Improving destination choice modeling using location-based big data

Joseph Molloy

ETH Zürich

May 2017

STRC | **17th Swiss Transport Research Conference**
Monte Verità / Ascona, May 17 – 19, 2017

ETH Zürich

Improving destination choice modeling using location-based big data

Joseph Molloy
Institute for Transport Planning and Systems
(IVT)
ETH Zürich
Stefano-Franscini-Platz 5, 8093 Zurich
phone: +41 44 633 31 51
fax: +41-44-633 10 57
joseph.molloy@ivt.baug.ethz.ch

May 2017

Abstract

Citizens are increasingly sharing their location and movements through‘check-ins’ on location based social networks (LBSNs). These services are collecting unprecedented amounts of big data that can be used to study how we travel and interact with our environment. This paper will present the development of a destination choice model for Ontario, Canada which uses data from Foursquare, the largest LBSN to model destination attractiveness. Models are estimated for leisure, visit and business long distance travel purposes separately. A methodology to collect, process and aggregate historical check-in counts has been developed, allowing the utility of each destination to be calculated based on the intensity of different activities performed at the destination. Destinations such as national parks and ski areas are very strong attractors of leisure trips, yet do not employ many people, and have few residents. Trip counts to such destinations are therefore poorly predicted by models based on population and employment. Traditionally, this has been remedied by extensive manual data collection. The integration of Foursquare data offers an alternative approach to solve this problem that has not been deeply explored until now.The Foursquare based destination choice model is evaluated against a traditional model that is estimated only with population and employment. The results demonstrate that data from LBSNs can be used to improve destination choice models, particularly for leisure travel.

Keywords

Destination Choice, Discrete Choice, Foursquare, Big Data, Location Based Social Networks

1 Introduction

Destination Choice modeling using multinomial logit models allows for much more sophisticated models than the aggregate gravity models that have persisted in the field since the 1950's (Train, 2009). Despite the opportunities for much more advanced representations of destination utility, the models still rely on socio-economic indicators such as the population and employment at the destination. The choice of the destination made by a traveler is not necessarily made based on how many people live and work there. A simple example demonstrates where such traditional metrics fall short: National parks have no population and little employment, but are large attractors of leisure trips. Ski areas are another example of this effect.

While it is desirable to better represent the zonal utility, destination choice modeling is often characterized by a large set of alternatives (Ben-Akiva and Lerman, 1985). As such, the the acquisition of detailed data for each alternative can often be prohibitive within the context of the development of a transport model. Simma *et al.* (2001) explored such variables in detail for long distance leisure travel in Switzerland, and reported that the data collection work was indeed particularly onerous. This paper presents an alternative approach, using aggregated data from the location based social network (LBSN), Foursquare, to represent components of destination attractiveness in the utility function of a multinomial logit model.

Big data, such as that collected by Foursquare, is a 'topic du jour' in transport modeling. Foursquare enable users to 'check-in' to a point of interest (POI), known as a 'venue', and provide tips, ratings and reviews. With 50 million monthly active users and over 7.8 billion check-ins to date (Issac, 2015), foursquare is the largest LBSN. This enormous amount of data can be used in a multitude of ways to explore mobility patterns. In recent relevant research using Foursquare, 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. 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 .Noulas *et al.* (2012) used Foursquare data to design gravity model based on Stouffer's theory of intervening opportunities (Stouffer, 1940).

Whereas even comprehensive travel surveys such as the TSRC in Canada often have a sample size of around only 50,000 records per year, big data sources can record the movements of millions of individuals at unprecedented spatial and temporal accuracy (Beyer and Laney, 2012).

It is important to note that the high temporal and spatial resolution of geolocated big data comes with its own trade-offs. Often social-demographic attributes are not available, making it extremely difficult to correctly weight the sample. Furthermore the data available publicly to researchers can be limited or highly aggregated, and the collection and sampling methodologies are normally not available (Morstatter *et al.,* 2013). Clearly, there is a need to investigate new ways of combining both 'traditional' (travel surveys and census data) and 'new' (big) data sources to harness the best attributes of both in the context of transport modeling.

This paper explores how data from Foursquare can be combined with traditional data sources in the development of a long distance destination choice model for Ontario, Canada. Section 2 presents a methodology to develop a zoning system for the destination choice model, and enrich it with the foursquare data. Section 3 presents the results of the model estimation and a scenario analysis. Section 4 provides a discussion of the results, including limitations and areas for future work, and section 5 concludes.

2 Methodology and Data

2.1 Applying the Travel Survey of Residents of Canada

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. In this paper, the TSRC provides the 'traditional' data source for the estimation and calibration of the destination choice model. All spatial data points, namely those for trip origins and destinations and stopovers are provided in the microdata at three resolutions (from lowest to highest): Province or Territory, Census Division, and Census Metropolitan Agglomeration (CMA).

The TSRC trip files provide trip records for all of Canada. However, as a model for Ontario, trips are removed that do not interact with Ontario, namely:

- Trips by air that do not arrive or depart Ontario.
- Ground based trips where the shortest path does not contain Ontario.

In total 69,328 individual trip records remain from the TSRC dataset for model estimation after filtering (see Table 1).

Table 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	22,909
Visit	9,057	8,777	7,962	7,618	33,414
Total	18,694	18,016	1,6547	16,071	69,328

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

For this particular destination choice model, a domestic zone system was already provided by the project partner, consisting of 6,495 Traffic analysis zones (TAZs) for Ontario and 48 representing the rest of Canada. Sociodemographic data was provided for each TAZ. However, the trip origins and destinations in the TSRC are only defined at broader spatial resolutions, namely province, Census Division and CMA. Hence, a new internal zone system for Ontario is defined for this destination choice model, based on the TSRC.

The provided external TAZs were previously defined by the project partner from the TSRC Census Divisions and selected CMAs of interest to the model, and as such, can be transferred directly to the new model. The Internal TAZs were not comparable to the TSRC resolution, as they were allocated 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 $0.879 km^2$ to $3600 km^2$, with smaller cells defined for more populous areas, and larger cells for regional areas.

Since CMAs often overlap multiple Census Divisions, zones are defined by the union of the Census Divison and CMA geometries. This fully reflects the number of destination choices discernable in the TSRC schema. The resulting zone system has 69 internal zones for Ontario. Using this approach, the distinction between urban and rural areas is encoded into the zone system. It is worth noting that 51.5% of trips in the filtered TSRC survey originated in a CMA, and 48.3% had destination recorded as a CMA. Both urban and regional areas contributing to long distance travel, and CMAs more likely to be origins than destinations. The final zone system used is presented in figure 1

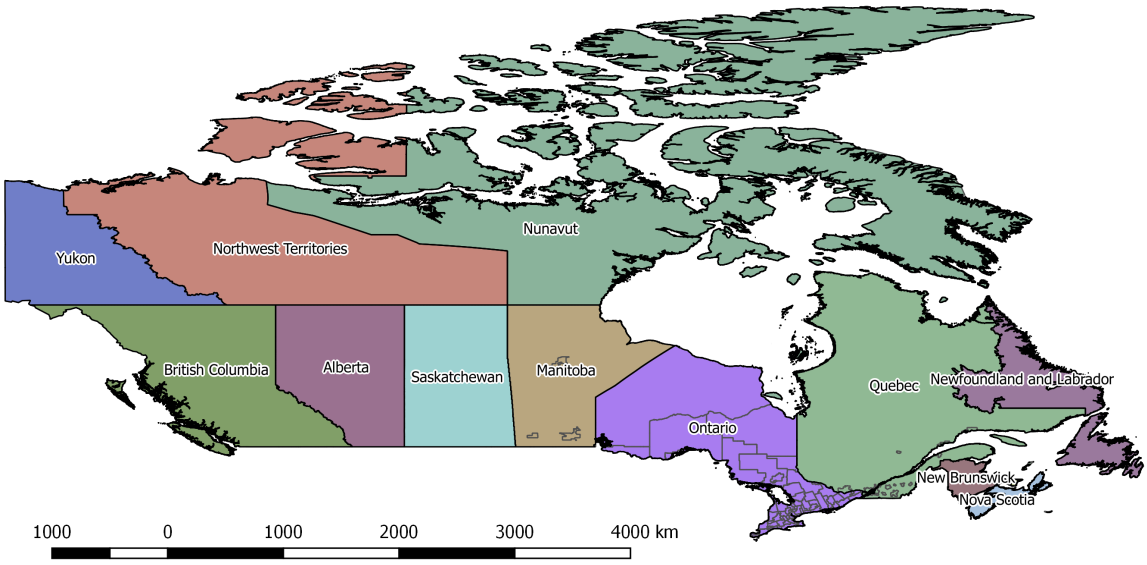

Figure 1: Zones by province for Canada

The travel time, population and employment data was provided at the TAZ level by the project partner. Each TAZ was assigned to the respective TSRC zone, and the Socioeconomic variables, namely population and employment, summed for each zone over the grouped TAZs. To aggregate the auto travel times between all TAZs to the zonal level, the travel time t between each child origin-destination pair was weighted by the multiplied populations p of the origin and destination.

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

2.3 Foursquare

Foursquare collects a wealth of data, on where and when users check-in. Data was collected using the Foursquare public venue API[1]. The API returns a list of venues in JSON format. Each venue record provides the following relevant information:

- Name
- Venue category

[1] more details on the API can be found at `https://developer.foursquare.com/overview/venues.html`

- 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. However, in this paper, the foursquare metrics are only used for identifying the intensity of activity in zones that may not be reflected by socioeconomic variables.

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

1. The maximum search area allowed is smaller than most external zones, so a search grid of one degree raster cells was generated for the study area.
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, for which the search was performed, to exclude Foursquare subcategories such as 'States & Municipalities'.
4. The Foursquare API is queried for each cell and category, returning the top 50 venues, while adhering to the rate limit of 5000 requests per hour.
5. Resulting individual venues are stored in the PostGIS database, and the number of check-ins for each category and zone calculated (see Table 5 in Appendix A). In total, 34,041 unique venues and 7,981,458 check-ins were collected for the different categories.

2.4 Other Model Variables

Metropolitan areas are not homogeneous in land use patterns. Within urban areas there are certain residential areas and central business districts to which people are more likely 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. In order to simplify the further model equations, we assign a new variable for each destination

j:

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

with population as p_j and employment emp_j.

Mishra *et al.* (2013) found that interaction terms between the origin and the destination were significant for their destination choice model for Maryland. In a similar vein, three variables are presented here to control for intra- and inter-zonal effects, where *metro* indicates that the zone is a CMA:

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

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

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

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 separately.

In discrete choice models that include distance or travel time terms in the form $exp(x)$, it is common to include an additional parameter $\alpha\ exp(\alpha\dot{x})$. However, such exponential parameters are not estimable with simple multinomial logit models.

To avoid the necessity of using more complex models from the GEV family (Train, 2009), or trial and error methods, the α for each model were taken from a previously designed gravity model previously, estimated with the same dataset as the discrete models in this paper (see the masters thesis of Molloy (2017) for further details). This method produces good results, with an improved model accuracy and a more significant travel time parameters than models tested without the α parameter.

$$trips_{ij} = \frac{A_j \cdot e^{-\alpha \cdot tt_{ij}}}{\sum_j^J Civic_j \cdot e^{-\alpha \cdot tt_{ij}}} \cdot P_i$$

Where
P_j is the number of trips produced in origin zone i
$civic_j$ is the attraction at destination zone j
α is the impedance factor, calibrated with the average trip travel time
tt_{ij} is the travel time between zones i, j

Table 2: Gravity Model calibration

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

3 Results

3.1 Estimation Results

This section discusses the estimation results of the destination choice model. The dataset was split into three categories, representing the three travel purposes: leisure, visit and business. In the first model iteration, model A, only the TSRC data is used to generate parameters.

The NRMSE considers the sample size of the estimation data by dividing the RMSE by the standard division of the observed values, which allows for the comparison of the model performance across the trip purposes, despite their varying samples. In terms of bpth the r^2 and normalized root mean square error (NRMSE), the results of this first model are good, particularly when compared to the a simple singly constrained gravity model estimated on the same dataset. However, the performance of the leisure sub-model is significantly weaker than the others.

Furthermore, all the parameters are highly significant, and have the expected signs. The parameter signs and magnitude vary strongly across trip purposes. Business is the only purpose where an urban destination is more likely to attract urban trips. On the other hand, leisure

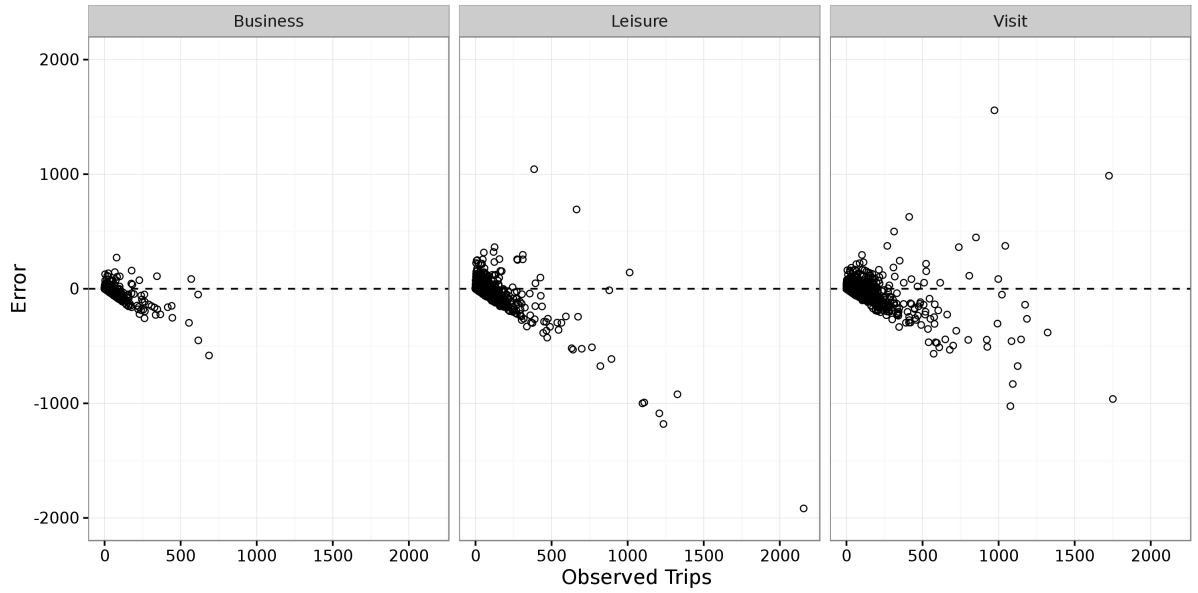

Figure 2: Model A: Errors by observed trip count for OD pairs by trip purpose

travelers are much more likely to head for destinations outside the city. For visitation, there is a weak positive effect toward urban areas.

For intra-zonal travel in urban areas, there is a strong negative effect for all trip purposes while for intra-zonal rural travel the effect across all trip purposes is positive. This is as expected. Urban areas are often smaller, making long distance trips (those over 40km) physically much more unlikely. In rural zones, since they are much larger, the power law of travel distance means that long distance trips crossing into other zones are less likely (Gonzalez *et al.,* 2008). Of note is the large negative coefficient for leisure intra-metro travel. This, combined with the other two origin-destination interaction parameters for leisure travel suggest a strong preference for leaving urban areas for leisure. This is supported by the TSRC data, where the key leisure travel reasons include the outdoor activities such as skiing, visiting national parks and camping.

On closer inspection, the residuals graph in Figure 2 illustrates a trend for the model to underestimate the demand of OD pairs with large numbers of trips, and overestimate some other smaller OD pairs. These sources of error fall into two categories:

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

Accounting for the attraction of zones such as Niagara is difficult, as the attractiveness is due to factors not observable from parameters used in model A. For the second, expanded model,

B, variables based on the foursquare data are included. Each of these variables represents the natural log of the check-in count collated for the respective category.

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

Certain categories were found to be significant for particular trip purposes. For example, the outdoor category is only significant for leisure trips, and the medical category is only significant for visit trips. As would be reasonably expected, the number of hotel check-ins is a significant variable across all trip purposes for long distance travel. It is logical that the presence of hotels and sightseeing venues is particularly important for leisure travel, and this is appropriately reflected in the coefficients in the model. 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 was found to be influential on the attractiveness of visit trip destinations. In model A, leisure trips to the zone containing Niagara Falls are underestimated by 85%. In model B, a *niagara* variable controls for this using the sightseeing category for leisure travel to the Niagara zone. Two variables *outdoors* and *skiing* were found to be significant only for leisure travel in the season in which the respective activity is normally performed.

Overall, model B performs better across all trip purposes than model A, demonstrating the benefit of including the foursquare based parameters. Particularly noticeable is the large improvement across all metrics for leisure travel. Figure 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 OD pairs have been reduced.

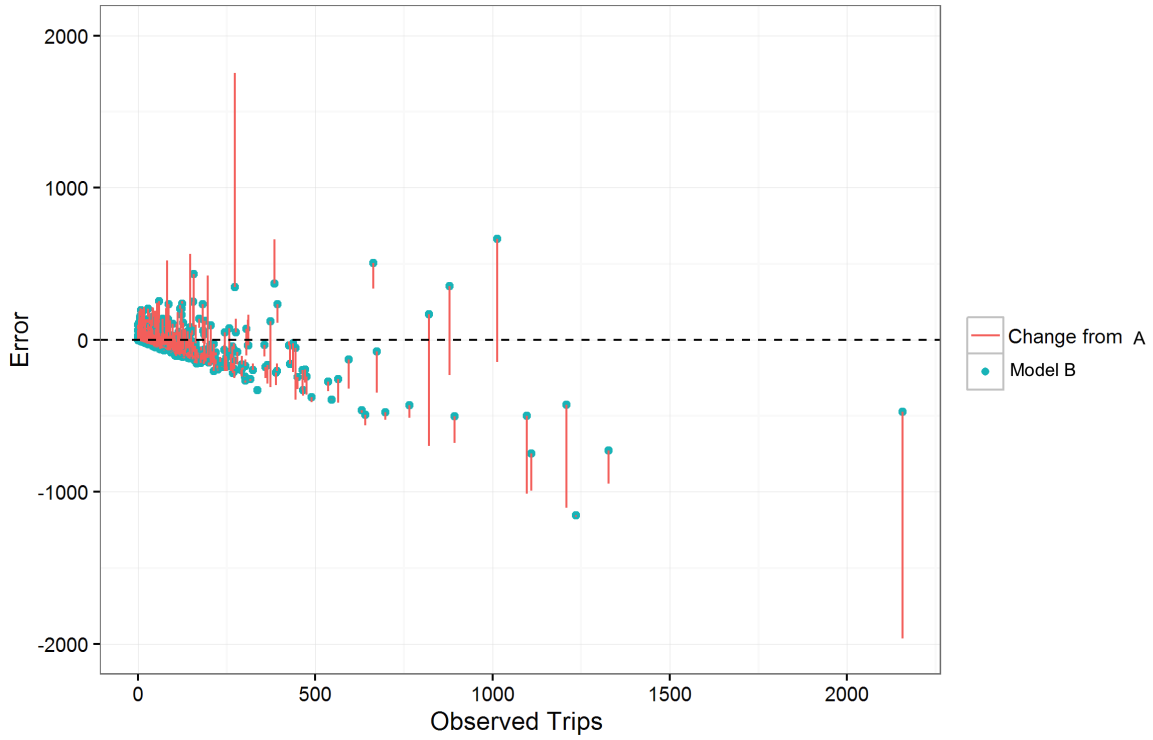

Figure 3: Effect of adding foursquare variables in model B on leisure trip counts

Table 3: Model Estimation Results (for all coefficients, $Pr(>|t|) < 0.001$)

Parameter	Visit		Leisure		Business	
	A	B	A	B	A	B
α	0.0013		0.0035		0.0030	
$e^{-\alpha \cdot tt_{ij}}$	4.83	5.00	4.75	5.35	4.19	4.37
$civic_j$	0.57	0.21	0.52	-0.15	0.76	0.36
$intermetro_{ij}$	-0.08		-0.87	-0.81	0.56	0.72
$intrametro_{ij}$	-1.68	-1.75	-2.56	-2.88	-0.89	-0.87
$intrarural_{ij}$	0.39	0.24	0.85	0.58	1.66	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				
# Coefficients	5	6	5	10	5	7
Loglikelyhood	- 115,666	-114,557	- 83,663	-78,038	- 20,596	-20,288
AIC	231,342	229,128	167,337	156,095	41,201	40,590
r^2	0.80	0.82	0.56	0.80	0.73	0.77
NRMSE (%)	0.62	0.59	0.84	0.61	0.70	0.66

3.2 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 4). 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.

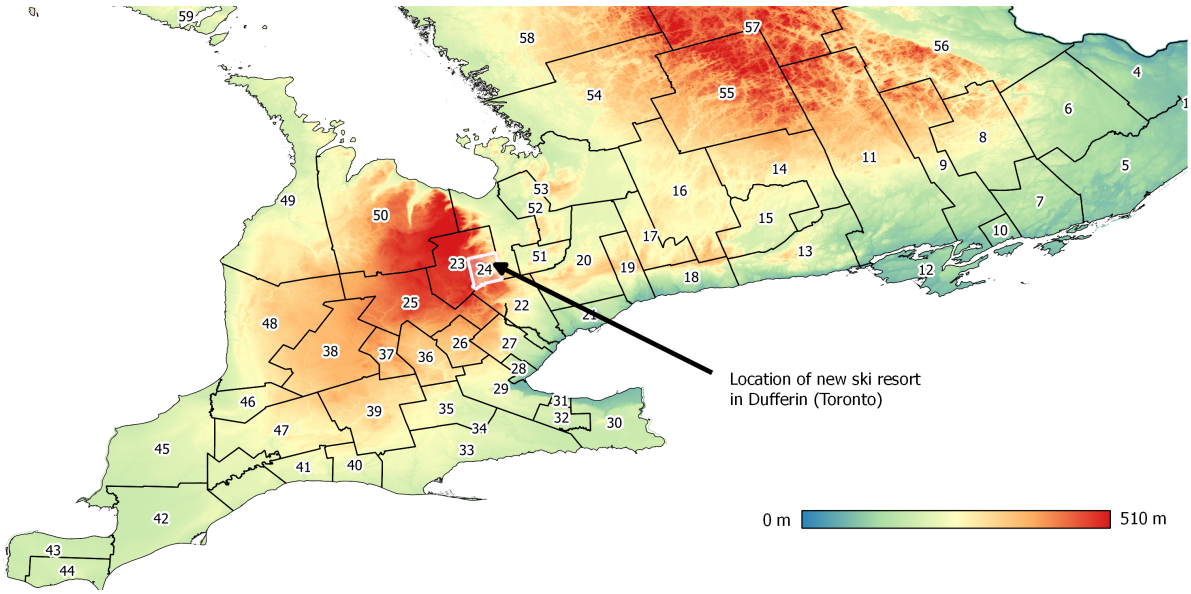

Figure 4: Scenario analysis: Location of a new ski resort on Dufferin (Toronto CMA). Elevation data from the Ontario Provincial Digital Elevation Model - Version 3.0

This scenario does not consider other policy and development considerations, such as site location and transport access. 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

Table 4: 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

developed in Section 2.3 is 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 is calculated. The following adjustments are made for the respective zones, and their values are displayed in Table 4.

- 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

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 using a calibrated version of Model B to account for the stochastic nature of destination choice. The calibration process is documented further by Molloy (2017). Figure 5 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 results show that the parameters 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 $civic_j$ variable in model A for leisure travel. This would clearly be unrealistic, and as such, this scenario gives a good example of why better representations of destination attractiveness are important, particularly for leisure travel.

4 Discussion

A closer inspection of the OD matrix generated by model B on the estimated data indicates 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.

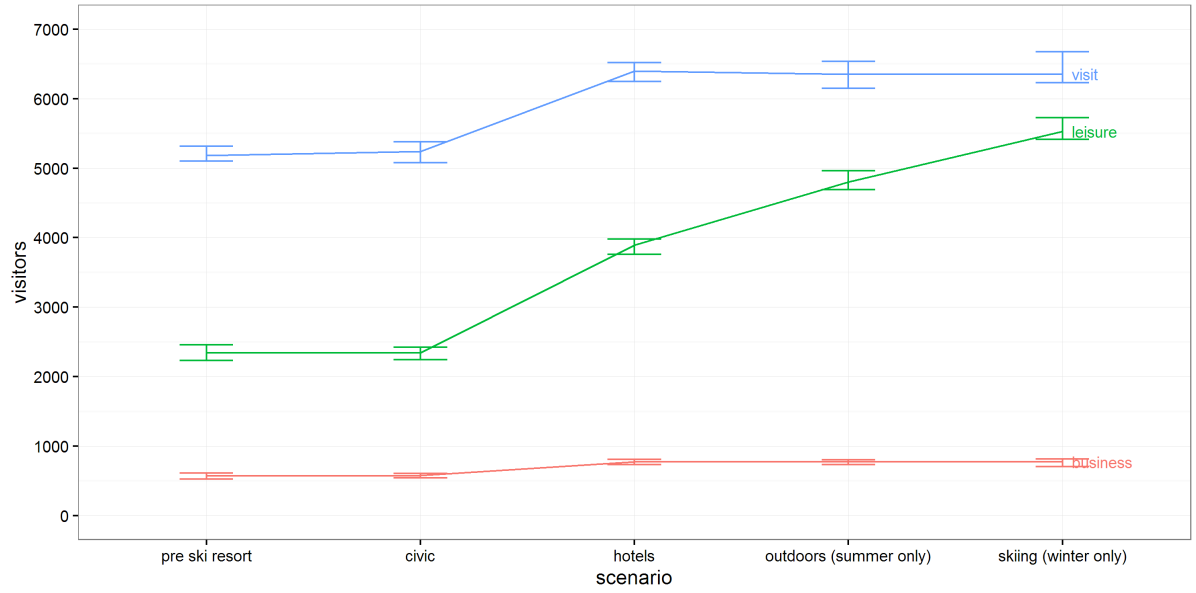

Figure 5: Scenario analysis: Impact of a new ski resort on Dufferin (Toronto CMA)

Figure 6 identifies the connections where the model falls short. The connections between the triangle of major cities, Toronto, Montreal and Ottawa, are underestimated. 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.

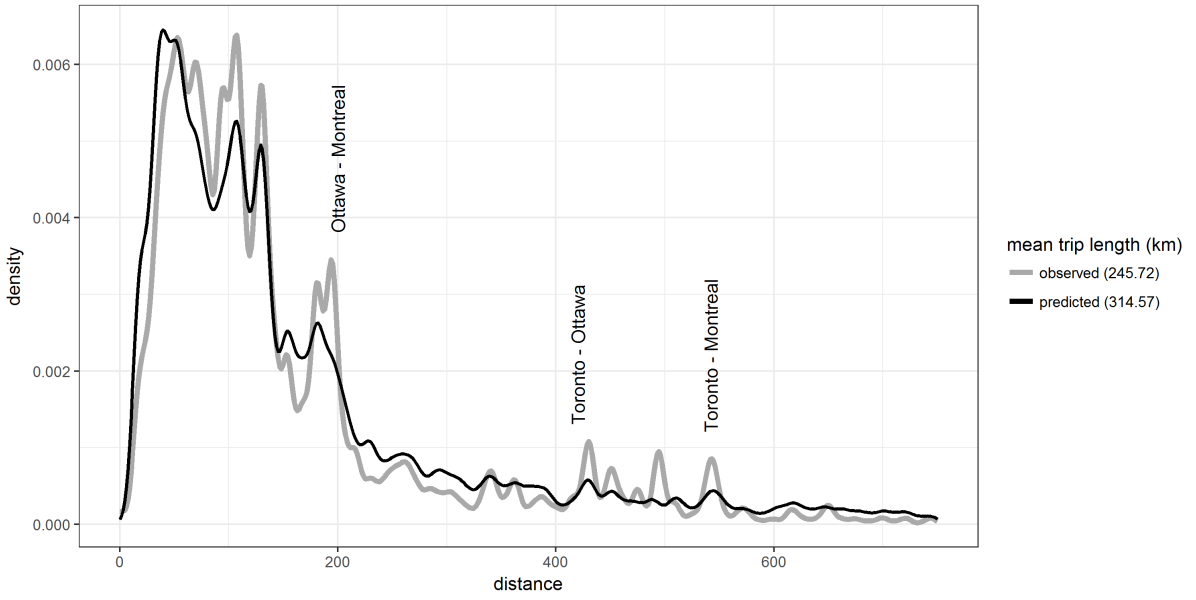

Figure 6: Trip length distribution of model (CITE) after calibration (0-750km window)

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.

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, check-ins measure the intensity of activity at each POI. A measure of importance is clearly beneficial in the model presented above, as not all POIs are equal; hotels are of different sizes, and 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. However, the data collection required is prohibitive for most large scale models. LBSN data provides an easily accessible data on the importance of individual POIs, and in turn, destination utility.

4.1 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. In future work utilizing more detailed Foursquare data, check-ins could be filtered for those performed only 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 are 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 or impute the intensity of activity at these locations, thereby avoiding the selective reporting behavior evident in Foursquare usage.

5 Conclusion

In conclusion, this paper confirms the hypothesis that aggregated geo-tagged big data can improve the modeling of destination choice when combined with traditional data sources. First, a zone system for Long distance travel in Ontario, Canada was presented. Then, a methodology for the aggregation of historical Foursquare checkins as measures of destination attractiveness for particular categories was developed. Two multinomial logit models were estimated to explore the potential of the foursquare check-ins for measuring destination attractiveness. The 'traditional' model based primarily on population,employment and zonal interactions was found to work well enough for visit and business travel, but not leisure travel. With the addition of alternative specific parameters based on the Foursquare check-in data, the model accuracy across all trip purposes improved significantly, particularly for leisure travel. A scenario analysis using the expanded model further reinforced the importance of properly measuring destination attractiveness for leisure travel.

Acknowledgement

The majority of the work contained in this paper was completed as part of the author's Masters thesis at the Technical University of Munich. The author would like to thank Prof. Rolf Moeckel for his guidance throughout the project.

6 References

Ben-Akiva, M. E. and S. R. Lerman (1985) *Discrete choice analysis: theory and application to travel demand*, vol. 9, MIT press, ISBN 0262022176.

Beyer, M. A. and D. Laney (2012) The importance of 'big data': a definition, *Stamford, CT: Gartner*, 2014–2018.

Cheng, Z., J. Caverlee, K. Lee and D. Z. Sui (2011) Exploring millions of footprints in location sharing services., *ICWSM*, **2011**, 81–88.

Erath, A. (2015) Transport Modelling in the Age of Big Data, Seoul.

Gonzalez, M. C., C. A. Hidalgo and A.-L. Barabasi (2008) Understanding individual human mobility patterns, *Nature*, **453** (7196) 779–782.

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

Issac, M. (2015) 11 2015.

Lindqvist, J., J. Cranshaw, J. Wiese, J. Hong and J. Zimmerman (2011) I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application, paper presented at the *Proceedings of the SIGCHI conference on human factors in computing systems*, 2409–2418.

Mishra, S., Y. Wang, X. Zhu, R. Moeckel and S. Mahapatra (2013) Comparison between gravity and destination choice models for trip distribution in maryland, paper presented at the *Transportation Research Board 92nd Annual Meeting*.

Moeckel, R. and R. Donnelly (2015) Gradual rasterization: redefining spatial resolution in transport modelling, *Environment and Planning B: Planning and Design*, **42** (5) 888–903.

Molloy, J. (2017) Development of a destination choice model for ontario, Master Thesis, Technical University of Munich, February 2017.

Morstatter, F., J. Pfeffer, H. Liu and K. M. Carley (2013) Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose, paper presented at the *Seventh International AAAI Conference on Weblogs and Social Media*.

Nagel, K. and K. W. Axhausen (2001) Workshop report: Microsimulation.

Noulas, A., S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo (2012) A tale of many cities: universal patterns in human urban mobility, *PloS one*, **7** (5) e37027.

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

SA, R., M. A. Karim, T. Z. Qiu and A. Kim (2015) Origin-destination trip estimation from anonymous cell phone and foursquare data, paper presented at the *Transportation Research Board 94th Annual Meeting*, no. 15-2379.

Simma, A., R. Schlich and K. W. Axhausen (2001) Destination choice modelling of leisure trips: The case of switzerland, *Arbeitsberichte Verkehrs-und Raumplanung*, **99**.

Statistics Canada (2014) Travel survey of residents of canada (tsrc), `http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3810`.

Stouffer, S. A. (1940) Intervening opportunities: a theory relating mobility and distance, *American sociological review*, **5** (6) 845–867.

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

A Appendix

Table 5: Foursquare venue categories

Category	Venue Categories	Venues	Check-ins
Medical	Dentist's Office	6,294	586,082
	Doctor's Office		
	Hospital		
	Medical Center		
	Veterinarian		
Ski Area	Ski Area	1,048	203,266
	Ski Chairlift		
	Ski Chalet		
	Ski Lodge		
	Ski Trail		
Hotel	Bed & Breakfast	7,268	1,502,248
	Hostel		
	Hotel		
	Motel		
	Resort		
Outdoors	National Park	7,262	709,274
	Campground		
	Nature Preserve		
	Other Great Outdoors		
	Scenic Lookout		
Sightseeing	Art Gallery	4,387	1,125,385
	Historic Site		
	Museum		
	Theme Park		
	Scenic Lookout		
Total		34,041	7,981,458