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 Transport Systems

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

December 2016

Declaration of Authorship

I, AUTHOR NAME, declare that this thesis titled, 'THESIS TITLE' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

Technical University of Munich

Abstract

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

Master of Science in Transport Systems

by Joseph Molloy

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor…

Contents

List of Figures

List of Tables

Abbreviations

LAH	**L**ist **A**bbreviations **H**ere

Chapter 1

Literature Review

Also known as Intercity models, long-distance transport 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, "either 50, 75, or 100 miles as the minimum threshold for trips to be considered long-distance."(Schiffer 2012)

According to 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 often hard to replicate, if they are published. 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)

1.1 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 (1962; 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.

1.2 Discrete Destination Choice Models

Despite their widespread acceptance and use in practice, the aggregate gravity model has some fundamental flaws as a modelling methodology. As an alternative, McFadden (1973) and M. E. 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 socio-economic data much better, avoiding 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 modelling has been continually refined, with important research done in both modelling 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 M. E. 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 used. Since the utility of every alternative must be calculated, this technique was prohibitive before the advent of modern computers for large scale problems. 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, AkivaLerman85 present a comprehensive discussion of destination choice models. They note that "destination choice is characterized by a very large number of alternatives", and that 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.

Discrete choice models have been repeatedly shown to provide better results than aggregate methods when modeling travel behavior (1984; 2013) compared the gravity model and multinomial logit destination choice model when integrated into a model for Maryland. They found that the destination choice model performed much better than the gravity model for a state-wide travel demand model.

1.3 Trip Chaining

Adler and M. Ben-Akiva (1979) were one 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 area, focus 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 socio-economic 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 isn't 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 that A to a trip maker when it is surrounded by destinations supporting other opportunities that the trip maker wishes to pursue.

1.4 Recent Intercity Transport Models

(M. Outwater et al. 2010) developed a state-wide model to model high speed rail. 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 combine some network data to calculate auto and non-auto accessibility, for peak and off-peak respectively. Destination was estimated 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 one of 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 model focuses includes 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 classical model. Unlike activity based models, it uses a temporal resolution of months and weeks, not days, and jointly predicts destination and mode choice together.

1.5 Mnlogit R Package

In this thesis, the R package *mnlogit* (Hasan, Zhiyu, and Mahani 2014) is used to estimate the multinomial model of destination choice. This package provides improvements over the classic *mlogit* package, by reducing computer memory usage to allow for more model parameters. It also performs the maximum likelyhood estimation in parallel for significant decrease in estimation runtime.

1.5.1 Data format

The model input must be provided in a long format. The format of the input can described as followed: 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, where

$$R_s = \{a | R_s \subseteq R \wedge \exists t \in S : destination(t) = a\}$$

The model input is then 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 that indicated whether a particular choice was chosen for that trip or not.

1.5.2 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 vary does not vary across the choice set for each trip, only between trips. Alternative specific variables can have coefficients 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 modelling; 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. Intersects can also be removed by adding $+0$ to the respective section. 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.

1.6 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 (LBSN), 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 'checkin' to a 'venue', such as a shop, tourist attraction or airport, and provide tips, ratings and reviews. When these services are used by millions of people around the world, in different countries and cities, a enourmous 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 checkins across 220,000 users to quantitatively assess human mobility patterns. 53% of their checkins came from foursquare, highlighting the dominance of foursquare in the LSBN space. They sampled location sharing status updates from the public Twitter feed, identifying users with geo-referenced tweets, and then collected the most recent 2000 geo-labeled tweets.

Noulas et al. (2012) used foursquare data to design gravity model based on Stouffer's theory of intervening opportunities. 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 analyse 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) introduces a framework for inferring activity travel given nearby land use information gathered from LSBNs. Their results suggest that daily activity travel can be automatically inferred from LSBN data, and they present a generic method for acquiring land use data from LSBN 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 and Origin Destination Demand matrix. They found that The cell phone and Foursquare data were consistent with OD pairs expected 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 O-D estimation errors caused by sampling bias when compared to a singly constrained model.

1.7 Objectives

Disaggregate models provide clear advantages over aggregate methods in modelling trip distribution. While aggregate methods still hold sway in the modelling of long distance travel, due to the availability of data and more powerful computers, the modelling of destination choice using logit models is becoming more popular. Destination choice models provide more flexibility in attribute selection, and more efficient use of data. Most models include the basic socio-economic variables and a description of zone attractiveness.

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 weren't available to modellers working on other regions, or be more restricted, forcing the modeler to be creative in designing parameters that can represent the travel behavior in the region. 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 reflecting was added to the model, not only would it be computationally infeasible, but there would be a high risk of overfitting in the model. The fact that the destination choice models already presented in the literature are individually unique supports this notion that modelling is both a science and an art, and that there is no "one size fits all" model.

For these reasons, the design of a destination choice model for Ontario in itself reflects a new contribution to the field. The field of transport modelling is advanced every time the process of a designing a new model is performed. 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 contribution of this thesis is the application of LSBNs to improve the utility modelling of destination choice. While work has been done on investigating mobility patterns, and the generation of OD matrices using LSBNs, their application to disaggregate models in transport has not been considered. This thesis will explore how foursquare check-in data can be used in the calculation of destination utility, and whether it can effectively replace other socio-economic variables such as employment. This is a significant contribution, as socio-economic data is not always available, especially at the high modelling resolutions made possible by the advances in microsimulation and computing technologies. check-ins data also provides an opportunity to model important traits of destination utility, such as the presence of national parks, that are not reflected in standard socio-economic variables.

Chapter 2

Data Acquisition and Analysis

2.1 Travel Survey of Residents of Canada

2.1.1 Introduction

The Transport Survey of Residents of Canada (TSRC) is a monthly, cross-sectional survey collected by Statistic 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 socio-economic 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 Organisation guidelines, and align the recorded activities with the International Travel Survey (ITS).

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

2.1.2 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.

2.1.3 Data

2.1.3.1 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 Metropolitain Areas (CMA) and Census Agglomerations (CA). CMAs and CAs represent certain clustered areas of population around an urban core. More specifically, to be defined as a CMA, an area must have a total population of at least 100,000, with half of those living in the 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. CMAs are only recorded for 51.5% of trip origins, and 48.3% of trip destinations. However, CMA's can be used to better allocate trips to transport zones in urban areas, as discussed in section (CITE).

2.1.3.2 Error Detection and Imputation

The computer-assisted nature of the survey allows for real-time error detection and consistency checking during the interview process. One example is that the program will inform the interviewer if the number of nights recorded for a trip does not match the number of nights recorded in various types of accommodations. 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.

2.1.3.3 Weighting

The weighting of records is particularly important when working with survey data. They allow the researcher to scale up the results for a sample to build an accurate representation of population, taking into account under- and over- represented groups within the survey. In total, four weightings are provided for the TSRC, with two relating to trip avriables: 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 recalibrated to reflect subsampling, non-response, and known control groups.

After the 2011 redesign, respondents are asked about same-day trips that ended in the previous month, but 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 treated for outliers and recall bias. 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 socio-economic characteristics, as well all trip and visit variables, excluding expenditures. Trip weight (WTTP) is then calculated by dividing the number of household members that went on the trip, is only used to calculate expenditures, and as such is not relevant to the model design.

2.1.4 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 don't 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, or 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.

2.1.4.1 Trip Datafile

Trips covered by the TSRC include same-day trips of more than 40km 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.

Socio-economic 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, 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

2.1.4.2 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 visit records from being used model trip chaining, as discussed in section (CITE).

2.1.5 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 2.1 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 2.1: 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 is an activity that people willing travel long distances for. In 2014, TSRC respondents reported participating in winter sports in 4% of overnight trips, and 2% of same day trips. These trips naturally only occur in the winter months, at certain destinations.

2.2 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 didn't fit one of these categories is also excluded from the trip dataset used to estimate the destination choice model. Internal, incoming and outgoing trips don't need to be filtered, and all trips in these categories are retained in the trip set. External trips, on the other hand, are filtered to remove trips that don't 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.

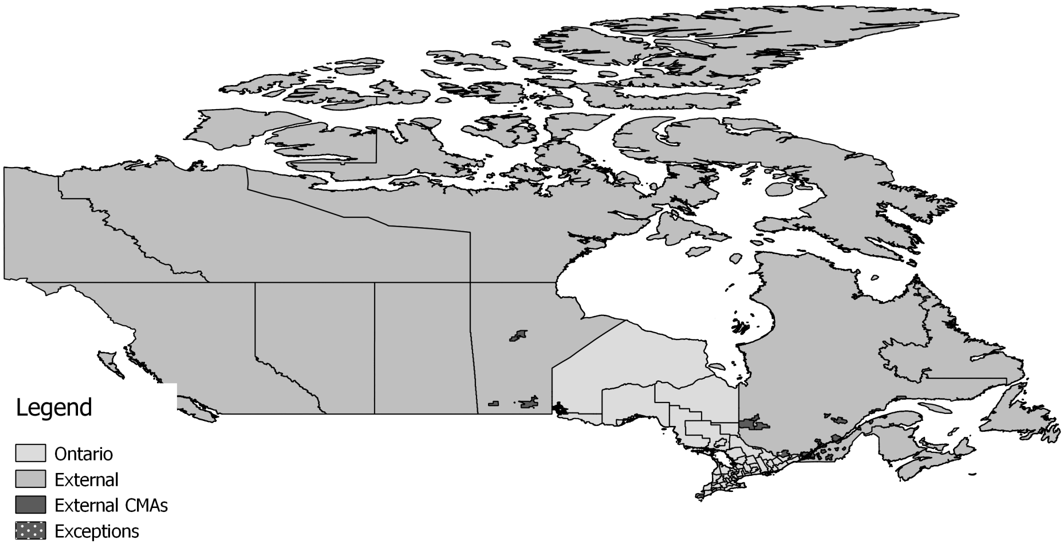


Figure 2.1: Dividing external zones into east and west

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 2.1, dividing the external zones into two groups, east and west. Trips originating in one group and arriving in another have to pass through Ontario. And the converse in true for trips within a group. Hence all trips that don't 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 2.2 illustrates a journey from Gatieau, Quebec to Montreal Airport takes around 2 hours when passing through Ontario, and 2 hours and 30 minutes otherwise. Trips between these two exception zones and all other zones in Quebec are therefore retained.

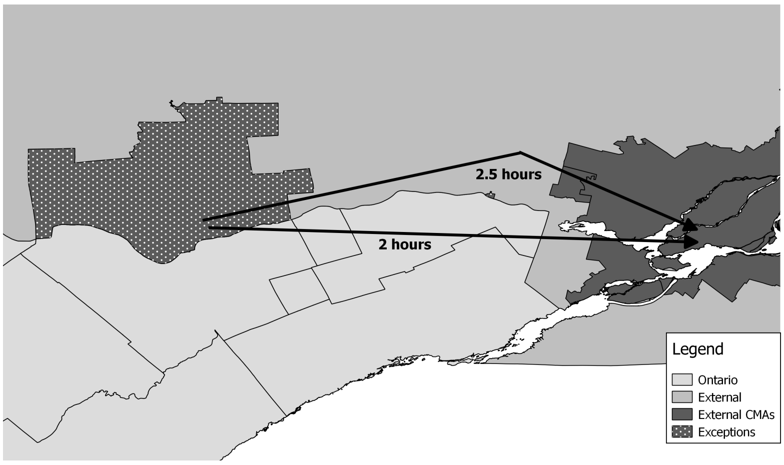


FIGURE 2.2: An example of an external origin-destination pair that passes through Ontario

Figure 2.3 illustrates how our the trip distribution of the estimated trips fits the observed distribution much better after undesired external trips are removed. In total 69,328 individual trip records remain from the TSRC dataset for model estimation, representing 40,177,841 weighted trips.

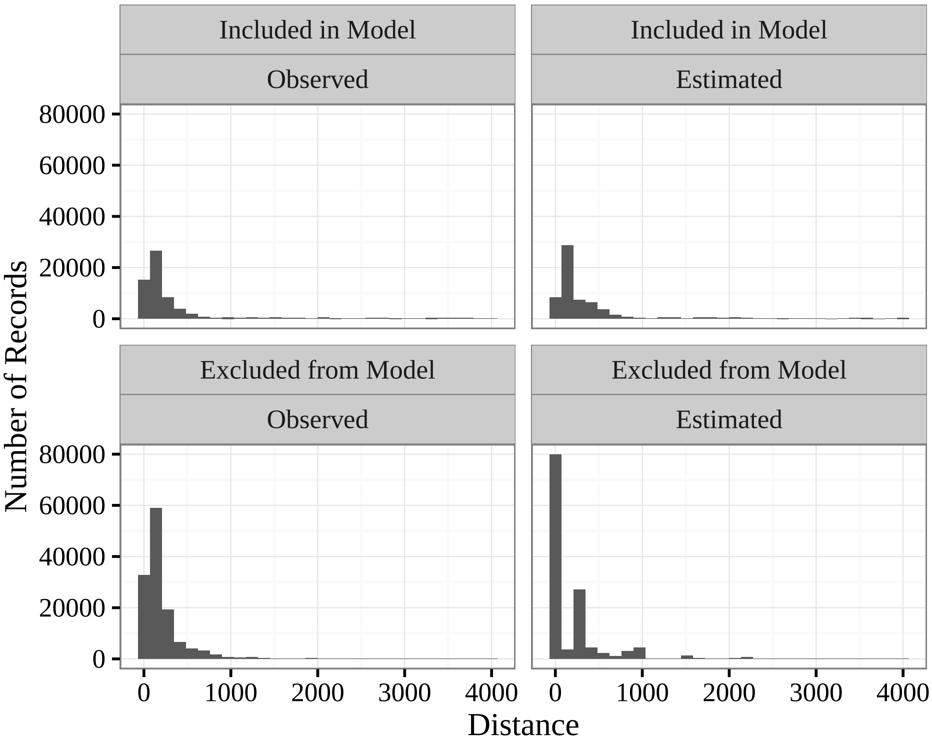


Figure 2.3: Estimated vs. Observed Trip Distance

2.3 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.

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

This chapter 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. The size of the choiceset needs to be considered. Large destinations choice sets 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 and validity of the coefficients.

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 thesis is only concerned with domestic travel within Canada, only the Internal zones for Ontario, and 48 external zones within Canada are considered. The external zones are not modified as TSRC origins and destinations are directly translatable to the external zones.

For the Internal zones (TAZs) within Ontario are allocated using a gradual raster based zone approach, based on the method presented 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.

However, the TSRC trip origins and destinations don't match this custom zone system for Ontario, being only available at a much broader resolution. For this thesis, a zone system is designed based on the TSRC spatial resolution for the design of a destination choice model. The allocation of trips to the TAZ level will be performed at a later stage in the transport model, and the proposed method is discussed further in section 4.3.

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

As discussed in 2.1.3, provinces and census divisions cover the national study area completely. hence as a first step, the zones are defined by the census divisions comprising Ontario, of which there are 49. However, even in rural areas, the TAZs are much smaller than the size of a Census Division. 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 (see figure 2.4.

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 modellers. By simply including CMAs as zones in the aggregated zoning model, the number of zones is increased to 57, allowing trip origins and destinations in urban areas to be more accurately assigned during disaggregation to TAZs. As seen in figure 2.4, this significantly lowers the mean number of TAZs per zone, and also reduces the spread of values, indicating a much improved zone system.

This approach has one drawback, as a large outlier, representing the Toronto CMA is now observable, consisting of over 2000 TAZs. This outlier corresponds to the CMA of Toronto, the most populous in Ontario, both a very large generator and attractor of trips. In 2014, Toronto represented 13.4% and 10% of trip origins and destinations respectively. It should be clear that this is a very undesirable occurrence. While it is hard to say whether this outlier would significantly affect the destination choice model, it would regardless provide significant challenges when allocating trips to individual TAZs, affecting the overall quality of the model. It is worth noting briefly that the choice of zone system can have cascading effects throughout the whole transport model, which need to be considered during its creation.

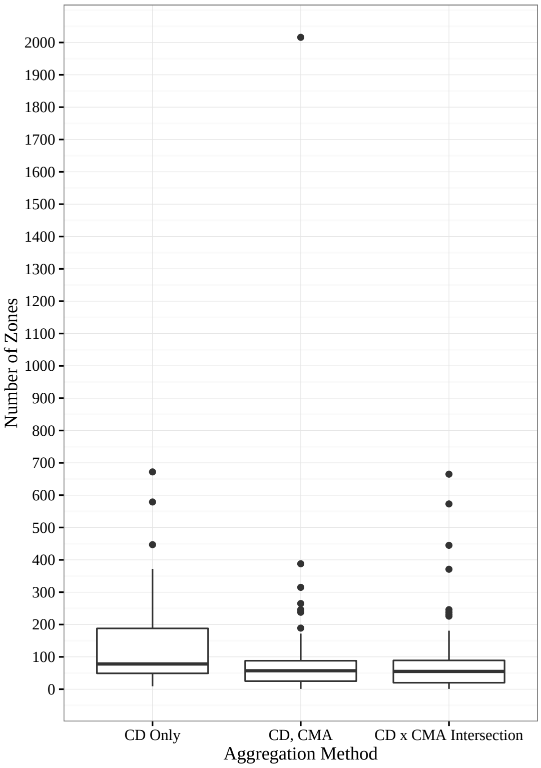


Figure 2.4: Different methods of aggregating internal zones to match the TSRC spatial resolution.

This lone 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, we can overlap the CDs and CMAs to fully reflect the number of destination choices available in the TSRC data. This is done as followed; the area of each CD that does not intersect with a CMA is selected as a zone. Then, each unique combination of CMA and CD is recorded as a new zone. An example of this process is shown in figure (CITE). This process gives us 69 zones for Ontario, at 41% increase over the simplest approach that only considers CDs.

Figure 2.4 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 better results when trips origins and destinations are disaggregated. Taking the intersection of CDs and CMAs has little effect on most zones, but still improves the overall result in a very significant way. 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, in that 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 that, interestingly, CMAs are more likely to be origins than destinations.

2.4 Aggregating of Zonal Data

All the data on distances, population and employment was provided at the TAZ level. This section describes how they were allocated to the zone system. The TAZs themselves were assigned to the zone which intersected their centroid. Where the centroid of the TAZ didn't intersect any zones, the first intersecting zone was used. When the TAZ did not intersect with any part of the Canadian census boundaries at all, it was assigned manually to the nearest zone.

Socio-economic variables, namely population and employment were aggregated from the TAZ level to the zone system using a summation.

A distance matrix was provided for the original set of TAZs. It was calculated without congestion using the Canadian road network and intra-zonal travel times were not included. This skim matrix needed to be aggregated to the zone system used in the destination choice model. This was done by taking a distance d between each pair of sub zones weighted by the multiplied populations p of the origin and destination.

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

2.5 Foursquare

Trip distribution models that consider only population and employment, while common, have a significant flaw. They fail to account for landmarks and attractions, such as National Parks that attract large numbers of people, yet have low population or employment. Destination choice modeling provides the opportunity to incorporate parameters that reflect these drivers of travel demand.

Leisure travel is a particular case where socio-economic metrics don't always reflect the attractiveness of a destination. Areas such as as lakes, national parks and ski resorts are popular long 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 of 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 doesn't 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.

2.5.1 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 prevented this method. Instead the public venue API was used, which is much more limited in its scope. 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 recommendations
- Number of total check-ins

Each request is limited to roughly 1 square degree 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 doesn't return check-in counts by date, so it can only be used to generate a total historical metric of activity for each venue. For the forecasting of trips to individual venues, this would present a significant obstacle. However, in this thesis, the foursquare metrics are only used for identifying the intensity of activity in zones for significant attractions that can't be reflected in socio-economic variables. Check-in counts also can't be filtered by origin country or state. This capability would, in a larger model, also allow us to identify US destinations that are commonly visited by Canadian travelers.

2.5.2 Demographic Bias

While the use of social networks is becoming more ubiquitous throughout the general population, LSBNs 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.

2.5.3 Methodology

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

1. A developer account was set up which allowed access to the API. The maximum search area allowed is smaller than most external zones, so a search grid of 1 degree raster cells was generated for all of Canada.

2. Using the activities specified in the TSRC as a reference, a selection of potentially important venue categories was curated.

3. 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 that such *States & Municipalities.*

4. 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. The request is 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.

5. Unique Venues were then stored in the PostGIS database, and tagged with the zone to which it geographically belongs.

6. Venue and check-in counts by category were aggregated for each zone.

In table 2.2, 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.

2.5.4 Summary

In conclusion, the foursquare API provides data at a higher spatial resolution than the TSRC microfile, but without the temporal detail. While more temporal detail could naturally improve the model, The ranking and total check-in counts for each venue still provide very useful indicators for the intensity of activity in many zones. These metrics are particularly useful for identifying destinations in non-metropolitan areas.

Table 2.2: 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 3

Gravity Model

As discussed in section 1, 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 in simpler models can aid 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 4.

3.1 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:

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

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.
k is the impedance factor, calibrated with the average trip distance.
d_{ij} is the distance between zones i, j.

3.2 Model Strata

It is common practice to design a transport model that is in reality, a collection of separate models for heterogeneous groups of travelers, and trips. Using dis-aggregate modeling greatly assists this process, as we can cluster trips by particular attributes into more homogeneous categories. The most common attribute to categorize by is trip purpose. The TSRC provides two attributes recording trip purpose, and we use first, more general categorization for our model. This consists of 4 Categories, Business, Visit, Leisure and Other, of which only the first three are used. The category other is excluded. In figure 3.1, the number of trips for each category by year in the TSRC is shown. Each purpose, especially when multiple years are combined in one dataset, has a suitably large sample size to support model estimation.

Table 3.1: 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	2,2909
Visit	9,057	8,777	7,962	7,618	3,3414
Total	18,694	18,016	1,6547	16,071	6,9328

3.3 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 3.2. The average observed trip distance is $\bar{d}$, the average predicted trip distance d, and the impedance factor k. As measurements of error, the root mean square error (RMSE) and model correlation (r^2) are provided.

Table 3.2: Gravity Model calibration

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

3.4 Results

Figure 3.1 presents an error plot, with the 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, 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 3.1 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 3.1, as the error Large y values are only of concern when the x value, namely absolute error, is also large.

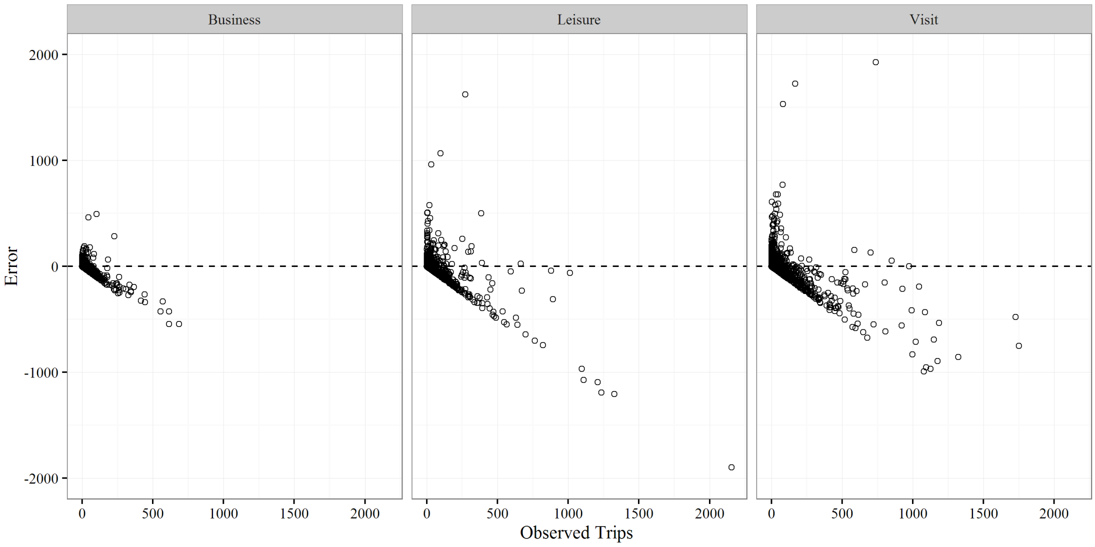


Figure 3.1: Gravity model errors by observed trip count for OD pairs by trip purpose

Large outliers are present for all three trip purposes in figure 3.1. 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 4

Destination Choice Model

4.1 Introduction

As mentioned multiple times in this thesis already, discrete choice models are much more powerful than gravity models. However, this power comes at the cost of additional complexity, both in their set-up, estimation and application. A destination choice model gives the probability that the traveller i will choose alternative j.

$$P_{ni} = \frac{e^{\beta x_{ni}}}{\sum_{k=1}^{K} e^{\beta x_{nk}}}$$

where

$$x_{nk} = \{\beta_{nk}^{1}, ..., \beta_{nk}^{K}\}\, and\, \{w_1, ..., w_K\}$$
$$K = \text{is the total number of alternatives}$$

These β, the coefficients of the model, need to be calibrated using some optimisation function. In this thesis, this is done using Maximum Likelyhood Esimation, through the mnlogit package described in section 1.5. The probabilities provided by an estimated model for each alternative and individual can then be used in a weighted selection of a destination for that individual. Furthermore, a trip distribution matrix can be created by summing the probabilities over the trip sample for each OD pair. This allows the model to directly be compared with aggregate models.

Firstly, the data must be transformed into the correct input structure, this process was described in section 1.5. The next step is estimation. The estimation of a discrete choice model for destination choice is much more involved than the construction of a gravity model, since modeler has almost infinite possible permutations 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. Rather than just present a final model, this next section elaborates on the model development process, covering the estimation, validation and implementation stages.

4.2 Estimation

In this section, the evolution of a destination choice model is presented. in *m1*, a simple model based on the gravity model is presented.*m2* and *m3* add further interaction variables between origin and destinations. *m4* explores the potential of LSBN data to improve the measure of destination utility. Finally, *m5* adds the important explanatory variable of income to the model.

4.2.1 Socio-economic 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 k estimated for the respective gravity model (see 3). This approach significantly improved the model, and provides a quick way to calibrate the distance coefficient without requiring the design of a more complicated GEV models.

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%. Hence, as in the gravity model, population and employment are summed together. This value is then log transformed, due to the long tail in the distribution (see figure A.1 in appendix (CITE). In order to simplify the further model equations, we assign a new variable for each destination

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

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

$$u_{ij} = \beta_1 \; e^{-kd_{ij}} + \beta_2 \; civic_j$$

where β_n are the coefficients to be estimated by the mnlogit package. This model forms the base for further models discussed in the next sections.

The employment data is classified by NAICS category, and models were explored where only a section of categories was included as employment. 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 inter-correlated (see table A.2 in appendix (CITE).

Table 4.1: *m1* model coefficients

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

The parameters of this model *m1* (see table 4.1encouraging. 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, suggesting a preference to "escape the city", 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 (CITE).

4.2.2 Origin-Destination Interactions

In this section, the model is extended with two additional variables that were introduced in section (CITE). Model *m2* is specified by the utility function

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

where

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

There are 4 possible combinations of a metropolitan flag for origin and destination pairs, however, only two were selected for inclusion in the model. The flag that identifies trips leaving metropolitan areas,mr_{ij}, results in an unsolvable model, 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 4.2 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 and rm_{ij} do not work as well. They are not statistically significant for visit trips, and the coefficient value 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, however the language coefficient for leisure is hard to justify. Finally, rm_{ij} is not significant for two trip purposes, despite working well for leisure trips.

Table 4.2: *m2* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-k\ d_{ij}}$	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 3.1 and 4.1, many outliers have been significantly brought back towards the origin, indicating an improvement in the model. However, there are still some significant outliers, with a sample of the largest in table A.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.

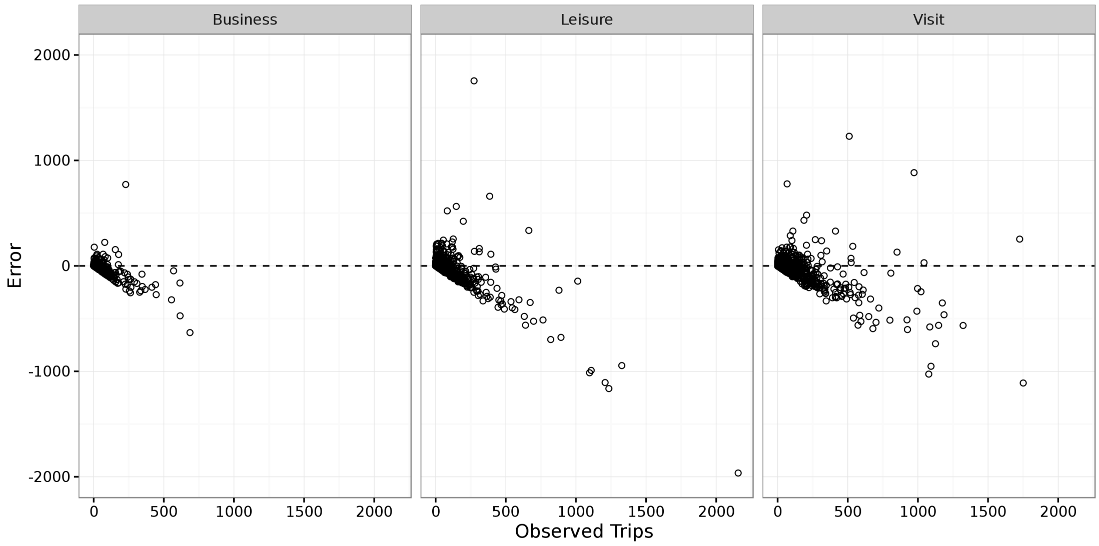


FIGURE 4.1: *m2* model errors by observed trip count for OD pairs by trip purpose

The large intra-zonal trip counts occur in small metropolitan zones, while in rural zones, intrazonal trip counts are underestimated (See figure A.3 in the appendix. Since we want to penalize intra-zonal travel in the metropolitan zones, but allow it in larger rural zones, mm_{ij} is replaced with three new variables:

$$intrametro_{ij} = \left\{ \begin{array}{ll} 1, & \text{if } metro(i) \wedge i = j \\ 0, & \text{otherwise} \end{array} \right\}$$

$$intermetro_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } metro(i) \wedge metro(j) \wedge i \neq j \\ 0, & \text{otherwise} \end{array} \right\}$$

$$intrarural_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } !metro(i) \wedge i = j \\ 0, & \text{otherwise} \end{array} \right\}$$

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. This may be a common choice for business travelers, but less likely for recreational trips. The third variable, $intrarural_{ij}$ allows the model to consider that in larger, rural zones, there are more likely to be sufficient opportunities within the zone, and that the distances to reach other zones are more significant. However, its inclusion significantly improves the model results, and as presented in 4.3, it has a strong influence, particularly for business travel.

The other zone interaction variables $language_{ij}$ and rm_{ij} are removed in this iteration. They weren't 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 showing in table 4.3. 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 4.3: *m3* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-k\ d_{ij}}$	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 4.6 we can see significant improvements throughout the model iterations across all metrics. In figure B.1, we can also see visually the significant improvements 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 B.1), and this problem is tackled in the next section.

4.2.3 Incorporating LBSN Data

The traditional socio-economic variables don't reflect why people travel to a particular destination. People don't travel to a location purely because many people live there, but because there are opportunities to perform certain activities in that location. Population and employment act as proxy variables for some of these opportunities, but not all. This section incorporates data from LSBNs 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 LSBN data from foursquare was covered in section 2.5. To summarize briefly, the venues were collected into the following categories for each destination:

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

Different multinomial logit 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 checkins 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 checkin 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 checkins 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, skiing, were represented though categories containing the types of venues commonly visited to perform these activities. Hence, two variables were added only 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%. The 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 multinomial logit model as the following.

$$\begin{aligned}
medical_j =& (purpose == \text{“}visit\text{”}) \cdot \log(medical_j) \\
hotel_j =& \log(hotel_j) \\
sightseeing_j =& \log(sightseeing_j) \\
naigara_j =& (purpose == \text{“}leisure\text{”}) \cdot (j == 30) \cdot \log(sightseeing_j) \\
outdoors_j =& (purpose == \text{“}leisure\text{”}) \cdot (season == \text{“}summer\text{”}) \cdot \log(outdoors_j) \\
skiing_j =& (purpose == \text{“}leisure\text{”}) \cdot (season == \text{“}winter\text{”}) \cdot \log(skiing_j)
\end{aligned}$$

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 4.6).

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 strongly influential on attractiveness of visit trip destinations.

TABLE 4.4: *m4* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-k\ d_{ij}}$	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, remain consistent after the addition of the the variables from model *m3*. The signs and magnitude of the variables from *m3* also change little.

Table 4.5: *m5* model coefficients

Parameter	Visit		Leisure		Business	
$e^{-k\ d_{ij}}$	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 noticable is the large improvement across all metrics for leisure travel. Figure 4.2 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.

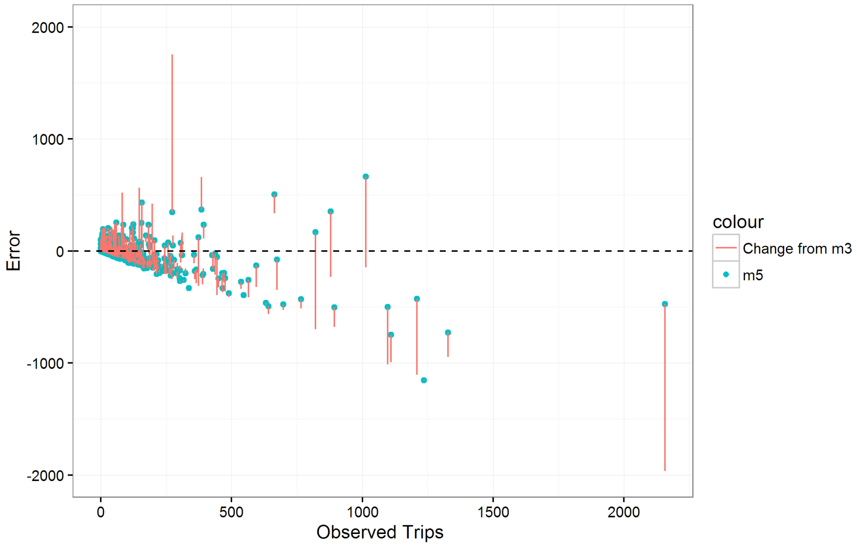


FIGURE 4.2: Effect of adding foursquare variables to model *m3* on leisure trips

4.2.4 Income Strata

4.2.5 Estimation Results

Table 4.6 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 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 4.6: Comparison of model iterations

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

4.3 Implementation

The Destination Choice model described in this thesis was designed as a component of a 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.

4.3.1 Algorithm

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. 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 of 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 *EnumeratedDistribution* from the Apache commons math library

    ```
    return new EnumeratedDistribution<>(probabilities).sample();
    ```

    (e) store the destination in the trip object

4.4 Calibration

4.5 Validation - Case study of a new Ski Resort

Chapter 5

Discussion

overview of results

what your findings might mean

how valuable they are

why

limitations

future work

Chapter 6

Conclusions

Appendix A

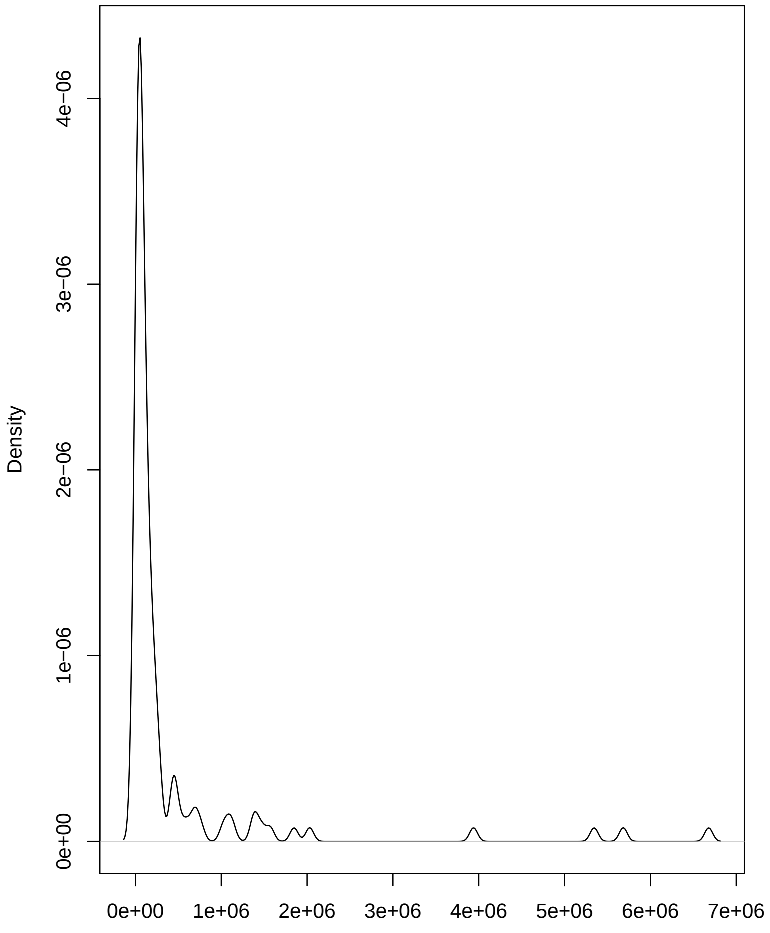
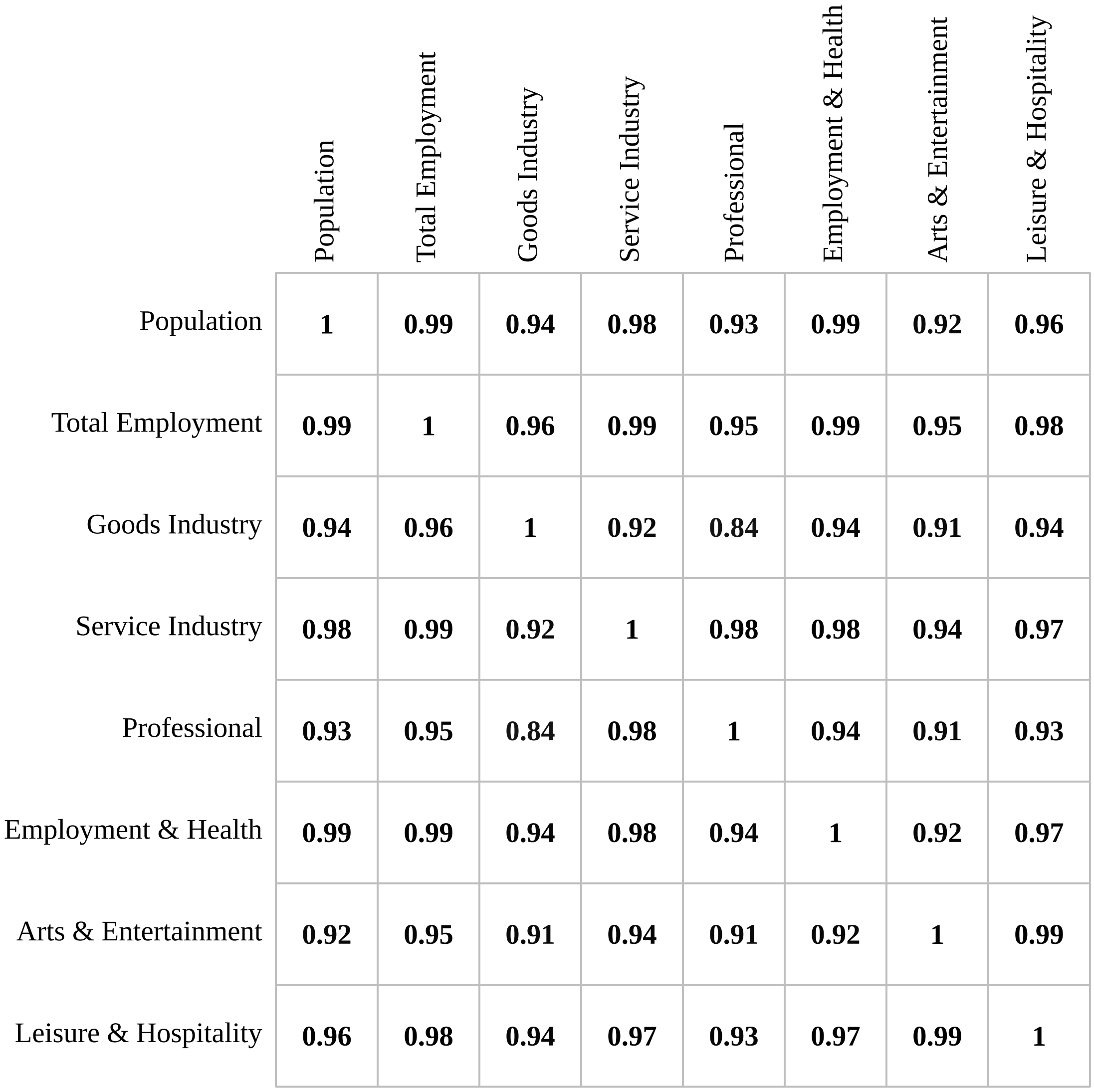


Figure A.1: Long Tail and right skew of (population + employment) for each destination



	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

Table A.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

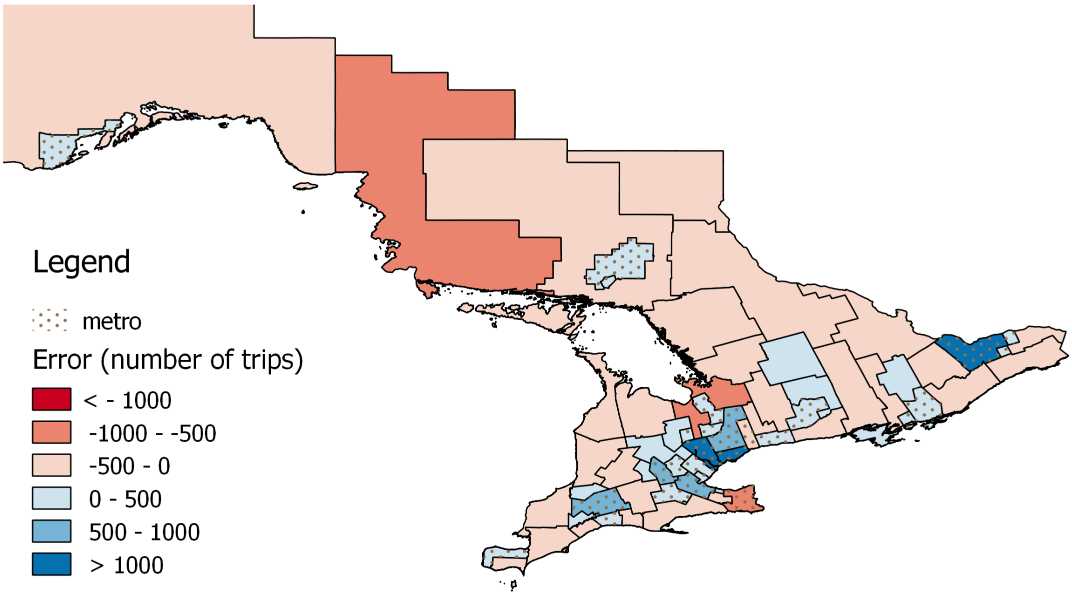


Figure A.3: Intrazonal errors produced by the *m2* model

Appendix B

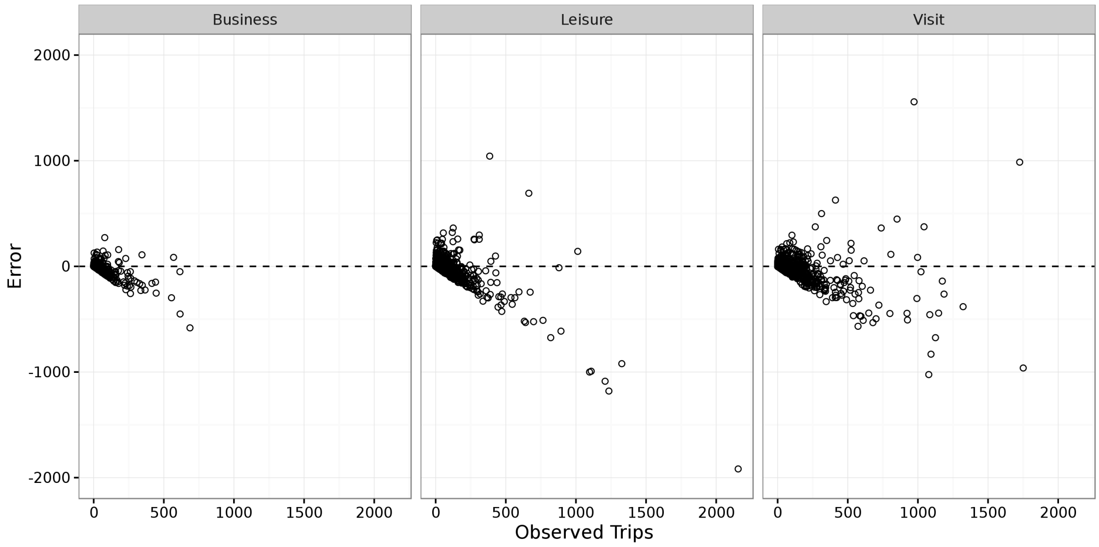


FIGURE B.1: *m3* model errors by observed trip count for OD pairs by trip purpose

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 and Moshe 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.

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.

Hasan, Asad, Wang Zhiyu, and Alireza S Mahani (2014). “Fast Estimation of Multinomial Logit Models: R Package mnlogit”. In: *arXiv preprint arXiv:1404.3177*.

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.

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.

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:

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.

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

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.

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.

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.

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.

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*.

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.

Stephanedes, Yorgos J, Vijaya Kumar, and Balakrishnan Padmanabhan (1984). "A fully disaggregate mode-choice model for business intercity travel". In: *Transportation Planning and Technology* 9.1, pp. 13–23. ISSN: 0308-1060.

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 TRANSPORTATION SERIES*.

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