Travel Behavior Classification: An Approach with Social Network and Deep Learning

Article *in* Transportation Research Record Journal of the Transportation Research Board · June 2018

DOI: 10.1177/0361198118772723

Yu Cui
University at Buffalo, The State University of New York
9 PUBLICATIONS 79 CITATIONS
SEE PROFILE

Qing He
Southwest Jiaotong University
70 PUBLICATIONS 1,456 CITATIONS
SEE PROFILE

Alireza Khani
University of Minnesota Twin Cities
26 PUBLICATIONS 399 CITATIONS
SEE PROFILE

Article

TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Transportation Research Record
1–13
© National Academy of Sciences:
Transportation Research Board 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0361198118772723
journals.sagepub.com/home/trr
SAGE

Travel Behavior Classification: An Approach with Social Network and Deep Learning

Yu Cui[1], Qing He[2], and Alireza Khani[3]

Abstract

Uncovering human travel behavior is crucial for not only travel demand analysis but also ride-sharing opportunities. To group similar travelers, this paper develops a deep-learning-based approach to classify travelers' behaviors given their trip characteristics, including time of day and day of week for trips, travel modes, previous trip purposes, personal demographics, and nearby place categories of trip ends. This study first examines the dataset of California Household Travel Survey (CHTS) between the years 2012 and 2013. After preprocessing and exploring the raw data, an activity matrix is constructed for each participant. The Jaccard similarity coefficient is employed to calculate matrix similarities between each pair of individuals. Moreover, given matrix similarity measures, a community social network is constructed for all participants. A community detection algorithm is further implemented to cluster travelers with similar travel behavior into the same groups. There are five clusters detected: non-working people with more shopping activities, non-working people with more recreation activities, normal commute working people, shorter working duration people, later working time people, and individuals needing to attend school. An image of activity map is built from each participant's activity matrix. Finally, a deep learning approach with convolutional neural network is employed to classify travelers into corresponding groups according to their activity maps. The accuracy of classification reaches up to 97%. The proposed approach offers a new perspective for travel behavior analysis and traveler classification.

As car ownership increases rapidly, together with increasing environmental concerns, there has been more interest in services that enable people to share their automobiles. Ride-sharing or carpooling services are the ultimate way to make better use of the empty seats in personal passenger cars in order to reduce fuel consumption, transportation cost, and emissions. Cici et al. found that if individuals were willing to carpool with others who live and work within 1 km, the traffic in the city of Madrid would decrease by 59% (*1*). There are many existing smartphone apps providing ride-sharing services. However, this kind of one-time sharing mode offers only limited benefits. Long-time ride-sharing services can provide more advantages. There are many kinds of travelers, including ones who depart at the same time of day, ones who do not have regular departure times, ones who start work later than others, and so forth. The assumption is that travelers with similar behaviors may enjoy their shared rides better given more common behaviors. Long-term ride sharing matched according to travel behaviors will provide steady benefits for transportation, environment, and society.

Household travel survey data, or travel itinerary data, is a major input to travel behavior modeling. It can be used in many areas, particularly in travel behavior research and activity-based travel demand forecasting. The traditional methods used for collecting these individual travel data are telephone-based or computer-assisted interviews and activity logs recorded from study participants. The typical drawbacks of these methods include high recruitment cost, low response and sampling rates, undersampling or oversampling on certain types of trips, inaccuracies in times, surrogate reporting and confusion of appropriate trip purpose (*2*). Nowadays time–location

[1] Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY
[2] Department of Industrial and Systems Engineering and Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY
[3] Department of Civil, Environmental, and Geo-Engineering, University of Minnesota, Minneapolis, MN

Corresponding Author:
Address correspondence to Qing He: qinghe@buffalo.edu

data becomes accessible with the development of new techniques. Travel activities can be traced by various sensors such as GPS, GSM, Wi-Fi, RFID, and Bluetooth that are commonly available in smartphones or cars. Such data are usually collected when an event is triggered such as making a phone call, passing a toll booth, or turning on Bluetooth devices. Conducting a household travel survey with GPS devices is a complementary way of collecting reliable and accurate data. GPS provides 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 spatial and temporal dimensions.

The essential goal of this study is to classify travelers based on the characteristics of their historical travel data. Therefore, travelers within the same category could be potentially paired and recommended with ride-sharing services. To pursue this goal, this paper constructs a social network of travelers based on the Jaccard similarity coefficient, and employs a community detection algorithm to cluster travelers into groups. In this paper, the social network concept is borrowed to describe people who have similar behavioral patterns rather than construct a real social network. After detecting travelers' groups, labels are manually assigned to each group according to trip and activity information. Further, an image is built of the activity map for each traveler and a deep-learning-based approach is employed to perform image classification. Therefore, the travelers are classified accordingly into different groups depending on their activity maps.

The rest of this paper is structured as follows: the literature review summarizes previous studies in travel behavior analysis, community detection and deep learning. Next the datasets and initial result of the data analysis are introduced. Then the methodology is presented. The results are demonstrated with some numerical examples. Finally, the conclusions and future research topics are presented.

Literature Review

Travel Behavior Data Analysis and Classification

Household trip data is crucial for travel demand forecasting and transportation system planning. The survey-based methods used for trip data collection went through the stages of paper and pencil interviews (PAPI), computer-assisted telephone interviews (CATI), and computer-assisted-self- interviews (CASI) (*3*). 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 (*4*). Recently, GPS and GIS technologies have been used to supplement the traditional survey data. GPS and GIS land use data can be used for trip identification, travel characteristics identification, trip end clustering, and trip purpose prediction (*3*, *5*, *6*). However, the accuracy is influenced by the dilution of precision of the GPS logs and inaccuracy in the GIS database (*3*). Kim et al. developed an activity travel data collection method facilitated by a smartphone application and an interactive web interface. The data collected by this method in Singapore was further implemented with an ensemble-learning-based classification method to recognize travel patterns (*7*). Schumpeter et al. devised a multi-stage hierarchical matching procedure to calculate a cluster center of stop ends by combining trip ends and identifying trips with obvious purposes with the sociodemographics of the respondents (*8*). Some researchers also used decision-tree-based classifiers to derive trip purposes and implemented the methods in C4.5, C5.0, or an adaptive boosting environment (*9*, *10*). Some of the previous studies above also require the socioeconomic characteristics of respondents (such as age, gender, and household income) for travel behavior analysis.

Researchers have also been making great efforts to classify travelers by using daily travel data and sociodemographic data. The criteria to select similarity measures depend on the analysts' importance ranking of various affecting attributes and the situations to be dealt with (*11*). Consequently, the resulting similarity measures could be subjective and case sensitive and thus derive quite inconsistent results. Hanson et al. divided individuals into five homogeneous travel behavior groups by using complex multi-day travel data and explained variability in individuals' daily travels (*12*). Shoval et al. implemented a sequence alignment method based on GPS data and clustered the data into three temporal–spatial time geographies (*13*). Kitamura and van der Horn showed that daily participation could be very stable in different types of activities (based on the categories of working, leisure, shopping and other activities) (*14*). Axhausen et al. collected six weeks' continuous travel diaries from about 300,000 inhabitants in Germany in Fall 1999 (*15*). Hazard models were used to analyze this high-quality data. A low degree of spatial variability of daily activities was also found from the analysis. Jiang et al. employed the K-Means algorithm via principal component analysis (PCA) to cluster daily patterns of human activities in Chicago (*16*). They separated more than 3,000 individuals who participated in a 1-day or 2-day survey conducted by "Travel Tracker Survey" from January 2007 to February 2008 into 8 groups on weekdays and 7 groups on weekends, respectively. The same methodology applied to kernel density estimation, allowed them to analyze and explore diverse urban spatial–temporal structures. This research indicated how individuals in different activity pattern clusters make use of different sub-regions for different activity types (*17*). Travel behavior classification not only differentiates individuals with different travel patterns, but also uncovers human mobility patterns. Gonzalez et al. found that travel trajectories show lévy flight or random walk pattern to a large extent. Individuals show a high probability of returning to a few highly frequented locations (*18*). This means humans are following highly predictable mobility patterns (*19*). Ponieman et al. also retrieved social phenomena information (e.g. commute and major sport events) from call detail records (CDR) data by just counting how many phone calls were made by users in two different time windows (from 21:00 to 05:00, and from noon to 16:00 during weekdays) from inside and outside Buenos Aires city. They found the average radius of commute (ROC) was approximately 7.8 km (*19*). Nevertheless, that study was limited by the accuracy of CDR data. With the combination of several machine learning algorithms, Ma et al. identified travel patterns for transit riders from smart transit card data in order to attract more users, retain loyal users, and finally improve overall transit services performance (*20*). Williams et al. developed a new method derived from the neural coding concept of synchrony and measured regularity of visiting a specific location for individuals (*21*). However, this research only explored three places. Investigating how location types will influence visit patterns of individuals will provide more information to understand people travel patterns.

Community Detection

Nowadays, lots of complex systems can be represented by networks. For instance, each user can be a node in the social network. Then the friendship can be represented by edges. Researchers aim to understand the network by finding community structures. A community is a collection of nodes that are homogeneous within the group and heterogeneous with other groups in the network, and this kind of network is known as a community structure. Newman and Girvan (*22*) employed centrality indices to find boundaries of communities. They tested this method on two networks which are collaboration networks and food web networks. Both cases retrieved significant and informative community segments. The same authors also developed a community detection algorithm and proposed a new community structure strength measurement (*23*). Their algorithm showed highly effective performance for both computer-generated and real-world networks while detecting communities. Radicchi et al. (*24*) developed a fully self-contained new local algorithm for community detection and tested it on both artificial and real-world network graphs. This new method demonstrated the potentials of implementing community detection algorithms in large-scale technological and biological applications. In a community structure, groups do not need to be necessarily mutually exclusive; they can overlap. Palla et al. (*25*) proposed an algorithm which can uncover the overlapping community structure of complex networks in nature and society. Community detection techniques can also be used for habitat preservation, animal genetics and wildlife corridors (*26*). Moreover, it can be implemented in the transportation field. For example, Lin et al. employed a community detection algorithm to study vehicle accident causative factors (*27*).

Compared with traditional clustering algorithms, the community detection algorithm provides several advantages. First, it is easy to implement, and steps are intuitive. Second, final networks can be decomposed into communities for different levels. Third, this algorithm runs fast even for large and high dimensional datasets.

Deep Learning

With technological innovation, artificial intelligence (AI) emerges and integrates into everyday life rapidly. From education to finance, from marketing to health care, and from communication to transportation, with the advent of AI, individuals can save plenty of time, reduce mistakes, relieve pressures, and stay safe (*28*).

Deep learning or deep neural network, as a branch of machine learning and AI, is an artificial neural network (ANN) that contains more than one hidden layer. This kind of algorithm has shown superior performance especially in automatic speech recognition, image recognition, natural language processing, and recommendation systems. Readers can refer to Schmidhuber et al. (*29*) for more details of deep learning. Deep neural networks typically can be categorized into two types, recurrent neural networks (RNN) and convolutional neural networks (CNN). Long short-term memory RNNs is a widely used algorithm in the field of speech recognition (*30*, *31*). However, in the field of image recognition, the CNN is the most prevailing algorithm. Krizhevshy et al. developed a CNN consisting of five convolutional layers, some of these followed by max-pooling layers and three fully-connected layers. This network achieved an error rate of 15.3% while implementing the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-2012 dataset (*32*). Simonyan and Zisserman improved CNNs with utilizing very small convolution filters and reached an error rate of 6.8% on the ILSVRC-2014 dataset. He et al. won the ILSVRC-2015 competition with 3.57% error using deep residual learning. Further, the winning team Trips-Soushen of ILSVRC-2016 produced a 2.99% error rate (*33*, *34*). This technique also has been applied in transportation, especially in the visual sensor data (images, videos) processing domain. Only a few studies implemented deep learning in travel behavior. Dong et al. employed deep learning to model driving behavior based on GPS data. They combined CNN and RNN to extract features to represent driver behaviors. A driver classification task was also conducted and they achieved significant outstanding performance compared with traditional machine learning algorithms (*35*).

Data Description and Preliminary Analysis

In this paper, the raw data is acquired from the California Household Travel Survey (CHTS) conducted by the California Department of Transportation (Caltrans) from February 2012 to January 2013. The CHTS is designed to collect household travel information across all 58 counties of California and three adjacent counties in Nevada by using computer-assisted telephone interviewing (CATI), website, and GPS devices. The entire household survey uses three types of GPS devices, wearable GPS device, in-vehicle GPS device, and in-vehicle GPS device plus an on-board diagnostic (OBD) unit. A total of 108,778 individuals belonging to 42,431 households participated in this survey, and 10,474 respondents from 5,460 households carried GPS devices.

Households who conducted the survey with GPS devices generated both survey and GPS data. The survey data includes activities they completed on the assigned travel date only. Each participant has one assigned travel date; this indicates that every participant only has one day of survey data for both non-GPS households and GPS households. However, besides survey data, GPS households also collect 7 days of GPS data as well.

There are 39 different trip purposes included in this household travel survey. These trip purposes are categorized into eight groups, which are home, work, school, transportation or transitions (transit), shopping or errands (shop), personal business (person), recreation or entertainment (rec), and other as shown in Table 1. Table 2 verifies the common knowledge that there are more work and school activities during weekdays, while people have more recreation, shopping, and personal business activities on weekends. It is found that the average time individuals spend at home is 13.98 hours per day during weekdays and 16.77 hours on weekends, respectively. The average time participants spend in the workplace is 7.11 hours during the weekday and 5.51 hours on the weekend. Also, the average time that participants spend at school is 5.18 hour during weekdays and 2.92 hours on weekends.

1. This study only utilizes data of participants who have both valid survey and GPS data. Several rules are created as follows in order to screen the feasible data. Both the first and last activity location should be 'home'.
2. Participants need to have more than one activity on the assigned travel date. If individuals stay at home for a whole assigned travel day, travel information cannot be retrieved. This indicates that participants should also make more than one trip on the assigned travel date.
3. From the household survey data, it is noticed that there might be more than one driver utilizing the same vehicles within a household. It is difficult to distinguish travelers who share the same car since GPS trips are recorded at the vehicle level, not the person level. To avoid sampling errors arising from multiple drivers when they were sharing the same vehicle, the records of the vehicles with multiple drivers are removed.

After these rules are applied, 8,849 unique individuals remain with 50,103 trips in this research. The radius of gyration (ROG) in transportation is a measure to describe the activity territory for each participant. According to Kang (*36*), the ROG for each participant's trajectory up to time t can be calculated by using the formula

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

Where coordinates x_i, y_i denote the ith $(i = 1, 2, \ldots, n_c^{\alpha}(t))$ position recorded for user α, and x_c and y_c represent the center of the mass of trajectories.

In Figure 1, the blue line is the ROG calculated for weekdays, and the red dashed line is the ROG for weekends. It can be seen that these two ROG distributions are almost the same. Therefore, individuals follow similar travel habits whether weekday or weekend. They seldom travel farther on weekends. Figure 1 also shows that there is a significant effect of distance decay. Moreover, based on the probability distribution of ROG, more than 95% of the participants have ROG values of less than 10 miles, and the mode is around 5 miles. This result is similar to the ROC calculated in Ponieman et al. (*19*) which is 7.8 km (4.85 miles).

Figure 2*a* and *b* depict the distribution of the first trip departure time and the last trip return time, respectively. As can be seen, there is a significant pattern of departure time in the morning, and return time in the evening during weekdays. For weekends, individuals depart in the morning and return home in the evening later than weekdays. Moreover, the distribution of departure times on

Table 1. Taxonomy of Activity Purposes

Activity purpose category	Activity purpose
Home	Personal activities (sleeping, personal care, leisure, chores) Preparing meals or eating Hosting visitors or entertaining guests Exercise (with or without equipment) or playing sports Study or schoolwork Work for pay at home using telecommunications equipment Using computer, telephone, cell, smartphone, or other communications device for personal activities All other activities at my home
Work	Work or job duties Training Meals at work Work-sponsored social activities (holiday or birthday celebrations, etc.) All other work-related activities at my work Work-related (meeting, sales call, delivery)
School	In school, classroom, or laboratory Meals at school or college After school, or non-class-related sports or physical activity All other after school or non-class related activities (library, band rehearsal, clubs, etc.)
Transportation or Transitions (Transit)	Change type of transportation or transfer (walk to bus, walk to or from parked car) Pickup or drop off passenger(s) Loop trip (for interviewer only-not listed on diary)
Shopping or Errands (Shop)	Routine shopping (groceries, clothing, convenience store, household maintenance) Shopping for major purchases or specialty items (appliance, electronics, new vehicle, major household repairs) Household errands (bank, dry cleaning, etc.)
Personal Business (Personal)	Volunteer work or activities Drive through other (ATM, bank) Service private vehicles (gas, oil, lube, repairs) Personal business (visit government office, attorney, accountant) Healthcare (doctor, dentist, eye care, chiropractic)
Recreation or Entertainment (Rec)	Non-work-related activities (social clubs, etc.) Exercise or sports Drive through meals (snacks, coffee, etc.) Eat meal at restaurant or diner Outdoor exercise (playing sports, jogging, bicycling, walking, waking the dog, etc.) Indoor exercise (gym, yoga, etc.) Entertainment (movies, watch sports, etc.) Social, or visit friends or relatives
Other	Other (specify) Note: listed on diary Don't know or refused

Table 2. Activity Purpose Statistics

	Weekday		Weekend	
Activity purpose	Number of activity	Proportion	Number of activity	Proportion
Home	199554	54.33%	85920	57.16%
Work	35424	9.64%	4051	2.70%
School	12410	3.38%	381	0.25%
Transit	42142	11.47%	11258	7.49%
Shop	24774	6.75%	14998	9.98%
Person	15299	4.17%	8440	5.62%
Rec	34766	9.47%	24036	15.99%
Other	2911	0.79%	1222	0.81%
Total	367280	100.00%	150306	100.00%

the weekend is flatter which means the departure times on the weekend involve more uncertainty than those on weekdays. The pattern of return time is similar. The mean first departure times on weekdays and weekends are 08:56 and 10:54, respectively. However, the last trip return times are 18:28 and 17:55 for weekdays and weekends, respectively. As shown in Figure 2*b*, there is a small hump near noon. This hump is more flat and larger on weekends than weekdays. Individuals have more flexible times on weekends, and some of them may have eating, recreation, or shopping activities and return home at noon then stay home until the next day. On the contrary, individuals' departure and return times are restricted by work or study on weekdays.

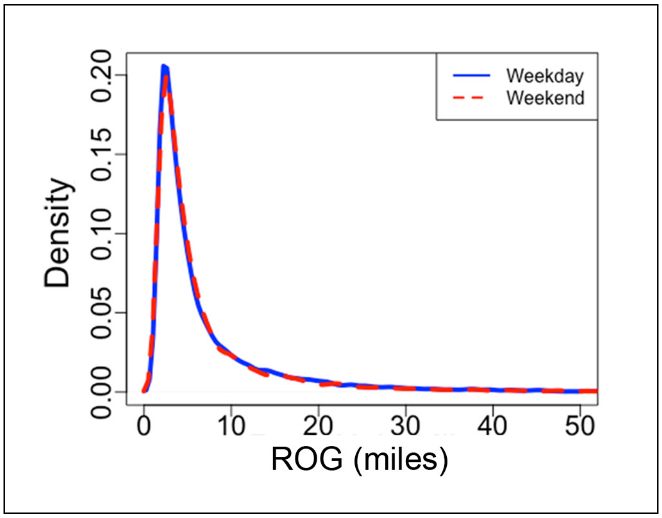


Figure 1. ROG plot.

Methodology

Matrix Similarities

In this paper, individuals' daily activities are represented using a matrix. In order to construct this matrix, all activity purposes are categorized into eight groups and 24 hours are divided into 288 five-minute bins. Therefore, the dimension of the matrix is 8 × 288. Each activity starts from the trip starting time for this activity and ends when the activity ends. For instance, an individual departs from home at 17:30 for the supermarket and arrives at the supermarket at 17:40. The individual finishes shopping at 18:20 and leaves the supermarket. Therefore, this shopping activity is from 17:30 to 18:20.

Individuals only can conduct one of eight activities shown in Table 3 at one time. So,

$$x_{a,t} = \begin{cases} 1 & \text{when activity a occurs at time interval } t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where a represents activity and $a \in \{\text{Home, Work, School, Trans, Shop, Person, Rec, Other}\}$, . t represents time bin number, and $t \in [1, 288]$. In this fashion, a binary matrix is generated to represent daily activities. The two plots in Figure 3 are examples of color-coded daily activity matrices for two individuals. Figure 3*a* is a daily activity map of a student. This participant departs from home to school at approximately 07:30 and stays there until 14:30. After going back home at 18:30, the student again attends a recreation activity, and returns home at 20:30. Figure 3*b* describes a full time employee's weekday activities. The employee departs from home at 06:00,

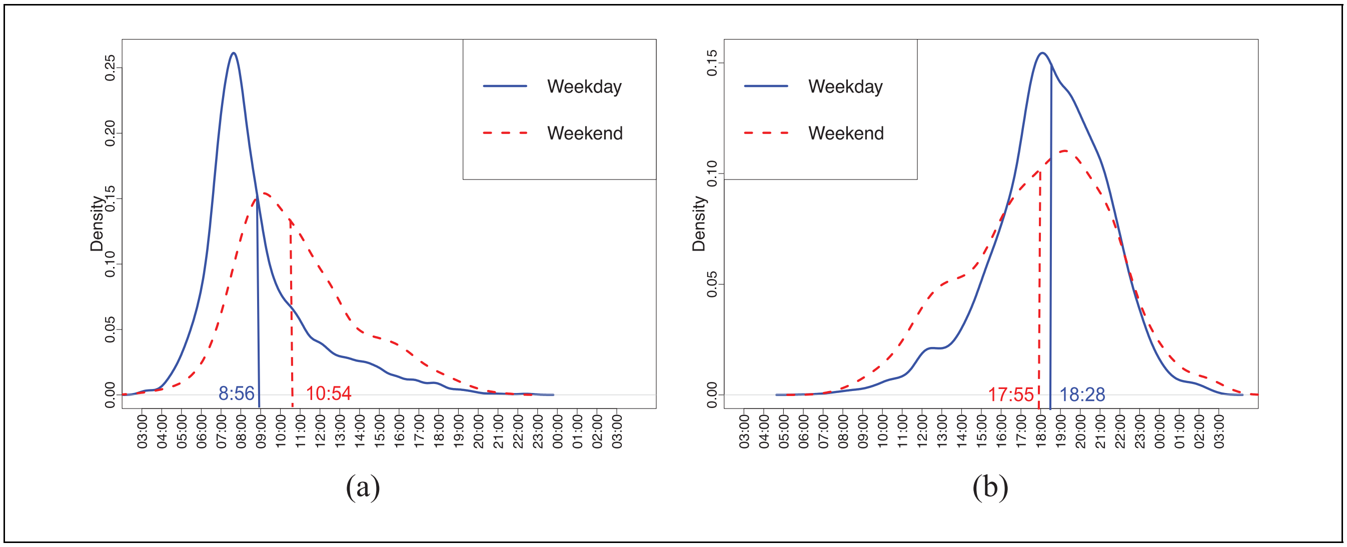


Figure 2. Distributions of (*a*) first trip departure time and (*b*) last trip return time.

Table 3. Statistics of Clusters

	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6	
Activity purpose	#	%	#	%	#	%	#	%	#	%	#	%
Home	2201	55.74	545	56.83	3426	48.53	548	54.69	377	56.61	311	56.75
Work	37	0.94	5	0.52	1936	27.43	285	28.44	183	27.48	1	0.18
School	1	0.03	1	0.10	1	0.01	0	0.00	0	0.00	155	28.28
Transit	351	8.89	51	5.32	696	9.86	61	6.09	35	5.26	36	6.57
Shop	650	16.46	71	7.40	347	4.92	46	4.59	24	3.60	12	2.19
Person	234	5.93	24	2.50	143	2.03	15	1.50	12	1.80	8	1.46
Rec	465	11.78	260	27.11	504	7.14	44	4.39	34	5.11	24	4.38
Other	10	0.25	2	0.21	6	0.08	3	0.30	1	0.15	1	0.18

Note: # = number of activities; % = percentage; Transit = Transportation/Transitions; Person = Personal Business.

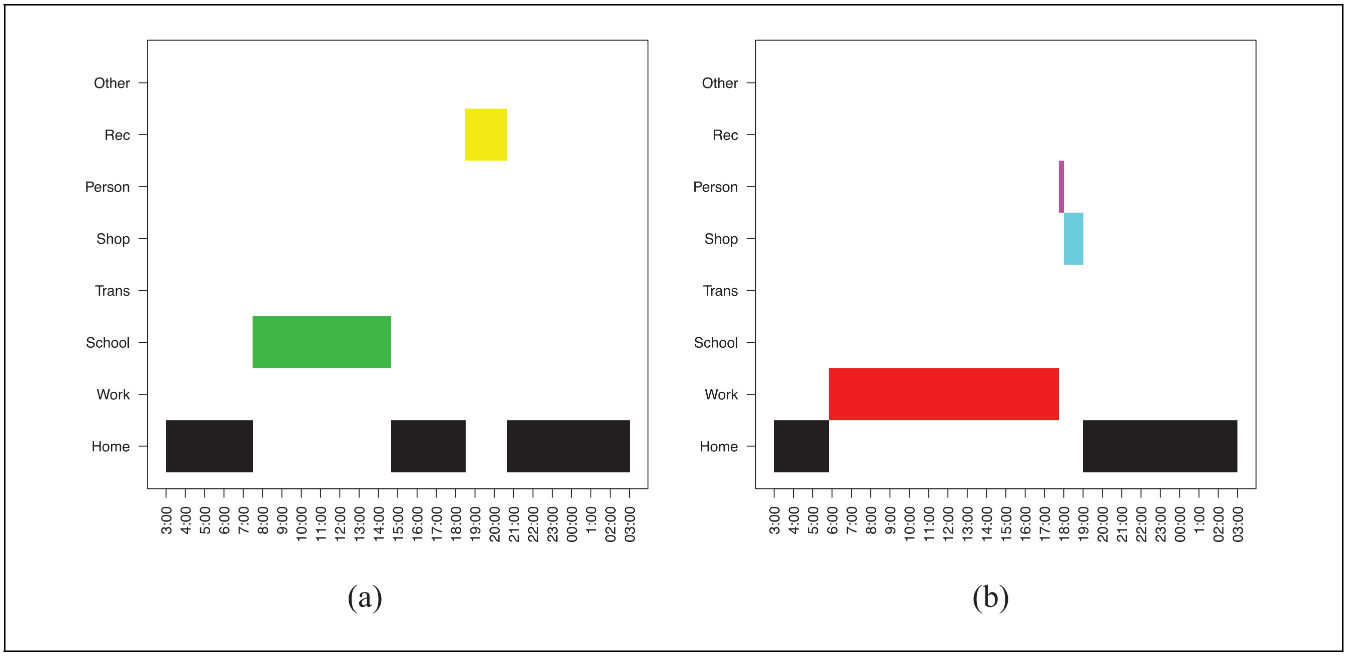


Figure 3. Activity maps.

conducts a short personal business activity and a shopping trip after work at 18:00.

Given an individual's activity matrix, the similarity between every pair of participants is now calculated for constructing the community structure. In this paper, the Jaccard similarity coefficient is implemented to measure similarities for binary matrices of individuals' daily trips.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{3}$$

where A and B represents n binary entries, respectively. M_{01} represents the total number of entries where the entries of A is 0 and the entries of B is 1; M_{10} represents the total number of entries where the entries of A is 1 and entries of B is 0; M_{11} represents the total number of entries where the value of entries of A and B are both 1. Jaccard similarity coefficient ranges from 0 to 1. If $J(A,B) = 0$, this means A and B are totally different. And if $J(A,B) = 1$, this means A and B are exactly the same. Moreover, if A and B are both empty, $J(A,B)$ is defined as 1.

Another similarity measure is the simple matching coefficient (SMC), shown in Equation 4. Unlike the Jaccard similarity coefficient, SMC includes M_{00} which represents the total number of entries where the entries of A and B are both 0. SMC is appropriate when 0 and 1 represent equivalent information, such as gender (*37*).

However, in this paper, 0 and 1 do not carry symmetrical information, and a majority of entries is 0 in the dataset. If SMC is used to measure similarity, M_{00} will dominate the similarity and force it close to 1. Therefore, the Jaccard similarity coefficient is more suitable to describe similarity for asymmetrical attributives than SMC.

$$\mathrm{SMC} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}} \tag{4}$$

Community Detection

In this subsection, first the data is converted into a social network. Each participant in the dataset can be considered as a node. Jaccard similarity coefficient is used to measure the similarity between two users. If Jaccard similarity coefficient is greater than a threshold θ, it indicates two users have similar travel behaviors. In consequence, an edge is created between these two users

$$\begin{cases} \mathrm{Edge}_{ij} = 1 & W_{ij} \geq \theta \\ \mathrm{Edge}_{ij} = 0 & W_{ij} < \theta \end{cases} \tag{5}$$

and the weight of the edge of user i and j can be represented by W_{ij}, this value equals to Jaccard similarity coefficient between these two users. After building the social network, the fast unfolding community detection method is employed to cluster nodes (*38*). This is a kind of modularity maximization algorithm which is one of the most widely used methods for community detection.

In community detection, the modularity of a partition measures the density of edges within communities while comparing with edges between communities and scales between –1 and 1. In the network where edges have weights, the formula of modularity is shown as

$$Q = \frac{1}{2m}\sum_{i,j}\left[W_{ij} - \frac{k_i k_j}{2m}\delta(c_i, c_j)\right] \tag{6}$$

where
W_{ij} is the weight of the edge between node i and j,
$k_i = \sum_j W_{ij}$ represents the summation of the weights for edges attached to node i,
c_i is the community index assigned to this node in this iteration,
$\delta(c_i, c_j)$ equals 1 if $c_i = c_j$ and 0 otherwise, and
$m = \frac{1}{2}\sum_j W_{ij}$.

This algorithm consists of two stages. First, each node is assigned to different communities so that each node is a community, and the initial number of communities is as many as nodes. Then, for each node i, the gains of modularity are measured when removing i from its community and adding in neighboring communities respectively. Then for each isolated node i, the gain of modularity is measured when removing i from the origin community and adding in neighboring communities respectively. And the node i will be allocated into the community with the maximal gain, but only if this gain is positive. If the maximum gain is not positive, node i will stay in the original community. This stage will proceed recursively until a local maximum of the modularity is achieved. The second stage of this algorithm is to build a new network with communities found during the first part. After the second stage is complete, the first stage of this algorithm should be reactivated to reconstruct weighted network. Then these two stages will run alternatively until there are no more changes and the global modularity maximum ($\max(q)$) is found.

Deep Learning

Clustering procedure assigns the same labels for individuals in the same groups, and these labels are considered as the true label in the classification task. In order to classify travelers according to their activity maps, the CNN is implemented to complete this task. CNN, which is a kind of feed-forward artificial neural network, is recognized as a powerful and prevalent tool in image reorganization. Compared with the traditional neural network, CNN not only contains more layers, but also learns from filters, represented by a vector of weights with which the input is convolved. Filters are implemented to slide across all the areas of images, and the size of the commonly used filter is 2 × 2 which is also utilized in this paper. A CNN is composed of input and output layers, as well as multiple hidden layers between input and output layers. There are three kinds of hidden layers, including convolutional layers, pooling layers, and fully connected layers. The function of convolution layers is applying a convolution task to the input and passing the result to the next layer. Pooling layers combine the output from one layer and pass it into a single neuron in the other layer. In the CNN, every convolutional layer is followed by a pooling layer. Moreover, the max pooling is implemented in the pooling layer which will pass the maximal value from the previous layer to the next layer. Fully connected layers aim to connect all neurons in the previous layer to all neurons in the next layer. Figure 4 illustrates a CNN with two convolutional layers and two fully collected layers. This CNN is constructed in the next section.

In the traditional ANN networks, a sigmoid function is utilized to process data. However, in this paper, Rectified Linear Units (ReLU) are implemented as

$$f(x) = \max(0, x) \tag{7}$$

where x is the input to a neuron.

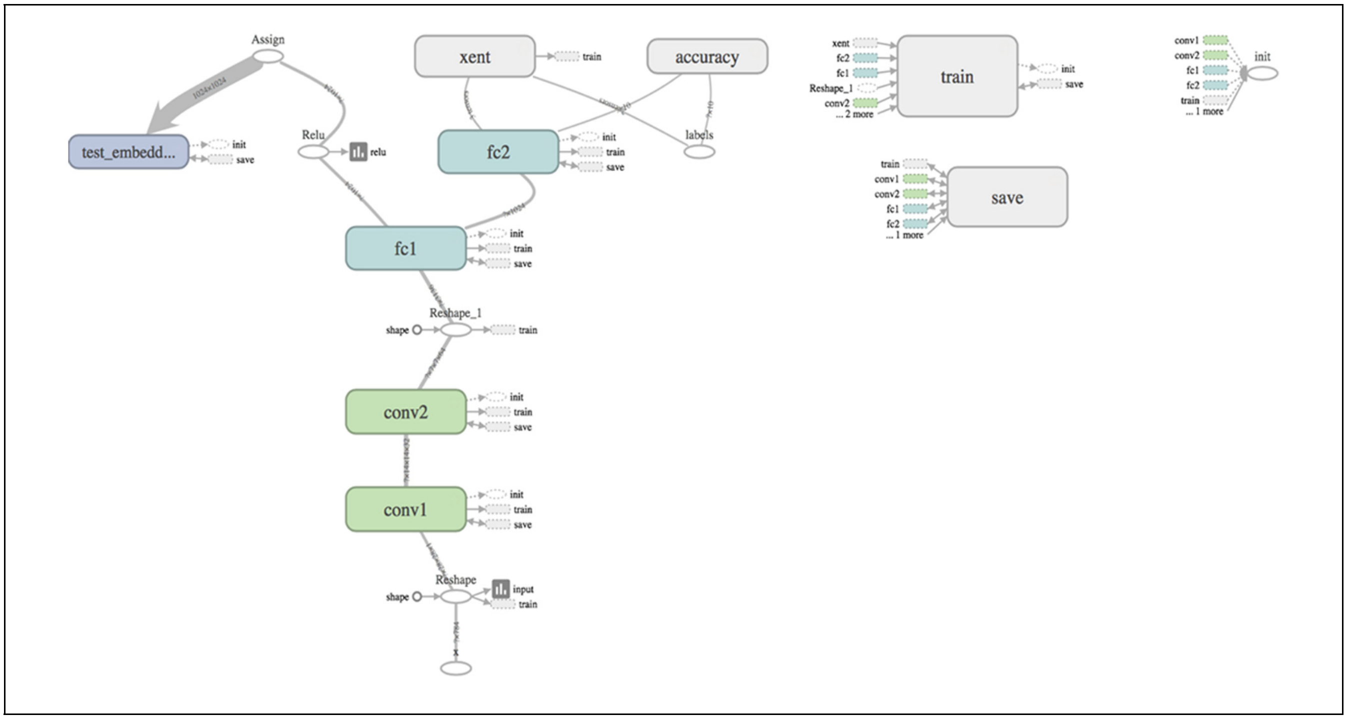


Figure 4. The layout of a CNN.

The sigmoid function contains three problems: the first one is that saturated neurons kill the gradients that might always be zero. Second, the sigmoid output is not zero-centered. The last one is that the computation of exponential is very expensive. However, ReLU will not saturate in positive region and it is very computationally efficient (approximately 6 times faster than sigmoid function) (*32*).

In this paper, the proposed CNN includes two fully connected layers, and this may result in overfitting since fully connected layer occupies most of the parameters. Dropout process is added between two fully connected layers in order to prevent overfitting. Moreover, this process also can speed up the training process. Finally, cross entropy is utilized as the loss function as shown in Equation 8.

$$\mathrm{H(p,q)} = -\sum_{x} p(x)\log q(x) \tag{8}$$

where $p(x)$ and $q(x)$ are two probability distributions over discrete variable x, and $q(x)$ is the estimate distribution for true distribution $p(x)$. $H(p,q)$ is the cross entropy for the distribution p and q.

Numerical Examples

In the first step, a 1 × 2,034 (= 288 × 8) binary vector is constructed to represent the activity chain on the assigned travel date for each participant. Then the Jaccard similarity coefficient is calculated for every pair of individuals according to their activity matrices. The Jaccard similarity coefficient becomes the link weight between a pair of individuals. The Jaccard similarity coefficient ranges from 0 to 1. In this paper, an edge is built between two individuals in the community structure when the matrix similarity is greater than 0.9 ($\theta = 0.9$). An undirected graph is constructed with 3,887 nodes and 98,026 edges, shown in Figure 5. An experiment is conducted to find an appropriate threshold θ of matrix similarity on 0.85, 0.9, and 0.95. When θ is 0.85, there are too many edges and the ratio of edge to node is too high. In this case, there is a very big cluster and several really small clusters. Therefore, the clustering result is not good for 0.85. When θ is set as 0.95, the number of the node is too low and the ratio of edge to node is too low as well. Each cluster contains not too many nodes. Moreover, most nodes are not clustered in any group. Therefore, the clustering result is not good for 0.95 either. Under 0.90, the number of node, edge and edge to node is reasonable. The number of nodes within clusters is reasonable and there are significant differences among each cluster. Moreover, the number of nodes which do not belong to any cluster is not too high. Therefore, the threshold θ is set as 0.9. In this graph, community detection algorithm is applied with Gephi, a graph visualization tool. Finally, 7 clusters in green (Cluster 1, 905 nodes), blue (Cluster 2, 229 nodes), pink (Cluster 3, 1583 nodes), orange (Cluster 4, 255 nodes), red (Cluster 5, 172 nodes), and yellow (Cluster 6, 147 nodes) are detected.

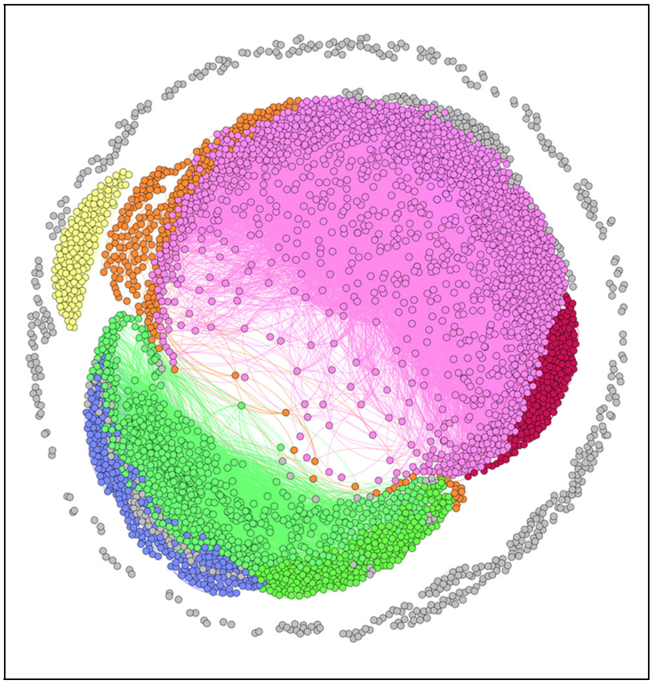
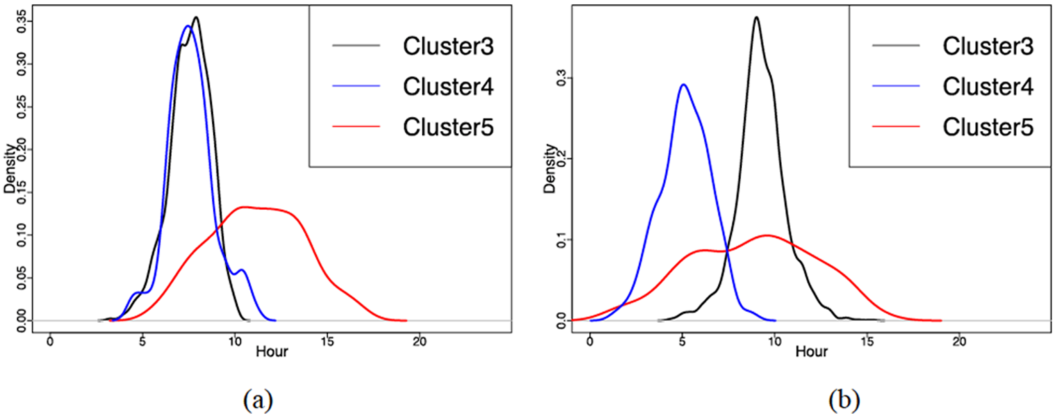


Figure 5. Clustering results.

The nodes isolated with others or not clustered in any groups are also displayed in the periphery in gray color, and they are considered as Cluster 7. Each node represents an individual, and nodes in same colors belong to the same clusters.

After identifying clusters, characteristics are summarized and concluded for each cluster. From Table 3, it can be seen that Cluster 1 and Cluster 2 represent individuals who do not go to work. On the contrary, individuals in Cluster 3, Cluster 4 and Cluster 5 need to work a lot. And participants in Cluster 6 take lots of trips to school. Furthermore, Cluster 1 and Cluster 2 can be differentiated from each other according to shopping and recreation activities in Table 3. From Table 3, it can be seen that individuals in Cluster 1 conduct more shopping activities, whereas individuals in Cluster 2 do many more recreation activities. Then Cluster 3, Cluster 4, and Cluster 5 can be distinguished with the support of Figure 6. Figure 6*a* shows the starting time (including travel time) of the first working activity. As can be seen, individuals in Cluster 3 and Cluster 4 go to work mostly at 07:00. However, the departure time of individuals in Cluster 5 is much later and shows more uncertainty than



(a)

(b)

Figure 6. Distributions of (a) the starting time of the first working activity, and (b) working duration.

Table 4. Accuracy for Traditional Machine Learning Algorithm and CNN

	SVM	KNN	RF	CNN
Cluster 1	97.5%	100.0%	97.5%	100.0%
Cluster 2	57.5%	87.5%	60.0%	95.0%
Cluster 3	95.0%	90.0%	92.5%	97.5%
Cluster 4	90.0%	97.5%	90.0%	100.0%
Cluster 5	97.5%	97.5%	100.0%	97.5%
Cluster 6	92.5%	100.0%	95.0%	95.0%
Cluster 7	17.5%	35.0%	30.0%	85.0%
Average accuracy	78.2%	86.8%	80.7%	95.7%

individuals in the other two clusters. Regarding Cluster 5, the departure time for the first working activity is distributed from 10:00 to 14:00 more evenly. To separate Cluster 3 and 4, refer to Figure 6*b* that presents the working duration distribution. It can be seen that Cluster 3 shows a high probability of working approximately 9 hours (including travel time to work) per day. However, for Cluster 4, the working duration, which is around 5 hours, is much shorter than Cluster 3. The working duration distribution for Cluster 5 is also flat and ranges from 4 to 12 hours.

In summary, individuals can be clustered into 7 groups:

- Cluster 1: Non-working individuals with more shopping activities;
- Cluster 2: Non-working individuals with more recreation activities;
- Cluster 3: Individuals with normal working start time and a full-time job;
- Cluster 4: Individuals with part-time job;
- Cluster 5: Individuals with late start working time;
- Cluster 6: Individuals who need to attend school; and
- Cluster 7: Individuals that are not in any of the first 6 clusters.

With identified travel behavior clusters, a travel behavior classification task is performed with CNN to classify new travelers according to their activity maps. TensorFlow (*39*), one of most powerful deep learning tools, is employed to construct CNN in this paper. According to the aforementioned clustering results, clusters have different sizes. To create balance training and testing datasets, 100 records are randomly selected from each cluster for training, and 40 records from each cluster for test. There are 7 classes in classification task, in addition to the 6 classes which are identified by community detection algorithm. Class 'other' is added which represents users that are not in any of the 6 clusters. Moreover, several traditional machine learning algorithms are also implemented as benchmarks.

After training a CNN, accuracy as high as approximately 95% is finally achieved on the test dataset, and detailed accuracy information for all algorithms is shown in Table 4. Therefore, CNN is an appropriate and efficient algorithm for classifying participants into groups with similar travel behavior according to their activity maps.

Conclusion

In this paper, CHTS data is used to analyze and classify travel behavior. After processing raw data, some interesting observations are made. First, travelers make more commute trips to work and school during weekdays but more recreation, shopping, and personal business activities on weekends. Second, most participants' travel territory is around 5 miles, and 95% of participants travel within 10 miles for each trip. Third, most individuals have a low degree of spatial variability.

Based on 5-minute interval and eight activity types, an activity matrix is further constructed for each participant. Jaccard similarity coefficient is employed to calculate the similarity between every pair of participants. The output similarities are utilized to construct a social network for participants. After this, a community detection algorithm is adopted to cluster individuals into groups with the same travel behavior. The algorithm produces seven clusters: 1) non-working people with more shopping activities, 2) non-working people with more recreational activities, 3) individuals with normal working start time and a full-time job, 4) Individuals with part-time job, 5) individuals with late start working time, 6) individuals who need to attend school, and 7) individuals that are not in any of the first 6 clusters. Then individuals are classified by using a CNN and approximately 95% is achieved in accuracy.

In future, a variety of data sources, including passive trip trajectory data, transit smart card data, and social media data can be collected to provide more information about individual travel behavior. Therefore, the analysis of similar driver behavior can go a step further.

Acknowledgments

This study was partially supported by National Science Foundation award CMMI-1637604 and Region 2 University Transportation Research Center faculty-initiated research project.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Qing He, and Alireza Khani; data collection: Yu Cui; analysis and interpretation of results: Yu Cui, Qing He, and Alireza Khani; draft manuscript preparation: Yu Cui, Qing He, and Alireza Khani. All authors reviewed the results and approved the final version of the manuscript.

References

1. Cici, B., A. Markopoulou, E. Frias-Martinez, and N. Laoutaris. Assessing the Potential of Ride-Sharing Using Mobile and Social Data: A Tale of Four Cities. *Proc., 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, Washington, ACM, New York, 2014. pp. 201–211.
2. Gong, L., T. Morikawa, T. Yamamoto, and H. Sato. Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia-Social and Behavioral Sciences*, Vol. 138, 2014, pp. 557–565.

3. Wolf, J., R. Guensler, and W. Bachman. Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2001. 1768: 125–134.
4. Wu, J., C. Jiang, D. Houston, D. Baker, and R. Delfino. Automated Time Activity Classification Based on Global Positioning System (GPS) Tracking Data. *Environmental Health*, Vol. 10, No. 1, 2011, p. 101.
5. Bohte, W., and K. Maat. Deriving and Validating Trip Destinations and Modes for Multiday GPS-Based Travel Surveys: Application in the Netherlands. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C., 2008.
6. Chen, C., H. Gong, C. Lawson, and E. Bialostozky. Evaluating the Feasibility of a Passive Travel Survey Collection in a Complex Urban Environment: Lessons Learned from the New York City Case Study. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 10, 2010, pp. 830–840.
7. Kim, Y., F. C. Pereira, F. Zhao, A. Ghorpade, P. C. Zegras, and M. Ben-Akiva. Activity Recognition for a Smartphone and Web Based Travel Survey. *arXiv preprint arXiv:1502.03634,* 2015.
8. Schönfelder, S., K. W. Axhausen, N. Antille, and M. Bierlaire. Exploring the Potentials of Automatically Collected GPS Data for Travel Behaviour Analysis-A Swedish Data Source. In *GI-Technologien für Verkehr und Logistik* (J. Möltgen, and A. Wytzisk, eds.), IfGIprints, Münster, Germany, 2002, pp. 155–179.
9. Deng, Z., and M. Ji. Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. *Proc., 7th International Conference on Traffic and Transportation Studies 2010*, ASCE, Kunming, China, 2010. pp. 768–777.
10. Griffin, T., and Y. Huang. A Decision Tree Classification Model to Automate Trip Purpose Derivation. *Proc., ISCA 18th International Conference on Computer Applications in Industry and Engineering*, Honolulu, Hawaii, ISCA, Honolulu, 2005. pp. 44–49.
11. Jones, P. Developments in Dynamic and Activity-Based Approaches to Travel Analysis. Oxford Studies in Transport, Avebury, Aldershot, 1990.
12. Hanson, S., and J. Huff. Classification Issues in the Analysis of Complex Travel Behavior. *Transportation*, Vol. 13, No. 3, 1986, pp. 271–293.
13. Shoval, N., and M. Isaacson. Sequence Alignment as a Method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers*, Vol. 97, No. 2, 2007, pp. 282–297.
14. Kitamura, R., and T. Hoorn. Regularity and Irreversibility of Weekly Travel Behavior. *Transportation*, Vol. 14, No. 3, 1987, pp. 227–251.
15. Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt. Observing the Rhythms of Daily Life: A Six-Week Travel Diary. *Transportation*, Vol. 29, No. 2, 2002, pp. 95–124.
16. Jiang, S., J. Ferreira, and M. C. González. Clustering Daily Patterns of Human Activities in the City. *Data Mining and Knowledge Discovery*, Vol. 25, 2012, pp. 478–510.
17. Jiang, S., J. Ferreira, Jr., and M. C. Gonzalez. Discovering Urban Spatial-Temporal Structure from Human Activity Patterns. *Proc., ACM SIGKDD International Workshop on Urban Computing*, Beijing, China, ACM, New York, 2012, pp. 95–102.
18. Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, Vol. 453, No. 7196, 2008, pp. 779–782.
19. Ponieman, N. B., A. Salles, and C. Sarraute. Human Mobility and Predictability Enriched by Social Phenomena Information. *Proc., 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagara, Ontario, ACM, New York, 2013, pp. 1331–1336.
20. Ma, X., Y.-J. Wu, Y. Wang, F. Chen, and J. Liu. Mining Smart Card Data for Transit Riders' Travel Patterns. *Transportation Research Part C: Emerging Technologies*, Vol. 36, 2013, pp. 1–12.
21. Williams, M. J., R. M. Whitaker, and S. M. Allen. Measuring Individual Regularity in Human Visiting Patterns. *Proc., Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, Amsterdam, Netherlands, IEEE, New York, 2012. pp. 117–122.
22. Girvan, M., and M. E. Newman. Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 12, 2002, pp. 7821–7826.
23. Newman, M. E., and M. Girvan. Finding and Evaluating Community Structure in Networks. *Physical Review E*, Vol. 69, No. 2, 2004, p. 026113.
24. Radicchi, F., C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and Identifying Communities in Networks. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 9, 2004, pp. 2658–2663.
25. Palla, G., I. Derényi, I. Farkas, and T. Vicsek. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature*, Vol. 435, No. 7043, 2005, pp. 814–818.
26. Soundarajan, S., and C. Gomes. Using Community Detection Algorithms for Sustainability Applications. *Proc., 3rd International Conference on Computational Sustainability*, Biocenter, University of Copenhagen, Copenhagen, Denmark, 2012.
27. Lin, L., Q. Wang, and A. Sadek. Data Mining and Complex Network Algorithms for Traffic Accident Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2014. 2460: 128–136.
28. Wikipedia. *Applications of Artificial Intelligence*. https://en.wikipedia.org/wiki/Applications_of_artificial_intelligence-Transportation.
29. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks*, Vol. 61, 2015, pp. 85–117.
30. Gers, F. A., N. N. Schraudolph, and J. Schmidhuber. Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research*, Vol. 3, 2002, pp. 115–143.

31. Fernández, S., A. Graves, and J. Schmidhuber. An Application of Recurrent Neural Networks to Discriminative Keyword Spotting. In *Artificial Neural Networks–ICANN 2007*, 2007, pp. 220-229.
32. Krizhevsky, A., I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, December 3–6, 2012. pp. 1097–1105.
33. He, K., X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *Proc., IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nev., IEEE, New York, 2016. pp. 770–778.
34. IMAGENET. *Large Scale Visual Recognition Challenge*, 2016. http://image-net.org/challenges/LSVRC/2016/results. Accessed May 8, 2018.
35. Dong, W., J. Li, R. Yao, C. Li, T. Yuan, and L. Wang. Characterizing Driving Styles with Deep Learning. *arXiv preprint arXiv:1607.03611*, 2016.
36. Kang, C., X. Ma, D. Tong, and Y. Liu. Intra-Urban Human Mobility Patterns: An Urban Morphology Perspective. *Physica A: Statistical Mechanics and its Applications*, Vol. 391, No. 4, 2012, pp. 1702–1717.
37. Wikipedia. *Jaccard index*. https://en.wikipedia.org/wiki/Jaccard_index.
38. Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, 2008, p. P10008.
39. TensorFlow. *TensorFlow*. https://www.tensorflow.org/. Accessed May 8, 2018.

The Standing Committee on Traveler Behavior and Values (ADB10) peer-reviewed this paper (18-01761).