Destination Choice Modeling using Location-based Social Media Data

Md Mehedi Hasnat[1], Ahmadreza Faghih-Imani[2], Naveen Eluru[3], Samiul Hasan[4]

[1]Graduate Research Assistant; Department of Civil, Environmental, and Construction Engineering; University of Central Florida; 12800 Pegasus Drive, Orlando, FL 32816; Email: hasnat@knights.ucf.edu

[2]Postdoctoral Fellow; Department of Civil Engineering, University of Toronto, 35 St. George Street, Toronto, Ontario, Canada M5S 1A4, Email: a.faghihimani@utoronto.ca

[3]Associate Professor, Department of Civil, Environmental, and Construction Engineering; University of Central Florida; 12800 Pegasus Drive, Orlando, FL 32816; Email: naveen.eluru@ucf.edu

[4]Assistant Professor; Department of Civil, Environmental, and Construction Engineering; University of Central Florida; 12800 Pegasus Drive, Orlando, FL 32816; Email: samiul.hasan@ucf.edu

1. Introduction

Travel surveys complemented by additional land use and socio-economic data have served as primary inputs for travel demand models. A complete household survey with all the required travel information costs about $200 per household (Zhang and Mohammadian, 2008). Although access to such individual level travel information is crucial for developing advanced travel behavior models, conducting such a survey is costly and time consuming (Flyvbjerg et al., 2005). With increased use of pervasive technologies, alternative approaches can potentially be used to collect/augment this information in a cost-effective way. Web-based surveys (including trip planning apps), social networking applications, smart phones, and personal health sensors have been explored to collect individual travel information. To gather travel behavior information, organizations have started using global positioning system (GPS) log data (NYMTC and NJTPA, 2014), smart phone based travel surveys (Greene et al., 2015) and web based surveys ("North Florida Travel Survey," 2017). Different countries around the world such as Singapore (Cottrill et al., 2013), New Zealand (Safi et al., 2015), Australia (Greaves et al., 2015), Netherlands (Geurs et al., 2015) etc. have resorted to smart phone based GPS travel data as a complementary approach to traditional travel surveys. These studies have found GPS based travel surveys through wearable devices as promising alternative or addition to traditional trip diaries. However, researchers are yet to fully explore their potential as well as identify all the limitations of these emerging technology-based data collection methods (Abbasi et al., 2015, Geurs et al., 2015).

1.1. Social media data

In addition to smartphone-based surveys, passively collected data can be used for travel behavior modeling. For instance, we can access a large volume of user generated content shared in various social media platforms (Chi, 2008; Kuflik et al., 2017). Social media can be defined as a collection of internet-based applications which allow users to create and share contents (Kaplan and Haenlein, 2010). About 80 percent of Americans use social media creating a unique opportunity to gather digital traces (Perrin, 2005). Analyzing the millions of user footprints, it is possible to extract travel behavior at a greater resolution (Hendrik and Perdana, 2014).

However, there are some challenges of using social media data in various transportation studies. For instance, in users' trip inference studies, it is difficult to accurately find out the trip start time, end time, and trip length (Zhang et al., 2017). In case of sampling, social media may over represent some groups of users (Zhang et al., 2017) and specific types of activities such as leisure and discretionary activities (Hasan and Ukkusuri, 2014; Rashidi et al., 2017). It has been found that smartphone and/or check-in service users are slightly over represented by young people (Comscore, 2011). A biased public participation may also result due to the difference in income, education, and place of residence (Wiersma, 2010). Lack of user socio-demographic attributes makes it difficult to correctly weigh the sample (Beyer and Laney, 2012). However, there have been efforts to infer demographic information through data mining approaches (Mislove et al., 2011). Thus, extracting meaningful travel information from social media data and inferring user demographic information are challenging issues (Rashidi et al., 2017; Zhang et al., 2017). Social media datasets also require appropriate filtering of noises (e.g., social bots) before extracting any meaningful information. Specialized algorithms need to be developed to extract information such as trip purpose, travel mode. In this regard, employing check-in and geo-tagged data (such as geo-tagged Twitter posts, Foursquare check-ins) will reduce the computational burden to analyze activity destinations as these records are associated with a location and/or activity (Beirão and Sarsfield Cabral, 2007).

Twitter is a very pervasive means of communication with 317 million monthly active users (67 million users from the USA) sending 500 million tweets per day ("Twitter Facts," 2017). Twitter data, accessed through simple web scraping, provides a wide range of information within each post (tweet). Also, despite being unstructured, tweets provide important clues about latent user attributes and activities- absent in GPS logs and mobile phone records (Cao et al., 2014). From Twitter, we can extract spatial (geo-tagged) and temporal (time-stamped) information for a longer period and large number of users without invading user privacy (Frias-Martinez et al., 2012; Hasan and Ukkusuri, 2015).

1.2. Destination choice modeling

Across the various travel demand dimensions analyzed, urban destination choice decisions are characterized by a large set of alternatives (theoretically any spatial unit within the study region). In traditional travel surveys, choice information available for the respondent sample is unlikely to offer a well sampled destination choice information due to inherently large number of origin-destination combinations available (characterized by the square of the number of zones).

Furthermore, collecting individual level destination choice data in an urban region is costly and time consuming, and therefore infeasible to gather on a frequent basis. In this context, the availability of location based social media data (LBSM) potentially offers a rapidly updated destination choice behavior in the urban region. LBSM data can be obtained more frequently while also providing a larger data sample enhancing the spatial and temporal coverage (Beyer and Laney, 2012).

Given these aforementioned benefits of LBSM data and availability of this information in Twitter, we present a methodological framework to model destination choice using Twitter data. Using web scripts, we have gathered an extensive sample of geo-tagged tweets from the Central Florida region. We have merged these geo-tagged tweets with different geographic databases collected from state level data libraries. We have identified resident profiles and extracted their home and visited destinations over the data collection period. For each destination, we recognize that all census tracts in the entire study region are potential destination alternatives. However, to reduce computational burden we have generated destination choice level alternative choice sets by randomly selecting a manageable choice set (of 30 census tracts). Our selection of the size of the choice set is consistent with previous studies (Nerella and Bhat, 2004; Pozsgay and Bhat, 2001; Faghih-Imani and Eluru, 2015, Faghih-Imani and Eluru, 2017). The destination choice behavior is explored within a random utility framework employing a multinomial logit (MNL) model. However, traditional multinomial logit models do not consider the presence of population heterogeneity. Specifically, in modeling destination choice behavior, varying preferences are likely to exist by gender (Faghih-Imani et al., 2016), activity purpose (Moscardo, 2004; Recker and Kostyniuk, 1978; Seddighi and Theocharous, 2002) and origin location (Waddell et al., 2007). A common approach to accommodate such potential variations is exogenous segmentation where the data are segmented by the exogenous variable of interest and separate models are estimated by segment (Bhat, 1997). However, these approaches are appropriate only for one or two variables. In cases where segmentation is desired by more number of variables, a latent segmentation approach is preferred (see Eluru et al., 2012 or Sobhani et al., 2013 for more discussion). To account for population heterogeneity in the data, we also develop a latent segmentation MNL or LSMNL. In addition, our data has multiple observations over many days from the same user, i.e. we have repeated observation or panel data. Hence, we have estimated a

Panel Latent Segmentation Multinomial Logit (PLSMNL) model capturing the features affecting individual destination choices.

Our paper makes three major contributions. *First*, it describes how to gather and merge emerging social media data with existing geographic databases enriching the set of variables available for modeling. *Second*, to study destination choices from social media data, we have developed a choice modeling framework based on a Latent Segmentation Multinomial Logit model. To the best of our knowledge, this is one of the first few papers that uses an advanced econometric modeling framework for social media data analysis. The developed model has added explanatory power compared to the existing data mining/machine learning approaches. *Third*, we present key insights on individual destination choices residing in a region. Such insights are hard to obtain using traditional survey-based data or using state-of-the-art machine learning models applied over social media data. Thus, this is a timely study showing the opportunities of emerging social media data and how effectively such data can be utilized in transportation planning studies. Such techniques will be useful in developing advanced travel demand models by complementing traditional survey-based travel behavior data with longitudinal activity data available in social media.

2. Earlier Studies and Current Work in Context

We organize our review along two broad streams. First, we briefly review earlier work examining destination choice in the transportation field. Second, we review earlier research employing social media data for travel behavior analysis, particularly the efforts that employed social media data for destination choice modeling. Subsequently, we identify the limitations of earlier work and position our current research.

2.1. Destination Choice

The area of destination modeling has received wide attention in the transportation field. Hence, an exhaustive review of earlier work is beyond the scope of this paper. With growing emphasis on activity based models in recent decades several research efforts have explored location decision process (Jonnalagadda et al., 2001; Koppelman and Sethi, 2005; McFadden, 1978; Shiftan and Ben-Akiva, 2011). Several studies examined activity purpose specific individual destination choice – such as shopping trips (Bekhor and Prashker, 2008; Horni et al., 2009) and recreational/leisure trips (Horni et al., 2009; Pozsgay and Bhat, 2001; Sivakumar and Bhat, 2007). Other analogous analysis of destination choices include railway station choice (Chakour and Eluru, 2014; Givoni and Rietveld, 2014), airport choice (Marcucci and Gatta, 2011) and vacation location choice (Hong et al., 2006). A number of research efforts also examined residential location and work place location choices (Sermons and Koppelman, 2001; Waddell et al., 2007). The multinomial logit model is the most common approach employed in these research efforts.

2.2. Social Media Data for Travel Behavior Analysis

Social media platform such as Twitter has been considered as an useful source of travel behavior information in various studies (Cao et al., 2014; Chang et al., 2012; Gal-Tzur et al., 2014; Maghrebi et al., 2015). The easy availability and wide range of applications have made the data valuable for researchers in multiple fields including social science, marketing, public health, computer science, and transportation science (Lian and Xie, 2011). The dimensions considered include finding mobility and activity choices (Chen et al., 2017; Hasan and Ukkusuri, 2014), classification of activity choice patterns (Cheng et al., 2011), role of friendship on mobility (Hasan et al., 2016; Sadri et al., 2017), and modeling activity sequence (Hasan and Ukkusuri, 2017). In transportation planning, researchers have used this data to estimate urban travel demand (Lee et al., 2017; Liu et al., 2014) and traffic flow (Liu et al., 2014; Wu et al., 2014). Thus, social media data has a significant potential for travel demand models, traffic operations and management and long term transportation planning purposes (Rashidi et al., 2017). Despite the increased interests to social media data, few studies have employed such data for destination choice analysis (Molloy and Moeckel, 2017)).

2.3. Current Research in Context

From the aforementioned review, it is evident that while traditional survey data have been widely employed for destination choice analysis, only one research effort employed social media data for destination choice analysis. Furthermore, this study adopted the traditional multinomial logit model and thereby did not consider population heterogeneity. As stated earlier, to capture the population heterogeneity in terms of several major variables, a latent segmentation approach is preferred (Eluru et al., 2012 or Sobhani et al., 2013 for more discussion). Faghih-Imani et al.

(Faghih-Imani et al., 2016) recently employed a latent segmentation multinomial logit model (LSMNL) to model bicyclists' destination preferences for the New York CitiBike system. We customize the LSMNL approach for analyzing destination choice with Twitter data by recognizing the presence of repeated observations in twitter data. To illustrate the value of the proposed model, we compare its performance with estimates of separate MNL models developed by activity purpose.

3. Data Preparation

Twitter data were collected using its streaming API from March 29, 2017 to October 10, 2017 within geographic boundary of Central Florida region (defined by the coordinates -82.059860, 27.034087 (lower left corner of De-soto County) and -81.153310, 29.266654 (a corner of Volusia County). However, collected data also included tweets without geo-tagged coordinates as the 'user locations' in their Twitter profiles mentioned places inside Florida; which is not unusual as explained in Twitter Developer Documentation ("Twitter Developer Documentation: Streaming API," 2006). The coordinates of the collected geo-tagged tweets were found to be spread across the whole state of Florida instead of remaining within the defined boundary of Central Florida region only.

We then filtered out BOTs and users with less than two geotagged tweets from the data. A social BOT is a software program which interacts like any human user on platforms like Twitter, Facebook, Reddit etc. (Woolley, 2016). Botometer provides the bot-likelihood scores of user profiles by analyzing the recent activities of user profiles i.e. content, sentiments, friends, networks etc. (Davis et al., 2016). BOT score ranges from 0 to 1 and a social BOT is likely to have higher BOT score ("Botometer," 2014). By collecting the BOT scores of each user profile and by placing a suitable threshold value, the social BOTs were cleared out of the data set. Details of this filtering process can be found in (Hasnat and Hasan 2018). After filtering, we collected the latest 3200 tweets of 4601 resident user accounts. To identify residents, we used self-declared locations ('user location') posted in their Twitter profiles. Within the data collection period (March 29, 2017 - October 10, 2017), we were able to extract 77,751 geo-tagged tweets from these 4601 resident users.

Next, we located resident users' home locations at census tract level inside Florida using Python's geohash ("Geohash 1.0," 2015) library. Geohash divides the geographical area into predefined rectangular boundaries (in our case we selected 152 meter by 152 meter geohash). We counted the number of coordinates that fell within each geohash and reported the geohash with the largest number of coordinates as the user's home location. Again, we set a minimum threshold of 3 geo-tagged posts within a geohash to consider the location as the user's home. In this method, we found home locations of nearly 400 users. But we were able to manually validate the home location (city level locations posted in the user profile) and extract the demographic information (age group and gender) of only 345 users from their Twitter profiles. Therefore, we conducted the subsequent analyses for the destinations of these 345 users.

We then spatially joined destinations with Florida's census tracts shapefile Using ArcGIS. We merged different data sources containing the number of offices, schools, entertainment centers, hospitals etc. in Florida and spatially joined them with the census tract shapefile. The files were gathered from different sources including the tigerline shape files (United States Census Bureau), Florida geographic files database ("Florida Geographic File Database") etc.

For this study, based on the destination types, trips are categorized as recreational trips, shopping trips, and others. In order to find the destination types, we found the locations of the geo-tagged tweets and merged them with the latest open street and land use map of Florida in ArcGIS environment. Using the land use map, we classified majority of the trips into different purposes. For example, the tweets posted from the recreational areas, national and state parks, reserve forests etc. can be easily associated with recreational trips. When locating the destinations, we excluded the tweets posted from roads. To do that, we created 10-meter buffers on both sides of a road in the road network shapefile and excluded the tweets that fell within that buffer area. We manually checked the higher density regions along the highways to avoid excluding the tweets posted from motels, hotels, and restaurants located close to the highways. Geo-tagged tweets were also found to be posted from locations such as shopping malls, airports, Amtrak stations, restaurants etc. All the locations, that could not be classified using the land use map, were manually classified after locating the geo-tagged tweets and extracting the corresponding location categories from Florida's latest Open Street Map. In case of large shopping malls, we separated the restaurants, bars, coffee shops, hotels inside or near the shopping malls and put them in the 'Others' category of trip purposes. One potential limitation of our approach is that if a user works in a shopping mall, his/her tweets can be mistakenly associated with a shopping trip instead of a work trip.

The data set contained 345 users with home in 199 different census tracts and 44,085 destinations in 1651 different census tracts. Some of the users' home locations and their travelled destinations inside Florida are shown in Figure 1.

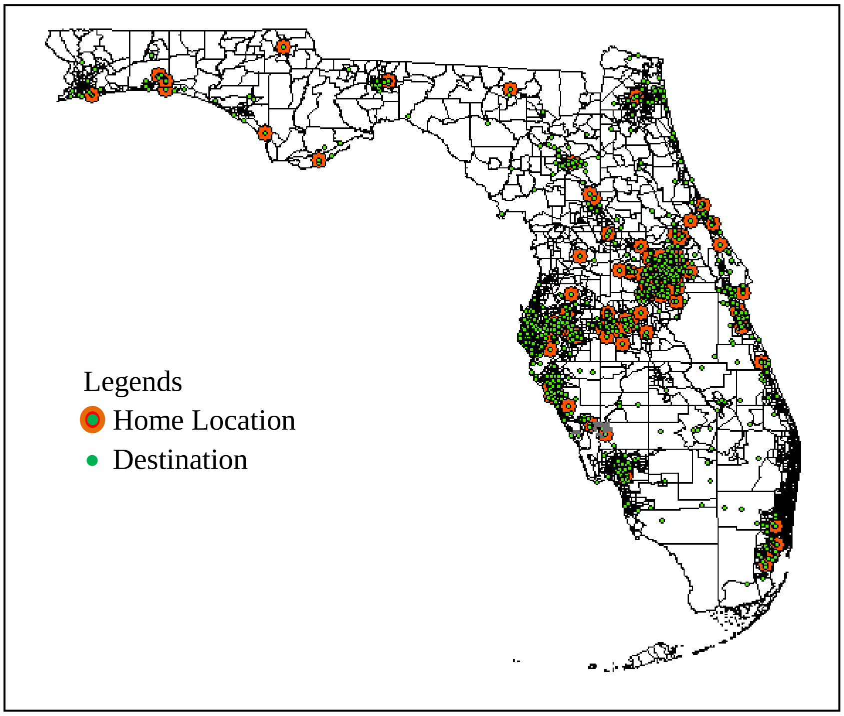

FIGURE 1: Merging user home and destination with census tracts.

We only retained the destinations that made sense based on timeline analyses. From a user's destination set, we excluded any location if he/she posted several tweets within a very short period of time from that location. Out of 44,085 trips, 34,000 trips are used for model estimation and the remaining 10,085 trips are kept for model validation. To reduce computational burden for estimating the discrete choice model, we employed a choice set size of 30 census tracts. For these purposes, for every destination choice record, the choice set is created by adding 29 randomly drawn census tracts as alternatives.

Several earlier studies have shown that a choice set of size 30 is adequate for sampling in MNL models (see Nerella and Bhat, 2004; Pozsgay and Bhat, 2001; Faghih-Imani and Eluru, 2015 for examples). In our context, the choice set size was determined based on the complexity of model estimation and model run times. In a recent paper, Faghih-Imani and Eluru (2017) tested the impact of sampling within a latent segmentation multinomial logit and found that the choice set size of 30 performed as well as the choice sets with larger number of alternatives (60 and 120). Based on these findings, we employed a choice set size of 30.

The variables we extracted include:

a) User age (divided into 5 Age groups: upto 15, 16-25, 26-40, 41-55, 56 and above), and user gender from Twitter profile pictures.

b) Per-capita income (individual mean, 5-year estimate) in 1000 USD.

c) Number of civic centers, schools, hospitals, government building in point shape files.

d) Land use types using the area of residential, industrial, institutional, recreational, office, and landuse mix of the destination and home census tracts.

e) Distance from the center of the home census tract to the center of the destination census tract in kilometers.

We collected the information in (a) by manually going through the profile of each user, and for the other (b to d) we used the geographical databases ("Florida Geographic File Database") and shapefiles (United States Census Bureau).

Table 1 lists the variables and their description used in the models.

TABLE 1: Description of Variables used in Choice Model

Variable	Description	Variable	Description
HINDUSTR	Industrial area in home	DINDUSTR	Industrial area in destination
HRECREAT	Recreational area in home	DRECREAT	Recreational area in destination
HOFFICE	Office area in home	DOFFICE	Office area in destination
HAGRICUL	Agricultural area in home	DAGRICUL	Agricultural area in destination
HRESIDEN	Residential area in home	DRESIDEN	Residential area in destination
HLANDMIX	Landuse mix in home	DLANDMIX	Landuse mix in destination
HHOSPITA	Number of hospitals in home	DHOSPITA	Number of hospitals in destination
HSCHOOL	Number of schools in home	DSCHOOL	Number of schools in destination
HCIVICCE	Number of civic centers in home	DCIVICCE	Number of civic centers in destination
HINCOME	Per-capita income in home	DINCOME	Per-capita income in destination
HGOVMNTB	Number of government buildings in home	DGOVMNTB	Number of government buildings in destination
DISTKM	Distance in kilometers	Weekend	Dummy variable for Weekend
PShop	Dummy variable for shopping trips	PRec	Dummy variable for recreational trips
POther	Dummy variable for other trips	Female	Dummy variable for gender(female=1)

Income in home census tract, age, gender etc. are the invariant alternatives (does not change for an individual, no matter what destination he/she chooses).

4. Methodological Approach

The decision process dictating the individual's destination choice is studied as a random utility maximization approach where the destination/alternative with the highest utility has the highest probability of being chosen (Faghih-Imani and Eluru, 2018). The exogenous variables i.e. the trip attributes, destination attributes which changes across the choices are considered in general MNL model estimation, while origin attributes, user attributes which remains the same across the choices can only be considered through the interaction with the exogenous variables (Faghih-Imani et al., 2016). The Latent Segmentation Multinomial Logit framework allows us to probabilistically classify trips into latent segments based on a host of characteristics including trips, origin, and destination attributes. The destination choice model with latent segmentation assumes that there are S relatively homogenous segments of trips, where the optimal number S has to be determined. The pattern of destination preferences and the sensitivity to the utilities are identical for each user within each segment. Therefore, separate segment specific destination choice models can be developed to present the understanding in an elaborate and clear fashion. Let the utility for assigning a trip j (1, 2, … J) made by individual i (1,2, …, I) to segment s is defined as:

$$U^*_{ijs} = \beta'_s z_{ij} + \xi_{ijs} \tag{1}$$

z_{ij} is a (M x 1) column vector of attributes that influences the propensity of belonging to segment s, β'_s is a corresponding (M x 1) column vector of coefficients and ξ_{ijs} is an idiosyncratic random error term assumed to be identically and independently Gumbel-distributed across trips j and segment s. Then the probability that trip j made by individual i belongs to segment s is given as:

$$P_{ijs} = \frac{\exp(\beta'_s z_{ij})}{\sum_s \exp(\beta'_s z_{ij})} \tag{2}$$

Now let us assume k (1,2, … K, in our study K=30) to be an index to represent the destination zone. When a trip is probabilistically assigned to a segment s and zone k is chosen as the destination, the random utility formulation takes the following form:

$$U_{ijk} | \ s = \alpha_s' x_{ij} + \varepsilon_{ijk} \tag{3}$$

x_{ij} is a (L x 1) column vector of attributes that influences the utility of destination choice model. α_s' is a corresponding (L x 1)-column vector of coefficients and ε_{ijk} is an idiosyncratic random error term assumed to be identically and independently Gumbel distributed across the dataset. Then the probability that trip j chooses zone k as destination within the segment s for individual i is given as:

$$P_{ij}(k) \mid s = \frac{\exp(\alpha_s' x_{ij})}{\sum_k \exp(\ \alpha_s' x_{ij})} \tag{4}$$

Within the latent segmentation framework, the overall probability of trip j by individual i to be destined to zone k is given as:

$$P_{ij}(k) = \sum_{s=1}^{S} (\ P_{ij}(k) \ |s)(P_{ijs}) \tag{5}$$

Therefore, the log-likelihood function for the entire dataset is:

$$LL = \sum_{i=1}^{I} \sum_{j=1}^{J} \log(P_{ij}(k_{ij}^*)) \tag{6}$$

where k_q^* represents the chosen zone for trip j by individual i. By maximizing this log-likelihood function, the model parameters β and α are estimated. GAUSS matrix programming language is used to code the maximum likelihood model estimation.

The model estimation approach begins with a model considering two segments. The final number of segments is determined by adding one segment at a time until further addition does not enhance intuitive interpretation and data fit. We have utilized Bayesian Information Criterion (BIC) to statistically measure the fit as it applies higher penalty on over-fitting and is the most common information criteria used to identify the suitable number of classes for latent segmentation based analysis (Nylund et al., 2007). We have estimated the model with 2, 3, and 4 segments and found the best intuitive results with 3 segments. It must be noted that our panel structure was unbalanced, meaning that the number of repeated observations for individuals (trips made by individuals) varies across the dataset (from 1 trip to 1823 trips with the mean of 98.6 and median of 31 trips).

In the presence of repeated observations, ignoring for such repetitions results in two major considerations for model estimation. *First*, the estimated standard errors are likely to be under-estimated i.e. parameters that are likely to be insignificant might appear as significant. In our study, we have explicitly accommodated for the potential error in standard errors by developing a panel based estimation process that recognizes the repetitions. *Second*, in data with repetitions, common unobserved factors specific to an individual might affect the choice process. However, in our choice context with unlabeled alternatives, given that individual attributes remain constant across all the alternatives, the impact of unobserved factors can only be accommodated across destination attributes or through interaction of destination attributes with demographic variables. Thus, the consideration of individual specific factors is not as direct as is the case in choice contexts with labelled alternatives. For example, in a mode choice context, impact of gender or employment on a particular mode can be considered as a random parameter. However, such an estimation is not possible in a destination choice model.

Further, any attempt to accommodate for these factors will require us to resort to simulation based approaches as closed approaches are not feasible. The estimation of latent segmentation model within a simulated log-likelihood context with large number of alternatives is quite complex and is not easy to arrive a stable specification. Hence, given the increase in model complexity and the relatively marginal benefit of considering unobserved effects, we did not accommodate for individual level preferences in the model.

5. Model Results and Interpretation

Prior to discussing the model results we present a brief comparison of various models we estimated. We developed four different MNL models: one model for all the trips combined and the other three models by activity purposes, i.e. one for recreational trips, one for shopping trips, and one for other trips. The Null Model log-likelihood for the estimation sample is N*ln(1/30). The log-likelihood values for these models were found to be -48,688.78 from the model with all the trips combined, -20,595.79 from the model for recreational trips, -2,078.21 from the model for shopping trips and -20,969.19 from the model for other trips (Table 2). The overall log-likelihood for all observations for trip purpose specific models was - 43,643.19 ((-20,861.4) + (-2,092.97) +(-20,978.34)). The log-likelihood for the PLSMNL model was -34,752.8 which is significantly higher than the overall MNL model or the trip purpose specific model suite. Therefore, it is clear that the PLSMNL model provides a superior fit. For the sake of brevity, from here on we restrict our discussion to the PLSMNL model results. The reader is referred to the appendix for the model for all trips and trip purpose model results. In the subsequent discussion of PLSMNL model, we present the segment membership component followed by discussion of segment specific destination choice models.

TABLE 2: Performance Measures of different MNL Models and PLSMNL Model.

	MNL (All purposes)	**MNL (Recreational)**	**MNL (Shopping)**	**MNL (Other)**	**PLSMNL**
Number of Observations	34,000	15,903	5,921	12,176	34,000
Number of Variables	11	11	9	9	31
LL- Null	-115,640.7	-54,089.2	-20,138.5	-41,413.0	-115,640.7
LL- Final	-48,688.8	-20,861.4	-2,093.0	-20,978.3	-34,752.8
BIC	97492.4	41829.2	4264.2	42041.3	69829.1

5.1.Segment Membership Component

The segmentation membership results are shown in Table 3 with the significant variables (at 90% confidence interval) that influence segment membership. The reader would note that the segment membership model provides a unique perspective on the characteristics of each segment.

TABLE 3: Segmentation Characteristics of PLSMNL

	Segment 1		Segment 2		Segment 3	
Segment Share	0.2029		0.5359		0.2612	
Variable	**Estimates**	**t-stats**	**Estimates**	**t-stats**	**Estimates**	**t-stats**
Constant	-1.0005	-2.038	0.9274	2.752		
WEEKEND	0.736	3.046	-0.573	-1.933	_	_
FEMALE	-1.0239	-1.917	-1.1573	-2.43	_	_
HAGRICUL	0.5064	2.527	_	_	_	_
HRESIDEN	-2.2669	-2.996	_	_	_	_
HOFFICE	0.219	4.069	_	_	_	_
PShop	_	_	5.1135	20.241	_	_

After introduction of continuous variables in the segment membership models, the constant terms do not have any substantive interpretation. The results for the weekend variable indicate a preferential sequence across the three segments. Specifically, destination choices made over the weekend are most likely to be allocated to segment 1 while they are least likely to be allocated to segment 2. In terms of individual gender variable, destination choices of female users are likely to be assigned to segment 3. The segment membership variables are also affected by land use variables. The individuals residing in census tracts with higher agricultural and office area are more likely to be assigned to segment 1 while individuals residing in census tracts with lower residential density are least likely to be allocated to segment 1. Trip purpose variables also influence segment membership. Shopping trips are most likely to be allocated to segment 2.

In addition to identifying various factors affecting segment membership, the PLSMNL model allows us to compute the shares of various segments. In our analysis, the segment shares are as follows: segment 1 – 20.3%, segment 2 – 53.6% and segment 3 – 26.1%. The PLSMNL model can also be employed to generate segment level means for the independent variables (see Table 4).

TABLE 4: Segment shares in PLSMNL

	Segment 1	**Segment 2**	**Segment 3**	**Variable Mean in Overall Sample**
Variables	**Mean of Independent Variables**			
PShop	0.00303	***0.32223***	0.00326	0.17415
PRec	***0.63685***	0.36170	0.55391	0.46774
POther	0.36011	0.31608	***0.44283***	0.35812
FEMALE	0.42563	0.26684	0.50669	0.36171
HAGRICULTURAL	0.09390	-0.01161	0.00780	0.01487
HRESIDENTIAL	**0.452862**	0.178777	0.118021	0.218535
HOFFICE	**11.65662**	1.639678	2.193297	3.817169
DISTKM	45.83628	**24.39721**	35.74828	31.71260
WEEKEND	**0.49882**	0.25468	0.34367	0.32747

An examination of the trip purpose variable means indicates that each segment is dominated by one activity purpose: (1) Segment 1 is likely to be recreational destinations, (2) Segment 2 is mostly shopping activity oriented destination and (3) Segment 3 is predominantly other activities. The reader would note that the segment membership allocation is probabilistic (not exclusive) and hence other activity purposes might exist within these segments. Overall, based on segment membership characteristics from Table 4, it is possible to label the various segments in the model. Segment 1 is predominantly a male weekend recreational activity segment. Segment 2 is geared toward shopping destinations on weekdays. Finally Segment 3 mainly represents female other activity destination trips.

5.2. Segment specific Destination Choice Models

Within a segment, all the destination choice records follow the same utility function (Bhat, 1997). The results of the three segment specific multinomial logit models (MNL) are presented in Table 5.

TABLE 5: Destination Characteristics from Segments specific MNL.

Variable	Segment 1		Segment 2		Segment 3	
	Estimates	***t-stats***	***Estimates***	***t-stats***	***Estimates***	***t-stats***
DISTKM	-0.0064	-4.327	-0.2161	-8.629	-0.0602	-7.060
DINDUSTR	-0.3572	-2.398	0.3424	2.707	-0.2095	-2.372
DRECREAT	0.0600	3.439	_	_	_	_
DOFFICE	_	_	0.1249	7.629	0.4253	4.824
DAGRICUL	_	_	_	_	0.5686	5.126
DLANDMIX	0.3623	4.37	0.2218	2.15	_	_
DSCHOOL	0.1168	2.167	0.2825	3.832	_	_
DCIVICCE	0.4525	15.562	_	_	0.4666	5.319
DINCOME	0.2031	2.605	0.287	2.836		
DGOVMNTB	_	_	_	_	0.3659	3.698
DHOSPITAL	_	_	_	_	0.2166	2.118

In the segment specific model estimation, we employed several destination characteristics. A cursory examination of the results clearly highlights how the variables (and parameter sign/magnitude) influencing the destination choice models across the various segments are quite different. The result provides strong support to our study hypothesis for the presence of population heterogeneity.

In all models, travel distance has a negative coefficient. While a direct comparison of the travel distance across segments needs to be judiciously conducted, a preliminary examination highlights intuitive trends. A low magnitude for the impact of destination is observed for weekend recreational destinations, indicating the higher spatial flexibility over weekends for such trips. A high negative magnitude is observed for the weekday shopping segment highlighting inherent preference for shorter distance trips on weekdays.

In segment 1 destination tract recreational area, land use mix, number of schools, number of civic centers and per-capita income are found to have significant positive impact on the destination alternative. On the other hand, the increased presence of industrial area is likely to reduce the preference for the destination.

In segment 2 industrial area, office area, land use mix, number of schools and income of the destination census are found to have significant positive impact on the individual choice of destination. The results are intuitive considering segment 2 is predominantly weekday shopping destinations. The positive impact of number of schools and office areas variables can be related to the fact that people on weekdays do not leave home only for shopping, rather they prefer shopping on their way to office or in some cases near schools.

For segment 3 we find the variables for office area, agricultural area, number of civic centers and government buildings in the destination census are found to have significant positive impacts (Table 5).

5.3. Validation

To further investigate the performance of the developed models, a validation exercise is undertaken on a hold-out sample. The validation sample has 10,085 trips made by 313 individuals. The same data processing and choice set generation approach are employed for the validation sample preparation. As an evaluation measure for prediction performance, the predictive log-likelihood is computed based on the estimation results of the proposed PLSMNL model as well as the MNL models for all trips. Further, the trip purpose specific models are used to predict for the trips in validation sample corresponding to that specific purpose. Table 6 presents the results of the validation exercise.

As expected, the trip purpose specific models perform better than the traditional MNL models. The PLSMNL model outperforms the traditional MNL models. The predictive log-likelihood for PLSMNL model is -10,248.8 while the corresponding value for the traditional MNL is -14,253.9. Only the shopping specific model slightly performs better than the PLSMNL model in predicting for shopping trips. Overall, the validation exercise exhibits that in addition to providing a richer explanatory power, the proposed PLSMNL model performs relatively well in terms of prediction.

TABLE 6: Model Validation Results

Predictive Log-likelihood	MNL (All purposes)	MNL (Recreational)	MNL (Shopping)	MNL (Other)	PLSMNL
Overall	-14,253.9	-	-	-	-10,247.8
Recreational Trips	-6043.2	-5826.3	-	-	-4833.9
Shopping Trips	-1884.5	-	-622.2	-	-638.6
Other Trips	-6326.3	-	-	-6224.67	-4775.

6. **Conclusions**

In this study, we present methods to extract and analyze data collected from Twitter for modeling travelers' destination choice behavior. We have adopted filtering steps to remove social bots from the dataset and prepare a reliable sample for analysis. We have created a dataset combining social media data with traditional census tract based socio-economic, land-use, and infrastructure data. To understand destination choice behavior from social media data, we propose a Panel Latent Segmentation Multinomial Logit (PLSMNL) model. The model has best fit with three segments and outperforms an overall MNL model and trip specific MNL models. The qualitative assessments of the models indicate that the proposed PLSMNL has intuitively assigned destinations by trip purpose (shopping, recreational, and other), gender, weekday (or weekend) and home zone land use measures. The segment specific destination choice models offer interesting insights on the impact of land use attributes on destination choice. The results highlight an application of social media data for destination choice analysis. Overall, the results indicate how we can augment traditional travel survey-based data collection efforts with social media data analytics.

To be sure, our study is not without limitations. We have considered all the trips anchored to the home i.e. travel distance is calculated from home to the destination. We have resorted to this approximation since trip origins and associated trip start times are not readily available from

Twitter data. Also, when selecting trip purposes based on tweet coordinates, our approach has some limitations. For instance, if a shopping mall employee tweets from his/her work place, we classify that as a shopping trip, not as a work trip. However, using a much larger data set, studies have identified user work locations (McNeill et al., 2017). Several studies have demonstrated significant similarities between the findings with social media based data set and the results from traditional survey data (Cheng et al., 2011; Zhu et al., 2014), and have successfully merged data sets from these two domains (i.e. social media data with traditional sensor data) (Zheng et al., 2015). We have not included any such validation analysis.

Future studies using Twitter data may follow several directions. It is possible to associate trips to particular travel modes by analyzing tweet content (Maghrebi et al., 2016). Given that the data is for the Central Florida region, this is unlikely to create any issue as automobile is the predominant alternative. While we employed manual approaches to determine age group and gender of the users, there are methods to find the demographic features of Twitter users such as age group, gender, ethnicity etc. (Longley et al., 2015; Mislove et al., 2011; Sloan et al., 2015). These methods can be employed for larger sample of users. Collecting data on a larger bounding box, for longer period, and finally finding better and accurate ways of filtering social BOTs will certainly increase the sample size.

Transportation agencies still rely on traditional household surveys for planning future development projects. Being costly and time consuming, these surveys can only be afforded once in every 5 to 10 years at a limited scale. Social media data can provide a potential solution to this issue. With limited resources, social media data can provide the most recent and longitudinal travel information for a large number of people. However, more research efforts are needed for utilizing social media data in practice. We believe, in future, such efforts will be made in several directions. Natural language processing techniques can be adopted to incorporate more content-based data (i.e. age, gender, travel mode, trip purposes etc.), making the most versatile use of travel information from social media data. Advanced machine learning approaches can be used to extract information from non-text based data (e.g., photos and videos) for using in travel behavior analysis. With millions of active users generating content in social media, it is anticipated to have a large enough and representative sample (i.e. consistent with the overall distribution of population by age and gender) in social media. However, econometric approaches should be developed to test this assumption and address potential sampling biases. Finally, novel fusion approaches combining large-scale noisy social media data and small-scale gold-standard survey data will be a major step towards utilizing social media data in practice.

REFERENCES

Abbasi, A., Rashidi, T.H., Maghrebi, M., Waller, S.T., 2015. Utilising Location Based Social Media in Travel Survey Methods. Proc. 8th ACM SIGSPATIAL Int. Work. Locat. Soc. Networks - LBSN'15 1–9. https://doi.org/10.1145/2830657.2830660

Beirão, G., Sarsfield Cabral, J.A., 2007. Understanding attitudes towards public transport and private car: A qualitative study. Transp. Policy 14, 478–489. https://doi.org/10.1016/j.tranpol.2007.04.009

Bekhor, S., Prashker, J.N., 2008. GEV-based destination choice models that account for unobserved similarities among alternatives. Transp. Res. Part B Methodol. 42, 243–262. https://doi.org/10.1016/j.trb.2007.08.003

Beyer, M.A., Laney, D., 2012. The importance of 'big data': a definition. Gartner: Stamford, CT, USA.

Bhat, C.R., 1997. An Endogenous Segmentation Mode Choice Model with an Application to Intercity Travel. Transp. Sci. 31, 34–48. https://doi.org/10.1287/trsc.31.1.34

Botometer [WWW Document], 2014. URL https://botometer.iuni.iu.edu/#!/ (accessed 12.26.17).

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., Soltani, K., 2014. A Scalable Framework for Spatiotemporal Analysis of Location-based Social Media Data. https://doi.org/10.1016/j.compenvurbsys.2015.01.002

Chakour, V., Eluru, N., 2014. Analyzing commuter train user behavior: A decision framework for access mode and station choice. Transportation (Amst). 41, 211–228. https://doi.org/10.1007/s11116-013-9509-y

Chang, H.W., Lee, D., Eltaher, M., Lee, J., 2012. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). pp. 111–118. https://doi.org/10.1109/ASONAM.2012.29

Chen, Y., Mahmassani, H.S., Frei, A., 2017. Incorporating social media in travel and activity choice models: conceptual framework and exploratory analysis. Int. J. Urban Sci. 0, 1–21. https://doi.org/10.1080/12265934.2017.1331749

Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z., 2011. Exploring Millions of Footprints in Location Sharing

Services. Icwsm 2010, 81–88. https://doi.org/papers3://publication/uuid/0C46BD5D-4908-4A8A-BD06-5BCB2F1DE282

Chi, E.H., 2008. The social web: Research and opportunities. Computer (Long. Beach. Calif). 41, 88–91. https://doi.org/10.1109/MC.2008.401

Comscore, 2011. Nearly 1 in 5 Smartphone Owners Access Check-In Services Via their Mobile Device [WWW Document]. Press Release.

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., Zegras, P., 2013. The Future Mobility Survey: Experiences in developing a smartphone-based travel survey in Singapore Caitlin. Transp. Res. Rec. J. Transp. Res. Board 59–67.

Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F., 2016. BotOrNot: A System to Evaluate Social Bots 4–5. https://doi.org/10.1145/2872518.2889302

Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accid. Anal. Prev. 47, 119–127. https://doi.org/10.1016/j.aap.2012.01.027

Faghih-Imani, A., Eluru, N., Forum, C.T.R., 2016. A Latent Segmentation Multinomial Logit Approach to Examine Bicycle Sharing System Users' Destination Preferences 1 PDF file, 578 KB, 7p.

Florida Geographic File Database [WWW Document], 2008. URL ftp://ftp1.fgdl.org/pub/state/ (accessed 12.15.17).

Flyvbjerg, B., Holm, M.S., Buhl, S.L., 2005. How (In) accurate Are Demand Forecasts in Public Works Project? The Case of Transportation. J. Am. Plan. Assoc. 71, 131–146. https://doi.org/10.1080/01944360508976688

Frias-Martinez, V., Soto, V., Hohwald, H., Frias-Martinez, E., 2012. Characterizing urban landscapes using geolocated tweets. ASE/IEEE Int. Conf. Soc. Comput. Soc. 2012 239–248. https://doi.org/10.1109/SocialCom-PASSAT.2012.19

Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I., 2014. The potential of social media in delivering transport policy goals. Transp. Policy 32, 115–123. https://doi.org/10.1016/j.tranpol.2014.01.007

Geohash 1.0 [WWW Document], 2015. URL https://www.elastic.co/guide/en/elasticsearch/guide/current/geohashes.html (accessed 2.20.18).

Geurs, K.T., Thomas, T., Bijlsma, M., Douhou, S., 2015. Automatic trip and mode detection with move smarter: First results from the Dutch Mobile Mobility Panel. Transp. Res. Procedia 11, 247–262. https://doi.org/10.1016/j.trpro.2015.12.022

Givoni, M., Rietveld, P., 2014. Do cities deserve more railway stations? The choice of a departure railway station in a multiple-station region. J. Transp. Geogr. 36, 89–97. https://doi.org/10.1016/j.jtrangeo.2014.03.004

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., Crane, M., 2015. A web-based diary and companion smartphone app for travel/activity surveys. Transp. Res. Procedia 11, 297–310. https://doi.org/10.1016/j.trpro.2015.12.026

Greene, E., Flake, L., Hathaway, K., Geilich, M., 2015. A Seven-Day Smartphone-Based Gps Household Travel Survey in Indiana, in: Transportation Research Board 95th Annual Meeting. https://doi.org/10.3141/2594-06

Hasan, S., Ukkusuri, S. V., 2017. Reconstructing Activity Location Sequences From Incomplete Check-In Data: A Semi-Markov Continuous-Time Bayesian Network Model. IEEE Trans. Intell. Transp. Syst. 1–12. https://doi.org/10.1109/TITS.2017.2700481

Hasan, S., Ukkusuri, S. V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. PLoS One 10, 1–19. https://doi.org/10.1371/journal.pone.0124819

Hasan, S., Ukkusuri, S. V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C Emerg. Technol. 44, 363–381. https://doi.org/10.1016/j.trc.2014.04.003

Hasan, S., Ukkusuri, S. V., Zhan, X., 2016. Understanding Social Influence in Activity Location Choice and Lifestyle Patterns Using Geolocation Data from Social Media. Front. ICT 3, 1–9. https://doi.org/10.3389/fict.2016.00010

Hasnat, M.M., Hasan, S., 2018. Analyzing Spatial Patterns of Tourist Destinations from Location-based Social Media Data: Filtering, Classification and Clustering Methods, in: Transportation Research Board 97th Annual Meeting.

Hendrik, H., Perdana, D.H.F., 2014. Trip Guidance: A Linked Data Based Mobile Tourists Guide. Adv. Sci. Lett. 20, 75–79. https://doi.org/https://doi.org/10.1166/asl.2014.5285

Hong, S. kwon, Kim, J. hyun, Jang, H., Lee, S., 2006. The roles of categorization, affective image and constraints on destination choice: An application of the NMNL model. Tour. Manag. 27, 750–761. https://doi.org/10.1016/j.tourman.2005.11.001

Horni, A., Scott, D., Balmer, M., Axhausen, K., 2009. Location Choice Modeling for Shopping and

Leisure Activities with MATSim. Transp. Res. Rec. J. Transp. Res. Board 2135, 87–95. https://doi.org/10.3141/2135-11

Jonnalagadda, N., Freedman, J., Davidson, W., Hunt, J., 2001. Development of Microsimulation Activity-Based Model for San Francisco: Destination and Mode Choice Models. Transp. Res. Rec. 1777, 25–35. https://doi.org/10.3141/1777-03

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. 53, 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Koppelman, F.S., Sethi, V., 2005. Incorporating variance and covariance heterogeneity in the Generalized Nested Logit model: An application to modeling long distance travel choice behavior. Transp. Res. Part B Methodol. 39, 825–853. https://doi.org/10.1016/j.trb.2004.10.003

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., Shoor, I., 2017. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. Transp. Res. Part C Emerg. Technol. 77, 275–291. https://doi.org/10.1016/j.trc.2017.02.003

Lee, J.H., Mcbride, E., Mcbride, E., Goulias, K.G., 2017. Exploring Social Media Data for Travel Demand Analysis : A comparison of Twitter , household travel survey and synthetic population data in California. 95th Annu. Meet. Transp. Res. Board 500.

Lian, D., Xie, X., 2011. Collaborative activity recognition via check-in history. Proc. 3rd ACM SIGSPATIAL Int. Work. Locat. Soc. Networks - LBSN ’11 1. https://doi.org/10.1145/2063212.2063230

Liu, Y., Sui, Z., Kang, C., Gao, Y., 2014. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS One 9. https://doi.org/10.1371/journal.pone.0086026

Longley, P.A., Adnan, M., Lansley, G., 2015. The geotemporal demographics of twitter usage. Environ. Plan. A 47, 465–484. https://doi.org/10.1068/a130122p

Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, S.T., 2015. Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC 2015–Octob, 208–213. https://doi.org/10.1109/ITSC.2015.43

Maghrebi, M., Abbasi, A., Waller, S.T., 2016. Transportation application of social media: Travel mode extraction. IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC 1648–1653. https://doi.org/10.1109/ITSC.2016.7795779

Marcucci, E., Gatta, V., 2011. Regional airport choice: Consumer behaviour and policy implications. J. Transp. Geogr. 19, 70–84. https://doi.org/10.1016/j.jtrangeo.2009.10.001

McFadden, D., 1978. Modeling the Choice of Residential Location. Transp. Res. Rec. 72–77.

McNeill, G., Bright, J., Hale, S.A., 2017. Estimating local commuting patterns from geolocated Twitter data. EPJ Data Sci. 6. https://doi.org/10.1140/epjds/s13688-017-0120-x

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., Rosenquist, J.N., 2011. Understanding the Demographics of Twitter Users. Artif. Intell. 554–557.

Molloy, J., Moeckel, R., 2017. Improving Destination Choice Modeling Using Location-Based Big Data. ISPRS Int. J. Geo-Information 6, 291. https://doi.org/10.3390/ijgi6090291

Moscardo, G., 2004. Shopping as a destination attraction : An empirical examination of the role of ... 10, 294–307.

North Florida Travel Survey [WWW Document], 2017. URL https://www.northfloridatravelsurvey.com/northfloridahtsweb/pages/privacy?locale=en-US

Nylund, K.L., Asparouhov, T., Muthén, B.O., 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Struct. Equ. Model. 14, 535–569. https://doi.org/10.1080/10705510701575396

NYMTC, NJTPA, 2014. Executive Summary: 2010 / 2011 Regional Household Travel Survey.

Perrin, A., 2005. Social Media Usage: 2005-2015: 65% of Adults Now Use Social Networking Sites--a Nearly Tenfold Jump in the Past Decade. Pew Res. Cent.

Pozsgay, M., Bhat, C., 2001. Destination Choice Modeling for Home-Based Recreational Trips: Analysis and Implications for Land Use, Transportation, and Air Quality Planning. Transp. Res. Rec. 1777, 47–54. https://doi.org/10.3141/1777-05

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp. Res. Part C Emerg. Technol. 75, 197–211. https://doi.org/10.1016/j.trc.2016.12.008

Rashidi, T.H., Mohammadian, A. (Kouros), Zhang, Y., 2010. Effect of Variation in Household Sociodemographics, Lifestyles, and Built Environment on Travel Behavior. Transp. Res. Rec. J. Transp. Res. Board 2156, 64–72. https://doi.org/10.3141/2156-08

Recker, W.W., Kostyniuk, L.P., 1978. Factors influencing destination choice for the urban grocery shopping trip. Transportation (Amst). 7, 19–33. https://doi.org/10.1007/BF00148369

Sadri, A.M., Hasan, S., Ukkusuri, S. V., 2017. Joint Inference of User Community and Interest Patterns in Social Interaction Networks.

Safi, H., Assemi, B., Mesbah, M., Ferreira, L., Hickman, M., 2015. Design and Implementation of a Smartphone-Based Travel Survey. Transp. Res. Rec. J. Transp. Res. Board 2526, 99–107. https://doi.org/10.3141/2526-11

Seddighi, H.R., Theocharous, A.L., 2002. A model of tourism destination choice: A theoretical and empirical analysis. Tour. Manag. 23, 475–487. https://doi.org/10.1016/S0261-5177(02)00012-2

Sermons, M.W., Koppelman, F.S., 2001. Representing the differences between female and male commute behavior in residential location choice models. J. Transp. Geogr. 9, 101–110. https://doi.org/10.1016/S0966-6923(00)00047-8

Shiftan, Y., Ben-Akiva, M., 2011. A practical policy-sensitive, activity-based, travel-demand model. Ann. Reg. Sci. 47, 517–541. https://doi.org/10.1007/s00168-010-0393-5

Sivakumar, A., Bhat, C., 2007. Comprehensive, Unified Framework for Analyzing Spatial Location Choice. Transp. Res. Rec. J. Transp. Res. Board 2003, 103–111. https://doi.org/10.3141/2003-13

Sloan, L., Morgan, J., Burnap, P., Williams, M., 2015. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLoS One 10, 1–20. https://doi.org/10.1371/journal.pone.0115545

Sobhani, A., Eluru, N., Faghih-Imani, A., 2013. A latent segmentation based multiple discrete continuous extreme value model. Transp. Res. Part B.

Twitter Developer Documentation: Streaming API [WWW Document], 2006. URL https://dev.twitter.com/streaming/overview/request-parameters#filter_level (accessed 12.30.17).

Twitter Facts [WWW Document], 2017. . Omnicore. URL https://www.omnicoreagency.com/twitter-statistics/.

United States Census Bureau, n.d. TIGER/Line® Shapefiles and TIGER/Line® Files [WWW Document]. URL https://www.census.gov/geo/maps-data/data/tiger-line.html (accessed 3.15.18).

Waddell, P., Bhat, C., Eluru, N., Wang, L., Pendyala, R., 2007. Modeling Interdependence in Household Residence and Workplace Choices. Transp. Res. Rec. J. Transp. Res. Board 2003, 84–92. https://doi.org/10.3141/2003-11

Wiersma, Y., 2010. Birding 2.0: Citizen science and effective monitoring in the Web 2.0 World. Avian Conserv. Ecol. 5, 1–9.

Woolley, S.C., 2016. Automating power: Social bot interference in global politics. First Monday 21. https://doi.org/http://dx.doi.org/10.5210/fm.v21i4.6161

Wu, L., Zhi, Y., Sui, Z., Liu, Y., 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. PLoS One 9. https://doi.org/10.1371/journal.pone.0097010

Zhang, Y., Mohammadian, A. (Kouros), 2008. Bayesian Updating of Transferred Household Travel Data. Transp. Res. Rec. J. Transp. Res. Board 2049, 111–118. https://doi.org/10.3141/2049-13

Zhang, Z., He, Q., Zhu, S., 2017. Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. Transp. Res. Part C Emerg. Technol. 85, 396–414. https://doi.org/10.1016/j.trc.2017.10.005

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., Yang, L., 2015. Big Data for Social Transportation. IEEE Trans. Intell. Transp. Syst. 17, 620–630. https://doi.org/10.1109/TITS.2015.2480157

Zhu, Z., Blanke, U., Tröster, G., 2014. Inferring Travel Purpose from Crowd-Augmented Human Mobility Data. Proc. First Int. Conf. IoT Urban Sp. https://doi.org/10.4108/icst.urb-iot.2014.257173

APPENDIX

TABLE 1: MNL for all Trips

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.0296063	0.000178	-166.27
DAGRICULTURAL	0.0428767	0.015523	2.76
DRESIDENTIAL	0.0948184	0.016194	5.86
DOFFICE	0.0669291	0.007265	9.21
DLANDMIX	0.2210507	0.00904	24.45
DGOVMNTB	0.1654784	0.010166	16.28
DHOSPITA	0.0878285	0.005896	14.9
DSCHOOL	0.1403953	0.006081	23.09
DCIVICCE	0.2403403	0.008739	27.5
DINCOME	0.208048	0.012313	16.9
Male_DAGRICULTURAL	0.0631316	0.017523	3.6
Male_DRESIDENTIAL	-0.060991	0.019041	-3.2
Male_DOFFICE	0.0272397	0.009159	2.97
Male_DGOVMNTB	-0.0836544	0.012501	-6.69
Male_DCIVICCE	-0.0573614	0.01061	-5.41
Male_DINCOME	-0.0755817	0.016081	-4.7

TABLE 2: MNL for Recreational Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.0235425	0.000217	-108.34
DLANDMIX	0.3192707	0.019527	16.35
DGOVMNTB	0.1364427	0.00835	16.34
DCIVICCE	0.3287527	0.007436	44.21
DRECREATION	0.0702356	0.012205	5.75
DINCOME	0.3261342	0.010863	30.02
Male_DOFFICE	0.1429704	0.007433	19.23
Male_DCIVICCE	-0.063596	0.011016	-5.77
Male_DLANDMIX	-0.061612	0.024924	-2.47

TABLE 3: MNL for Shopping Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.2566439	0.006075	-42.25
DINSTITUTIONAL	1.103844	0.4531562	2.44
DRESIDENTIAL	0.2005166	0.0658365	3.05
DOFFICE	0.2421377	0.0328572	7.37
DOINDUSTRIAL	0.1631876	0.08158	2
DGOVMNTB	0.422072	0.0652397	6.47
DSCHOOL	0.3211626	0.0273803	11.73
DINCOME	0.1959426	0.0368708	5.31
Male_DOFFICE	-0.2746261	0.04256	-6.45
Male_DLANDMIX	0.2798222	0.0435436	6.43
Male_DCIVICCE	0.1563664	0.0236936	6.6
Male_DGOVMNTB	-0.625961	0.0785387	-7.97

TABLE 4: MNL for Other Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.0293272	0.0002742	-106.97
DAGRICULTURAL	0.1026485	0.0187474	5.48
DRESIDENTIAL	0.0686423	0.0155368	4.42
DOFFICE	0.0298732	0.0127478	2.34
DRECREATION	0.0850152	0.0130115	6.53
DLANDMIX	0.2606018	0.013769	18.93
DGOVMNTB	0.2664115	0.0144137	18.48
DSCHOOL	0.2558833	0.0151766	16.86
DCIVICCE	0.1338829	0.0079945	16.75
DINCOME	0.07935	0.0204572	3.88
Male_DAGRICULTURE	0.1620221	0.0212213	7.63
Male_DOFFICE	-0.0385111	0.0138642	-2.78
Male_DGOVMNTB	-0.1036496	0.0179286	-5.78
Male_DRECREATION	-0.7799382	0.0677137	-11.52
Male_Dschool	-0.2083414	0.0189557	-10.99
Male_DINCOME	-0.1115122	0.0265897	-4.19

TABLE 5: Segment Shares for PLSMNL.

Variables	Mean of Independent Variables			Variable Mena in Overall Sample
	Segment 1	Segment 2	Segment 3	
AGE15	0.00590	0.00258	0.00712	0.00444
AGE1625	0.09010	0.05563	0.09502	0.07291
AGE2640	0.57565	0.62839	0.55796	0.59929
AGE4155	0.27527	0.24973	0.24511	0.25371
AGE56	0.05233	0.06343	0.09395	0.06915
FEMALE	0.42563	0.26684	0.50669	0.36171
PSHOP	0.00303	0.32223	0.00326	0.17415
PREC	0.63685	0.36170	0.55391	0.46774
POTHER	0.36011	0.31608	0.44283	0.35812
HAGRICULTURAL	0.09390	-0.01161	0.00780	0.01487
HINDUSTRIAL	0.80618	0.19245	0.19874	0.31865
HINSTITUTIONAL	-0.02250	-0.02226	-0.02332	-0.02258
HRECREATION	0.01137	-0.02817	-0.02666	-0.01975
HRESIDENTIAL	0.45286	0.17878	0.11802	0.21854
HOFFICE	11.65662	1.63968	2.19330	3.81717
HBUA	0.19221	0.00053	0.01760	0.04389
HLANDMIX	0.84659	0.33081	0.45246	0.46725
HGOVMNTBUILDING	0.58614	0.51493	0.57211	0.54432
HHOSPITAL	0.04214	0.12884	0.13467	0.11277
HSCHOOL	0.91779	1.00072	0.70210	0.90590
HCIVICCENTER	6.85813	1.78206	1.87492	2.83649
HINCOME	0.01268	0.03979	0.12637	0.05690
DAGRICULTURAL	0.10177	-0.04344	0.00111	-0.00233
DINDUSTRIAL	0.56183	0.20329	0.27183	0.29396
DINSTITUTIONAL	-0.01528	-0.01992	-0.01848	-0.01860
DRECREATION	0.01803	-0.00512	0.01657	0.00524
DRESIDENTIAL	0.42787	0.16326	0.20760	0.22854
DOFFICE	8.77714	3.11301	4.09335	4.51855
DBUA	0.17819	-0.00979	0.04542	0.04278

Variables	Mean of Independent Variables			Variable Mena in Overall Sample
	Segment 1	Segment 2	Segment 3	
DLANDMIX	0.64311	0.45145	0.48407	0.49887
DGOVMNTBUILDING	0.48698	0.32611	0.45881	0.39341
DHOSPITAL	0.07109	0.18085	0.17306	0.15654
DSCHOOL	0.78526	0.75808	0.72269	0.75435
DCIVICCENTER	5.26990	2.23893	2.74115	2.98521
DINCOME	0.04375	-0.02437	0.03959	0.00616
WEEKEND	0.49882	0.25468	0.34367	0.32747
DISTKM	45.83628	24.39721	35.74828	31.71260