Ontario. Excluding such external trips is important to make sure that the estimated model reflects the behavior of travel in Ontario, which could be different to the behaviors in other provinces.

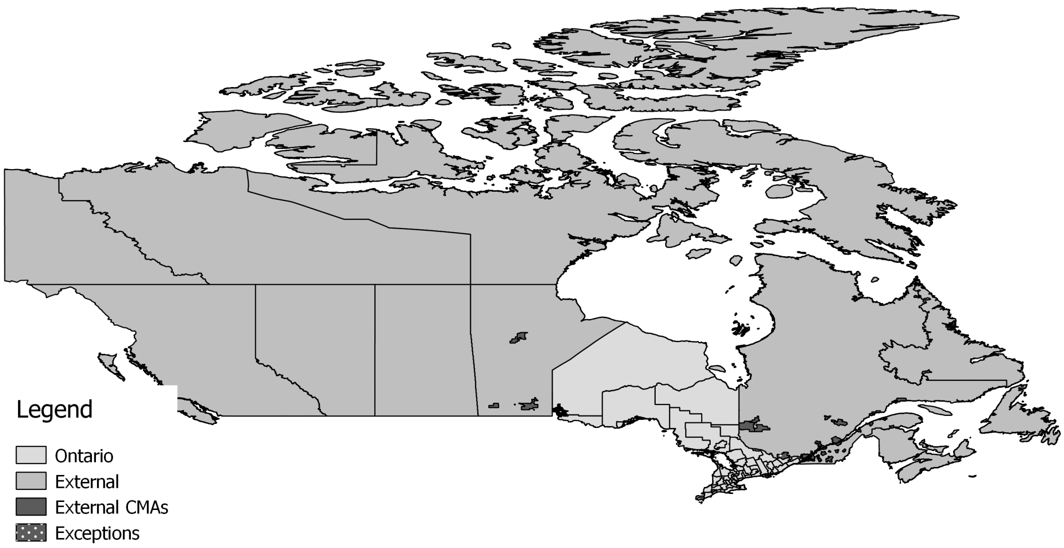

Figure 3.1: Dividing external zones into east and west

The unique geography of the Canadian provinces greatly restricts the number of external origin-destination pairs that need to be considered when excluding unwanted external trips. Ontario acts as land bridge between the eastern and western parts of Canada, see Figure 3.1, dividing the external zones into two groups, east and west. Trips originating in the east and arriving in west have to pass through Ontario, and vice-versa. For trips within a group, this is not true. Hence all trips that do not go between east and west can be removed. There are two zones which are the exception to this, zones 85 and 117 in western Quebec. Journeys between these zones and other zones in Quebec may pass through Ontario. For example, Figure 3.2 illustrates a journey from Gatieau, Quebec to Montreal Airport. a car journey takes around 2 hours when passing through Ontario, and 2 hours and 30 minutes otherwise. It is hence likely that travelers will chose the route passing through Ontario. Trips between these two exception zones and all other zones in Quebec are therefore retained.

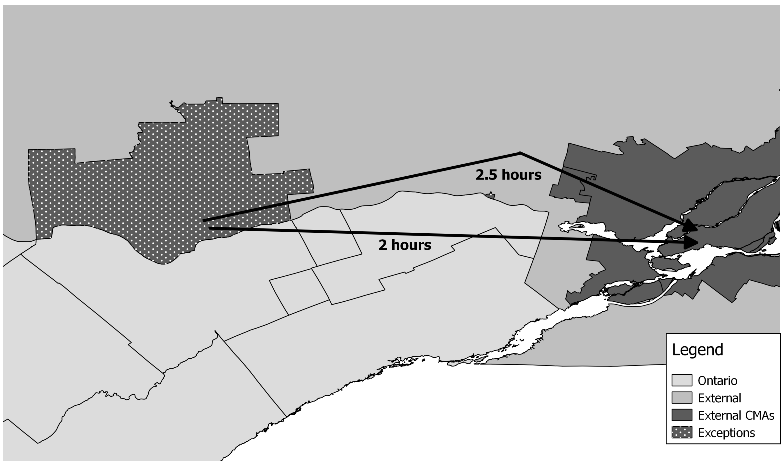

FIGURE 3.2: An example of an external origin-destination pair that passes through Ontario

In total 69,328 individual trip records remain from the TSRC dataset for model estimation (see Table 3.2).

TABLE 3.2: Sample size by trip purpose

	2011	2012	2013	2014	Total
Business	1,798	1,640	1,449	1,341	6,228
Leisure	5,939	5,878	5,515	5,577	22,909
Visit	9,057	8,777	7,962	7,618	33,414
Total	18,694	18,016	1,6547	16,071	69,328

3.3 Zone system

To avoid confusion throughout the rest of this thesis, the reader should be aware that there are two zone systems considered in the following chapter, and at various points throughout this document:

- **TAZs**, or traffic analysis zones, are the zones provided for the project, representing the proposed final spatial resolution of the transport model.
- **Zones**, representing the destinations in the destination choice model, referred to collectively as the *zone system* in the remainder of this thesis.

This section discusses the definition of the choice set of alternatives for the destination choice model. Numerous factors need to be considered when designing the choice set. Firstly, The sample size of the data available to estimate the model coefficients is an important restraint. With a small sample set relative to the size of the destination choiceset, not enough records are available to calculate the parameter coefficients with high confidence. Hence, the size of the choiceset needs to be considered carefully. Large destinations choice sets also lead to very long computation times when estimating the model coefficients. A balance needs to be found between the detail represented in the choice set, and the computability of the model.

For this particular destination choice model, a zone system was already provided, consisting of 6671 Traffic Analysis Zones (TAZ). The TAZs can be grouped into 4 categories; 6495 internal zones for Ontario, 48 and 121 external zones representing the rest of Canada and North America respectively, and 7 zones for remaining world-wide destinations. As this destination choice model is concerned only with domestic travel within Canada, the internal TAZs for Ontario, and the 48 external TAZs covering the other provinces of Canada are considered. The external zones are taken directly from the TAZs as TSRC origins and destinations are directly translatable to the external TAZs.

The Internal TAZs within Ontario were allocated by the team at Parsons Brinckerhoff using a gradual raster based zone approach, based on the method developed by Moeckel and Donnelly (2015). The 6495 generated TAZs vary in size from $0.879km^2$ to $3600km^2$, with smaller cells defined for more populous areas, and larger cells for regional areas. The gradual zone system is designed on the premise that it is desirable to have larger zones in rural areas where there is less population, and hence, less activity. This method reduces the number of TAZs, and hence, the complexity of the model, while only removing detail where it is least required.

Unfortunately, the TSRC trip origins and destinations are only recorded at the resolution of CMAs and CDs, which are much broader than the raster based TAZs created for Ontario. A zone system is designed based on the TSRC spatial resolution for the design and estimation of the destination choice model. The allocation of trips origins and destinations to TAZs will be performed at a later stage in the transport model. This allocation is discussed further in Section 5.2.

3.3.1 Defining a zone system for Ontario based on the TSRC data

Provinces and census divisions cover the national study area completely. Hence as a first step, the zones are defined by the census divisions within Ontario, of which there are 49. However, the TAZs are much smaller than their encompassing CDs, even in rural areas.

When the zone system is defined purely using the Census Divisions within Ontario, over 50% of Census Divisions have more than 75 TAZs, with a large spread of values (see Figure 3.3.

Although CMAs are defined only for selected urban areas of Canada, they can be considered alongside the CDs when allocating zones to improve the spatial resolution of the zone system. The concept of a CMA aligns closely with the objective behind the gradual raster cell size of the provided zone system for Ontario. CMAs identify areas of denser population around an urban core that may be of particular significance to geographers and modelers. By including CMAs as zones in the aggregated zoning model, the number of zones is increased to 57. This will also aid the disaggregation of trip origins and destinations to TAZs in urban areas.

However, there is still further detail to be extracted from the combination of CDs and CMAs. The previous approach creates a large outlier that consists of over 2000 TAZs. This outlier corresponds to the CMA of Toronto, the most populous in Ontario, a very large generator and attractor of trips. In 2014, Toronto represented 13.4% and 10% of trip origins and destinations respectively. The disaggregation of trip origins and destinations between 2000 child TAZs would undoubtedly affect the overall quality of the model.

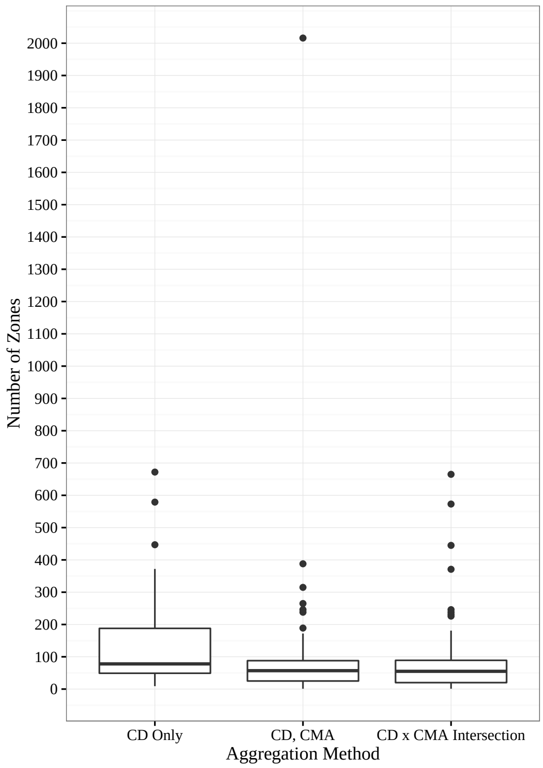

Figure 3.3: Different methods of aggregating internal zones to match the TSRC spatial resolution.

This outlier was not present when only the CDs were considered as destination alternatives. Since CMAs often overlap multiple CDs, rather than simply including CMAs and CDs independently, the union of the CD and CMA can be taken to fully reflect the number of destination choices available in the TSRC data. This results in 69 zones for Ontario, a 41% increase over the simplest approach that only considers CDs.

Figure 3.3 illustrates the difference between these methods. When only Census Divisions are used, a significant number of CDs have a large number of assigned TAZs. When the CMAs are considered, the results are clearly better. A lower average number of TAZs per aggregate zone will give improved results when trips origins and destinations are disaggregated. The CMA of Toronto overlaps with 7 separate CDs, and can with this method be divided into seven smaller zones.

This third method has another advantage. The a distinction between urban and rural areas is now encoded into the zone system. This will be important in the estimation process as 51.5% of trips in the filtered TSRC survey originated in a CMA, and 48.3% had destination recorded as a CMA. Not only is it clear that urban areas are important drivers of long distance travel, but also, CMAs are more likely to be origins than destinations.

3.4 Aggregating zonal data

All the data on distances, population and employment was provided at the TAZ level. This data was aggregated to the zone system using the following approach. The TAZs themselves were assigned to the zone which intersected their centroid. Where the centroid of the TAZ did not intersect any zones, the first intersecting zone was used. if the TAZ did not intersect with any part of the Canadian census boundaries at all, it was assigned manually to the nearest zone using GIS software.

Socioeconomic variables, namely population and employment, were aggregated from the TAZ level to the zone system using a simple summation.

The aggregation of the distance skim matrix was more involved. The matrix, in OMX format (Stabler et al. 2014), contained the auto travel times between all TAZs. It was calculated without congestion using the Canadian road network by the team at Parsons Brinckerhoff. Intra-zonal travel times were not included. To aggregate the TAZ skim matrix to the zone system, the distance d between each child OD pair was weighted by the multiplied populations p of the origin and destination.

$$d_{ij} = \frac{\sum_{k \in i, l \in j} d_{kl} \cdot p_k \cdot p_l}{\sum_{k \in i, l \in j} p_k \cdot p_l}$$

3.5 Foursquare

Trip distribution models that consider only population and employment have a significant flaw. They fail to account attractions, such as national parks that draw large numbers of people, yet exist in areas of low population and employment. Discrete choice modeling provides the opportunity to incorporate parameters that reflect these drivers of travel demand.

Leisure travel is a particular case where socioeconomic metrics do not always reflect the attractiveness of a destination. Areas such as as lakes, national parks and ski resorts are popular long distance recreational destinations in the summer and winter respectively, yet have small populations and employment.

The TSRC microfile records the activities performed on each recorded trip and destination visited during a trip. When aggregated by trip destination, these activities give an indication of which zones provide particular attractions such as national parks or ski areas. The number of trips with each recorded activity can also be used to identify the importance of a particular activity for each zone. However, there are two key problems with this approach:

- When implementing the model, the spatial resolution of the TSRC microfile means that the location where an activity was performed can only be determined at a zone level. Another data source is still needed to identify key points of interest such as hospitals and tourist attractions to predict trip destinations at the TAZ level.
- As a domestic survey, the TSRC does not cover the US, meaning a different method would need to be used to identify key attractions across the border.

In this section, we describe how the collection and processing of foursquare check-in data was performed in order to build destination utility variables.

3.5.1 Foursquare venue search API

Foursquare collects a wealth of data, on where and when users check-in. Users and their behavior can be tracked over time using twitter as a proxy, however the time-frame for this thesis made this method unfeasible. Instead, the public venue API was used, which is much more limited in its scope, only providing a cross-sectional dataset of POIs within the search area. For a search area and criteria, the API returns a list of venues in JSON format. Each venue record provides the following relevant information:

- Name
- Venue category
- Geo-referenced location
- Number of unique visitors
- Number of total check-ins

Each request is limited to roughly 1 square degree of longitude and latitude in search area, and only the top 50 venues for that search are returned. A limit of 5000 requests per hour is also enforced. Search results are returned based on the popularity of the venues. How the rank of returned venues is determined by foursquare is not specified.

The API does not return check-in counts by date, so it can only be used to generate a total metric of activity for each venue, up to the time of the search. For the forecasting of trips to individual venues, this would present a significant obstacle. In this thesis, however, the foursquare metrics are only used for identifying the intensity of activity in zones that may not be reflected in socioeconomic variables. Check-in counts also can not be filtered by origin country or state. This capability would, in a larger model, also allow us to identify U.S. destinations that are commonly visited by Canadian travelers, as opposed to all users of foursquare.

3.5.2 Demographic bias

While the use of social networks is becoming more ubiquitous throughout the general population, LBSNs such as foursquare still have a particular user demographic, which should be taken into account when working with social network based data. The data retrieved from the foursquare API does not provide any demographic information that can be used to weight the retrieved data.

In this thesis, the potential impact of bias is minimized, as only the intensity of activity for each category in a zone is measured as a variable. There is also no stratification of these variables by age, gender or education level in the model estimation. Such stratification would cause concerns with demographic bias, for example with older groups of travelers. One concern is that certain venue categories could be under-represented in the data, such as aged-care services, or those where a check-in might be taboo such as a place of worship. This is considered by grouping venues into broad categories, which are then considered as model parameters.

3.5.3 Methodology

To collect the venue data from the Foursquare API, the following procedure was followed:

1. A developer account was set up to access the API.
2. The maximum search area allowed is smaller than most external zones, so a search grid of 1 degree raster cells was generated for the study area.
3. Using the activities specified in the TSRC as a reference, a selection of potentially important venue categories was curated (see Table 3.3).
4. Each category was mapped to at most 5 main foursquare venue categories, on which the search was performed. This is necessary to exclude venue categories such as "States & Municipalities".
5. A python script was written to query the foursquare API for each raster cell and category, returning the top 50 venues, while adhering to the rate limit of 5000 requests per hour. Calls to the api had to be structured as:

   ```
   https://api.foursquare.com/v2/venues/search?intent=browse
       &limit=50&sw={sw}&ne={nw}& categoryId={categories}
   ```

   where

   - sw, ne are the bottom-left and top-right corners of the search area
   - categories is a comma separated list of venue ids.
6. Unique venues were then stored in the PostGIS database, and each tagged with the zone to which it geographically belongs.
7. Two attributes were then calculated for each zone and category; the number of venues of that category in the zone, and the number of check-ins for each category and zone.

A correlation analysis of the created variables shows a high correlation between them (see Figure 3.4). However, this would not be the case if aggregation was performed at the TAZ level. It is still worth investigating the usefulness of these variables in the model estimation, as the number of variables increases the applicability of the model to scenario analysis. The scenario analysis in Section 5.4 demonstrates that a model based on these variables produces reasonable results, despite the correlation between the variables.

The foursquare API provides data at a higher spatial resolution than the TSRC microfile, but without the temporal detail. In Table 3.3, the number of venues and check-ins per category are presented. In total, 34,041 unique venues and 7,981,458 check-ins were collected for the different categories.

	airport	hotel	medical	nightlife	outdoors	sightseeing	skiing
airport	1	0.59	0.67	0.58	0.61	0.44	0.54
hotel	0.59	1	0.97	0.98	0.92	0.9	0.73
medical	0.67	0.97	1	0.98	0.89	0.88	0.67
nightlife	0.58	0.98	0.98	1	0.89	0.91	0.68
outdoors	0.61	0.92	0.89	0.89	1	0.73	0.91
sightseeing	0.44	0.9	0.88	0.91	0.73	1	0.48
skiing	0.54	0.73	0.67	0.68	0.91	0.48	1

Figure 3.4: Correlation of foursquare categories between destination zones

Table 3.3: foursquare venue categories

Search Category	Venue Categories					Venues	Check-ins
Medical	Dentist's Office	Doctor's Office	Hospital	Medical Center	Veterinarian	6,294	586,082
Ski Area	Ski Area	Ski Chairlift	Ski Chalet	Ski Lodge	Ski Trail	1,048	203,266
Airport	Airport	Airport Gate	Airport Lounge	Airport Terminal	Plane	1,882	1,919,050
Hotel	Bed & Breakfast	Hostel	Hotel	Motel	Resort	7,268	1,502,248
Nightlife	Bar	Brewery	Dive Bar	Pub	Sports Bar	5,900	1,936,153
Outdoors	National Park	Campground	Nature Preserve	Other Great Outdoors	Scenic Lookout	7,262	709,274
Sightseeing	Art Gallery	Historic Site	Museum	Theme Park	Scenic Lookout	4,387	1,125,385
					Total	34,041	7,981,458

Chapter 4

Gravity Model

As discussed in Section 2, the gravity model is still the standard approach to estimating OD trip distribution matrices. Its simplicity and low computational complexity makes it attractive to modelers. In modeling, it is always a good idea to develop the simplest model first. Firstly, it may be good enough, and a more complex model not required. Secondly, the errors observed in simpler models can inform the development of more complicated models. As such, this model will be used as a baseline to compare the destination choice models presented in Section 5.

4.1 Design

The gravity model is singly constrained on the origin, with the size of each zone being the sum of population and employment. The gravity model was implemented in JAVA, and is specified as followed:

$$T_{ij} = \frac{A_j \cdot e^{-\alpha \cdot d_{ij}}}{\sum_j^J A_j \cdot e^{-\alpha \cdot d_{ij}}} \cdot P_i$$

Where
T_{ij} is the number of trips between zones i, j.
P_j is the number of trips produced in origin zone i.
A_j is the attraction at destination zone j.
α is the impedance factor, calibrated with the average trip distance.
d_{ij} is the distance between zones i, j.

4.2 Model Strata

It is common practice to design a transport model as a collection of separate models for heterogeneous groups of travelers. The most common attribute to categorize by is trip purpose. The categories recorded for each trip by the TSRC, our sample of trip records is split into 4 categories, business, visit, leisure and other, of which only the first three are used. A model is then estimated for each trip purpose, and the results of the model strata combined into the final model, *m0*.

4.3 Calibration

A separate model was created for each trip purpose, and calibrated to match the expected average trip distance $\bar{d}$, calculated from the trip distances recorded in the TSRC, to within 1%. The results of the calibration are presented in Table 4.1. The average observed trip distance is $\bar{d}$, the average predicted trip distance d, and the impedance factor α. As measurements of model accuracy, the model correlation (r^2), root mean square error (RMSE) and normalized root mean square error (NRMSE) are provided.

TABLE 4.1: Gravity Model calibration

Model	Trips	$\bar{d}$	d	α	r^2	RMSE	NRMSE
Business	34,229.43	244	243.20	0.0013	0.42	53.45	0.94
Leisure	83,357.94	149	148.13	0.0035	0.36	100.72	1.03
Visit	129,843.18	163	164.77	0.0030	0.52	103.65	0.93

4.4 Results

Figure 4.1 presents an residuals plot, with the number of observed trips on the x axis, and difference between the observed and predicted on the y axis. While the three purposes cannot be compared with each other in the graph, due to the differing sample sizes, it is clear that all three models have serious errors, and are almost unusable. The predicted values should fall roughly above and below the dotted line. There is a definite pattern in the observed data, indicating that important OD pairs, ones with large numbers of trips, are strongly underestimated. The numerous OD pairs with small numbers of trips dominate the calibration to the observed average trip distance. However, this comes at the expense of model accuracy for large, important connections.

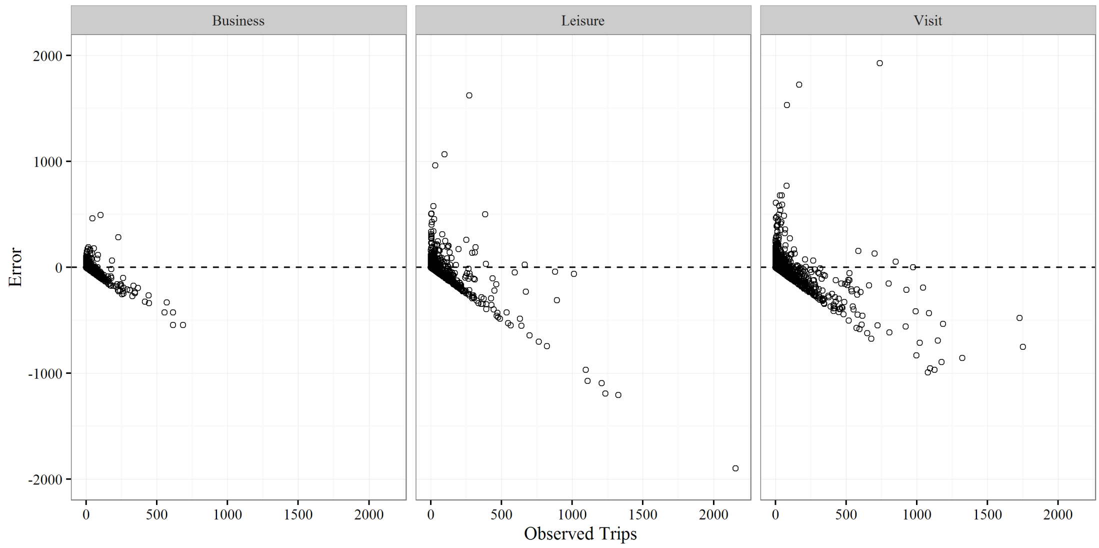

FIGURE 4.1: Gravity model errors by observed trip count for OD pairs by trip purpose

Figure 4.2 gives a better indication of how the model fits these important zones. On the x axis is the absolute error $|x - E(x)|$, and on the y axis, a variant of the relative error, which we call maximum relative error is plotted.

$$\frac{|x - E(x)|}{\min(x, E(x))}$$

In the standard relative error $\frac{|x-E(x)|}{E(x)}$, only one term, $E(x)$ is present in the denominator, meaning that large underestimations produce very small relative errors, reducing the visibility of such errors in the chart. In contrast, the maximum relative error treats overestimations and underestimations equally. For this model, it is also more useful than the error plot in figure 4.1, as the error Large y values are only of concern when the x value, namely absolute error, is also large.

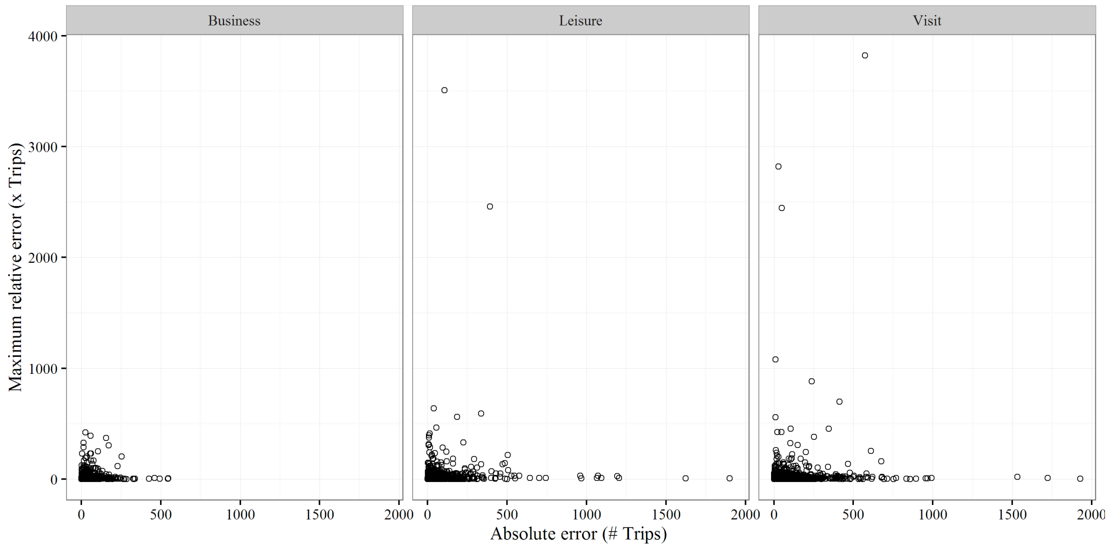

FIGURE 4.2: Maximum relative error chart for OD pairs by trip purpose

Large outliers are present for all three trip purposes in figure 4.2. A clear weakness of the gravity model can be seen by further examining some of these outliers for the leisure purpose. The number of leisure trips originating from zones in the Toronto region to tourist destinations such as Niagara Falls and Muskoka are strongly underestimated. By its nature, the gravity model is limited in how well it can model such zone interactions, as it only takes into account one attraction factor and one impedance factor.

The propensity for leisure travelers to visit destinations with tourist attractions is clearly determined by factors other than the population and employment of the destination. The multinomial logit model of destination choice discussed in the next section adds such factors, and explores how they can be modeled.

Chapter 5

Destination Choice Model

The design of a destination choice is much more involved than the construction of a gravity model, since modeler has almost infinite possible combinations of variables at his or her disposal. Divining the best variables is part art, part science. Some variables may be statistically significant, while adding little useful explanatory power to the model. Others may only be significant when paired with certain other variables. In this chapter, the design process of the destination choice model is presented. The estimation, implementation and calibration of a multinomial logit (MNL) model are covered in detail. The completed model is then applied to a simple scenario to gauge the usefulness of the model, and explore the impact of different parameters.

The MNL model is one of the simplier discrete choice models, where the unobserved errors between alternatives are not correlated and assumed to fit a Gumbel distribution. The representative utility of each alternative is specified as linear in the parameters: $V_{nj} = \beta' x_{nj}$ where x_{ni} is a vector of observed variables relating to alternative j (Train 2009). The probability that a certain alternative i will be selected by person n is

$$P_{ni} = \frac{\exp^{\beta' x_{ni}}}{\sum_j \beta' x_{nj}}$$

The correct value of these β' need to be estimated to give a utility function V_{nj} that best fits the observed data. This performed using a maximum likelihood estimation through R *mnlogit* package. This package is covered in more detail in Section 2.

5.1 Estimation

Rather than just present a final model, this thesis elaborates on the model development process, covering the important estimation iterations. In *m1*, a simple model based on the gravity model is presented. *m2* and *m3* add further interaction variables between origin and destinations. *m4* and *m5* explore the potential of LBSN data to improve destination choice models when incorporated into the calculation of destination utility. Finally, *m6* makes some adjustments to fit the model to the observed average trip length. The model summaries from the final *m6* are available in the appendix. The estimation code and input data are available on the CD accompanying this thesis.

5.1.1 Socioeconomic variables

For the first model, the same inputs as for the gravity model are used, namely the exponential of distance $e^{-d_{ij}}$, and the combined population p and employment emp.

The distance factor for each trip purpose was adjusted by the impedance factor α estimated for the respective gravity model (see Chapter 4). This approach significantly improves the model, and provides a quick way to calibrate the distance coefficient.

Metropolitan areas are not homogeneous in land use patterns. There exists residential areas and central business districts to which people may choose to travel. However, at the spatial resolution of our zone system these differences are hidden, resulting in a very high correlation between population and employment across the destination choice set of 98.95%. Therefore as with the gravity model, population and employment are summed together. This value is then log transformed, to account for the long tail in the distribution (see figure A.1 in the appendix. In order to simplify the further model equations, we assign a new variable for each destination

$$civic_j = \log\left(p_j + emp_j\right)$$

The resulting model *m1*, is defined by the utility u of destination j for a traveler in origin i:

$$u_{ij} = \beta_1\ e^{-\alpha \cdot d_{ij}} + \beta_2\ civic_j$$

where β_n are the coefficients to be estimated by the mnlogit package.

The employment data is classified by NAICS category, and models were tested that considered different combinations of employment categories. This investigation found that filtering the categories of employment did not improve the model. The individual employment categories were also not considered separately as unique variables, as they were highly correlated (see Table A.2 in the appendix.

TABLE 5.1: *m1* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-\alpha \cdot d_{ij}}$	4.29	***	3.86	***	4.21	***
$civic_j$	0.51	***	0.35	***	0.76	***

The parameters of this model *m1* (see Table 5.1) are encouraging. All the signs are as expected, and differences in the coefficients across trip purposes are evident. A leisure trip is less likely to go towards areas of civic importance than visits or business, and the trip distance is naturally less important for business travelers. For each trip purpose, the basic multinomial logit model already performs better than the gravity model, as evident in the higher correlation and lower RMSE values in Table 5.9.

5.1.2 Origin-Destination interactions

In this section, the model is extended to reflect the relationships between the origin and destination that might affect the choice of destination. Model *m2* is specified by the utility function

$$u_{ij} = \beta_1 \; e^{-\alpha \cdot d_{ij}} + \beta_2 \; civic_j + \beta_3 \; language_{ij} + \beta_4 \; mm_{ij} + \beta_5 \; rm_{ij}$$

where

$$\begin{aligned} language_{ij} &= language(i) \neq language(j) \\ mm_{ij} &= metro(i) \;\wedge\; metro(j) \\ rm_{ij} &= !metro(i) \;\wedge\; metro(j) \end{aligned}$$

The variable $language_{ij}$ reflects if the traveler is traveling to a zone where the main spoken language is different, namely somewhere in Quebec, where the main spoken language is French. Potentially, travelers may be more likely to choose a destination where the same language is spoken. mm_{ij} and rm_{ij} are designed to model the tendency to travel towards metropolitan areas. A zone is classified as metro when it is part of a CMA, and rural otherwise. There are 4 possible combinations of the metropolitan flag for origin and destination pairs. However, only two were selected for inclusion in the model. The flag identifying trips leaving metropolitan areas towards rural areas, mr_{ij}, results in an unsolvable model, and all other combinations, other than the one selected, $\beta_4\ mm_{ij} + \beta_5\ rm_{ij}$, also result in unsolvable models.

The use of these two parameters add small improvements to the model, as can be seen in the lower AIC. The RMSE is almost the same between the models. Table 5.2 presents the estimated parameters for this model. The new parameters vary strongly between trip purposes. mm_{ij} works well for each trip purpose, with visit and leisure trips more likely to leave metropolitan areas, and business travel more likely to be inter metropolitan. However, $language_{ij}$ and rm_{ij} do not work as well. They are not statistically significant in the visit model, and the coefficient is at least an order of magnitude smaller than for the other trip purposes. Business dealings normally require a common language, and hence it is not surprising to see a negative coefficient for language in this category. Finally, rm_{ij} is not significant for two trip purposes, despite working well for leisure trips.

TABLE 5.2: *m2* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-\alpha \cdot d_{ij}}$	4.35	***	4.57	***	3.81	***
$civic_j$	0.52	***	0.48	***	0.73	***
$language_{ij}$	0.05	*	0.58	***	-0.44	***
mm_{ij}	-0.10	***	-0.99	***	0.55	***
rm_{ij}	0.06	*	-0.39	***	-0.09	.

Comparing Figures 4.1 and 5.1, many outliers have been significantly brought back towards the axes, indicating an improvement in the model. However, there are still some significant outliers, with a sample of the largest in Table C.1 in the appendix. These outliers fall into two categories:

- Overestimation of intra-zonal trips within metropolitan zones such as Toronto.
- Underestimation of leisure and visit trips from metropolitan centers to tourist attractions such as Niagara Falls.

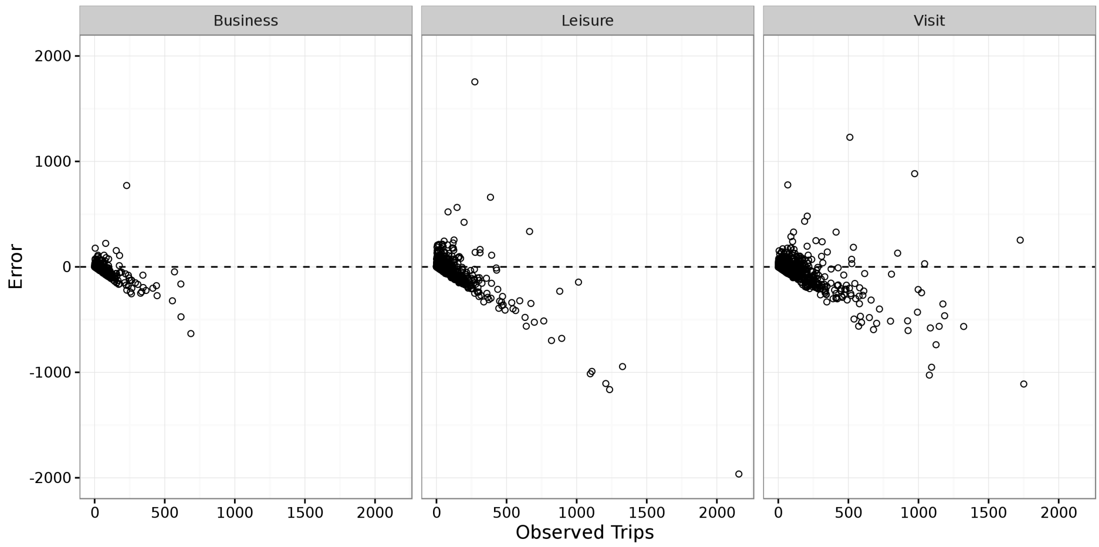

FIGURE 5.1: *m2* model errors by observed trip count for OD pairs by trip purpose

The large intra-zonal trip counts occur in small metropolitan zones, while in rural zones, intrazonal trip counts are underestimated (See Figure C.1 in the appendix. To penalize intra-zonal travel in the metropolitan zones, but allow it in larger rural zones, mm_{ij} is replaced with three new variables:

$$intrametro_{ij} = \begin{Bmatrix} 1, & \text{if } metro(i) \land i = j \\ 0, & \text{otherwise} \end{Bmatrix}$$

$$intermetro_{ij} = \begin{Bmatrix} 1 & \text{if } metro(i) \land metro(j) \land i \neq j \\ 0, & \text{otherwise} \end{Bmatrix}$$

$$intrarural_{ij} = \begin{Bmatrix} 1 & \text{if } !metro(i) \land i = j \\ 0, & \text{otherwise} \end{Bmatrix}$$

The first variable $intrametro_{ij}$ identifies trips within the same zone, where that zone is a metropolitan zone. This allows the model to reflect the propensity of a traveler to leave a metropolitan zone when they travel. The second, $intermetro_{ij}$ is 1 when the traveler is traveling from one metropolitan zone to another and 0 otherwise. This may be a common pattern for business travelers, but less likely for recreational trips. The third variable, $intrarural_{ij}$ allows the model to consider the intra-zonal behavior in larger, rural zones. The inclusion of these variables significantly improves the model results, and as presented in 5.3, particularly for business travel.

The other zone interaction variables $language_{ij}$ and rm_{ij} are removed in this iteration. They were not suitable in the previous model, and the significance of their coefficients did not improve in this iteration when combined with the new variables $intrametro_{ij}$, $intrametro_{ij}$ and $intrarural_{ij}$.

The parameters for the *m3* model are shown in Table 5.3. They are all significant, with the three new variables having differing magnitudes and signs, that each make logical sense for the different purposes. Business shows a strong preference for traveling to other metropolitan areas, as expected. Leisure travel is also very strongly influenced by metropolitan connections, but with a negative sign. This replicates the observed explanatory power of the rm_{ij} variable for leisure travel in *m2*, while also working for visit and business trips as well.

TABLE 5.3: *m3* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-\alpha \cdot d_{ij}}$	4.83	***	4.75	***	4.19	***
$civic_j$	0.57	***	0.52	***	0.76	***
$intermetro_{ij}$	-0.08	***	-0.87	***	0.56	***
$intrametro_{ij}$	-1.68	***	-2.56	***	-0.89	***
$intrarural_{ij}$	0.39	***	0.85	***	1.66	***

In Table 5.9, significant improvements throughout the model iterations are evident across all metrics. When compared with Figure 4.1, Figure 5.2 highlights the significant improvements of the destination choice model over the gravity model. The errors on OD pairs with small numbers of observed trips are drastically reduced, particularly for visit trips. The trend to under-estimate OD pairs with large numbers of observed trips is still evident (see Figure 5.2), and this problem is tackled in the next section.

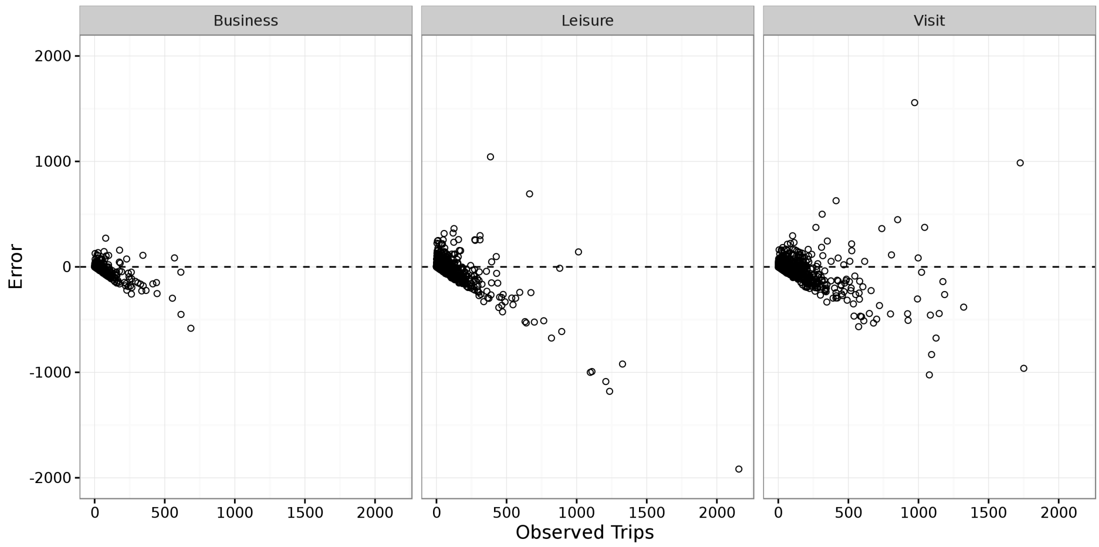

FIGURE 5.2: *m3* model errors by observed trip count for OD pairs by trip purpose

5.1.3 Incorporating LBSN data

The traditional socioeconomic variables of a destination do not completely reflect why people travel to a particular destination. People do not travel to a location purely because many people live there, but because there are opportunities to perform certain activities at that location. Population and employment act as proxy variables for some of these opportunities, but not all. This section incorporates data from LBSNs to improve the destination choice model, particularly for leisure trips.

The TSRC data show that activities such as skiing and visiting national parks are commonly performed on long distance trips. Areas where these outdoor activities are performed often have a low population and employment, while still providing attractive features to the traveler.

The collection and processing of LBSN data from foursquare was covered in section 3.5. To summarize briefly, the venues were collected into the following categories for each destination:

- Medical
- Ski Area
- Airport
- Hotel
- Sightseeing
- Nightlife
- Outdoors

Different MNL models utilizing the foursquare data were created, to explore the suitability of different categories, and whether they had different explanatory effects for different trip purposes. For each destination, two metrics were available for representing destination attractiveness; the number of venues, and the total number of check-ins across all venues. It was found that the best approach involved taking the natural log of the check-in count for each category. This gave the highest level of significance, as it corrected for the long right-hand tail present in the check-in counts for each category.

Certain categories were found to be significant for particular trip purposes. For example, the outdoor category was only significant for leisure trips, and the medical category was only significant for visit trips. As would be reasonably expected, the number of hotel check-ins was a significant variable across all trip purposes for long distance travel.

After exploring different combinations of the foursquare categories as parameters in the model, a model was settled on that was simple, yet powerful, using the most effective categories. The following categories were included in the model; hotels, sightseeing, medical, outdoors and skiing.

The main objective of including foursquare data was to investigate how the modeling of leisure destination choice can be improved. To this end, the common summer leisure activity, outdoor recreation, and the classic winter activity of skiing were represented though categories containing the types of venues commonly visited to perform these activities. Hence, two variables were added just for the leisure model strata. One for outdoor venues, and one for skiing areas. These two variables were found to be significant only when they were estimated for trips occurring in the season in which their respective activities are normally performed.

In model *m3*, it was observed that leisure trips to the zone containing Niagara Falls were underestimated by 85%. This particular important case was additionally controlled for by the addition of an extra variable using the sightseeing category, that is only considered for trips of the leisure purpose to the Niagara zone. The sightseeing category was also evaluated across all trip purposes as its own variable.

The foursquare variables were included into the MNL model using the following parameters:

$$\begin{aligned}
medical_j &= (purpose == \text{“}visit\text{”}) \cdot \log(medical_j) \\
hotel_j &= \log(hotel_j) \\
sightseeing_j &= \log(sightseeing_j) \\
naigara_j &= (purpose == \text{“}leisure\text{”}) \cdot (j == \text{“}niagara\text{”}) \cdot \log(sightseeing_j) \\
outdoors_j &= (purpose == \text{“}leisure\text{”}) \cdot (season == \text{“}summer\text{”}) \cdot \log(outdoors_j) \\
skiing_j &= (purpose == \text{“}leisure\text{”}) \cdot (season == \text{“}winter\text{”}) \cdot \log(skiing_j)
\end{aligned}$$

Below, two models are presented that apply foursquare data, *m4* and *m5*. *m4* illustrates the explanatory power of the foursquare data alone, by excluding the classic measure of attraction. For practical purposes, such a model is unfeasible, as population and employment are important variables for predicting the impact of socioeconomic changes on travel patterns. However, even on its own, the foursquare data still performs equivalently to the *m3* model for business and visit trips, and significantly better for leisure trips (see Table 5.9).

We can see that the popularity of hotels and sightseeing venues is particularly important for leisure travel. Business conferences are often located in areas of tourist significance as a way of promoting the event, supporting the large coefficient for sightseeing in the business category. The presence of medical facilities is also influential on attractiveness of visit trip destinations.

TABLE 5.4: *m4* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-\alpha \cdot d_{ij}}$	4.41	***	4.11	***	4.43	***
$hotel_j$	0.09	***	0.21	***	0.20	***
$sightseeing_j$	0.08	***	0.02	***	0.24	***
$niagara_j$			0.12	***		
$outdoors_j$			0.04	***		
$skiing_j$			0.09	***		
$medical_j$	0.16	***				

m5 re-includes all the variables from *m3*. In this model, $intermetro_{ij}$ and $intrametro_{ij}$ were found to be no longer significant for for the visit trip purpose, and were therefore excluded for this model strata. They were retained for both leisure and business trips. The combination of *m3* and *m4* to form *m5* gives the best model so far, with noticeably higher correlation and lower normalized RMSE for both business and leisure trips. The AIC metric also improves dramatically, even despite the increased number of parameters. The value of the foursquare variables, except for sightseeing, remains consistent after the addition of the the variables from model *m3*. The signs and magnitude of the variables from *m3* also change little.

Table 5.5: *m5* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-\alpha \cdot d_{ij}}$	5.00	***	5.35	***	4.37	***
$civic_j$	0.21	***	-0.15	***	0.36	***
$intermetro_{ij}$			-0.81	***	0.72	***
$intrametro_{ij}$	-1.75	***	-2.88	***	-0.87	***
$intrarural_{ij}$	0.24	***	0.58	***	1.51	***
$hotel_j$	0.11	***	0.27	***	0.17	***
$sightseeing_j$	0.04	***	0.13	***	0.08	***
$niagara_j$			0.13	***		
$outdoors_j$			0.03	***		
$skiing_j$			0.10	***		
$medical_j$	0.07	***				

Overall, this model performs better across all trip purposes than the *m3* model without variables based on foursquare data. Particularly noticeable is the large improvement across all metrics for leisure travel. Figure 5.3 shows impact of the foursquare variables for leisure travel. While it is hard to see the impacts for smaller OD pairs, the graph does illustrate how the errors for major outliers have been reduced.

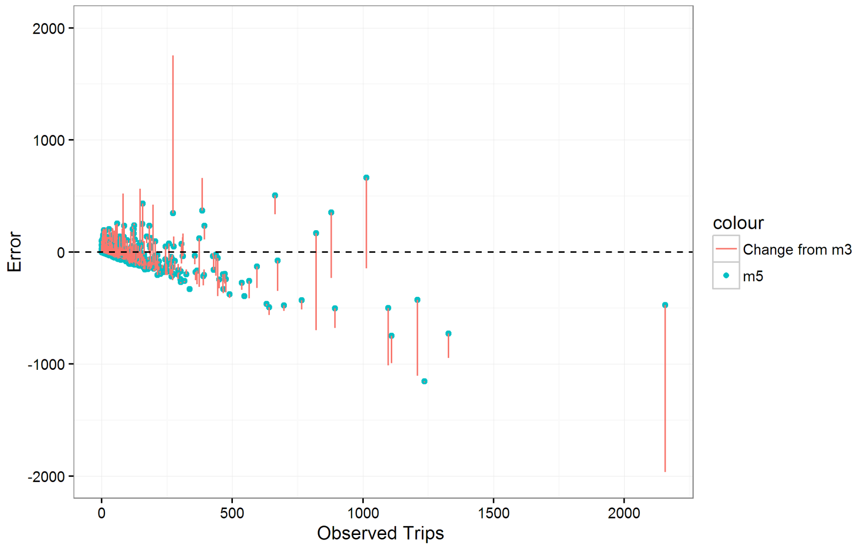

FIGURE 5.3: Effect of adding foursquare variables to model *m3* on leisure trips

5.1.4 Income strata

Socioeconomic factors such as income are important explanatory variables (Kitamura, Mokhtarian, and Laidet 1997). Since income is a characteristic of the individual, it does not vary between alternatives. However the income of the traveler may influence his or her perception of the utility of each alternative. Including income as a variable in the destination choice model requires at a minimum, one separate coefficient for each of the 117 alternatives. Since the income is stored as a ordinal variable, realistically, a dummy variable for each income category would be required. Tried in the model, this resulted in significant coefficients for some zones, but not others, with no appreciable pattern. This approach was discarded due to the lack of overall significance, and large number of coefficients required.

An alternative approach is to strata the destination choice model by income categories. This approach was tried with numerous permutations. First with the 4 individual categories, secondly, with two categories: low income (1,2) and high income(3,4), and third, with the income categories grouped as 1,2,3 and 4. Table 5.6 shows the parameters of the models for the groupings of 1,2 and 3,4 against the non-strata model. A visual inspection shows that the parameters are mostly consistent between the models. The *distance* variable does show variation between the two strata, but other variables are mostly consistent across the strata. Where a coefficient is different between the strata, such as *intermetro* for visit travel for the low income strata, it is no longer significant. In all scenarios, the performance of the models with income strata, according to the r^2 and $NRMSE$ metrics, is almost identical compared to the non-strata model (see Table 5.7).

Although there are surely model permutations where variables based on income are more significant, based on the lack of difference between the strata coefficients, income was not investigated futher for this destination choice model.

Table 5.6: Income-strata model coefficients (high & low income groupings)

	Visit				Leisure				Business			
Income-strata	low		high		low		high		low		high	
$e^{-\alpha \cdot d_{ij}}$	5.23	***	4.83	***	5.41	***	5.39	***	5.95	***	3.87	***
$civic_j$	0.23	***	0.23	***	-0.17	***	-0.14	***	0.31	***	0.4	***
$intermetro_{ij}$	0		-0.11	***	-0.68	***	-0.91	***	0.83	***	0.71	***
$intrametro_{ij}$	-1.66	***	-2.29	***	-2.62	***	-3	***	-1.49	***	-1	***
$intrarural_{ij}$	0.25	***	0.07		0.51	***	0.47	***	1.57	***	1.38	***
$hotel_j$	0.1	***	0.13	***	0.28	***	0.25	***	0.23	***	0.18	***
$sightseeing_j$	0.06	***	0.01		0.15	***	0.14	***	0.02		0.08	***
$niagara_j$					0.13	***	0.13	***				
$outdoors_j$					0		0.04	***				
$skiing_j$					0.09	***	0.1	***				
$medical_j$	0.02	.	0.12	***								

Table 5.7: Income-strata model results

Model	***m5***	***income-strata***
$\boldsymbol{r^2}$		
Business	0.77	0.78
Leisure	0.80	0.80
Visit	0.82	0.81
NRMSE (%)		
Business	0.66	0.65
Leisure	0.61	0.60
Visit	0.59	0.59

5.1.5 Refinement after considering trip length

A distribution of the trip lengths produced by the implemented model was used to evaluate the accuracy of the estimated destination choice model. The observed average trip length is different to those in the gravity model, as for the gravity model, the recorded distances from the TSRC were used. For the destination choice model, the length of each trip was calculated using the zonal skim matrix.

Figure 5.4 shows an excellent fit to the observed data, however, the average trip length of 318 km is high. To compensate for this, the logarithm of distance was tried as a parameter. While the inclusion of the logarithm did not affect the mean trip length, it did improve the accuracy of the estimation across trip purposes, most noticeably for visit trips. The $intrametro_{ij}$ parameter for business travel was removed as it was no longer significant. This model is presented as the final estimated model *m6*.

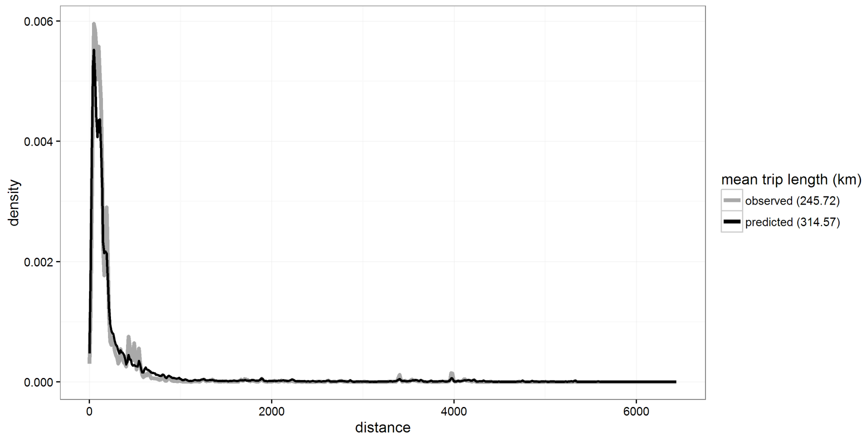

FIGURE 5.4: Trip length distribution of model *m5*

TABLE 5.8: *m6* (final) model coefficients

Parameter	Visit		Leisure		Business	
$e^{-\alpha \cdot d_{ij}}$	8.95	***	7.36	***	7.75	***
$log(d_{ij})$	0.90	***	0.48	***	0.65	***
$civic_j$	0.22	***	-0.17	***	0.45	***
$intermetro_{ij}$			-0.76	***	0.62	***
$intrametro_{ij}$	-0.96	***	-2.38	***		
$intrarural_{ij}$	0.64	***	0.66	***	2.00	***
$hotel_j$	0.10	***	0.26	***	0.15	***
$sightseeing_j$	0.05	***	0.14	***	0.07	***
$niagara_j$			0.12	***		
$outdoors_j$			0.04	***		
$skiing_j$			0.10	***		
$medical_j$	0.08	***				

5.1.6 Estimation results

Table 5.9 contains various statistical measures that measure the the iterative improvements throughout the model estimation process. An increase in the loglikelyhood indicates an higher probability that the model reflects the reality, assuming that the input data remains the same. r^2 is the correlation between the predicted and observed trip counts for each OD pair. Likewise, RMSE, or root mean square error is another measure of the differences between predicted an observed values. In this case, lower is better. Finally, the NRMSE is an alternative measure of the RMSE, normalized by the standard deviation of the observed trip counts. This last measure allows for a better comparison of model performance between trip purposes, as they have different sample sizes in the observed data.

TABLE 5.9: Comparison of model iterations

Model	***m0***	***m1***	***m2***	***m3***	***m4***	***m5***	***m6***
# Coefficients	1	2	4	5	7	11	12
Loglikelyhood							
Business		- 21,053	- 20,930	- 20,596	-21,071	-20,288	-20,091
Leisure		- 86,054	- 84,705	- 83,663	-81,192	-78,038	-77,808
Visit		- 117,463	- 117,441	- 115,666	-116,862	-114,557	-113,189
AIC							
Business		42,110	41,870	41,201	42,148	40,590	40,195
Leisure		172,113	169,420	167,337	162,396	156,095	155,638
Visit		234,931	234,893	231,342	233,731	229,128	226,393
r^2							
Business	0.43	0.62	0.62	0.73	0.56	0.77	0.77
Leisure	0.36	0.47	0.49	0.56	0.63	0.80	0.82
Visit	0.52	0.69	0.69	0.80	0.65	0.82	0.84
RMSE							
Business	53.45	45.95	45.75	39.53	49.76	37.26	36.98
Leisure	100.72	90.05	88.85	82.29	79.66	59.61	58.68
Visit	103.65	87.32	87.34	69.07	94.23	65.95	61.35
NRMSE (%)							
Business	0.94	0.81	0.80	0.70	0.88	0.66	0.65
Leisure	1.03	0.92	0.91	0.84	0.81	0.61	0.60
Visit	0.93	0.78	0.78	0.62	0.85	0.59	0.55

5.2 Implementation

The Destination Choice model described in this thesis was designed as a component of a larger long distance model for the Ministry of Transportation, Ontario. This long distance model is being developed in the JAVA programming language as a traditional 4-step model.

In the trip generation phase, a list of trips without destinations is generated for a synthetic population of households and persons. These trips are then passed into the destination choice model, which assigns a destination for each trip. For each trip the destination choice model is run, returning a predicted destination for that trip. For the calibration and scenario analysis, the observed dataset was used. The full expanded trip records numbers 362 million trips, and the model implementation currently cannot hold that many trips in memory. Therefore a sample of 1,000,000 trips is made by performing random draws from the observed dataset, based on the trip weight.

The algorithm works as followed, with step 2 being performed across the list of trips in parallel.

1. A Destination Choice Model is initialized with the following:
   - Coefficients for each model strata
   - Destination zones and their attributes
   - The distance matrix between zones
2. For each trip:
   (a) Calculate the utility u_j for each destination j, using the relevant stored coefficients.
   (b) calculate the denominator of the logit equation $q = \sum_{j=1}^{J} e^{u_j}$
   (c) Calculate the probability of each destination j, $P(j) = e^{u_j}/q$
   (d) Choose a destination based on the probabilities using an *EnumeratedIntegerDistribution* from the Apache commons math library

   ```
   return new EnumeratedIntegerDistribution(alternatives, probabilities).sample
   ```

   (e) store the destination in the trip object

5.3 Calibration

The implemented model was calibrated for each trip purpose against the observed average trip length. Through trial and error, the coefficient of the exponential distance term was multiplied by a factor k to adjust the predicted trip length. The results of the calibration are presented in Table 5.10. Figure 5.5 shows the results of the calibration for the overall model, as displayed in a trip length distribution. The two peaks after 3,000 km, representing connections to British Columbia and Alberta are visible in the predicted trip length distribution, although they are underestimated.

TABLE 5.10: Calibration coefficients and on average trip length

		Average trip length (km)			
Trip purpose	k	observed	estimated	calibrated	Δ
Business	1.05	377	408	391	3.6%
Leisure	1.15	212	264	214	0.9%
Visit	1.15	232	316	236	1.7%
Total		245	318	249	1.6%

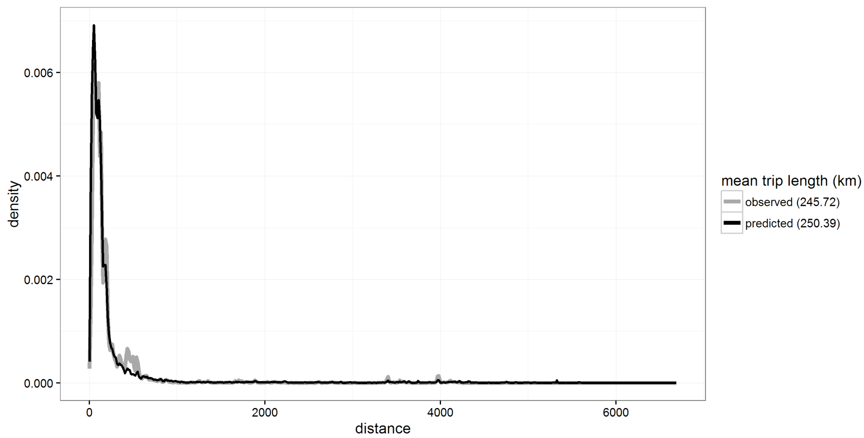

FIGURE 5.5: Trip length distribution of model *m6* after calibration

Trip lengths from 150 - 750 km are also underestimated. While calibration was required to match the overall trip length, it did so at the expense of the overestimation of shorter trips. The individual trip length distributions for business, leisure and visit purposes are available in the appendix. Figure 5.6 identifies the connections where the model falls short. The connections between the triangle of major cities, Toronto, Montreal and Ottawa, are underestimated. However, as a percentage of travel distance estimated by the model, it is very encouraging to see the more important larger peaks modeled so accurately.

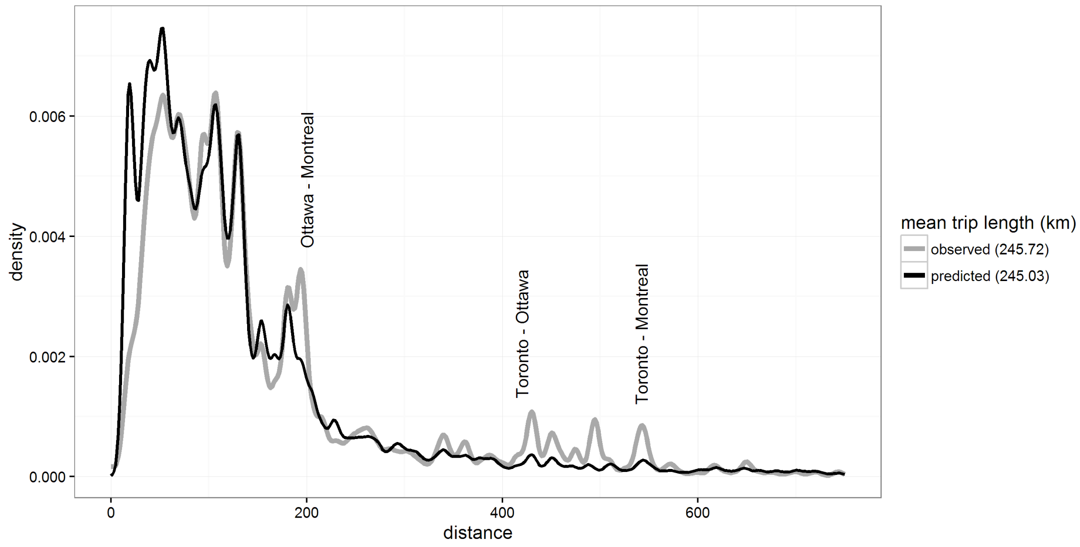

FIGURE 5.6: Trip length distribution of model *m6* after calibration (0-750km

5.4 Scenario analysis - Case study of a new ski resort

This section presents a hypothetical application of the developed destination choice model. For any large scale land use planning or development, it is important to model the impacts that such development will have on the transport network. As an example of this, a hypothetical scenario of the development of a large new ski resort is presented. Such resorts not only provide infrastructure for skiing and other snow-based activities, but require the development of multiple new hotels, employee housing, and retail infrastructure. In the winter months, ski resorts can generate significant demands on the transport network, and this needs to be taken account when considering such a development.

In the hypothetical scenario, a new resort is proposed for the highlands area north of Toronto in Dufferin (Toronto CMA) (see Figure 5.7). The higher elevation ensures good snowfall, and the elevation difference makes for exciting riders for snow sports enthusiasts. Two sites are being considered, one to the west of the range, and one to the east, closer to Ottawa. While this resort can naturally not be the size of mega resorts in British Columbia or Alberta, its development is expected to bring similar numbers of visitors as other large resorts in Ontario. Three average sized hotels will also be built at the base of the resort to accommodate guests. In the summer, the resort will attract visitors by providing mountain biking facilities and hiking. Additional housing for 400 new residents will be required to support 300 jobs.

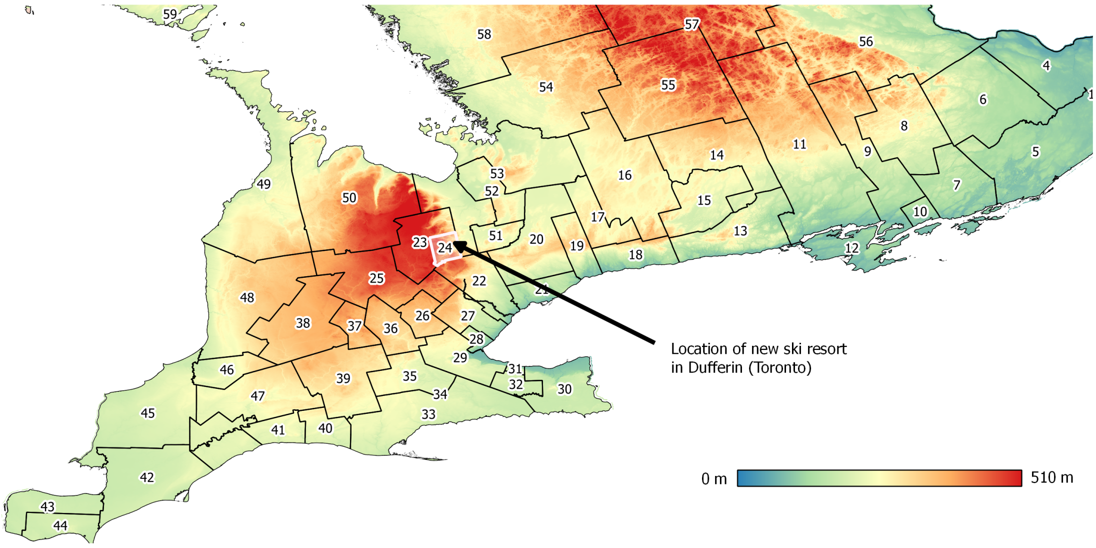

Figure 5.7: Scenario analysis: Location of a new ski resort on Dufferin (Toronto CMA). Elevation data from the Ontario Provincial Digital Elevation Model - Version 3.0

This somewhat contrived scenario assumes that other policy and development considerations, such as site location and transport access have all been dealt with. The design of the scenario presents the opportunity to investigate the sensitivity of the variables based on the foursquare data. The variables concerned are hotels, skiing and outdoor. The impact of the new development is estimated through adjusting these variables for the zone in which the development will take place. The foursquare POI database developed in Section 3.5 was used to estimate adjustments for each of the categories. Taking all venues in Ontario, the average number of check-ins per venue for each search category was calculated. The following adjustments are made for the respective zones, and their values are displayed in Table 5.11.

- Skiing: The average number of check-ins for ski areas
- Hotel: Twice the average number of check-ins for hotels
- Outdoor: The average number of check-ins per outdoor venue

Table 5.11: Inputs for scenario analysis

Parameter	Old Value	Adjustment	New Value
$civic_{ij}$	42,216	700	42,916
$hotel_j$	1,393	8,304	9,697
$outdoors_j$	1	3,389	3,390
$skiing_j$	40	3,550	3,590

The trips from the TSRC data used for estimation were inputted to the scenario, with $w/(365 * 4)$ copies of each record added to the trip table, where w is the trip weight of the record. The weighted TSRC data represents the total trips over 4 years, and for simplicity, the weights are scaled to give the approximate number of daily trips. 20 iterations of the scenario were performed to account for the stochastic nature of destination choice. Figure 5.8 shows the increase in incoming trips to Dufferin due to the new ski resort. The impacts of each input is presented from left to right, with the most right column being to total impact of the combined parameters. The table of results is available in the appendix in Table C.2. The results show that the parameters each behave reasonably. In particular the attractive effect for leisure travel is well modeled. Without the foursquare based parameters, the number of leisure trips would actually decrease with the addition of a new ski resort, due to the negative coefficient of the *civic* variable in model *m3* for leisure travel. This would clearly not be realistic, and this simply scenario gives a good example of why better representations of destination attractiveness are important, particularly for leisure travel.

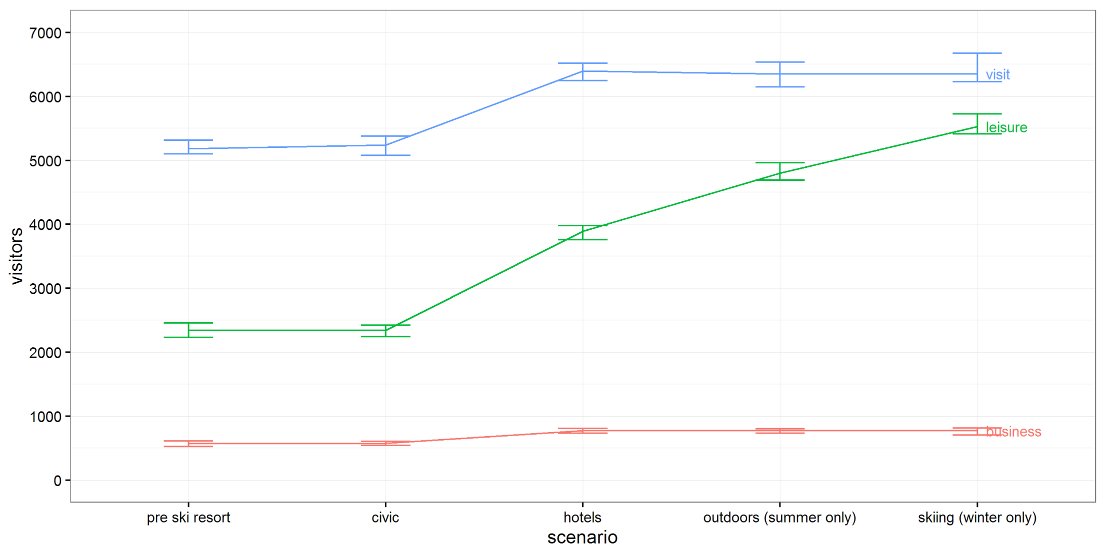

FIGURE 5.8: Scenario analysis: Impact of a new ski resort on Dufferin (Toronto CMA)

5.5 Remaining implementation work

This thesis presents an operational destination choice model for domestic travel to and from Ontario. The U.S.-Ontario border is an important source of incoming trips to Ontario and external trips that pass through to other parts of Canada. Further work is needed to extend the model to include continental travel to the United States, and then also intercontinental travel. The Canadian International Travel Survey (ITS) provides trip records that can be used to estimate such a model, although it does not include socioeconomic data for travelers.

For a fully functional transport model, the trip ends of the completely specified trips will have to be allocated to the higher resolution TAZs before route assignment can take place. This can be performed before or after mode choice. Train (2009) discusses the issue of geographical aggregation, specifically in regards to destination choice in Chapter 3 of *Discrete Choice Methods with Simulation*, "to specify a destination choice model that is not sensitive to the level of zonal aggregation, representative utility needs to be specified with parameters inside a log operation." As all parameters that actually represent destination utility (population and employment, and foursquare categorical variables) have been transformed with a logarithmic function, The model should be insensitive to the resolution of the zone system, and the estimated coefficients transferable to a an choice set of alternatives based on the TAZs of the larger transport model.

Chapter 6

Discussion

In this section, a discussion of the results is presented, and framed in the context of other work in the field. Firstly, a brief overview of the different models is presented. Chapter 5.1 on estimation explores the models in more detail. The importance and relevance of the work is then discussed. Finally, the limitations of the work are presented and areas where future work is required are detailed.

6.1 Overview of results

The output from the calibrated gravity model was unrealistic to the point of being almost unusable. The correlation of between 0.36 and 0.52 across trip purposes was very low. The calibrated impedance coefficients from the gravity model were retained to build a simple discrete choice model incorporating distance, population and employment. Essentially using the same variables as the gravity model, this first destination choice model performed significantly better across all trip purposes. Further models, incorporating interactions between the origin and destination, such as metropolitan/regional connections further improved the model results. Not all interactions were found to be significant. In particular, modeling the interaction between English and French speaking parts of Canada had only a limited impact.

The next iteration improved the estimation of intra-zonal trips in smaller metropolitan zones. To properly account for intra-zonal behavior, separate parameters were needed for rural and metropolitan origins. The coefficients for these two parameters were of opposite sign, logically indicating that intra-zonal long distance travel is more likely in larger rural zones, but unlikely in metropolitan zones that are barely large enough to support an internal long distance trip. The accuracy of the model for leisure trips was noticeably weaker than for the other two models, and it was hypothesized that better modeling of the destination attractiveness for leisure activities would improve the results.

The LBSN-based models demonstrated how venue check-in data from Foursquare could be incorporated into a destination choice model. As hypothesized, the Foursquare data dramatically improved the estimation of the leisure trips model. It also performed well for visit and business trips. Even with all other variables excluded except for distance, the model (*m4*) performed comparably to the basic *m1* model. When Foursquare data were combined with the variables from previous iterations, a very agreeable model (*m5*) was produced. In the final model *m6*, a second distance term was added to improve the accuracy of the model. Income was not included in the parameters. However, based on the literature, it should still be included in other steps in the model. Limtanakool, Dijst, and Schwanen (2006) found gender and imcome significant for trip generation, and it is widely recognised that income affects mode choice, especially through auto ownership (Ben-Akiva and Lerman 1974; Miller and Ibrahim 1998; Raphael and Rice 2002).

When the model was calibrated to the average trip length, the trip distribution suffered. On closer inspection, the model still overestimates the number of intra-zonal trips within Toronto, and underestimating the inter-zonal trips between large population centers, such as Toronto, Ottawa and Montreal. The car journey from Toronto to Ottawa takes over 4 hours, while flying takes only 55 minutes. For this thesis, only a skim matrix for car travel was available. The incorporation of travel times for all modes, and the inclusion of feedback from the mode choice model, when available, would improve the estimation of these connections.

The scenario analysis of a new ski resort demonstrated how the use of LBSN data to represent destination attractiveness modeled impacts that would otherwise not have been observable with the model. The sensitivity of each parameter in the scenario is also visible. Despite the high correlation between the Foursquare variables at this spatial resolution, the varying impact of different categories is still evident, and shows that parameters need to be considered in the context of the spatial resolution of the zone system.

6.2 Contributions

This thesis provides multiple contributions to the field. As a completed and calibrated domestic destination choice model, it presents a significant amount of work towards a completed long distance transport model for Ontario. The superior performance of the destination choice models over the gravity model validates the work of others in this area (Mishra, Wang, et al. 2013), and adds further weight to the argument that such disaggregate methods, despite the additional effort required, result in better transport models (Sbayti and Roden 2010; Lemp, McWethy, and Kockelman 2007). Bhatta and Larsen (2011) found that the inclusion of intra-zonal trips is important in model estimation, and the investigations in this paper into origin-destination interactions highlight some novel ways to adjust for these in long distance modeling.

Big data already presents both exciting opportunities and daunting challenges to transportation modeling. It is predicted that by 2020, 6.1 billion people will own and use a smartphone (*Ericsson Mobility Report* 2016). With GPS already standard on mobile devices, data is already being collected, in real time and at an unprecedented spatial resolution, that tracks individuals as they travel and interact with their environment. People are also choosing to share more about their behavior, and Foursquare is just one one example of a platform that enables this. The results of the estimation process and sensitivity analysis in this thesis show how even a limited application of such data can dramatically improve destination choice models for leisure travel. The capability to model many aspects of destination attractiveness is particularly useful for leisure travel.

While there has been a "virtual explosion of data availability" Nagel and Axhausen (2001), (Horni and Axhausen 2012) note that the collection of big data such as GPS and GSM data "is generally associated with privacy, cost and technical issues". These challenges go against the ideal of general models that are flexible and transferable (Patriksson 2015). None the less, big data undoubtedly has a role to play in the future of transport modeling. Erath (2015) suggests further research into probabilistic models based on big data and the blending of big data with data from travel diaries.

Traditional mobility surveys such as the TSRC still have an important role to play in transportation modeling. They track the same individuals over time, and provide the socioeconomic characteristics of the individual, which have been repeatedly shown to be important determinants of travel behavior (Pas 1984; Hanson 1982). However, they also have shortcomings; they normally rely on the recall ability of the participant, and are limited in both their spatial and temporal resolutions. The TSRC data exhibited this second shortcoming, which influenced the decisions around spatial resolution in the destination choice model. Most sources of big data collect information from the individual in real time (GPS), or rely on location tracking services to verify the check-in (Foursquare) to an accuracy of meters.

A higher spatial resolution in the observed trip records would allow for a more detailed destination choice model, but would also require a more detailed consideration of another issue, the choice set selection. With only 69 destinations, as used in thesis, it is not essential to restrict the choice set for an individual. However, should the number of alternatives be very large, the choice set presented to each individual needs to be reduced for each individual. Realistically, an individual is not capable of evaluating thousands of possible alternatives when selecting a destination. And any model that assumes this would be unrealistic. In particular, Ben-Akiva and Lerman (1985) found that the the size and selection of the choice set impacts the model performance.

The venue data for each zone essentially acts as database of the points of interest (POI) at a particular destination. POI data is available from many sources, such as Open Street Maps. However, LBSNs such as Foursquare take this POI database one step further, by measuring the popularity of each POI. In the case of Foursquare, as discussed in Section 3.5, when summed together, check-ins measure the intensity of activity at each POI. A measure of importance was clearly beneficial in the model. Not all POIs are equal. Hotels are of different sizes, some national parks are more visited than others. Of course, the importance of each POI can be measured based on attributes such as the number of hotel beds or recorded visitors per year. The data collection required is prohibitive, particularly for a large scale model. LBSN data provides an easily accessible metric the importance of POIs, and in turn, destination utility.

6.3 Limitations and future work

One of the benefits of models based on socioeconomic variables is the ability to run the model for future years and model the impacts of demographic change. Forecasting the Foursquare check-in counts for different categories presents challenges to the modeler. Not only is it hard to predict the how the popularity of certain venues will grow or decline in future years, but the quantity of check-ins depends on uptake of the Foursquare platform and the potential emergence of competing platforms. Further study of the demographics of Foursquare users would help to define the statistical limitations of LBSN-based models.

Many dimensions of the data which were not explored in this thesis, particularly the temporal aspect. Through public services such as twitter, panel data can be collected by associating check-ins over time with an individual. In future work utilizing more detailed Foursquare data, check-ins could be filtered for those performed by residents of Canada, or grouped by season to further improve the modeling of different trip purposes.

In study on why people use Foursquare, Lindqvist et al. (2011) found that “participants expressed reluctance to check-in at home, work, and other places that one might expect them to be at”. This suggests that there at limits to how effectively Foursquare can model travel behavior. A potential alternative would be to use Foursquare or a similar

LBSN as a POI database, and use GPS traces to identify the intensity of activity at these locations, thereby avoiding the selective reporting behavior evident in Foursquare usage.

Further work is still needed to allocate the predicted destinations of the model to the TAZs for Ontario. Some considerations towards this were discussed in Section 5.2, and the resolution of the Foursquare data will facilitate this process by enabling the easy measuring of destination attractiveness at the TAZ level.

6.4 Conclusions

In conclusion, this thesis presents an estimated, calibrated and implemented multinomial logit model for long distance destination choice in Ontario, Canada. The challenges of zonal aggregation were considered, and care taken to specify an appropriate zone system for the model. Models based primarily on population and employment were found to work well for visit and business travel, but not leisure travel. A POI database was built from millions of Foursquare check-ins to test the hypothesis that destination attributes based geo-tagged big data can improve the modeling of destination choice. Alternative specific parameters based on the check-in data did indeed improve the model accuracy across all trip purposes, particularly leisure travel, supporting this hypothesis. The results of the scenario analysis using the fully implemented model also reinforced the importance of properly measuring destination attractiveness for leisure travel. Further work is still needed to incorporate feedback from mode choice into the model, and to allocate destinations within Ontario to the finer resolution zone system of TAZs. The application of Foursquare data showed promising results, and invites further research into utilize big data in destination choice modeling.

Appendix A

Further data analysis

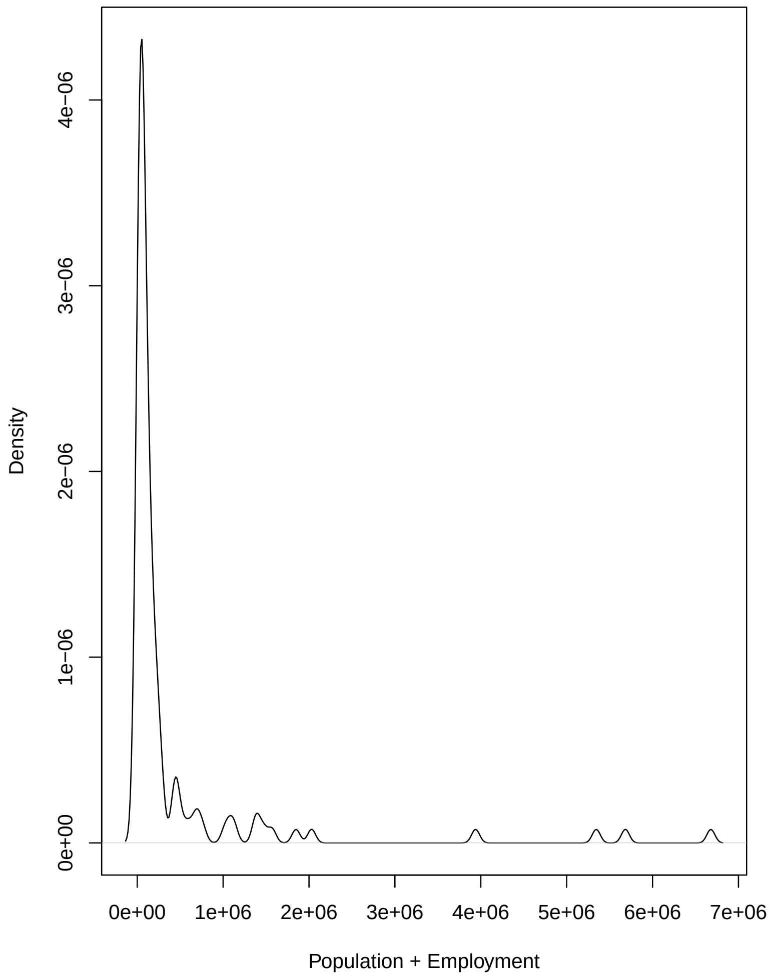

Figure A.1: Long tail and right skew of (population + employment) for each destination

	Population	Total Employment	Goods Industry	Service Industry	Professional	Employment & Health	Arts & Entertainment	Leisure & Hospitality
Population	**1**	**0.99**	**0.94**	**0.98**	**0.93**	**0.99**	**0.92**	**0.96**
Total Employment	**0.99**	**1**	**0.96**	**0.99**	**0.95**	**0.99**	**0.95**	**0.98**
Goods Industry	**0.94**	**0.96**	**1**	**0.92**	**0.84**	**0.94**	**0.91**	**0.94**
Service Industry	**0.98**	**0.99**	**0.92**	**1**	**0.98**	**0.98**	**0.94**	**0.97**
Professional	**0.93**	**0.95**	**0.84**	**0.98**	**1**	**0.94**	**0.91**	**0.93**
Employment & Health	**0.99**	**0.99**	**0.94**	**0.98**	**0.94**	**1**	**0.92**	**0.97**
Arts & Entertainment	**0.92**	**0.95**	**0.91**	**0.94**	**0.91**	**0.92**	**1**	**0.99**
Leisure & Hospitality	**0.96**	**0.98**	**0.94**	**0.97**	**0.93**	**0.97**	**0.99**	**1**

FIGURE A.2: High correlation between population, employment, and various employment categories across destinations

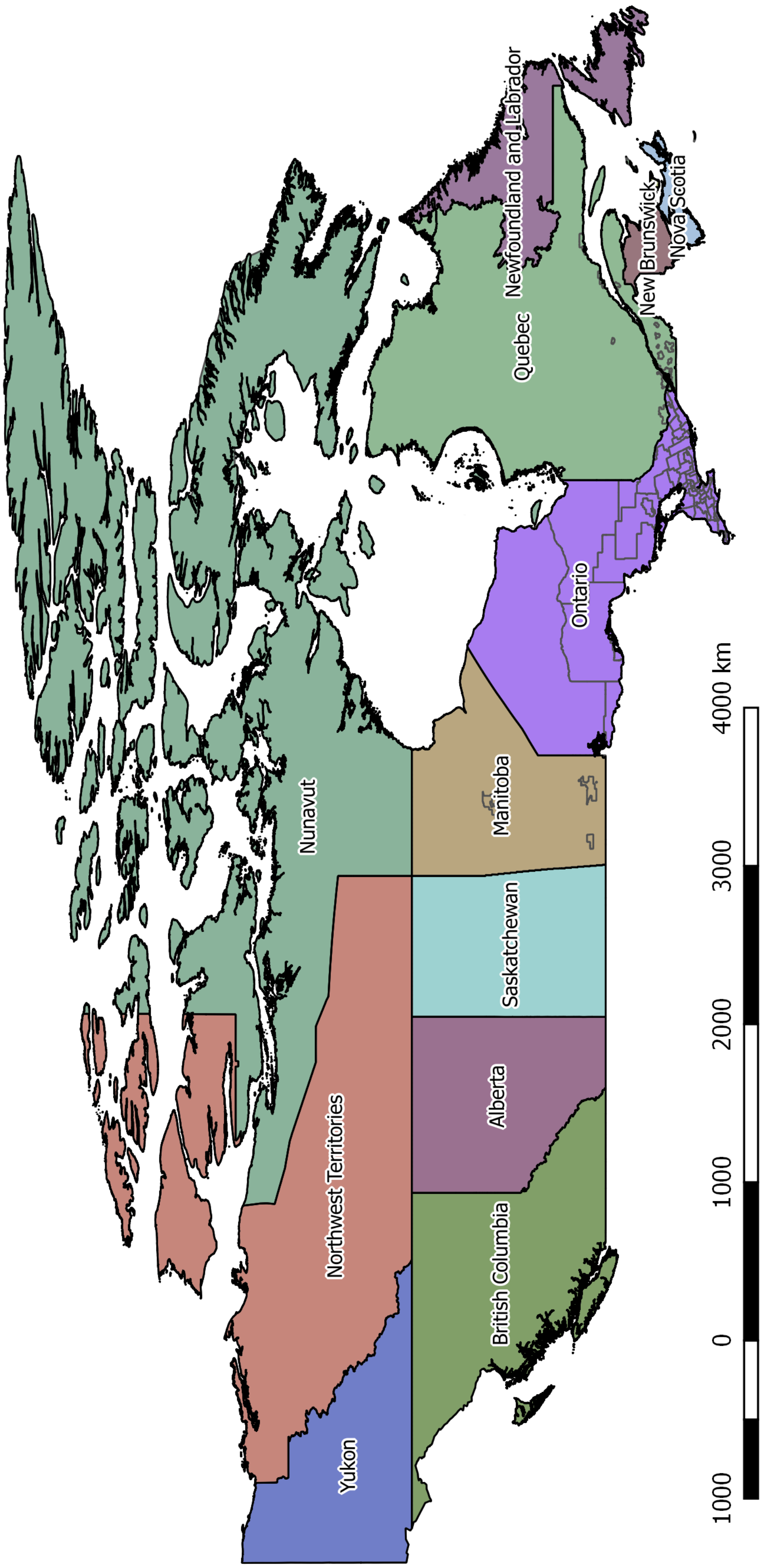

Figure A.3: Zones by province for Canada

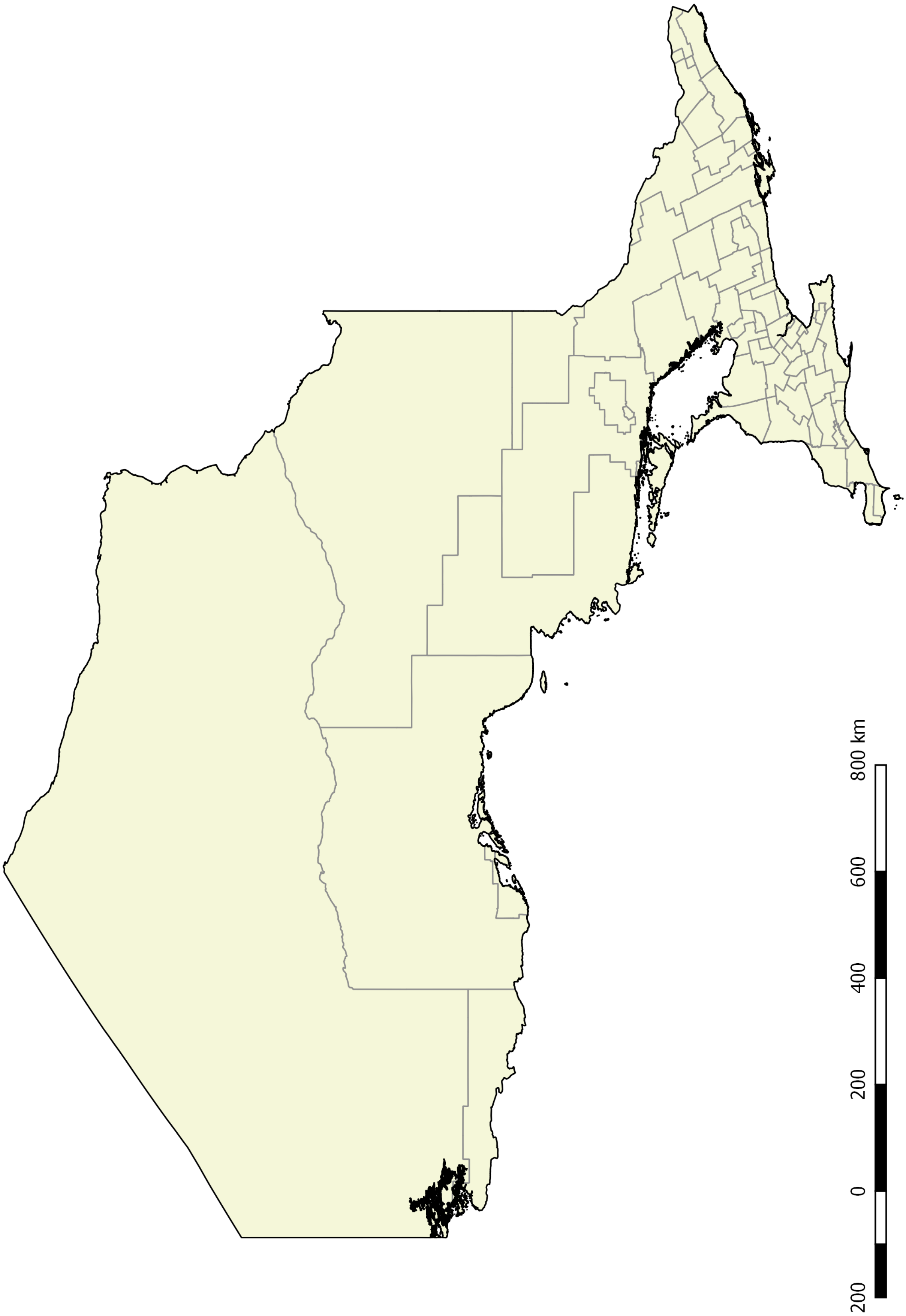

FIGURE A.4: Ontario internal zones

Appendix B

Final model estimation summaries

B.1 Business model strata

```
Call:
mnlogit(formula = f, data = model.inputs[[class]], choiceVar = "alt",
    weights = trips[[class]]$daily.weight, ncores = 8)

Frequencies of alternatives in input data:
...

Number of observations in data = 6228
Number of alternatives = 84
Intercept turned: OFF
Number of parameters in model = 7
  # individual specific variables = 0
  # choice specific coeff variables = 0
  # individual independent variables = 7

-------------------------------------------------------------
Maximum likelihood estimation using the Newton-Raphson method
-------------------------------------------------------------
  Number of iterations: 6
  Number of linesearch iterations: 7
At termination:
```

```
  Gradient norm = 1.65e-06
  Diff between last 2 loglik values = 0
  Stopping reason: Succesive loglik difference < ftol (1e-06).
Total estimation time (sec): 0.762
Time for Hessian calculations (sec): 0.108 using 8 processors.

Coefficients :
                     Estimate Std.Error t-value  Pr(>|t|)
dist_exp             7.794471  0.186216 41.8570 < 2.2e-16 ***
dist_log             0.661833  0.031626 20.9269 < 2.2e-16 ***
civic                0.444640  0.020771 21.4064 < 2.2e-16 ***
mm_inter_no_visit    0.619549  0.041236 15.0244 < 2.2e-16 ***
r_intra              2.003272  0.084420 23.7297 < 2.2e-16 ***
log_hotel            0.151913  0.013159 11.5440 < 2.2e-16 ***
log_sightseeing      0.076344  0.013138  5.8108 6.217e-09 ***
mm_intra_no_business       NA        NA      NA        NA
visit_log_medical          NA        NA      NA        NA
niagara                    NA        NA      NA        NA
summer_log_outdoors        NA        NA      NA        NA
winter_log_skiing          NA        NA      NA        NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -20084, df = 7
AIC:  40182
```

B.2 Leisure model strata

```
Call:
mnlogit(formula = f, data = model.inputs[[class]], choiceVar = "alt",
    weights = trips[[class]]$daily.weight, ncores = 8)

Frequencies of alternatives in input data:
...

Number of observations in data = 22909
```

```
Number of alternatives = 86
Intercept turned: OFF
Number of parameters in model = 11
  # individual specific variables = 0
  # choice specific coeff variables = 0
  # individual independent variables = 11

-------------------------------------------------------------
Maximum likelihood estimation using the Newton-Raphson method
-------------------------------------------------------------
  Number of iterations: 7
  Number of linesearch iterations: 11
At termination:
  Gradient norm = 1.656e-05
  Diff between last 2 loglik values = 0
  Stopping reason: Succesive loglik difference < ftol (1e-06).
Total estimation time (sec): 5.856
Time for Hessian calculations (sec): 1.357 using 8 processors.

Coefficients :
                       Estimate  Std.Error  t-value  Pr(>|t|)
dist_exp              7.3263663  0.1004878  72.9080 < 2.2e-16 ***
dist_log              0.4784102  0.0222526  21.4991 < 2.2e-16 ***
civic                -0.1741501  0.0111768 -15.5814 < 2.2e-16 ***
mm_inter_no_visit    -0.7638604  0.0213704 -35.7439 < 2.2e-16 ***
mm_intra_no_business -2.4007117  0.0524988 -45.7289 < 2.2e-16 ***
r_intra               0.6512536  0.0426148  15.2823 < 2.2e-16 ***
log_hotel             0.2657799  0.0068486  38.8081 < 2.2e-16 ***
log_sightseeing       0.1430539  0.0066258  21.5906 < 2.2e-16 ***
niagara               0.1245882  0.0028645  43.4932 < 2.2e-16 ***
summer_log_outdoors   0.0390384  0.0056098   6.9589 3.429e-12 ***
winter_log_skiing     0.0997631  0.0042804  23.3069 < 2.2e-16 ***
visit_log_medical            NA         NA       NA        NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -77812, df = 11
AIC:  155650
```

B.3 Visit model strata

```
Call:
mnlogit(formula = f, data = model.inputs[[class]], choiceVar = "alt",
    weights = trips[[class]]$daily.weight, ncores = 8)

Frequencies of alternatives in input data:
...

Number of observations in data = 33414
Number of alternatives = 86
Intercept turned: OFF
Number of parameters in model = 8
  # individual specific variables = 0
  # choice specific coeff variables = 0
  # individual independent variables = 8

-------------------------------------------------------------
Maximum likelihood estimation using the Newton-Raphson method
-------------------------------------------------------------
  Number of iterations: 8
  Number of linesearch iterations: 30
At termination:
  Gradient norm = 4.156e-05
  Diff between last 2 loglik values = 0
  Stopping reason: Succesive loglik difference < ftol (1e-06).
Total estimation time (sec): 14.288
Time for Hessian calculations (sec): 1.401 using 8 processors.

Coefficients :
                      Estimate  Std.Error  t-value  Pr(>|t|)
dist_exp             8.9093637  0.0864434 103.0658 < 2.2e-16 ***
dist_log             0.8964084  0.0185310  48.3734 < 2.2e-16 ***
civic                0.2180569  0.0103484  21.0715 < 2.2e-16 ***
mm_intra_no_business -0.9981920  0.0348092 -28.6761 < 2.2e-16 ***
```

```
r_intra               0.6328019  0.0375596  16.8479 < 2.2e-16 ***
visit_log_medical     0.0854261  0.0058964  14.4878 < 2.2e-16 ***
log_hotel             0.1051519  0.0048881  21.5117 < 2.2e-16 ***
log_sightseeing       0.0464822  0.0048149   9.6538 < 2.2e-16 ***
mm_inter_no_visit            NA         NA       NA        NA
niagara                      NA         NA       NA        NA
summer_log_outdoors          NA         NA       NA        NA
winter_log_skiing            NA         NA       NA        NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -113190, df = 8
AIC:  226410
```

Appendix C

Further model results

TABLE C.1: *m2* Results.
Toronto: zones 20-22, Niagara: zone 30

	Origin	Destination	Type	Predicted	Observed	Absolute Error	Max Rel. Error
1	21	30	II	877.21	3695.07	2817.86	3.21
2	85	72	EE	123.67	2407.73	2284.06	18.47
3	21	20	II	4507.84	2251.48	2256.36	1.00
4	21	22	II	4844.63	2680.21	2164.42	0.81
5	36	21	II	1346.76	3085.23	1738.46	1.29
6	103	4	EI	541.86	2061.78	1519.92	2.80
7	103	21	EI	198.15	1529.83	1331.68	6.72
8	21	53	II	821.51	2115.53	1294.01	1.58
9	64	64	II	209.47	1423.02	1213.55	5.79
10	21	54	II	215.05	1346.02	1130.97	5.26
11	20	30	II	261.47	1365.27	1103.80	4.22
12	22	30	II	352.63	1420.03	1067.40	3.03
13	30	30	II	157.40	1178.94	1021.54	6.49
14	21	52	II	804.06	1818.06	1014.00	1.26
15	21	4	II	264.90	1238.33	973.43	3.67
16	29	21	II	1165.79	2124.10	958.31	0.82
17	29	30	II	428.45	1353.10	924.65	2.16
18	4	21	II	403.14	1318.43	915.28	2.27
19	47	21	II	631.84	1535.96	904.12	1.43
20	4	85	IE	1660.31	809.08	851.23	1.05

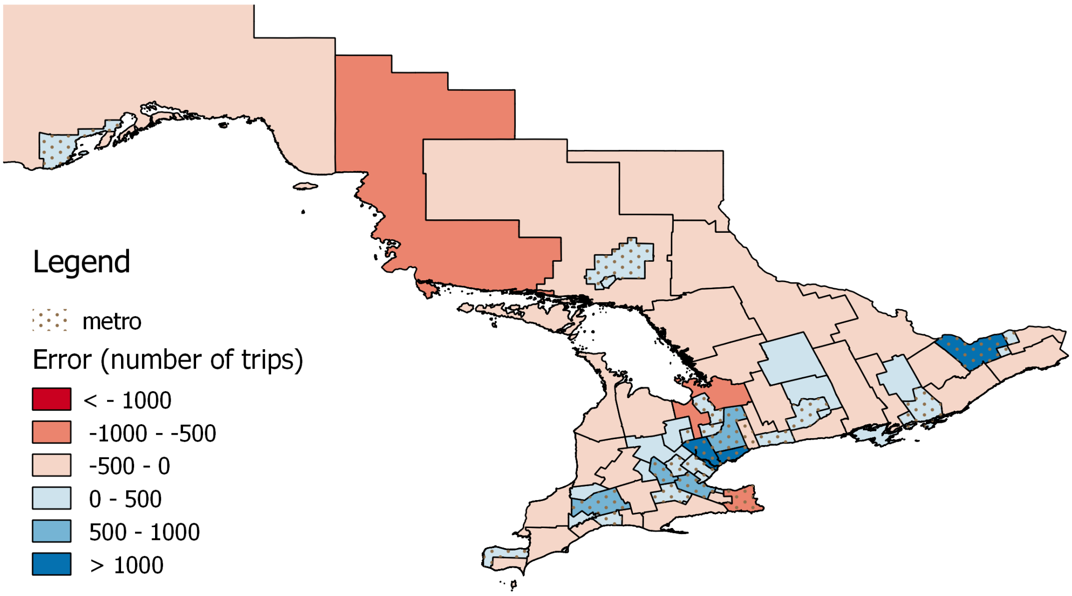

Figure C.1: Intrazonal errors produced by the *m2* model

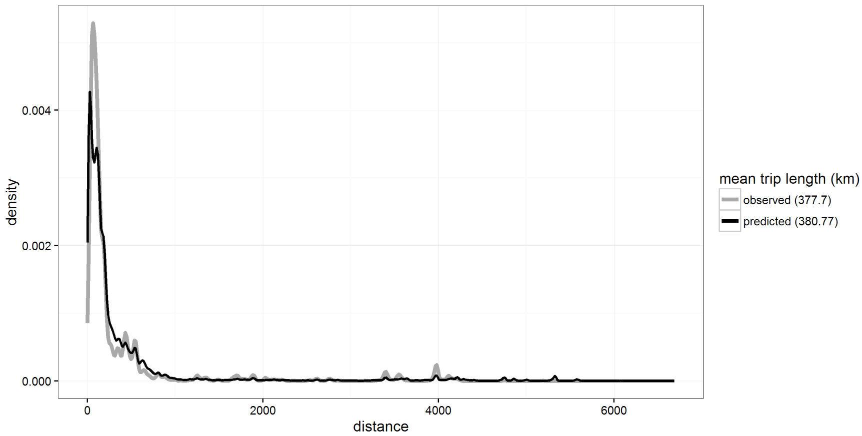

Figure C.2: Trip length distribution of calibrated model for business travel

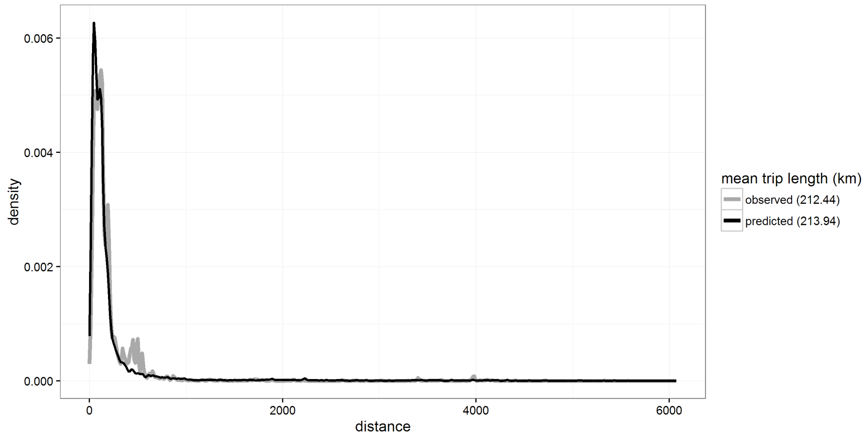

Figure C.3: Trip length distribution of calibrated model for leisure travel

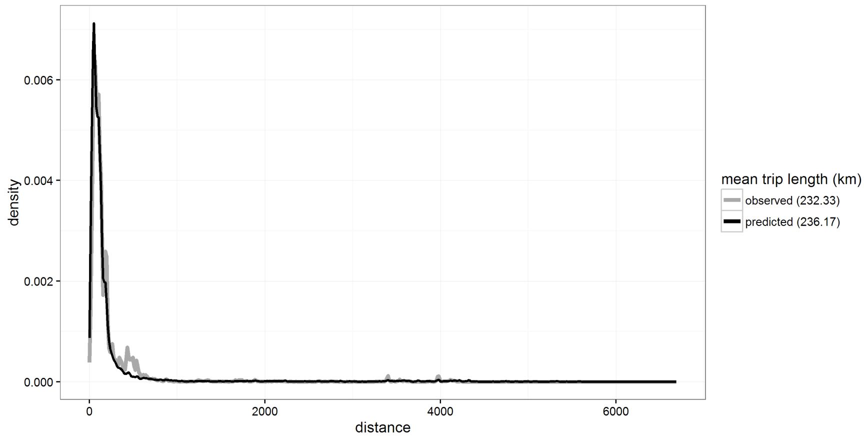

Figure C.4: Trip length distribution of calibrated model for visit travel

Table C.2: Scenario Analysis results, incoming trips to zone 24: Dufferin (Toronto)

Scenario	business	leisure	visit	total
pre ski resort	573	2345	5183	8101
civic	577	2341	5241	8159
hotels	775	3895	6396	11066
outdoors (summer only)	776	4802	6351	11929
skiing (winter only)	773	5532	6357	12662

Bibliography

Abdulazim, Tamer et al. (2015). "Framework for Automating Travel Activity Inference Using Land Use Data: The Case of Foursquare in the Greater Toronto and Hamilton Area, Ontario, Canada". In: *Transportation Research Record: Journal of the Transportation Research Board* 2526, pp. 136–142.

Adler, Thomas J and Moshe E Ben-Akiva (1976). "Joint-choice model for frequency, destination, and travel mode for shopping trips". In: *Transportation Research Record* 569.

Adler, Thomas and Moshe E Ben-Akiva (1979). "A theoretical and empirical model of trip chaining behavior". In: *Transportation Research Part B: Methodological* 13.3, pp. 243–257. ISSN: 0191-2615.

Anas, Alex (1983). "Discrete choice theory, information theory and the multinomial logit and gravity models". In: *Transportation Research Part B: Methodological* 17.1, pp. 13–23. ISSN: 0191-2615.

Ben-Akiva, Moshe E (1974). *Structure of passenger travel demand models.* ISBN: 0309023734.

Ben-Akiva, Moshe E and Steven R Lerman (1974). "Some estimation results of a simultaneous model of auto ownership and mode choice to work". In: *Transportation* 3.4, pp. 357–376.

– (1985). *Discrete choice analysis: theory and application to travel demand.* Vol. 9. MIT press. ISBN: 0262022176.

Beyer, Mark A and Douglas Laney (2012). "The importance of 'big data': a definition". In: *Stamford, CT: Gartner*, pp. 2014–2018.

Bhatta, Bharat P and Odd I Larsen (2011). "Are intrazonal trips ignorable?" In: *Transport policy* 18.1, pp. 13–22.

Boyce, David E et al. (1983). "Implementation and computational issues for combined models of location, destination, mode, and route choice". In: *Environment and Planning A* 15.9, pp. 1219–1230.

Broach, Joseph, Jennifer Dill, and John Gliebe (2012). "Where do cyclists ride? A route choice model developed with revealed preference GPS data". In: *Transportation Research Part A: Policy and Practice* 46.10, pp. 1730–1740.

Canadian Transport Commission (1971). “Intercity passenger transport study”. In:

Casey, HJ (1955). “Applications to traffic engineering of the law of retail gravitation”. In: *Traffic Quarterly* 9.1, pp. 23–35.

Cheng, Zhiyuan et al. (2011). “Exploring Millions of Footprints in Location Sharing Services.” In: *ICWSM* 2011, pp. 81–88.

Daly, Andrew (1982). “Estimating choice models containing attraction variables”. In: *Transportation Research Part B: Methodological* 16.1, pp. 5–15. ISSN: 0191-2615.

Dios Ortúzar, Juan de, Luis G Willumsen, et al. (1994). *Modelling transport.* Wiley New Jersey.

Erath, Alex (2015). “Transport Modelling in the Age of Big Data”. In: Seoul: University of Seoul.

Ericsson Mobility Report (2016). Tech. rep. Ericsson. URL: `https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf`.

Hanson, Susan (1982). “The determinants of daily travel-activity patterns: relative location and sociodemographic factors”. In: *Urban Geography* 3.3, pp. 179–202.

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

Horni, Andreas and Kay W Axhausen (2012). *How to improve MATSim destination choice for discretionary activities?* Eidgenössische Technische Hochschule Zürich, IVT, Institute for Transport Planning and Systems.

Jin, Peter et al. (2014). “Location-Based Social Networking Data: Exploration into Use of Doubly Constrained Gravity Model for Origin-Destination Estimation”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2430, pp. 72–82.

Jonnalagadda, Nageswar et al. (2001). “Development of microsimulation activity-based model for San Francisco: destination and mode choice models”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1777, pp. 25–35.

Kitamura, Ryuichi (1984). “Incorporating trip chaining into analysis of destination choice”. In: *Transportation Research Part B: Methodological* 18.1, pp. 67–81. ISSN: 0191-2615.

Kitamura, Ryuichi, Patricia L Mokhtarian, and Laura Laidet (1997). “A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area”. In: *Transportation* 24.2, pp. 125–158.

Lemp, Jason, Laura McWethy, and Kara Kockelman (2007). “From aggregate methods to microsimulation: Assessing benefits of microscopic activity-based models of travel demand”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1994, pp. 80–88.

Limtanakool, Narisra, Martin Dijst, and Tim Schwanen (2006). "On the participation in medium-and long-distance travel: A decomposition analysis for the UK and the Netherlands". In: *Tijdschrift voor economische en sociale geografie* 97.4, pp. 389–404.

Lindqvist, Janne et al. (2011). "I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pp. 2409–2418.

McFadden, Daniel (1973). "Conditional logit analysis of qualitative choice behavior". In:

Menghini, Gianluca et al. (2010). "Route choice of cyclists in Zurich". In: *Transportation research part A: policy and practice* 44.9, pp. 754–765.

Miller, Eric (2004). "The trouble with intercity travel demand models". In: *Transportation Research Record: Journal of the Transportation Research Board* 1895, pp. 94–101. ISSN: 0361-1981.

Miller, Eric and Amal Ibrahim (1998). "Urban form and vehicular travel: some empirical findings". In: *Transportation Research Record: Journal of the Transportation Research Board* 1617, pp. 18–27.

Mishra, Sabyasachee, Yanli Wang, et al. (2013). "Comparison between gravity and destination choice models for trip distribution in Maryland". In: *Transportation Research Board 92nd Annual Meeting*.

Mishra, Sabyasachee, Xin Ye, et al. (2011). "A functional integrated land use-transportation model for analyzing transportation impacts in the Maryland-Washington, DC Region". In: *Sustainability: Science, Practice, & Policy* 7.2.

Moeckel, Rolf and Rick Donnelly (2015). "Gradual rasterization: redefining spatial resolution in transport modelling". In: *Environment and Planning B: Planning and Design* 42.5, pp. 888–903.

Moeckel, Rolf, Rhett Fussell, and Rick Donnelly (2015). "Mode choice modeling for long-distance travel". In: *Transportation Letters* 7.1, pp. 35–46. ISSN: 1942-7867.

Nagel, Kai and Kay W Axhausen (2001). "Workshop report: Microsimulation". In:

Newman, Jeffrey P and Vincent L Bernardin Jr (2010). "Hierarchical ordering of nests in a joint mode and destination choice model". In: *Transportation* 37.4, pp. 677–688.

Noulas, Anastasios et al. (2012). "A tale of many cities: universal patterns in human urban mobility". In: *PloS one* 7.5, e37027.

Oi Walter; Schuldiner, Paul (1962). *An Analysis of Urban Transportation Demand*. Evanston, IL: Northwestern University Press.

Oreskes, Naomi (2004). "The scientific consensus on climate change". In: *Science* 306.5702, pp. 1686–1686.

Outwater, Maren L et al. (2015). "Tour-Based National Model System to Forecast Long-Distance Passenger Travel in the United States". In: *Transportation Research Board 94th Annual Meeting*.

Outwater, Maren et al. (2010). “California statewide model for high-speed rail”. In: *Journal of Choice Modelling* 3.1, pp. 58–83. ISSN: 1755-5345.

Pan, Changxuan et al. (2006). “Cellular-based data-extracting method for trip distribution”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1945, pp. 33–39.

Pas, Eric I (1984). “The effect of selected sociodemographic characteristics on daily travel-activity behavior”. In: *Environment and Planning A* 16.5, pp. 571–581.

Patriksson, Michael (2015). *The traffic assignment problem: models and methods*. Courier Dover Publications, p. 6.

Raphael, Steven and Lorien Rice (2002). “Car ownership, employment, and earnings”. In: *Journal of Urban Economics* 52.1, pp. 109–130.

Rohr, Charlene et al. (2010). “Modelling Long-Distance Travel in the UK”. In: *European Transport Conference, 2010.*

SA, Rokib et al. (2015). “Origin-Destination Trip Estimation from Anonymous Cell Phone and Foursquare Data”. In: *Transportation Research Board 94th Annual Meeting*. 15-2379.

Sbayti, Hayssam and David Roden (2010). “Best practices in the use of micro simulation models”. In: *Prepared for: American Association of State Highway and Transportation Officials (AASHTO) Standing Committee on Planning. AECOM, Arlington, VA.*

Schiffer, RG (2012). “NCHRP Report 735: Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models”. In: *Transportation Research Board of the National Academies, Washington, DC.*

Schönfelder, Stefan et al. (2006). *Analysis of commute Atlanta instrumented vehicle GPS data: Destination choice behavior and activity spaces*. IVT, ETH, Zürich.

Simma, Anja, Robert Schlich, and Kay W Axhausen (2001). “Destination choice modelling of leisure trips: The case of Switzerland”. In: *Arbeitsberichte Verkehrs-und Raumplanung* 99.

Spear, Bruce D (1977). *Applications of new travel demand forecasting techniques to transportation planning. A study of individual choice models*. Report. URL: `http://ntl.bts.gov/DOCS/SICM.html`.

Stabler, Ben et al. (2014). “Collaborative Development of an Open Matrix Standard”. In: *Innovations in Travel Modeling Conference*. Baltimore, Maryland.

Statistics Canada (2014). *Travel Survey of Residents of Canada (TSRC)*. English. URL: `http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3810`.

Stouffer, Samuel A (1940). “Intervening opportunities: a theory relating mobility and distance”. In: *American sociological review* 5.6, pp. 845–867.

Train, Kenneth E (2009). *Discrete choice methods with simulation*. Cambridge university press. ISBN: 1139480375.

Warner, Stanley Leon (1962). "Stochastic choice of mode in urban travel: A study in binary choice". In: *Metropolitan Transport Series*.

Wilson, Alan G (1967). "A statistical theory of spatial distribution models". In: *Transportation research* 1.3, pp. 253–269. ISSN: 0041-1647.