Predicting Tour Patterns Derived from Ubiquitous Data Sources

Md. Shahadat Iqbal  
Department of Civil Engineering.  
Bangladesh University of Engineering and Technology  
Shahadat.buet05@gmail.com

Md. Abu Bakkar Siddique  
Department of Civil Engineering.  
Bangladesh University of Engineering and Technology  
sohag_024@yahoo.com

Md. Mozahidul Islam  
Department of Civil Engineering.  
Bangladesh University of Engineering and Technology  
mozahidul_islam_ce@yahoo.com

Charisma F. Choudhury*  
Department of Civil Engineering.  
Bangladesh University of Engineering and Technology  
cfc@alum.mit.edu

Word Count  
Tables and Figures  
7

Word Count  
4493

Total  
6243

*Corresponding Author
Abstract

Ubiquitous data sources (e.g. GPS traces, mobile phone call records, smart card entries etc.) are increasingly being popular for deriving mobility patterns. These mobility patterns depend on socio-economic characteristics of the traveler as well as associated situational, contextual and environmental factors. These patterns therefore vary substantially with age, gender, employment status, income level and other demographic factors. Habitual markers (propensity to use the internet, social networking websites, mobile phones or media for example) also provide useful indications about these patterns. In this research, the day-to-day tour patterns of travelers have been extracted from GPS and WLAN records and a discrete choice modeling framework has been proposed to predict these patterns using demographic factors, habitual markers and other travel related attributes collected from travelers of Lausanne, Switzerland. The model parameters are estimated by maximum likelihood technique using the software BIOGEME. Estimated model parameters confirm that demographic factors (e.g. age, occupation and gender) combined with habitual markers (e.g. habit of listening to music on smart phone) and other contextual attributes (e.g. day of the week) can be used to predict the tour pattern of an individual on a certain day. The developed model demonstrates how information obtained from ubiquitous data sources can be successfully used as a tool for transportation planning and management.

Keywords: Tour pattern, GPS data, Mobile Data Challenge
1. Background

Predicting mobility patterns are critical for transportation and activity space planning. It is also important for other general planning purposes: modeling the spread of epidemics for example (1,2). Mobility pattern of an individual, which includes his/her spatial and temporal travel choices, depends on his/her socio-economic characteristics as well as associated situational, contextual and environmental factors. These patterns therefore vary substantially with age, gender, employment status, income level and other demographic factors (3,4,5). Habitual markers (propensity to use the internet, social networking websites, mobile phones for example) also provide useful indications about the mobility pattern (6,7,8).

Data available from ubiquitous sources (e.g. GPS trajectories, WLAN access records, mobile phone call records, etc.) have expanded both the quality and quantity of information available to model mobility patterns (9). These data sources are increasingly being used in travel behavior modeling because of several advantages. Firstly, compared to other data sources, it is relatively easy to get panel data over long durations using these sources which makes it possible to model travel choices of individuals at relatively lower costs (10,11). This is particularly useful since there have been evidences that there are often little repetition of tours from one day to the next and therefore models estimated with cross-sectional data can lead to serious errors in prediction(12). Secondly, unlike traditional travel diaries, ubiquitous sources do not have reporting errors and fatigue effects and have smaller non-response bias (13,14,15,16,17,18). However, extraction of trajectories from these ubiquitous data sources is extremely challenging because of many issues: the coarse spatial precisions, missing data points and absence of place labels (e.g. home, workplace, shops) for identifying trip purposes to name a few. This has substantially limited the usage of the full potential of ubiquitous data sources and the previous researches based on such data have been mostly limited to visualization (e.g.19,20,21), aggregate level analyses (e.g.10,11,22,23,24), route choice modeling (e.g.25,26,27,28,29,30,31,), origin-destination flows estimation (e.g. 32,33) and traffic model calibration (e.g.34,35,36,37,38).

In this research, the day-to-day tour patterns of travelers have been extracted from their GPS and WLAN records collected as part of the Mobile Data Challenge (MDC) of Nokia Research Center (39) and a discrete choice model has been developed to predict these tour patterns using demographic data and media usage records supplied with the data. The model parameters are estimated by maximum likelihood technique using the software BIOGEME (40).

The rest of the paper is organized as follows: the details of the data and the methodology to extract the tour patterns are presented first. The details of the discrete choice model developed to predict the patterns are presented next along with the estimation results. The key findings and the directions of further enhancement are presented at the end.
2. Tour Pattern Extraction

Data used in this research is a part of the data collected by a Data Collection Campaign in the Lake Geneva region by Nokia Research Center Lausanne and its Swiss academic partners (39). The data was collected using smart phones from nearly 200 individuals living in Lake Geneva region in the course of more than one year. For research community the data was released as the Mobile Data Challenge (MDC). Demographic data was however available for only 29 individuals. For a subset of days, media usage data was also available for these individuals.

The location data collected as part of MDC included the following:

- GPS coordinates (every 10 sec)
- Mobile phone usage records (time and duration of activity, destination number)
- WLAN data

The longitude and latitudes of the mobile phone towers were however not available and therefore, only the GPS and WLAN data were used to retrace the trajectories of individuals and extract their tour patterns. An example of plotted trajectory in the freeware GPS Visualizer (41) is presented in Figure 1.

![Figure 1: GPS data of a random individual on a random day](image)

It may be noted that in this research, tours are defined as a sequence of trip segments over a day. Initially the goal of this research was to find detailed tour patterns like Home -Work -Home (H-W-H), Home – Work – Shop – Home (H-W-S-H), Home -Work –Home- Work- Home (H-W-H-
W-H) etc. But place labels (e.g. tags of workplace, friend’s place, theaters, hospitals, shops etc.) were not available in the data and because of distortions associated with anonymization of the GPS data (slight shifts from actual coordinates due to privacy issues), it was not possible to identify place labels from GIS database or maps. It was however possible to identify the durations when an individual is at home (based on the frequency, duration and timing of occupancy). For others locations, only the intermediate stop points could be identified (based on the duration of occupancy). The tour patterns were hence classified into 145 types based on number of times home is visited and the number of intermediate stops in between. For example, if the individual was at home at beginning and end of day and made 2 stops in between, his/her tour was labeled as an H-2-H tour, if he/she was at home at beginning, end and middle of day and made 2 stops in between the first interval and 1 stop in the second, his/her tour was labeled as an H-2-H-1-H tour, etc. (Figure 2).

The methodology used for extracting the tour pattern has been summarized in Figure 3. It may be noted that part of the data had many missing observations (e.g. GPS signals lost or malfunctioned). Part of these data could be supplemented by WLAN records. The remaining part of data (which had substantial part of missing values for the day), were discarded.
The provided data was mainly manipulated by MATLAB. Using this software the data were extracted, interpreted, and sorted out. The tour patterns were also identified by this software.

Firstly, the GPS data of an individual was split into day wise GPS data using MATLAB. Due to the huge data size, 40% of the days were randomly selected for further processing. The selected data were used as input in GPS Visualizer to plot the day-wise travel trajectories. From these plots the stop points were identified and their latitude and longitudes were noted. For each individual, the most frequently made stops were counted and ranked. The stop which has been visited most (including weekends) and had the longest duration of stays was assumed as home (H). The timings of these stays were also cross checked and except small exceptions, they were reasonable for home stay. The others places were labeled intermediate stop. The tour patterns of an individual were encoded considering two aspects. First of all the number of stops in a trip which starts and ends at home (H) were counted and recorded. Secondly, the number of times the individual returned home in a day was counted for tour of each day and was recorded as well.
For simplicity it was assumed that the individual always starts its trip from home (H) and end at home (H) whether it is reflected from the processed data or not.

The aggregate level analysis of tour patterns revealed significant variations in tour patterns. For modeling purposes, only the tour patterns with frequency of 40 or higher among all individuals on all days were included as ‘choices’ in the model. Tour patterns with frequency less than 40 but higher than 10 were kept in the data, but assigned to their closest match in the choice-set (e.g. H-3-H-3-H was merged with H-3-H-2-H, H-5-H was merged with H-4-H, etc.). At the end of this process 15 patterns were included in the choice set. Their aggregated choice frequencies are presented in Table 1.

**Table 1: Aggregate choice frequencies**

<table>
<thead>
<tr>
<th>Tour Pattern</th>
<th>Aggregate Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>153</td>
</tr>
<tr>
<td>H-1-H</td>
<td>772</td>
</tr>
<tr>
<td>H-2-H</td>
<td>453</td>
</tr>
<tr>
<td>H-3-H</td>
<td>254</td>
</tr>
<tr>
<td>H-4-H</td>
<td>246</td>
</tr>
<tr>
<td>H-1-H-1-H</td>
<td>235</td>
</tr>
<tr>
<td>H-1-H-2-H</td>
<td>147</td>
</tr>
<tr>
<td>H-1-H-3-H</td>
<td>72</td>
</tr>
<tr>
<td>H-1-H-4-H</td>
<td>44</td>
</tr>
<tr>
<td>H-2-H-1-H</td>
<td>85</td>
</tr>
<tr>
<td>H-2-H-2-H</td>
<td>78</td>
</tr>
<tr>
<td>H-3-H-1-H</td>
<td>129</td>
</tr>
<tr>
<td>H-3-H-2-H</td>
<td>57</td>
</tr>
<tr>
<td>H-1-H-1-H-1-H</td>
<td>235</td>
</tr>
<tr>
<td>H-1-H-1-H-1-H-1-H</td>
<td>46</td>
</tr>
</tbody>
</table>

The tour pattern data extracted and selected in the above mentioned methodology was then used to estimate discrete choice models.

**3. Model**

**3.1 Model Structure**

A discrete choice model structure for panel data (42,43,44) was used to establish the relationship between the chosen tour patterns and the explanatory variables. The probability of individual $n$ to choose tour pattern $j$ (j=H, H-1-H, H-2-H, etc.) depends on the utility of that tour pattern which can be expressed with the following equation:

$$U_{jn} = \alpha + \beta X_n + \gamma A_j + \epsilon_{jn} + \vartheta_n, \vartheta_n \sim (0, \sigma^2)$$  (1)
Where, $U_{jn}$ is the utility of tour pattern $j$ for individual $n$, $X_n$ is the characteristics of the individual $n$, $A_j$ is the attributes associated with tour pattern $j$, $\vartheta_n$ is the individual specific error term representing unobserved individual specific characteristics (e.g. income, household composition, etc.), $\varepsilon_{jn}$ is the random error term.

The candidate characteristics of the individual available in the MDC dataset included demographic information (gender, age and occupation of the individuals), media usage records (whether or not the individual played any media on a certain time), mobile phone usage records (frequency of making calls, durations, destination number, whether or not the other number is in the contact list, etc.), internet usage records (time and duration), social network usage record (time and duration), mobile application usage record (time and duration) and general transportation mode availability.

Due to the anonymization of the data, the actual locations of the trips and hence the trip purposes could not be inferred from the data. Among tour attributes, only the day of the week and the frequency of home visit were available in the data.

The expected effects of the candidate variables on the choice of the tour pattern are presented in the Table 2.

**Table 2: Candidate variables and their expected effect**

<table>
<thead>
<tr>
<th>Attributes and Characteristics</th>
<th>Expected Casual Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>The number of home based trips are expected to be greater for females compared to males. Moreover, females (assuming some of them have children or dependants to take care of) are more likely to conduct majority of their trips at later parts of the day and show decreased propensity of trip chaining (which will involve more time out of home during each sub-tour).</td>
</tr>
<tr>
<td>Age</td>
<td>Young persons are expected to be more mobile and make more number of stops. They are expected to have less attachment to home and less likely to return home in between the tours. Older persons on the other hand are likely to be busier and show increased propensity to trip chain.</td>
</tr>
<tr>
<td>Attributes and Characteristics</td>
<td>Expected Casual Relationship</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Employment type</td>
<td>The probability of conducting no trips on a certain day is very low (especially during the weekday) for people who are employed full-time. They are also likely to show increased propensity to trip chain.</td>
</tr>
<tr>
<td>Media usage</td>
<td>Individuals are less likely to use media if they have a lot of stops in a sub-tour.</td>
</tr>
<tr>
<td>Internet usage</td>
<td>Individuals who use the internet more are more likely to do online shopping, online bill-payment, etc. which can reduce their need to travel (leading to decreased number of stops).</td>
</tr>
<tr>
<td>Usage of social-networking sites (e.g. Facebook, Twitter, Google+)</td>
<td>The correlation between usage of social networking sites and tour type can be positive or negative. Increased usage of social networking sites may be an indication that the individual has more friends and makes more social trips. On the other hand, it can also imply that since they are fulfilling their social interaction needs online, there is a reduced need to travel (leading to decreased number of stops).</td>
</tr>
<tr>
<td>Usage of mobile-applications</td>
<td>The correlation between usage of mobile applications and tour type can also be positive or negative depending on the type of application in use (which is unknown). For example, increased use of applications related to advanced traveler information may be an indication that the individual is more mobile and making complex tours. On the other hand, increased use of e-shopping applications may be an indication that the individual is less mobile and making simple tours.</td>
</tr>
<tr>
<td>Day of the week</td>
<td>In case of weekdays, people are more likely to have increased numbers of sub-tours. The behavior may be different on weekends and Fridays.</td>
</tr>
</tbody>
</table>
3.2. Results

The effects of the candidate variables presented in Table 2 were tested using different model specifications and the parameters were estimated using the software package BIOGEME (42). Different multinomial, nested and mixed model specifications were tested and a multinomial logit (MNL) model with individual specific random coefficients (to take into account the panel nature of the data) was found to be the best in terms of goodness-of-fit and values of estimated parameters. The definitions of the variables having a significant effect are presented in Table 3 and the estimation results are presented in Table 4.

Table 3: Variable Definition with Parameters

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Definition</th>
<th>Parameter in the Discrete Choice Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young female dummy</td>
<td>1 if individual is less than 39 years old and female, 0 otherwise</td>
<td>$\beta_{\text{young-female}}$</td>
</tr>
<tr>
<td>Mature dummy</td>
<td>1 if individual is more than 44 years Old, 0 otherwise</td>
<td>$\beta_{\text{mature}}$</td>
</tr>
<tr>
<td>Media usage dummy</td>
<td>1 if individual uses media, 0 otherwise</td>
<td>$\beta_{\text{media}}$</td>
</tr>
<tr>
<td>Weekend dummy</td>
<td>1 if the associated day is a weekend (Saturday and Sunday), 0 otherwise</td>
<td>$\beta_{\text{weekend}}$</td>
</tr>
<tr>
<td>Friday dummy</td>
<td>1 if the associated day is Friday, 0 otherwise</td>
<td>$\beta_{\text{friday}}$</td>
</tr>
<tr>
<td>Female dummy</td>
<td>1 if individual is female, 0 otherwise</td>
<td>$\beta_{\text{female}}$</td>
</tr>
<tr>
<td>Full time dummy</td>
<td>1 if individual is working or studying fulltime, 0 otherwise</td>
<td>$\beta_{\text{fulltime}}$</td>
</tr>
</tbody>
</table>
Table 4: Estimation results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Affected Alternatives</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
</tr>
<tr>
<td>ASC_ONE</td>
<td>H-1- H,H-2- H</td>
<td>1.20</td>
</tr>
<tr>
<td>ASC_TWO</td>
<td>H-3- H,H-4- H,H-1-H-1- H,H-1-H-1-H-1-H</td>
<td>0.316</td>
</tr>
<tr>
<td>$\beta_{mature}$</td>
<td>H,H-1- H,H-2- H,H-3- H,H-4- H</td>
<td>-0.464</td>
</tr>
<tr>
<td>$\beta_{media}$</td>
<td>H-1- H</td>
<td>0.617</td>
</tr>
<tr>
<td>$\beta_{friday}$</td>
<td>H-1- H</td>
<td>0.473</td>
</tr>
<tr>
<td>$\beta_{full time}$</td>
<td>H</td>
<td>-0.959</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>All</td>
<td>1.16</td>
</tr>
</tbody>
</table>


The value of the coefficient for young-female dummy, $\beta_{young-female}$ denotes that female travelers below 39 years are likely to return home frequently in between tours and have smaller number of stops in each sub-tour which is in agreement with the initial hypothesis.

The value of the coefficient for mature dummy, $\beta_{mature}$ is negative indicating that tours having higher number of home returns is not preferred by mature travelers (individuals older than 44 years).

Individuals are less likely to use media in their smart phones if they have a lot of stops in a sub-tour that is individuals using media are more likely to conduct simplest tours (H-1-H).
The negative value of the coefficient of the weekend dummy, $\beta_{\text{weekend}}$, indicates that in case of weekends, individuals are less likely to return home in-between the daily tour. The positive value of $\beta_{\text{friday}}$ on H-1-H tour states that individuals usually prefer simpler tours on these days.

The negative value of $\beta_{\text{female}}$ on tours with lower number of stops on later part of the day indicates that females prefer higher number of stops on later part of the day.

As hypothesized, individuals who work or study full time were found to have lower probabilities of conducting no trips (tour type H).

The variance of the individual specific error term $\sigma^2$ is highly significant, indicating there are significant correlations among the observations of the same individual.

The effect of internet, mobile application and social-networking site usage were found not to have any significant effect on tour types.

5. Conclusions and Further Research

In this research, the day-to-day tour patterns of travelers have been extracted from GPS and WLAN records and a discrete choice modeling framework has been proposed to predict these patterns. Estimation results indicate that demographic factors (age, gender, employment type), habitual markers (media usage) and tour related attributes (day of the week) have significant explanatory power for predicting the tour pattern of an individual on a certain day.

Though there have been other studies attempting to establish relationships between the socio-economic characteristics and contextual attributes with the choice of tour pattern, the richness of the data, in particular, more reliable data on chosen tour pattern, makes this research unique from other similar studies.

Nevertheless, the model can be enriched by more detailed data. In particular, availability of place labels can help in formulation of more detailed activity patterns. Moreover, availability of the geographic locations of the mobile phone towers can make it possible to use the call detail records as a supplementary data source, especially when the GPS signals are missing. More detailed demographic data and larger number of participants can also enrich the model and improved its prediction capability.

The research results however serve as a proof-of-concept for successful usage of information obtained from ubiquitous data sources (with no additional effort by the respondent other than carrying the smart phone in this case) in enriching travel behavior models.

Acknowledgment

The data of this research was provided by the Nokia Mobile Data Challenge (MDC). The findings of the study are solely of the authors and do not reflect the view of MDC.
References:


30. Schlaich, J., Otterstätter, T., Friedrich, M. (2010), Generating Trajectories from Mobile Phone Data, TRB 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies, Washington, D.C., USA.


