Behavior-based Traveler Classification Using High-Resolution
Connected Vehicles Trajectories and Land Use Data

by

Yu Cui

August 8, 2016

A thesis submitted to the

Faculty of the Graduate School of the

State University of New York at Buffalo

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Department of Civil, Structural and Environmental Engineering

University at Buffalo, SUNY, USA
Acknowledgement

First and foremost, I would like to show my deepest gratitude to my advisor, Dr. Qing He, a respectable, responsible and resourceful professor who has provided me with valuable guidance in every stage of this thesis. Without his enlightening instruction, impressive kindness, and patience, I could not have completed my thesis. His keen and vigorous academic observation enlightens me not only in this thesis but also in my future study.

I shall extend my thanks to Dr. Qian Wang for all her kindness and valuable advices. I would also like to thank all my teachers who have helped me to develop the fundamental and essential academic competence.

I also want to thank Dr. Adel Sadek, who also provided valuable comments to this thesis and taught me fundamental knowledge of travel demand modeling in the past year.

Last but not least, I would like to thank my parents and friends for their encouragement and support.
# Table of Contents

List of Tables ........................................................................................................................................... 3
List of Figures ............................................................................................................................................... 4
Abstract ....................................................................................................................................................... 5
Chapter 1 Introduction ............................................................................................................................... 6
    1.1 Background and Motivation ............................................................................................................ 6
    1.2 Objective ......................................................................................................................................... 8
    1.3 Thesis Organization ....................................................................................................................... 9
Chapter 2 Literature Review .................................................................................................................... 10
    2.1 Travel Behavior Data Analysis and Traveler Classification ......................................................... 10
    2.2 Distribution-based Similarity Measures ....................................................................................... 13
Chapter 3 Methodology ............................................................................................................................ 16
    3.1 DBSCAN ....................................................................................................................................... 16
    3.2 Bhattacharyya Distance ............................................................................................................... 17
    3.3 K-means ....................................................................................................................................... 19
Chapter 4 Data Description and Preliminary Analysis ........................................................................... 21
Chapter 5 Clustering Results .................................................................................................................. 37
Chapter 6 Conclusions and Future Research ......................................................................................... 42
    6.1 Conclusions ................................................................................................................................... 42
    6.2 Future Research ........................................................................................................................... 42
References ................................................................................................................................................... 44
List of Tables

Table 1 Sample records for BSM part 1 data (Booz, 2015) .................................................................................. 22
Table 2 Sample records of BSM part 1 data summary (Booz, 2015) .................................................................. 23
Table 3 Label rules of trip purpose .................................................................................................................. 25
Table 4 Trip purposes ......................................................................................................................................... 29
Table 5 Behavioral features for modeling ......................................................................................................... 33
Table 6 Highly similar probability distribution pairs finding results ................................................................. 36
Table 7 Cluster means of four parameters ......................................................................................................... 41
List of Figures

Figure 1 Framework of this thesis ................................................................. 8
Figure 2 p and q are density-reachable .......................................................... 16
Figure 3 p and q are density-connected .......................................................... 16
Figure 4 BSM data components ........................................................................ 22
Figure 5 The map of land use data with all trip ends in Ann Arbor, MI .................. 27
Figure 6 Daily CV trajectories during a week ....................................................... 28
Figure 7 Multiple one-day CV trajectories ......................................................... 28
Figure 8 Trip purpose comparison between weekdays and weekends .................. 30
Figure 9 ROG probability distributions .............................................................. 31
Figure 10 Departure time probability distributions in weekdays and weekends .......... 32
Figure 11 Visiting duration (for all data) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99) ................................................................. 34
Figure 12 Trip duration (for all data) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99) ........................................................................... 34
Figure 13 Departure time (only on weekdays) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99) ........................................................................... 34
Figure 14 Return time (only on weekdays) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99) ........................................................................... 35
Figure 15 Cumulative Probability distribution---Visit duration .............................. 38
Figure 16 Cumulative probability distribution---Number of visits .......................... 39
Figure 17 Cumulative probability distribution---Trip duration .............................. 39
Figure 18 Cumulative probability distribution---Departure time ............................ 40
Figure 19 Cumulative probability distribution---Return time ............................... 40
Abstract

Recent deployments of Connected Vehicles (CV) open new opportunities to travel behavior research. CV technologies passively generate massive travel data that significantly increase the sample size of travelers. In addition, CV-based data provides high-resolution trip information updated as often as every 0.1 second, which crucially enhances the richness of travel behavior data. By taking advantage of the CV data, this thesis analyzes a 200 GB CV dataset, generated by 2,204 individual drivers during a two-month period in the Safety Pilot testbed at Ann Arbor, MI. This study first extracts each traveler’s home and work locations using DBSCAN algorithm. Then traveler behavior patterns, including number of visits to each location, trip purpose, trip duration, departure and return time are identified by joining the CV data and the land use data. Moreover, probability distributions for these attributes are derived for each traveler. Furthermore, we use Bhattacharyya distance to measure the similarity between two probability distributions. Finally, this thesis develops an augmented K-means algorithm with Bhattacharyya distance to cluster drivers based on the identified similarities of travel patterns. The major clusters identified include commute workers with near working places, workers with far working places, and individuals who do not have full-time jobs or have flexible working hours and working places but traveling for non-work-related activities. This study provides useful inputs for activity-based modeling in travel demand analysis. Also, travelers identified with similar travel behavior are potential candidates for ride-sharing applications in future.

Keywords: Traveler classification; Activity-travel pattern; Travel behavior; Connected Vehicles; Bhattacharyya distance; K-means
Chapter 1 Introduction

1.1 Background and Motivation

Understanding people’s travel behavior is critical in transportation planning and large-scale transportation project investment. Such behavior usually shows up in particular patterns or clusters. Individuals in the same cluster have similar trip information, including departure time, number of visits, trip purposes, trip length, etc. One can group travelers in different clusters according to their behavior patterns. According to the clustering results, transportation planning agencies can better understand the daily travel demand. Also, they can use this information in activity-based modeling which is one kind of travel demand forecasting methods. Further, with traveler classification results, one can develop decision-making tools for determining the investment of large-scale infrastructure projects. Moreover, it will be beneficial to leverage travelers in the same cluster for potential ride-sharing opportunities. With ride sharing, individuals within same travel behavior cluster can share empty seat spots in individual cars to reduce traffic congestion and travel costs. However, we cannot obtain accurate travel classification results without accurate trip data.

Household trip data and travel itinerary data are a major input to travel behavior modeling and activity-based travel demand forecasting. The traditional methods used for collecting these individual travel data are telephone-based or computer-assisted interviews or activity logs recorded from study participants. The typical drawbacks of these methods include high recruitment and survey cost, low response and sampling rates, undersampling or oversampling of certain types of trips, inaccuracies in times, surrogate reporting, and confusions of appropriate trip purposes (Gong et al., 2014).
Nowadays time-location data becomes more accessible with the development of new communication techniques. Travel activities can be traced by various sensors (such as GPS, GSM, Wifi, and Bluetooth) that are commonly available in smartphones or cars. Such data is usually collected by triggering an event such as making a phone call, passing a toll booth, or turning on the Bluetooth devices. However, these data reporting methods are not continuous and thus not accurate enough for obtaining detailed trip information. For example, a mobile phone call detail record (CDR) data entry is generated when a user tries to communicate with a network (Gonzalez et al., 2008). However, the resolution of CDR data could vary from a few hundred meters in urban areas to more than three kilometers in rural areas (Cici et al., 2014), which makes location tracking very difficult. Passively generated CDR data could be massive and contains a massive amount of travelers’ activities. However, due to the low resolution, CDR data cannot provide meaningful trip information, such as accurate departure time (Chen et al., 2016), trip purpose, trip ends and so on. In comparison to CDR, GPS provide high-resolution time-space data. The main advantages of the high-resolution GPS data include near-continuous location tracking, high temporal resolution, and minimum report burden for participants, which may significantly improve the understanding of travel activities in both the spatial and temporal dimensions.

Most recently, the advent of Dedicated Short Range Communications (DSRC) technologies enable vehicles to be connected (USDOT, 2016). Connected Vehicles (CV) proves to greatly improve driving safety and mobility by taking advantages of vehicle to vehicle (V2V), vehicle to infrastructure (V2I) and vehicle to everything (V2X) communication. They also generate high-resolution (e.g., 0.1 second) time-space data with detailed vehicle information, such as steering,
brake, and so on. The new data source provides more opportunities to study travel behavior in much larger spatial, temporal and population scales.

1.2 Objective

The goal of this study is to group travelers based on the similarities of their travel behavior patterns identified by CV data and land use data. To pursue this goal, first, we employ DBSCAN algorithm to identify the home and work locations, and then validate cluster results with GIS land use data. Second, we develop an augmented K-means clustering algorithm with Bhattacharyya distance to cluster travels into different clusters, by taking advantage high-resolution CV data. Travelers within the same clusters could be potentially paired and offered with ride-sharing programs to reduce travel demand and travel costs. Therefore, the findings of this study are very meaningful for the future transportation system. Figure 1 shows the framework of this thesis.

Figure 1 Framework of this thesis
1.3 Thesis Organization

The rest of thesis is structured as follows: Chapter 2 conducts a comprehensive literature review. Chapter 3 describes the methodology. Chapter 4 summarizes the basic statistics of the datasets used for the case study. Chapter 5 demonstrates the numerical results. Moreover, Chapter 6 concludes with the major findings and the future research.
Chapter 2 Literature Review

2.1 Travel Behavior Data Analysis and Traveler Classification

Household trip data are crucial travel behavior data for travel demand forecasting and transportation system planning. The survey-based methods used for trip data collection experienced the stages of paper and pencil interviews (PAPI), computer-assisted telephone interviews (CATI), and computer-assisted-self- interviews (CASI) (Wolf et al., 2001). Although the computer-assisted interviews tried to help respondents to understand questions and recall trips they had during a day, these methods are restricted by the accuracy of recall, reliability, and compliance (Wu et al., 2011).

Recently, GPS and GIS technologies have been used to supplement the traditional survey data. Wolf et al. (2001) used GPS data and GIS land use data to successfully identify 151 out of 156 trips provided by 13 participants. Bohte et al. (2008) also successfully identified travel characteristics of individuals by using GPS and GIS data. Chen et al. (Chen et al., 2010) combined GPS and GIS technologies to develop procedures and models for trip end clustering and trip purpose prediction. However, the accuracy is influenced by dilution of precision of GPS logs and inaccuracy in GIS database(Wolf et al., 2001). Kim et al. (2015) used Future Mobility Survey (FMS) as an activity travel data collection method facilitated by a smartphone application or an interactive web interface. It acquires time-space data by utilizing sensors such as GPS, Wifi and Mobile Communication Systems (GSM, CDMA, and UMTS), and accelerometers. They used ensemble-learning-based classification to recognize travel patterns from the FMS data collected in Singapore. As found, more training days will help improve the overall model classification
accuracy for the seen users whose travel activity histories were used to train models. For the unseen users, the classification performance improves as the training days accumulate more than those used for seen users. Moreover, the classification accuracy for seen users is better than unseen users, which indicates that learning from users’ own histories does help improve the classification performance. Schonfelder et al. (2002) used a multi-stage hierarchical matching procedure to calculate a cluster center of stop ends by combining trip ends, identifying trips with obvious purposes with the socio-demographics of the respondents.

The method of deriving trip purposes based on GPS and GIS data was also incorporated into a machine learning algorithm. Some researcher used decision tree based classifiers to derive trip purposes. Lu et al. (2012) implemented a decision tree approach to identify trip purpose by using GPS-based travel surveys. They found out that trip end locations can affect trip purpose classification most and can improve more than 20% classification accuracy. Moreover, next and previous trip information can improve more than 10% classification accuracy. However, social-demographic of respondents only can improve less than 1% classification accuracy. Researchers also implemented decision tree methods in C4.5, C5.0, or an adaptive boosting environment (Deng and Ji, 2010, Griffin and Huang, 2005). Some of the research above require the social-economic characteristics of respondents (such as age, gender and household income).

Researchers have also been making great efforts to classify travelers by using daily travel data and socio-economic data. The criteria to select similarity measures depend on the analysts’ importance ranking of various affecting attributes and the situations to be dealt with (Jones, 1990).
Consequently, of the resulting similarity measures could be subjective and case sensitive, and thus derive quite inconsistent results. Hanson et al. (1986) divided individuals into five homogeneous travel behavior groups by using complex multi-day travel data and explained variability in individuals’ daily travels. Shoval et al. (2007) implemented a sequence alignment method based on GPS data and clustered the data to three temporal-spatial time geographies. Kitamura and van der Horn (1987) showed that daily participation could be very stable in different types of activities (based on the categories of working, leisure, shopping and other activities). Axhausien et al. (2002) collected six weeks continuous travel diaries from about 300,000 inhabitants in Germany in fall 1999. Hazard models were used to analyze this high-quality data. A low degree of spatial variability of daily participation of activities was also found from the analysis.

The typical similarity indices rely on Euclidean distance or some measures used in signal processing. Signal processing measures have been found vulnerable to process choice data. As for the Euclidean-distance based measurement, instead of using traditional methods that purely rely on conventional Euclidean distance, Joh et al. (2002) employed the alignment method to distinguish different travel patterns. This breakthrough overcomes the problem that the traditional Euclidean distance cannot deal with interdependent attributes. However, it still cannot capture similarities and relations among sequential activities. As an improvement, we developed a Bhattacharyya distance with K-means algorithm to measure similarities among sequential activities.
Travel trajectory clustering becomes more and more prevalent recently. Kim et al. (2015) used Longest Common Subsequence (LCS) to measure the similarity between two trajectories of vehicles. Then they incorporate LCS with DBSCAN to distinct traffic flow clusters. Finally, they employed Cluster Representative Subsequence (CRS) to allocate new trajectories into similar clusters. Besse et al. (2016) used a different way to cluster vehicle trajectories. They developed a new distance to compare trajectories which called Symmetrized Segment-Path Distance (SSPD).

2.2 Distribution-based Similarity Measures

Researchers mainly used tradition partitioning clustering methods such as k-means or KNN to cluster uncertain data according to geometric distances between objects. Such methods cannot handle uncertain objects that are geometrically distinguishable, such as datasets with the same mean but with different variances.

Fortunately, measuring the similarity between different probability distributions of uncertain objects can shed light on identifying similar uncertain objects. Researchers developed lots of methods to assess the similarity between two distributions. Kullback-Leibler divergence is the most prevalent one, which is a measure of the difference between two probability distributions P and Q. By minimizing the Kullback-Leibler divergence, Bunte et al. (2010) dimensionally reduced and visualized images from exploratory observation machine. However, Kullback-Leibler divergence is never a real distance or metric. It does not follow the triangle inequality; it does not have a boundary, and $D_{KL}(P||Q)$ does not equal to $D_{KL}(Q||P)$ in most scenarios. Moreover, it is hard to compare, because $D_{KL}(P||Q) > D_{KL}(R||Q)$ does not mean that P is more similar to Q than
R. The infinitesimal form, specifically its Hessian, gives a metric tensor called Fisher information metric. It is an information metric that can be used to similarities between uncertain objects. Frank (2009) maximized the amount of Fisher information about environment captured by the population and measured how natural selection influences evolutionary dynamics. Fisher information metric also has a relationship with Jensen-Shannon divergence when being used to measure actions and curve lengths. Jensen-Shannon is a prevailing method of determining the similarity between two probability distributions. It is built on the Kullback-Leibler divergence.

However, it is symmetric and always returns a finite value. Grosse et al. (2002) analyzed symbolic sequences using the Jensen-Shannon divergence, and applied this method to DNA sequences. Nevertheless, Jensen-Shannon divergence requires probability distributions following a multinomial distribution type. Wasserstein metric is a distance function defined between probability distributions on a given metric space M. Ni et al. (2009) employed Wasserstein metric to measure the distance between two histograms by comparing many pointwise distances. In addition, it is robust to noisy data, and it is insensitive to oscillations. Similar as Kullback-Leibler divergence, Hellinger distance is likewise a type of f-divergence. It is symmetric and has the boundary as well as Jensen-Shannon. Sengar et al. (2008) developed a VoIP floods method using the Hellinger distance successfully.

Bhattacharyya distance can also measure the similarity of two probability distributions that could be discrete and continuous. It can be used to determine the relative closeness of two samples according to the Bhattacharyya coefficient that measures the amount of overlapping between two
samples. The Bhattacharyya distance is widely used in research on feature extraction and selection (Choi and Lee, 2003, Narendra and Fukunaga, 1977, Xuan et al., 2006), image processing (Goudail et al., 2004), signal selection (Kailath, 1967), speaker recognition (You et al., 2009, Salvi, 2003) and phone clustering (Mak and Barnard, 1996). In this thesis, we leverage the Bhattacharyya distance to measure the distance between two distributions of travel attributes from two travelers.
Chapter 3 Methodology

3.1 DBSCAN

The Density-based spatial clustering of applications with noise (DBSCAN) is a kind of density-based algorithm which can identify clusters of arbitrary shape in large longitudinal data sets by looking at the local density of database elements. We can use this algorithm to identify home and work locations in this thesis. This algorithm only uses one input parameter and can also determine which points should be considered to be outliers or noise. There are two parameters defined in this algorithm, Eps (maximum radius of the neighborhood) and MinPts (minimum number of points in the Eps-neighborhood of a point). This algorithm consists of six definitions and two lemmas (Ester et al., 1996). The most important definitions are density-reachable and density-connected which are illustrated in Figure 2 and Figure 3 below. Moreover, DBSCAN algorithm process is also showed below.

![Diagram of DBSCAN algorithm process](image)

**Figure 2** p and q are density-reachable

**Figure 3** p and q are density-connected

---

**Algorithm 1 DBSCAN**

**Input:** points
Output: clusters

1 for each \( p \in D \) do
2 if \( p \) is not yet classified then
3 if \( p \) is a core point then
4 collect all objects density-reachable from \( p \) and assign them to a new cluster
5 else
6 assign \( p \) to noise
7 end
8 end
9 end

3.2 Bhattacharyya Distance

Bhattacharyya distance is a statistics method which measures the similarity of two discrete or continuous probability distribution. It is closely related to the Bhattacharyya coefficient which can measure the overlap area between two statistical samples or populations. This coefficient indicates the relative similarity of two probability distribution. For two probability distribution \( p \) and \( q \) with same parameter \( X \), the Bhattacharyya distance is defined as follow:

\[
D_B(p, q) = \ln(BC(p, q))
\]  

(1)

where \( BC(p, q) \) is the Bhattacharyya coefficient. Calculating the Bhattacharyya coefficient involves a rudimentary form of integration of the overlap of the two samples. The interval of the values of the two samples is split into a chosen number of partitions, and the number of members of each sample in each partition is used in the following formula,
\[ BC(p, q) = \begin{cases} 
\sum_{i=1}^{N} \sqrt{p(x)q(x)} & \text{for discrete probability distribution} \\
\int \sqrt{p(x)q(x)} dx & \text{for continuous probability distribution} 
\end{cases} \] (2)

Where considering the sample \( p \) and \( q \), \( n \) is the number of partitions, and \( p(i), q(i) \) are the number of members of samples \( p \) and \( q \) in the \( i \)th partition.

From the formula above, partitions with large overlaps will result in larger values. However, choice of the number of partitions can influence accuracy a lot. Too few numbers of partitions will reduce accuracy because overestimating the overlap area, and too many partitions will create individual partitions without any points even though there is in a densely population sample space. This will result in low accuracy also.

The Bhattacharyya distance measure has a simple geometric interpretation as the cosine of the angle between the \( N \)-dimensional vectors \((\sqrt{p(1)}, \sqrt{p(2)}, ..., \sqrt{p(N)})^T\) and \((\sqrt{q(1)}, \sqrt{q(2)}, ..., \sqrt{q(N)})^T\). Thus, if the two distributions are identical, we have:

\[ \cos(\theta) = \sum_{i=1}^{N} \sqrt{p(i)q(i)} = \sum_{i=1}^{N} \sqrt{p(i)p(i)} = \sum_{i=1}^{N} p(i) = 1 \] (3)

And consequently \( \theta = 0 \). Moreover, according to Jensen’s inequality, we can derive the following relationship

\[ 0 \leq BC(p, q) = \sum_{i=1}^{N} \sqrt{p(i)q(i)} \leq \sum_{i=1}^{N} \sqrt{q(i)} = 1 \] (4)
Therefore, $0 \leq \text{BC}(p, q) \leq 1$ and $0 \leq D_B(p, q) \leq \infty$. The Bhattacharyya coefficient will be 0 if there is no overlap at all due to the multiplication by zero in every partition. This means the distance between fully separated samples will not be exposed by this coefficient alone. The Bhattacharyya coefficient will be 1 if these two probability distributions are exactly the same.

### 3.3 K-means

K-means is a prevalent unsupervised learning algorithm that resolves the clustering problem. The key idea is define k centers, one for each cluster. Then minimize the squared error function for points for all the cluster.

$$\arg \min_S \sum_{l=1}^k \sum_{x \in S_l} ||x - \mu_l||^2$$  \hspace{1cm} (5)

Where $\mu_l$ is the centroid or mean points in cluster $S_l$. Time-space data has huge amount of information. If using points as centroids of clusters, we will lost information. Therefore, not like traditional K-means algorithm, centroid of clusters are points. We developed an Augmented K-means algorithm that centroid of clusters are probability distributions. In this thesis, instead Euclidean distance for points, we employed Bhattacharyya coefficient in Bhattacharyya distance to measure distance for probability distributions. Therefore, we rewrite the distance measurement equation as follow,

$$\arg \min_S \sum_{l=1}^k \sum_{x \in S_l} \sum (1 - BC(x, \mu_l))$$  \hspace{1cm} (6)

In order to minimize the object function, we write Bhattacharyya coefficient as $\sqrt{1 - BC(p, 1)}$ which is also called Hellinger distance. The centroids for each cluster will be recalculated after cluster. In traditional K-means algorithm, a centroid is a point. In order to incorporate with
probability distribution, instead of points we use medial distributions. The medial distribution is represented by mean of 15 quantiles, respectively. The process of augmented K-means algorithm is show as follows.

---

**Algorithm 2 Augmented K-means algorithm with Bhattacharyya distance**

**Input:** \( \{x_1, x_2, \ldots, x_n\} c, K \),

**Output:** \( \{\mu_1, \mu_2, \ldots, \mu_k\} \)

1. \((S_1, S_2, \ldots, S_K) \leftarrow \text{Select random seeds} \ ((\{x_1, x_2, \ldots, x_n\}, K) \)

2. **for** k in 1: K

3. \( \text{do } \mu_k \leftarrow S_k \)

4. **while** stopping criterion has not been met

5. **do for** k in 1:K

6. **do** \( \omega_k \leftarrow \{\} \)

7. **for** n in 1:N

8. **do** \( j \leftarrow \arg \min_j \sum (1 - BC(x, \mu_j)) \)

9. \( \omega_j \leftarrow \omega_j \cup \{x_n\} \) (reassignment of vectors)

10. **for** k in 1:K

11. **do** \( \mu_k \leftarrow \frac{\sum_{x \in \omega_k} x_p}{|\omega_k|} \) (recomputation of centroids, p is quantile)

12. **end**

13. **end**

14. **end**

15. **end**
Chapter 4 Data Description and Preliminary Analysis

In this thesis, raw trip data was acquired from a recently released CV dataset on Federal Highway Administration (FHWA) research data exchange (RDE) website (FHWA, 2013). The dataset was generated by safety pilot program, hosted by University of Michigan Transportation Research Institution (UMTRI) in Ann Arbor, MI. Basic safety message (BSM) was extracted from CV dataset. The BSM is a kind of ‘heartbeat’ message which transmits messages frequently (usually at approximately 10HZ, the same frequency of data used in this thesis). A BSM includes two parts; one part is a binary large object (blob) that includes every BSM. It consists of fundamental data elements that describe a vehicle’s position (latitude, longitude, and elevation) and motion (heading, speed and acceleration). The other part of BSM data is optional one which contains an extension of vehicle safety information (path history, path prediction and event flags) and pertains to the status of a vehicle’s components (lights, wipers, and brakes). Figure 4 shows all the component of BSM data. We only need the first part of BSM data in our research and data range is the whole month of April 2013. Table 1 is the sample records of first part of BSM data, and Table 2 is the sample records of a summary of first part of BSM data. The total size of these two datasets is 204 GB, and it consists of 216670 trips that made 2204 participants.
Figure 4 BSM data components

Table 1 Sample records for BSM part 1 data (Booz, 2015)

<table>
<thead>
<tr>
<th>RXDevice</th>
<th>TXDevice</th>
<th>GEvent</th>
<th>TXRandom</th>
<th>MSeqCount</th>
<th>DSecond</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Elevation</th>
<th>Speed</th>
<th>Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638030109</td>
<td>8895</td>
<td>15</td>
<td>15456</td>
<td>42.22497</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638030109</td>
<td>8895</td>
<td>15</td>
<td>15457</td>
<td>42.22497</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638126564</td>
<td>8895</td>
<td>16</td>
<td>15458</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638126564</td>
<td>8895</td>
<td>16</td>
<td>15459</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638126564</td>
<td>8895</td>
<td>16</td>
<td>15460</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638126564</td>
<td>8895</td>
<td>16</td>
<td>15461</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638229642</td>
<td>8895</td>
<td>17</td>
<td>15462</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638229642</td>
<td>8895</td>
<td>17</td>
<td>15463</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
<tr>
<td>13108</td>
<td>1290-0</td>
<td>13108</td>
<td>292773638229642</td>
<td>8895</td>
<td>17</td>
<td>15464</td>
<td>42.22496</td>
<td>-83.6839</td>
<td>223.7</td>
<td>17.1</td>
</tr>
</tbody>
</table>

Table continued.....

<table>
<thead>
<tr>
<th>Ax</th>
<th>Ay</th>
<th>Az</th>
<th>YawRate</th>
<th>PathCount</th>
<th>RadiusOfCurve</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.409999996</td>
<td>20.01</td>
<td>0.02</td>
<td>-1.54</td>
<td>9</td>
<td>-1132.8</td>
<td>80</td>
</tr>
<tr>
<td>0.409999996</td>
<td>20.01</td>
<td>0.02</td>
<td>-1.54</td>
<td>9</td>
<td>-1132.8</td>
<td>80</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-1054.5</td>
<td>83</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-1054.5</td>
<td>83</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-1054.5</td>
<td>83</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-1054.5</td>
<td>83</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-987.1</td>
<td>86</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-987.1</td>
<td>86</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-987.1</td>
<td>86</td>
</tr>
<tr>
<td>0.360000014</td>
<td>20.01</td>
<td>0</td>
<td>-1.59</td>
<td>9</td>
<td>-987.1</td>
<td>86</td>
</tr>
</tbody>
</table>
Table 2 Sample records of BSM part 1 data summary (Booz, 2015)

<table>
<thead>
<tr>
<th>DeviceID</th>
<th>TripID</th>
<th>Epoch Start Time</th>
<th>Start Date</th>
<th>Start Time</th>
<th>Epoch End Time</th>
<th>End Date</th>
<th>End Time</th>
<th>Total Trip Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1115</td>
<td>2912143</td>
<td>1357389975</td>
<td>4/12/2013</td>
<td>0:12:55</td>
<td>135741356</td>
<td>4/12/2013</td>
<td>0:35:56</td>
<td>38.463114</td>
</tr>
</tbody>
</table>

Table continued....

<table>
<thead>
<tr>
<th>Distance Traveled w/ Speed &gt;= 25mph</th>
<th>Trip Duration</th>
<th>Average Speed</th>
<th>Maximum Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.051528</td>
<td>1069</td>
<td>8.589875</td>
<td>20.68</td>
</tr>
<tr>
<td>3.662298</td>
<td>268</td>
<td>14.07644</td>
<td>22.68</td>
</tr>
<tr>
<td>2.864534</td>
<td>291</td>
<td>11.046584</td>
<td>19.540001</td>
</tr>
<tr>
<td>0.550884</td>
<td>435</td>
<td>4.81572</td>
<td>13.44</td>
</tr>
<tr>
<td>4.737656</td>
<td>310</td>
<td>16.312865</td>
<td>24.440001</td>
</tr>
<tr>
<td>3.232056</td>
<td>214</td>
<td>16.62929</td>
<td>27.620001</td>
</tr>
<tr>
<td>5.105162</td>
<td>599</td>
<td>10.109112</td>
<td>21.84</td>
</tr>
<tr>
<td>2.6404</td>
<td>297</td>
<td>9.951024</td>
<td>22.860001</td>
</tr>
<tr>
<td>0.68403</td>
<td>69</td>
<td>11.703855</td>
<td>19.860001</td>
</tr>
<tr>
<td>38.136406</td>
<td>1381</td>
<td>27.851639</td>
<td>36.52</td>
</tr>
</tbody>
</table>

Traditional travel survey data usually contain not only GPS time-space data but also participants’ travel activity information obtained from questionnaires or daily online surveys such as activity location, trip duration, trip purpose, and weekday or weekend. However, in this thesis, the CV data we have only GPS positions in deciseconds and time for each trip. We have to acquire date, weekday or weekend, trip duration and activity location through raw data.

Total 200 GB BSM data was collected in this thesis. BSM is a kind of high-resolution messages generated in 10 HZ (every 0.1 second). The initial data analysis shows that the total number of participants is 2204 and the total number of trips is 216,670. This large dataset makes data preprocessing process more difficult. In order to make use of this large dataset, we divided it into
100 partitions. Further, according to the summary of BSM part 1 data, we find out trip information for each trip of each DeviceID (participant). As the base of each trip, we extracted position information from BSM part 1 dataset with the same DeviceID and within traveling time window (Gentime of BSM part 1 data is between Start Time and End Time of each trip) and assigned the same TripID as the base trip. Since the rule to get Gentime in Table 1 is not same as Epoch Start/End Time in Table 2. Gentime is the number of milliseconds has elapsed since midnight, January 1, 2004 in UTC time with 35 seconds offset (Booz, 2015). Whereas, Epoch Start/End Time is Unix time which is the number of seconds has elapsed since midnight, January 1, 1970. So we used different rules to translate timestamp variables into the same scale. After this stage, we merged BSM part 1 data and summary of BSM part 1 data together and obtained every 0.1 second position for each trip of each participant. Also, trip start and end locations can detect from the big new merged table. We considered the first position of one trip is the trip start location and the last position is the trip end location after sorted trip data by Gentime.

Then we could use machine learning methods to mine trip purpose according to trip end for each trip. It is generally believed that it is not necessary for everyone having a job, but everyone should have a home location. Therefore, this thesis assumes that the location that individuals most frequently reported is home, followed by working place. Hence, we used Density-based spatial clustering of applications with noise (DBSCAN) which is the most prevalent density-base clustering method to identify home and work location. The location with the highest density of trip end was configured as home followed by workplaces. After identifying the home and work location for each participant, we used land use data in geographic information system (GIS) to verify the accuracy for home position clustering (There should be at least one land use parcel with
residential land-use within 500m area of the clustered home position. We loosen the home location area to 500m region in order to avoid dilution of precision of GPS system and driveway parking.

After using DBSCAN algorithm, we found both home and work locations for 667 participants and only home locations for 397 participants. We cannot identify home and work locations for 166 individuals.

Land use purposes are categorized into 9 different groups according to GIS land use data. They are commercial, industrial, mixed use, office, public, recreation, residential, transportation and vacant as showed in different colors in Figure 5. After identifying home and work locations, we labeled and divided trip purposes into 7 categories by using both home and work locations and GIS land use data. Label rules are shown in Table 3. If the trip end point is within more than one 500m land use buffer, we consider the nearest one is the land use or trip purpose.

### Table 3 Label rules of trip purpose

<table>
<thead>
<tr>
<th>Trip Purpose</th>
<th>Label Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>Within 500m of home location</td>
</tr>
<tr>
<td>Work</td>
<td>Within 500m of work location, industrial and office land use parcels</td>
</tr>
<tr>
<td>Visit</td>
<td>Within 500m of Residential parcels but not home location</td>
</tr>
<tr>
<td>Recreation</td>
<td>Within 500m of Recreation land use parcels</td>
</tr>
<tr>
<td>Transit</td>
<td>Within 500m of Transportation land use parcels</td>
</tr>
<tr>
<td>Shopping</td>
<td>Within 500m of Commercial and Mixed Use land use parcels</td>
</tr>
<tr>
<td>Others</td>
<td>Within 500m of Public and Vacant land use parcels</td>
</tr>
</tbody>
</table>
Individual travel trajectories can be utilized to uncover human behavioral characteristic. Figure 5 illustrates all trip ends for all participants during one month. Figure 6 shows daily trajectories reported by a vehicle during a week. As can be seen from the daily trajectory plots, a similar route between home and the workplace was repeatedly used during weekdays while very different trajectories were reported during weekends. Figure 7 shows the hourly trajectories. The departure time is still consistent and return time is almost the same. In this figure, purple lines mean trip purposes are industry, and yellow lines mean home based trips. The green lines at the bottom of both figures are the projection of all trajectories. Levy Flight pattern is a pattern characterized by many short-distance trips connected by long-distance relocations. Short-distance trips are mainly composed of commuting to work or supermarket shopping. On the contrary, long-distance trips mostly comprise rare and infrequent events. There are three widely accepted and used indicators: the trip distance distribution \( p(r) \), the radius of gyration (ROG) \( r_g(t) \) and the number of visited locations \( S(t) \) over time. We use these three indicators to identify different travel behavior. During the working day, some people just have to commute trip between home and work place. However, some people have social activities after work. On the weekend, someone likes to stay at home. Whereas, someone prefers to go out and enjoy their weekend in the city or even have long-distance intercity trips.
Figure 5 The map of land use data with all trip ends in Ann Arbor, MI
Figure 6 Daily CV trajectories during a week

Figure 7 Multiple one-day CV trajectories
Initially, there are 2,204 unique participants extracted from the raw dataset. However, not all of them are valid for the analysis. We created several rules to filter the feasible data.

1. Trip duration should be longer than 3 minutes;
2. Records for each participant should have trip information for at least include 14 days (10 weekdays and 4 weekends);
3. A feasible data entry should have a valid home location at least. Data with no home location would be eliminated.
4. After this, we validated home location clustered by the DBSCAN algorithm with GIS land use data. Participants with correctly clustered home locations remain in the dataset.

With above rules, 962 participants were sorted out for the analysis. For these valid data points, we then labeled trip purposes by using GIS land use data. Table 4 shows trip purpose probabilities for all trips made by all participants on weekdays and weekends, respectively. Moreover, Figure 8 verifies the general knowledge that there are more work based travels during weekdays. On the contrary, people have more recreation, visit (friends, families) and shopping related trips on weekends.

Table 4 Trip purposes

<table>
<thead>
<tr>
<th>Location ID</th>
<th>Trip Purpose</th>
<th>Weekday Number of Trips</th>
<th>Probability</th>
<th>Weekday Number of Trips</th>
<th>Weekend Number of Trips</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Home</td>
<td>32,833</td>
<td>43.06%</td>
<td>8,196</td>
<td>39.52%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Work</td>
<td>14,445</td>
<td>18.94%</td>
<td>1,979</td>
<td>9.54%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Visit</td>
<td>9,285</td>
<td>12.18%</td>
<td>3,080</td>
<td>14.85%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Recreation</td>
<td>8,109</td>
<td>10.63%</td>
<td>4,416</td>
<td>21.30%</td>
<td></td>
</tr>
</tbody>
</table>
The radius of gyration (ROG) in transportation is a measure to describe the activity territory for each participant. According to Kang (2012), we can calculate ROG for each participant’s trajectory up to time $t$ by using following formula:

$$r_g^\alpha(t) = \sqrt{\frac{1}{n_c^\alpha(t)} \sum_{i=1}^{n_c^\alpha(t)} (x_i - x_c)^2 + (y_i - y_c)^2}$$  \hspace{1cm} (7)$$

Where coordinates $x_i, y_i$ denote the $i$th ($i = 1, 2, ..., n_c^\alpha(t)$) position recorded for user $\alpha$. And $x_c, y_c$ represent the center of mass of trajectory.

**Figure 8 Trip purpose comparison between weekdays and weekends**
In Figure 9, the blue dashed line is the ROG calculated when the geometric center of all trip ends is considered the center of a trajectory, and the red line is the ROG when the location of home is treated as the center of trajectory. We found that these two ROG distributions are almost the same. Because ROG mostly depends on the frequent location, and home-based trips are the most frequent trips in the daily travel, blue dashed line and red line are almost overlapped with each other. The peak value for this ROG distribution is approximately 8 kilometers. The right skewness of ROG distribution with trip reached around 40 kilometers, or less indicates that a large portion of participants have their trips concentrated within a small activity territory, whereas only rare individuals travel a longer distance in a daily base. We also found more information from this figure that there is a significant effect of distance decay. Moreover, based on probability distribution above, more than 95% of the participants have ROG values less than 40 kilometers, and most of them are around 8 kilometers.
Given the home and work location identified by using DBSCAN algorithm, we found that the average time for individual spending at home is 8.8 hours per day and 8.5 hours during the weekday and 9.9 hours at weekends, respectively. Also, the average time for participants spending in working place is 8.8 hours per day during the weekday and 3.1 hours per day at the weekend.

![Figure 10 Departure time probability distributions in weekdays and weekends](image)

**Figure 10 Departure time probability distributions in weekdays and weekends**

From Figure 10 above we can learn that it is a significant departure time during the working day in the morning. Whereas, the distribution of departure time at the weekend is more flat which means the probability of travel at the weekend is more evenly than that on weekdays. The pattern of return time is in a similar way.
Data pre-processing generates many usable features to describe participants’ daily drive characteristics. People always follow the same pattern to conduct a one day trip diary. For example, David leaves home at 8 am and drove 10 minutes to work. After 9 hours, he returns and arrives home from the workplace at 5 pm. As shown in Table 5, we can extract five types of important behavioral features, including visit duration, number of visits, trip duration, departure time (for both weekday and weekend) and return time (for both weekday and weekend).

### Table 5 Behavioral features for modeling

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit duration</td>
<td>The duration staying at an activity location</td>
</tr>
<tr>
<td>Number of visits</td>
<td>Number of times to visit a location</td>
</tr>
<tr>
<td>Trip duration</td>
<td>The time spent in traveling to an activity location</td>
</tr>
<tr>
<td>Departure time</td>
<td>The time in hour when the first trip starts for a day (for both weekday and weekend)</td>
</tr>
<tr>
<td>Return time</td>
<td>The time in hour when the last trip ends for a day (for both weekday and weekend)</td>
</tr>
</tbody>
</table>

Then we use the Bhattacharyya distance and coefficient to calculate similarities between attributes such as visit duration, number of visits, trip duration, departure time (as a whole, weekday and weekend separately) and return time (as a whole, weekday and weekend separately).
Figure 11 Visiting duration (for all data) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99)

Figure 12 Trip duration (for all data) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99)
Figure 13 Departure time (only on weekdays) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99)

Figure 14 Return time (only on weekdays) probability distributions of two travelers with high similarity (Bhattacharyya distance=0.99)
Figure 11, 12, 13, and 14 show high similarities of travel behavior between two travelers. For departure time and return time, we only measured the similarity during weekdays. Consistent with what we discussed before, departure times and return times are more consistent during weekdays while very different on weekends, which makes it hard to capture similarity.

Table 6 Highly similar probability distribution pairs finding results

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Number of high similar (B-distance &gt; 0.9) probability distribution pairs</th>
<th>Number of Unique Drivers Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit duration</td>
<td>981</td>
<td>427 (44.39%)*</td>
</tr>
<tr>
<td>Trip duration</td>
<td>4,275</td>
<td>697 (72.45%)</td>
</tr>
<tr>
<td>Departure time</td>
<td>75,626</td>
<td>920 (95.63%)</td>
</tr>
<tr>
<td>Return time</td>
<td>16,1627</td>
<td>940 (97.71%)</td>
</tr>
<tr>
<td>Number of visits</td>
<td>176,736</td>
<td>922 (95.84%)</td>
</tr>
<tr>
<td>Total</td>
<td>162</td>
<td>121 (12.58%)</td>
</tr>
</tbody>
</table>

*: Percentage of drivers involved in the entire driver population

There are 962 participants who remain in the dataset after data preprocessing. Table 7 is the results of highly similar probability distribution pairs. From Table 6, we found out that 121 participants can find others who have high similar (with the threshold of Bhattacharyya-distance equal to 0.90) probability distributions satisfying all five features. The probability to find individuals who are similar from almost all aspects is approximate 13% which is very high.
Chapter 5 Clustering Results

In order to cluster individuals with high similarities, we develop an augmented K-means algorithm to cluster travelers according to visit duration, number of visits, trip duration, departure time, and return time. We try a different number of clusters and found out that 3 is the best number of clusters. This makes all features are most homogenous within each cluster and heterogenic among all the clusters.

Figure 15 shows a cumulative probability distribution for visit duration. In this figure, cluster 2 has a longer visit duration than other two clusters. Figure 16 is a cumulative probability distribution for number of visits. This figure contains lots of information. First, cluster 1 has more visit location types than other two clusters, which mean people in cluster 1 may have more activities other than home and work. Moreover, from lines of cluster 2 and cluster 3 we find out that home and work based trips can cover approximate 70% of total trips. This result verifies that the finding in Section 2 that most individuals have low degrees of spatial variability. From Figure 17, we can get that cluster 2 has the longest trip duration, and then follows cluster 3. Cluster 1 has the shortest trip duration. Figure 18 presents a cumulative probability distribution for departure time. We can see that cluster 1 always depart late and most departure time even at afternoon. Cluster 2 and cluster 3 almost depart at the same time, but people cluster 2 depart little earlier than individuals cluster 3. Figure 19 depicts the cumulative probability distribution for return time. Return time for cluster 2 and cluster 3 is similar. However, return time for individuals in cluster 1 is later than other 2 clusters.
Figure 15 Cumulative Probability distribution--- Visit duration
Figure 16 Cumulative probability distribution---Number of visits

Figure 17 Cumulative probability distribution---Trip duration
Figure 18 Cumulative probability distribution---Departure time

Figure 19 Cumulative probability distribution---Return time
Table 7 Cluster means of four parameters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Visiting duration (hr)</th>
<th>Trip duration (min)</th>
<th>Number of visiting location type</th>
<th>Departure time</th>
<th>Return time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.73</td>
<td>12.83</td>
<td>6</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>9.65</td>
<td>85.78</td>
<td>4</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>6.70</td>
<td>31.57</td>
<td>4</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

From Figure 15, 16, 17, 18, 19 and Table 7, we can figure out differences for each cluster and discover some useful information. According to departure time, return time and location visit times, individuals in cluster 1 may not have jobs or have flexible working hours and working places but have more other activities. Figure 16 indicates they visit more types of locations than other two clusters. And the departure time for their first trip is late, and they even depart at afternoon. Moreover, they return later than people in other clusters. Individuals in cluster 2 have full-time jobs. Their departure time from home was at 8 am, and they stayed in working place until returned home at 4 pm. Cluster 3 is also a typical individuals’ cluster. In this cluster, people have jobs and commuted regularly. However, different with individuals in cluster 2 who have far working places, travelers in cluster 3 have near working locations. One can verify this information in Figure 18, since the red line is at right-hand side of blue one. Due to near working locations, individuals in cluster 3 can depart later than people in cluster 2.
Chapter 6 Conclusions and Future Research

6.1 Conclusions
In this thesis, we use the time-space trajectory data collected by Connected Vehicles (CV) and land use data to analyze travel behavior. We find some interesting observations after processing the raw data and using the DBSCAN clustering method to identify home and work locations. First, the results confirm some common sense of travel behavior. Individual travelers make more commute trips to work during weekdays but more recreation, shopping and visiting trips on weekends. Second, most participants’ travel territory is around eight kilometers, and 95% participants travel within 40 kilometers for each trip. Third, most individuals have a low degree of spatial variability, indicating that home and work based trips can cover about 70% of total trips. Fourth, we employ Bhattacharyya distance and coefficient to find similar travel trajectories by using travel attributes such as visit duration, number of visits, trip duration, departure time and return time. Based on the analysis, we identify travelers who have similar travel patterns and thus may be paired in ride-sharing applications. Finally, we develop an augmented k-mean algorithm to classify drivers into three clusters, such as represent commute workers with near working places, workers with far working places and individuals who do not have jobs or have flexible working hours and working places but travel for other activates, respectively.

6.2 Future Research
In future, travel information and social-demographics data will be taken into account to enhance the result and improve accuracy. Moreover, some existing drawbacks of time-space data, for example, buildings or heavy foliage can block satellite signals, losing signal when going through
tunnels and dilution of precision of GPS data may be overcome in near future with the development of GPS and the Connected Vehicles technology. Therefore, we can acquire more passive, accurate time-space data. Furthermore, the analysis of similar driver behavior can go one step further. We can cluster similar drivers by their daily travel trajectories that are a crucial part in ride-sharing research. Moreover, we also can incorporate this information in the activity-based model to conduct accurate travel demand forecasting.


BOHTE, W. & MAAT, K. Deriving and Validating Trip Destinations and Modes for Multiday GPS-Based Travel Surveys: Application in the Netherlands. Transportation research board 87th annual meeting, 2008.


