TransInfo
Tier I University Transportation Center

FINAL REPORT

Travel Behavior Analysis with Social Media Data and Smartphone GPS Data

Date of report: February, 2019

Yu Cui, PhD, Graduate Research Assistant, University at Buffalo, The State University of New York
Qing He, PhD, Morton C Frank Endowed Associate Professor, University at Buffalo, The State University of New York

Prepared by:
Organization
Address line 1
Address line 2
Address line 3

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Travel Behavior Analysis with Social Media Data and Smartphone GPS Data		**5. Report Date** February, 2019 **6. Performing Organization Code**
7. Author(s) Yu Cui and Qing He		**8. Performing Organization Report No.**
9. Performing Organization Name and Address University at Buffalo, The State University of New York 313 Bell Hall, Buffalo, NY, 14260, USA		**10. Work Unit No. (TRAIS** **11. Contract or Grant No.** DTRT13-G-UTC48
12. Sponsoring Agency Name and Address US Department of Transportation Office of the UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		**13. Type of Report and Period Covered** 09/2017-12/2018 **14. Sponsoring Agency Code**
15. Supplementary Notes		

16. Abstract

This project consists of two research components. The objective of the first component is to develop an approach to resample social media data in order to reduce biases and errors through estimation of socio-demographics. Several machine learning models are proposed for predicting socio-demographics, including gender, age, ethnicity and education levels. Afterward, this study resamples social media data and compares the results with the 2009 California Household Travel Survey data. The resampled data shows comparable characteristics to the survey data. Moreover, since social media is a kind of long-term data, it shows several advantages in research over survey data. This research sheds light on tackling sampling bias issues when social media data is used for travel behavior analysis.

The second study shapes a sustainable and long-term travel survey with 7-month low-frequency smartphone GPS data with imperfect activity information. The essential goal is to develop a daily synthetic trip chain simulator. This research develops a new probabilistic method to handle imperfect activity data, and three different levels of trip chain generation models are proposed. The first model handles only known activities, the second model treats all unknown activities as a single category, and the third one models each unknown location separately. These models are able to generate trip chains in different levels of details for activity-based.

17. Key Words	18. Distribution Statement

Travel Behavior Analysis; Social Media Data; Smartphone GPS Data; Sampling Bias; Travel Survey		No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161	
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 41	**22. Price**

Insert your own project cover page here

Acknowledgements

We thank Prof. Ling Bian from University at Buffalo, who provided smartphone GPS data for this study.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Table of Contents

1. Problems

1.1 Inferring Twitters' Socio-Demographics to Correct Sampling Bias of Social Media Data for Augmenting Travel Behavior Analysis

Social media plays a very important role and has influence in virtually every aspect of our lives. It has tremendously changed the way people interact and carry on with their everyday lives. Social media becomes a necessity everywhere and people are willing to spend time on social media sites. Despite people access to different types of social media for different purposes, we still can take advantage of the passive datasets provided by social media for transportation applications (Zhang and He, 2019). For example, Twitter data contains tweet text, hashtags, and geo-location for geo-tagged tweets. The geo-tagged tweets are considered as check-in data which includes the tweet posted locations (Rashidi et al., 2017). And the attached locations indicate that the user used to be these places for certain activities. Hashtags, together with tweet text, can provide useful information related to traffic events. For example, researchers utilized social media data to detect traffic incidents (Zhang et al., 2018), uncover travel activity types (Meng et al., 2017, Cui et al., 2018b), model the impacts of inclement weather on freeway traffic speed (Lin et al., 2015), and predict subway ridership (Ni et al., 2017), etc.

Comparing to the Household Travel Survey (HHTS), the traditional data source in transportation area, social media data has several advantages. The first advantage of social media data response faster than HHTS data, and it even provides real-time information. A HHTS is typically conducted once a decade, and it provides elaborate and reliable data for many long-term transportation research areas including traffic safety (NHTS, 2011), transportation planning (Ouimet et al., 2010), travel behavior (Polzin et al., 2008), travel trend analysis (Cui et al., 2018a). However, HHTS data is not appropriate utilized in short-term and real-time research, whereas social media responses faster than HHTS. Information about traffic and events can be extracted from social media in a real-time fashion. The second obvious advantage of social media data is the long-term coverage. Longitudinal social media data can track individual's travel behavior in multi-years (Picornell et al., 2015), while HHTS usually covers only 1 day's travel. Although one obvious problem of social media data is the low daily post rate, the long-term coverage of social media data could potentially make up for its inherent sparsity.

Besides the existing virtues of social media data, nevertheless, the potential of social media data for augmenting travel behavior research is still not to be fully explored. A fundamental limitation is that the activity patterns cannot be interpreted from social media clearly due to the shortage of socio-demographic information (Rashidi et al., 2017). On one hand, some social media platforms that contain detailed personal information, such as Facebook, are forbidden to any automatic information retrieving methods. On the other hand, some social media platforms without too

many restrictions, for instance, Twitter, contains hardly any personal information. It is also well known that social media data suffers from sampling bias errors. Twitter users are not a representative sample of the overall population, tending to skew towards young, urban, minority individuals (Mislove et al., 2011, Lee et al., 2016), Given wide acknowledgment of this challenge, there exist some past studies to address it. However, little has been done from the standpoint view of travel behavior analysis.

In sum, understanding the structure and characteristics of social media users is essential to move forward to further measurements and analyses. Moreover, this helps correct sampling biases of social media data as well. This paper aims to infer a variety of demographics (including gender, age, ethnicity, and education level) of Twitter users as an applicable example, resample the tweets according to real population distribution, and compare with existing HHTS data. The key contributions of this paper lie in as follows:

- Examine how to use demography inference to correct sampling bias in social media data.
- Divide emojis into different categories according to gender, contrary and culture separately.
- Employ the feature of auto-defected languages of Twitter for the first time.
- Implement deep-learning to classify socio-demographic attributions.
- Validate the behavior characteristics of the resampled data using the survey data.

1.2 Generating a Synthetic probabilistic daily activity sequence using long-term and low-frequency smartphone gps data with imperfect activity information

Understanding travel behavior and obtaining activity sequences become new and prevalent approaches to estimate travel demand as the development of activity-based models. When the microsimulation requires daily travel patterns of individuals as input data, the synthetic daily travel chains or travel diaries are needed. However, traditional household travel survey usually collects a very short period of travel (e.g. 1 day). This short period of travel diaries restricts and neglects the variation of individuals' travel behaviors. As a result, the individual's longitudinal travel behavior can be barely captured with traditional survey method. A growing body of literature suggests that longer data collection periods are warranted to provide improved data for modeling purposes and understanding travel variations. Therefore, there is a pressing need to conduct longer periods of data collection and devise a new approach which avoids the possibly imposed respondent burden and survey costs.

With emerging information and communication technology (ICT) tools, the collection of passive datasets for travelers' real-time trajectory becomes available. Trajectory data can be divided into two categories, explicit trajectory data and implicit trajectory data (Kong et al., 2018). Explicit trajectory data is recorded in succession at constant intervals, such as data from GPS devices and smartphone GPS collection applications. On the country, implicit trajectory data is recorded with a random and relatively large time interval. Data sources of implicit trajectory data are sensor-based data (monitor), network-based data (social media check-in data) and signal-based data (Wi-Fi, Bluetooth, RFID, and mobile data).

High penetration of smartphones guarantees that collecting GPS trajectory becomes trivial since most smartphones have both GPS and accelerometer sensors (Shen and Stopher, 2014).

Smartphone GPS survey apps have emerged to be a popular tool for conducting household travel surveys. There are existing some surveys using smartphones as GPS devices to record GPS data (Bierlaire et al., 2013, Hudson et al., 2012, Reddy et al., 2010, Xiao et al., 2012). Comparing to GPS devices, using a smartphone to record GPS information for the survey has several advantages. First, smartphones reduce study cost since no additional GPS is needed. Second, it decreases the chance that participants forget to bring or charge wearable GPS devices. If GPS sensor is an in-vehicle GPS device, it stops recording after car stalling and the GPS records from the car parking places to activity places are missing. In contrast, smartphones record these trajectories as well. Even though the sampling frequency and accuracy of GPS points of smartphones are lower than the dedicated GPS devices, it is proved that data quality and information are comparable with GPS devices and sufficient for research (Montini et al., 2015). Moreover,

However, the smartphone-based travel survey also holds several intrinsic limitations and challenges. If the travel app keeps sampling high frequent GPS points as GPS devices, the smartphone battery drains quickly and may be damaged as well (Patterson and Fitzsimmons, 2016). Users are very sensitive to battery consumption, and high battery consumption makes people are unwilling to participate in experiments. Therefore, the trade-off between GPS sampling frequency and battery consumption needs to be accounted during the travel design.

There is also another concern when conducting travel survey with smartphone. This concern also appears GPS device-based travel survey, that travel mode and purpose cannot be recorded (Zhou et al., 2017). In GPS devices assisted travel survey, participants fill one- or two-day travel diaries. In smartphone-based travel survey, we still can request participants to validate or input these data. However, this increases the burden of participants as well, and this burden increases as the number of days need to be validated. Therefore, the trade-off also needs to be addressed between the amount of labeled activity information and the amount of required intervention of participants while one designs the smartphone app (Liao et al., 2017).

To address abovementioned two concerns, a low-frequency GPS data sampling method with few activity labels is introduced in this paper. This mechanism naturally resolves these two problems, battery drain, and reporting and intervention burden, and increases the ease of recruitment. Further, this smartphone-based survey approach enables a long-term participant period. Long-term data can capture more travel information, thus, this will reduce unknown information caused by low reporting and validating frequency. Moreover, long-term data also enable analysis of longitudinal travel behaviors which is difficult to be derived from traditional household travel surveys. In addition, the proposed method can account for the variation of peoples travel patterns and provide better results in simulation and traffic demand forecasting.

he proposed method also poses additional challenges. Our GPS dataset is collected by a smartphone app running in the background. The time interval between the two records are not uniform and the frequency is low. The traditional method to process GPS trajectory or implicit trajectory data is not appropriate for this study. Therefore, the first challenge during data preprocessing is how to identify stop/trip ends from this random interval and low-frequency data. More details are discussed in Session3.1.2. The other challenge is how to handle imperfect

activity information. In this paper, the imperfect activity information indicates the unreported activity location given an existing trip. And this is discussed in Session 2.3.

2. Approach and Methodology

2.1 Features for Twitter Demographic Analysis

There are four types of features that can be extracted from a Twitter user, which include profile, tweet behavior, tweet text, and connection.

(1) Profile is the feature about who the Twitter user is, including username, user-id, profile pictures, account creation time, the content shown in biofield, and location information if any.

(2) Twitter behavior contains the number of posts a day, the total number of tweets, number of replies, and post time.

(3) Tweet text is the post content. Traditional tweets are up to 140 characters including emojis, and Twitter doubled this limit to 280 characters on November 7, 2017. Moreover, each tweet can attach up to 2 hashtags, 4 photos or 1 video or 1 gif picture. Tweets also include a language feature, which is automatically detected by Twitter.

(4) Connection is the reachable people or social connections, including followers, followings, number of followers and number of followings.

Most papers only implement one or two types of features. However, in this study, we examine features of all four types. Moreover, we also explore the importance of these features for the prediction of each socio-demographic variable.

2.2 Models for Twitter Demographic Analysis

In this study, we examine several traditional machine learning methods, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF). As one can see later, RF models outperform other machine learning models. Since there are more than one thousand tweeting text features, we also implement a deep learning (DL) based approach, which is a strong natural language-mining tool. Table 1 illustrates models and significant features used in different socio-demographics.

Table 1 Models and significant features used in different socio-demographics

Socio-Demographics	Models	Significant features
Gender	KNN, SVM, ANN, ***RF****, DL	hashtags, Twitter behaviors, connections, emojis
Age	KNN, SVM, ANN, RF, ***DL****	Twitter contents
Ethnicity	KNN, SVM, ANN, ***RF****, DL	hashtags, Twitter behaviors, connections, emojis, language mastered

Education Level	KNN, SVM, ANN, ***RF****, DL	hashtags, Twitter behaviors, connections, number of languages mastered

*: indicates the model with the best performance

2.3 Methodology of Activity Sequence Simulator

Inspired by mobile Apps usage behavior predicting methods (Liao et al., 2013, Liao et al., 2012), a new probabilistic method is developed to handle imperfect activity data. This research develops daily synthetic activity sequence simulator in three levels. The first level only considers reported frequent places, the second level, the second model treats all unknown activities as a single category, and the third model treats unknown locations in a more detailed way. Figure 1 illustrates modeled places for each level.

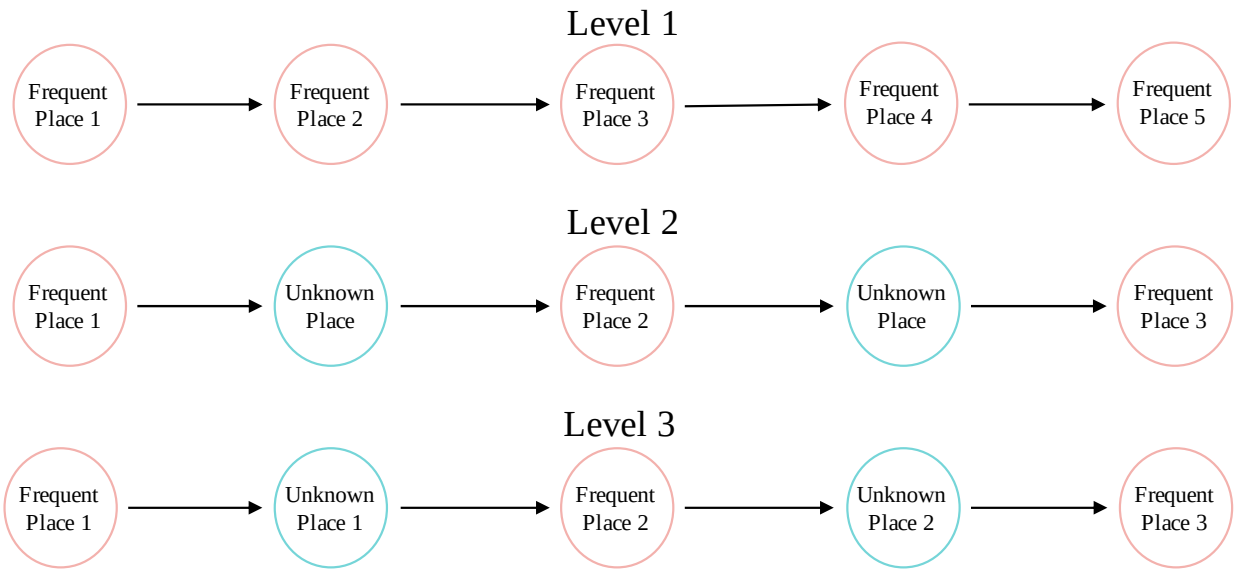


Figure 1 Modeled Places for (a) Level 1: consider frequently visited places only, (b) Level 2: consider unknown places as one category and (c) Level 3: consider unknown places in a more detailed way

There are three different scores to measure the probability of visiting a place, which are global visit score (GVS), temporal visit score (TVS) and periodical visit score (PVS).

2.3.1 Global Visit Score

GVS can describe the probability of the places where are visited in a global view regardless the time. However, the GVS of current place is still influenced by the previous visited place. We only consider the previous visited place, not all previous visited places. If all previous visited

places, data may not be enough to derive the score, since our activity sequence simulator is the more detailed location level, not activity level.

If it is the first place:

$$\mathrm{P}_{\mathrm{Global}}(p_1) = \sum_{p_j \in P} \frac{v(p_1)}{v(p_j)} \tag{1}$$

If it is not the first place:

$$\mathrm{P}_{\mathrm{Global}}(p_i|p_{i-1}) = \sum_{p_j \in PLACE} \frac{v(p_i|p_{i-1})}{v(p_j|p_{i-1})} \qquad i = 2, \dots, n \tag{2}$$

where p_i represents ith place in a activity sequence. $p_j \in P$ where P is the potential visited places list, and p_j represents jth place in potential visited places list. $v(p)$ represents the place's visit time count in the entire training set.

2.3.2 Temporal Visit Score

TVS is employed for those places are regularly visited at a specific time. First, we divided a day into 24-temporal time window (from 0 to 23 hour). Then we calculate the probability of visiting a place in each time window.

If it is the first place:

$$\mathrm{P}^t_{\mathrm{Temporal}}(p_1) = \sum_{p_j \in} \frac{v_t(place)}{v_t(place_j)} \qquad t = 0, 1, \dots, 23 \tag{3}$$

If it is not the first place:

$$\mathrm{P}^t_{\mathrm{Temporal}}(p_i|p_{i-1}) = \sum_{p_j \in P} \frac{v_t(p_i|p_{i-1})}{v_t(p_i|p_{i-1})} \qquad t = 0, 1, \dots, 23, \ i = 2, \dots n \tag{4}$$

where $v_t(place) = \frac{\sum_1^M v_t^m(place)}{M}$, M is the number of time periods.

2.3.3 Periodical Visit Score

PTS measures probabilities for those places where have significant visiting period. For example, employers visit their work places every 24 hours on weekdays. This measurement is assumed not influence by the previous visited place.

$$\mathrm{P}^q_{\mathrm{Periodical}}(p) = \sum_{p_j \in P} \frac{v_q(place)}{v_q(place_j)} \qquad q = 1, 2, \dots, Q \tag{5}$$

where $v_p(place) = \frac{\sum_1^M v_p^m(place)}{M}$, M is the number of time periods, and Q is the frequency of visiting a place. We employ Power Spectral Density (PSD) to detect period. It calculates power for each frequency, and the power with the highest power is the most appropriate frequency for this place. Then $Q = \frac{1}{frequency}$. For more information about PSD, please refer (Vlachos et al., 2005).

2.3.4 Minimum Entropy Selection Method

We use a minimum entropy selection method (MESM) to determine where an individual will be at time t. Since entropy measures the uncertainty of a random variable. The less uncertainty the random variable is, the smaller entropy is, and vice versa. The random variable with low entropy will provide more certain choose. Therefore, we utilize the score with the lowest entropy as indicator and the place with the highest probability in this indicator as staying place. Equation 6 and Figure 2 shows the algorithm of MESM.

$$Entropy = -\sum_i \mathrm{P}_i \ln \mathrm{P}_i \tag{6}$$

Algorithm 2 MESM

1 Calculate $E_{\mathrm{Global}} = \sum_{p_j \in P} -\mathrm{P}_{\mathrm{Global}}(p_j) \ln \mathrm{P}_{\mathrm{Global}}(p_j)$

$E_{Temperal} = \sum_{p_j \in P} -\mathrm{P}_{Temperal}(P_j) \ln \mathrm{P}_{Temperal}(p_j)$

$E_{Periodical} = \sum_{p_j \in P} -\mathrm{P}_{Periodical}(P_j) \ln \mathrm{P}_{Periodical}(p_j)$

2 Select score group with the lowest entropy,

$$s = \arg\min(E_{Global}, E_{Temperal}, E_{Periodical})$$

3 Select place with max probability in group s,

$$p = \arg\max_{p_j \in P} \mathrm{P}_s(p_j)$$

Figure 2 Algorithm of MESM

2.3.5 Level 1 and Level 2 Activity Sequence Simulator

This session introduces a activity sequence simulator for level 1 and level 2 activity sequence. According to several literature and household travel surveys (Kitamura et al., 2000, Cui et al., 2018a), a simulated day start at 3:00 am and end at 2:59 am on next day. Level 1 handles only known activities, and second model treats unknown activities as a single new category. The algorithm of level 1 and level 2 activity sequence simulator is shown in Figure 3.

Algorithm 3 Level 1 and Level 2 Activity Sequence Simulator

1 Initial: $t = 3$

2 While $t < 27$ (3 am of next day):

3 calculate $P_{s_g}, p_{s_t}^t, p_{s_p}^t, E_{f_i}$

4 $p \leftarrow MESM\ (\mathrm{P}_{\mathrm{Global}}, \mathrm{P}_{\mathrm{Temperal}}, \mathrm{P}_{\mathrm{Periodical}})$

5 determine staying duration at given place, select duration with max(density) from history duration

6 $t = t + duration$

Figure 3 Algorithm of Level 1 and Level 2 Activity Sequence Simulator

2.3.6 Level 3 Activity Sequence Simulator

Level 3 activity sequence simulators can handle unknown activities in a more detailed way. We first employ DBSCAN to group locations which near to each other into groups, then name them

as “Unknown 1”, “Unknown 2”, etc. The frequent places will update accordingly. If an unknown location does not group with other locations, it consists of a group with only one location. Figure 4 shows the algorithm of level 3 activity sequence simulator.

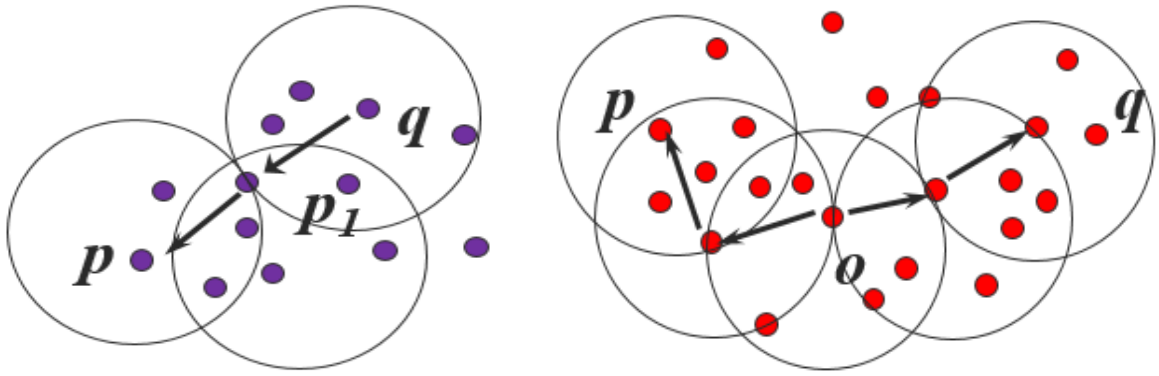
Algorithm 4 Level 3 Activity Sequence Simulator

Initial: $t = 3$
While $t < 27$ (3 am of next day):
    use DBSCAN to group unknow locations
    update place list accordingly
    calculate $P_{s_g}, p_{s_t}^t, p_{s_p}^t, E_{f_i}$
    $p \leftarrow MESM\ (\mathrm{P_{Global}}, \mathrm{P_{Temperal}}, \mathrm{P_{Periodical}})$
    determine staying duration at given place, select duration with max(density) from history duration
    $t = t + duration$

Figure 4 Algorithm of Level 3 Activity Sequence Simulator

2.3.7 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 5 below. Moreover, DBSCAN algorithm process is also showed below.



(a) (b)

Figure 5 (a) p and q are density-reachable; (b) p and q are density-connected

Algorithm 5 DBSCAN

Input: points

```
Output: clusters
for each p ∈ D do
    if p is not yet classified then
        if p is a core point then
            collect all objects density-reachable from p and assign them to a new cluster
        else
            assign p to noise
        end
    end
end
```

Figure 6 Algorithm of DBSCAN

3. Findings: Documentation of Data Gathered, Analyses Performed, Results Achieved

3.1 Data Description

3.1.1 Twitter Data

To gather the Twitter data, we employed Twitter Streaming API. Twitter Streaming API can push the near real-time tweets which match a set of criteria including predefined keywords and locations. The raw dataset contains 6.9 million tweets within the Bay Area of California for more than 4 years, from 01/31/2013 to 02/16/2017. The collected tweets include users who used to visit the Bay Area at least once within in this time period. Figure 7 shows the heatmap of the location of geo-tagged tweets.

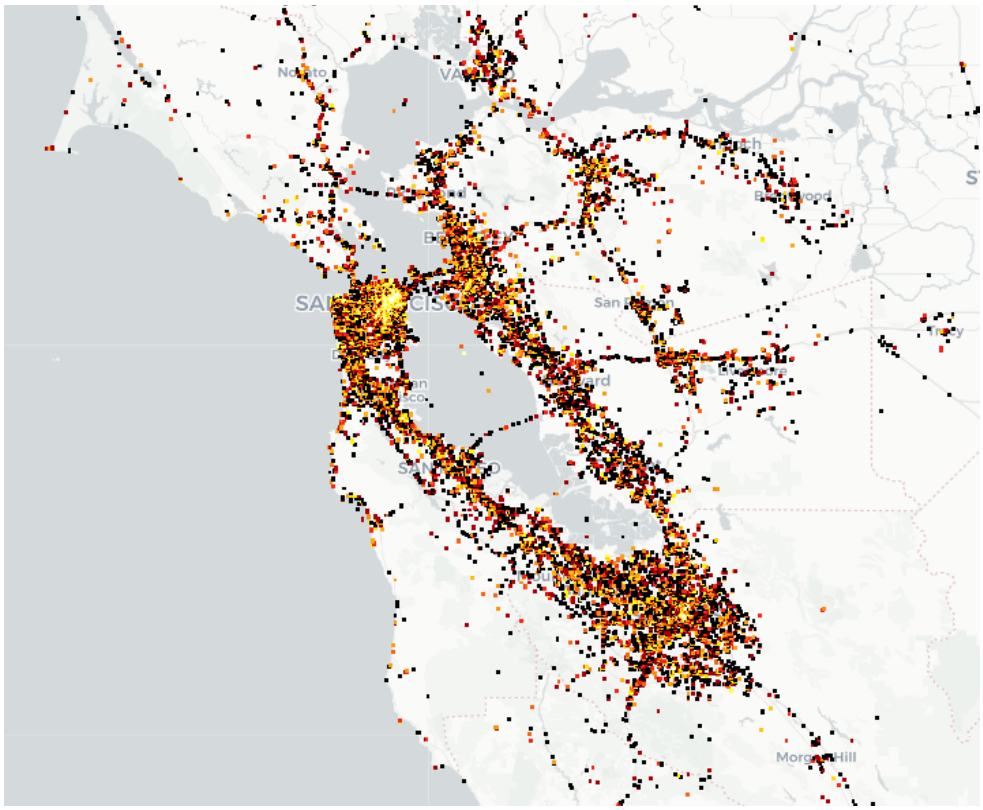


Figure 7 Visualization of posted locations of geo-tagged tweets in Bay Area, CA, USA

After we obtained the entire population of Twitter users in the Bay Area, we used Twitter REST API to obtain users' information according to their user-id or username. Such information includes the number of tweets, number of followers, number of followings, language, and the time of creating account. Afterward, we attempted to distinguish bot users from real Twitter users. There is research that shows 93.11% of users have less than 1,000 followers, and the average number of Twitter followers is 707 (KickFactory, 2016). Therefore, we excluded users with followers larger than 1,000.

These procedures narrow down potential users. However, we still did not have users' socio-demographics. It is a fact that one's Twitter accounts can connect with his/her Facebook accounts and display URL of Facebook pages at users' profiles. Assuming most of the Facebook users' information is real, we can acquire one's socio-demographics from associated Facebook accounts. Unlike Twitter API, Facebook Graph AIP v2.0 does not allow developers to get data from Facebook users even if the data is public. Therefore, we randomly extract 1500 active Twitter user profiles which have linked Facebook accounts. Out of 1500 Facebook pages, we found 987 valid pages. Non-valid pages include pages which do not exist, fan pages, and commercial pages.

From these valid Facebook pages, we obtained much personal information, including gender, age, education, occupation, language, race, status, and place of residence. Gender can be easily inferred from profile pictures and profiles. Regarding age, the best situation is that the profile includes the user's year of birth. If not there, hopefully, it contains the education information,

such as the year of graduation from high school (when we assume users were 18 years old), and year of graduation from university or college (when we assume users were 22 years old). Then we used this information to infer the ages of users. We considered education as the highest education degree shown in the profile. Occupation was classified into 4 categories, student, part-time, employee, and self-employed. This is according to the information regarding the latest work and education experience. Language was found in the "Details About You" section if any. Moreover, Twitter can detect language from some tweets. According to the user profile pictures, the race was divided into 3 categories, Asian, Black or Africa American, and other. There are 6 categories for status, including single, in a relationship, engaged, married, divorced, and widowed. This information is given in "family and relationships" pages. Finally, with regard to the living place, most users fill the current city. If there is no information about the current city, we consider it is the same as current working or studying place if any.

Afterward, we implemented Twitter REST API to obtain up to most recent 3000 tweets. However, some accounts are private and we cannot access them. Figure 2 shows the number of users we can access for each feature. Since we can recognize gender and race from profile pictures, the availability rate is high. However, since other information such as age and status tend to be more private, the number of users who provide this information is much less. In addition, age is mostly inferred from users' educational experience, and users seldom provide the exact birth year.

In this study, we aim to infer the user's gender, age, education, and race. There are 884 valid user profiles containing gender information, with 295 (33.37%) females and 589 (66.63%) males. The gender information we obtained is not as balanced as the distribution of Twitter users in the United States as of January 2017, which is 53% male and 47% female (Statista, 2018). However, this phenomenon also appears in other research. For example, Fink et al. also collected Twitter information from Facebook, and labeled 4,023 (36.16%) females and 7,119 (63.82%) males (Fink et al., 2012). Maybe this is because females are not willing to share their personal information on social media due to privacy issues. From these valid users, we obtained 2,130,004 tweets in total, and 491,439 (23.07%) are geo-tagged tweets. The number of available users for each category is shown in Figure 8. There are numbers of Twitter users who disable the access from public. Therefore, the number of users with Twitter access is less than the original number of users.

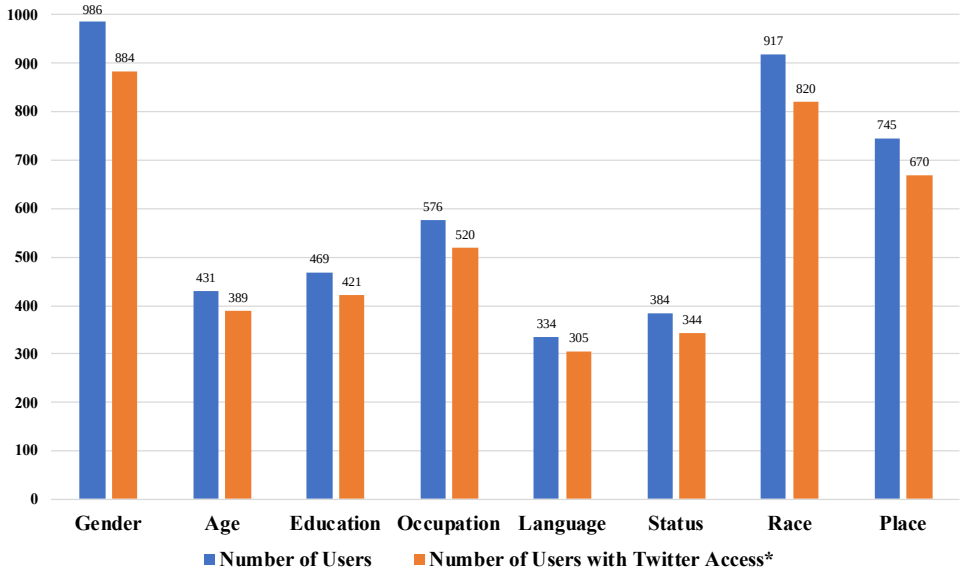


Figure 8 Number of users collected for each feature from Facebook

*: Twitter access means that their Twitter accounts allow public access.

3.1.2 National Institutes of Health (NIH) Data

The dataset utilized in this paper is from an influenza surveillance survey support by the National Institutes of Health. The purpose of this survey is to understand the travel mobility behavior at the individual level and discover the relationship between individuals' travel pattern and influenza incidences. There are more than 2200 participants were recruited from the urbanized areas of Western New York region. The survey was conducted from October 2016 to May 2017 which is the influenza season in the Western New York region, and participants provided three kinds of information. The first information is the socio-demographic information, including home and work places, gender, age, race and ethnicity, and number of people in the household. The second one is that each participant reported up to five most frequently visited places (by choosing the exact street address on Google Map) and if they were ill every week. The third kind of data is GPS trajectory data recorded by a smartphone app. Unlike the explicit GPS collect apps which need to run in front of the screen and battery consuming, the smartphone app can run at the background of the smartphone and powering saving. However, this compromises the frequency of data collection. There are two versions of smartphone apps, for Android and IOS phones, separately. For the app for Android phone, it records a point for every two hours. For the app for IOS phones, it records a point when it detects a significant location changing (when the user's position changes for 500 meters or more (Apple, 2018)). Since the GPS trajectory points are two spares for Android phones which recorded every two hours, we only utilize GPS trajectory points collected by IOS phones (1445 participants are using IOS phones).

This data source is similar to Household Travel surveys with GPS data, however, obvious differences are also existing. First, instead of complete one-day or multi-day travel diary, this dataset contains up to 5 frequent places every week. Second, the GPS trajectory points data is not as dense and uniform as explicitly GPS trajectory data. When participants have significant

movement detected, the GPS points are roughly uniform, and the interval between two adjacent moving GPS points is approximately 5 minutes. The time difference between two significant movements is random. This is the reason why our data set is in between explicitly and inexplicitly data as talked in Session 1.2 and the uniqueness this dataset.

The first challenge of this research is to identify trip ends for each trip from GPS trajectory data. Since the uniqueness of our dataset, the existing density-based methods (Hariharan and Toyama, 2004, Ye et al., 2009, Gong et al., 2012) are not eligible, and the rule-based methods (Tang and Meng, 2006, Palma et al., 2008, Thierry et al., 2013) are not appropriate, we proposed our own method to detect trip ends. "Haccuracy" represents horizontal accuracy of GPS data. If iPhone is connected to WIFI, the range of "haccuracy" is from 65 to 165m. It is typical to get poor accuracy (1000m) at the beginning of a trip, the hardware takes some time to get the accuracy. Then it gets better in a few seconds or more, and it can get as good as 5m accuracy.

Algorithm 1 Trip Ends Identify Algorithm

GPS points: $l_i = (lat_i, log_i)$, recording time: t_i, haccuracy: h_i
For participant in all PARTICIPANTS
    For day in all TRAVELED DAYS
        SORT GPS points according to recording time
            $TripID = 1$
            IF $t_i - t_{i-1} > 10\ min$
                $TripID = TripID + 1$
            ELSE IF $h_i > 1000$
                $TripID = TripID + 1$

Figure 9 Trip Ends Identify Algorithm

After we obtained all trip ends, we need to match these trip ends with reported frequent places. Trip ends is assigned to the nearest five frequent places and the distance needs smaller than $(0.5 \times haccuracy + 50)$m, if any. There are 3,139,453 trip ends are detected, and there are 1,733,538 (55.22%) out of all trip ends are matched with frequent places including home and work. There are 51.31% trip ends are 'Home' and 11.54% trip ends are 'Work' on weekdays, and 40.78% trip ends are 'Home' and '1.5%' trip ends are 'Work' on weekends. This indicates more work-related trips on weekdays and more home-related trips on weekends which is the same as common sense.

Figure 10 shows distributions and basic statistics of number unique visit places for reported frequent places and unreported places, and 'Home' and 'Work' places are excluded from frequent places. The red histogram shows that people visited frequent places for each individual not vary too much, and most people have around 20 frequently visiting places. However, people visit a higher number of unreported places than frequent places within 20 weeks.

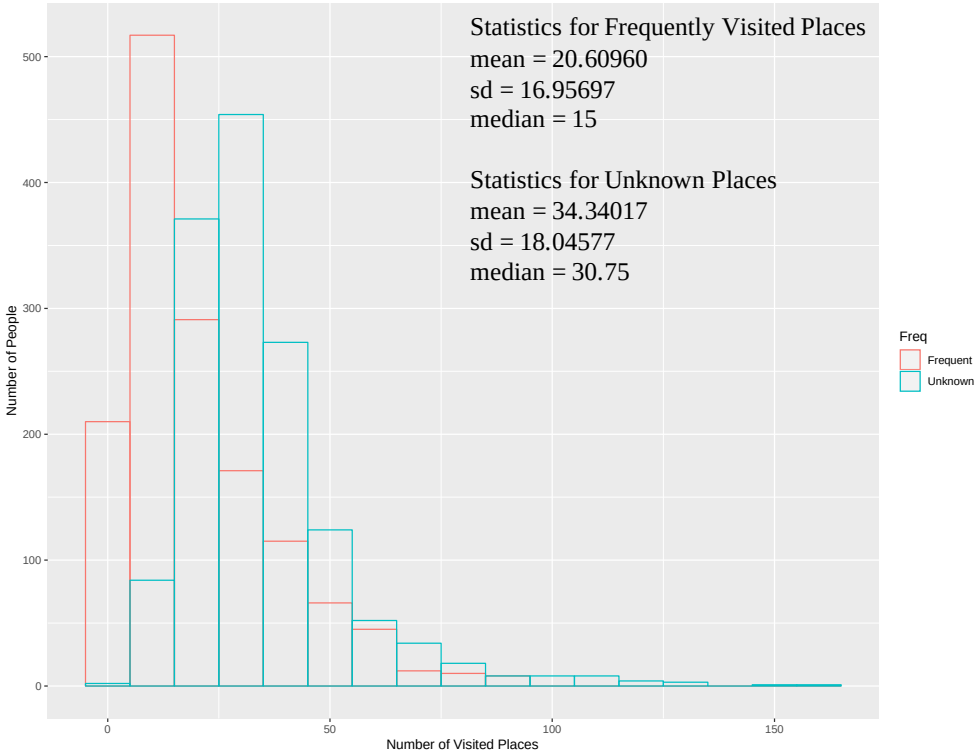


Figure 10 Distributions of Number of Unique Frequently Visited Places (reported) and Unique Unreported/Unknown places.

Distribution of weekly visit times for reported frequent places and unknown places is shown in Figure 11. If each unknown place is treated as a unique place, the weekly frequency of visiting that place is 1. There we employed density-based spatial clustering of applications with noise (DBSCAN) to group locations which are near to each other. From the plot, most unknown places are visited less than or equal to 1 time a week. There still some unknown places are visited more than 1 times so that these are maybe unreported frequently visited places. For reported frequent places, most of the places are visit 1 to 2 times a week, and there are a lot of people visit some frequent places more than 3 times a week. These places maybe schools where parents pick up and drop off their children, restaurant where people buy breakfast or lunch regularly, etc.

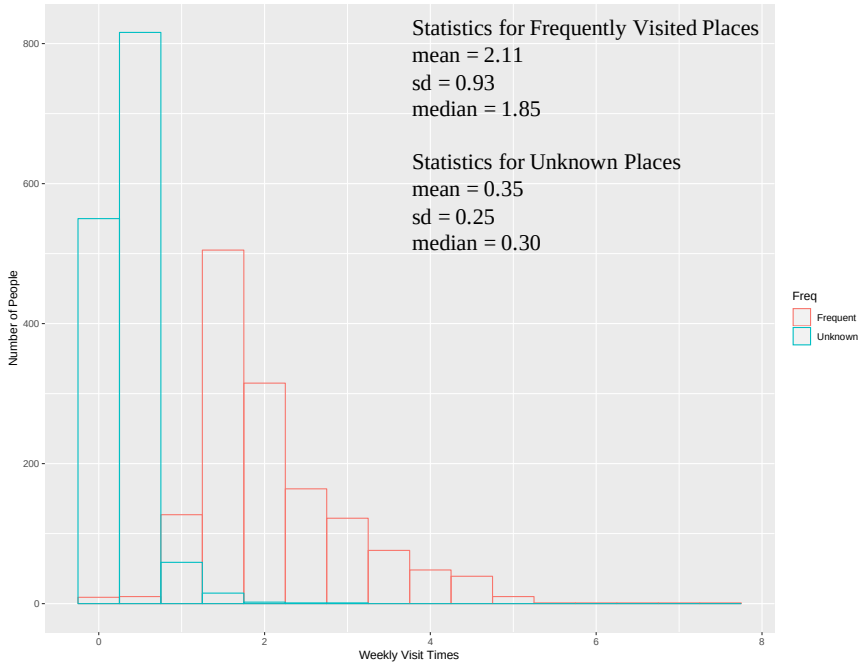


Figure 11 Distribution of Average Weekly Visit Times for Reported Frequent Places and Unknown Places

Table 2 Activity Type Distribution by Gender

Activity Type Distribution	Gender (Total)		Gender (per Person)	
	Male (451)	Female (992)	Male (per person)	Female (per person)
Home	117,593 (51%)	252,094 (52%)	260.74	254.13
Work	30,637 (13%)	52,532 (11%)	67.93	52.96
School	5,216 (2%)	13,958 (3%)	11.57	14.07
Shopping	12,943 (6%)	39,210 (8%)	28.70	39.53
Recreation	47,925 (21%)	97,996 (20%)	106.26	98.79
Personal Business	11,133 (5%)	25,821 (5%)	24.69	26.03
Transportation	385 (0.2%)	988 (0.2%)	0.85	1.00
Other	5,208 (2%)	6,817 (1%)	11.55	6.87

Table 3 Activity Type Distribution by Age

Activity Type Distribution	Age (Total)				Age (per Person)			
	13-17 (53)	18-35 (476)	35-65 (789)	66 or older (110)	13-17	18-35	35-65	66 or older
Home	12,927 (52%)	118,405 (51%)	208,260 (51%)	30,095 (55%)	243.91	248.75	263.95	273.59

Work	609 (2%)	29,496 (13%)	51,047 (13%)	2,017 (3%)	11.49	61.97	64.70	18.34
School	2,855 (12%)	6,993 (3%)	8,601 (2%)	725 (1%)	53.87	14.69	10.90	6.59
Shopping	1,128 (5%)	15,316 (6%)	30,719 (8%)	4,990 (9%)	21.28	32.18	38.93	45.36
Recreation	6,111 (25%)	49,050 (21%)	78,356 (19%)	12,404 (22%)	115.30	103.05	99.31	112.76
Personal Business	388 (2%)	10,917 (5%)	22,194 (5%)	3,455 (6%)	7.32	22.93	28.13	31.41
Transportation	39 (0.1%)	465 (0.2%)	781 (0.2%)	88 (0.2%)	0.74	0.98	0.99	0.80
Other	570 (2%)	2,480 (1%)	8,370 (2%)	605 (1%)	10.75	5.21	10.61	5.50

Table 4 Activity Type Distribution by Race

Activity Type Distribution	Race (Total)				
	American Indian or Alaska Native (9)	Asian (38)	Black or African American (24)	White (1341)	Other Race (14)
Home	1,971 (59%)	7,941 (55%)	5,434 (50%)	351,011 (51%)	3,330 (54%)
Work	390 (12%)	2,328 (16%)	742 (7%)	79,178 (12%)	531 (9%)
School	114 (3%)	452 (3%)	217 (3%)	18,024 (3%)	367 (6%)
Shopping	211 (6%)	917 (6%)	906 (8%)	49,596 (7%)	523 (8%)
Recreation	521 (16%)	1,821 (13%)	2,852 (26%)	139,586 (20%)	1,141 (18%)
Personal Business	114 (3%)	740 (5%)	296 (3%)	35,543 (5%)	261 (4%)
Transportation	1 (0.03%)	25 (0.1%)	176 (2%)	1,166 (0.2%)	5 (0.08%)
Other	1 (0.03%)	264 (2%)	180 (2%)	11,551 (2%)	29 (5%)
Activity Type Distribution	Race (per Person)				
	American Indian or Alaska Native	Asian	Black or African American	White	Other Race
Home	219.00	208.97	226.42	261.75	237.86
Work	43.33	61.26	30.92	59.04	37.93
School	12.67	11.89	9.04	13.44	26.21
Shopping	23.44	24.13	37.75	36.98	37.36
Recreation	57.89	47.92	118.83	104.09	81.50
Personal Business	12.67	19.47	12.33	26.50	18.64
Transportation	0.11	0.66	7.33	0.87	0.36

Other	0.11	6.95	7.50	8.61	2.07

Activity type distribution is also analyzed in this research. Activity types only focus on types of reported frequent places. We employed Google Places API to retrieve categories (Cui et al., 2018b) of places and group detailed types into activity categories, including “School”, “Shopping”, “Recreation”, “Personal Business”, “Transportation” and “Other”. The detailed categories dividing rule is shown in the Appendix. Table 2, 3 and 4 show activity type distribution for gender, age, and race, separately. The number in parentheses indicates the number of people in this category. The statistic for each socio-demographic is reasonable. For example, females have more “Shopping” activities, people younger than 18 years old conduct more “School” activities and less “Work” and “Personal Business” activities.

An example of travel pattern of a participant is illustrated in Figure 12. The darkness of red represent the frequent visiting a place, it indicates the high probability of stay at home before 10 am. This person remained at workplace from 10 am to 18:30 pm. Except go back to home, two residential places also have high chance to be visited after work, maybe these relatives’’ house or friends’ home. Other recreation activities, personal business, and shopping activities are also frequently conducted by this participant. As a typical day, we cannot expect an individual can visit all the high frequent places, however, we can generate a location chain for a typical day. This generates typical day may have different location chains since individuals travel patterns vary time of day and day of week. Moreover, people travel patterns and frequently visited locations are evolved over time (Habib and Miller, 2008).

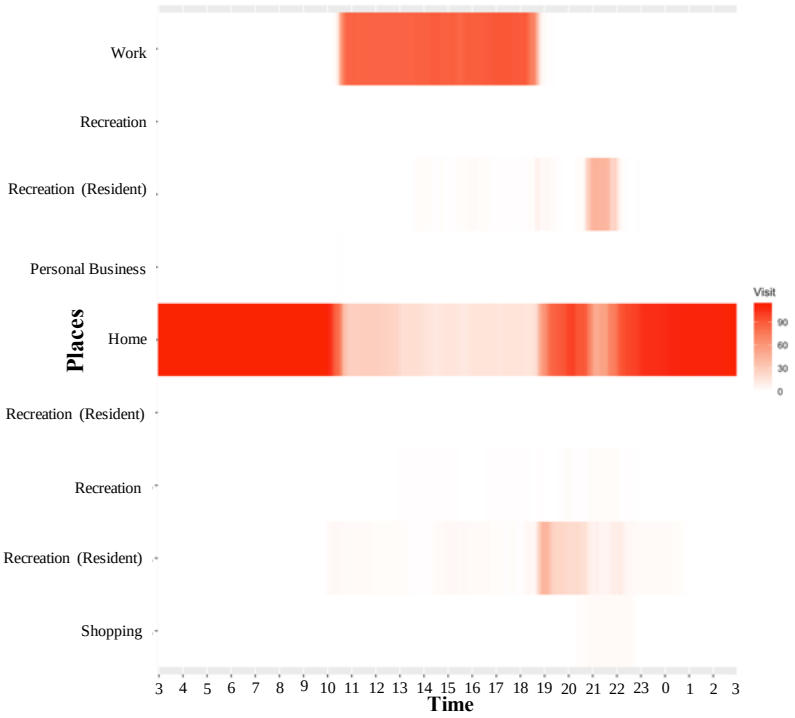


Figure 12 Heat map of an individual’s reported frequent places by visiting counts in the whole survey period.

3.2 Numerical Examples

3.2.1 Twitter Demographics - Gender

In this section, we examined several feature combinations, and their accuracies are shown in Figure 3. For the tweet contents, we first tokenized all the words. Words appearing more than 10,000 times were removed, and words with the frequency less than 100 times were removed as well. Words which appear too frequently are common words such as personal pronouns, link verbs, conjunctions and prepositions, such as “San Francisco”, “he”, “him”, “on”, “at”, etc. In addition, words which appear scarcely are usually not meaningful for models, since they are too rare to have representation. Figure 13 shows that tweet contents are not performing as well as hashtags as features. The reason is that hashtags can describe users’ interests on a specific topic. Even though including connection as features, we cannot improve the accuracy much.

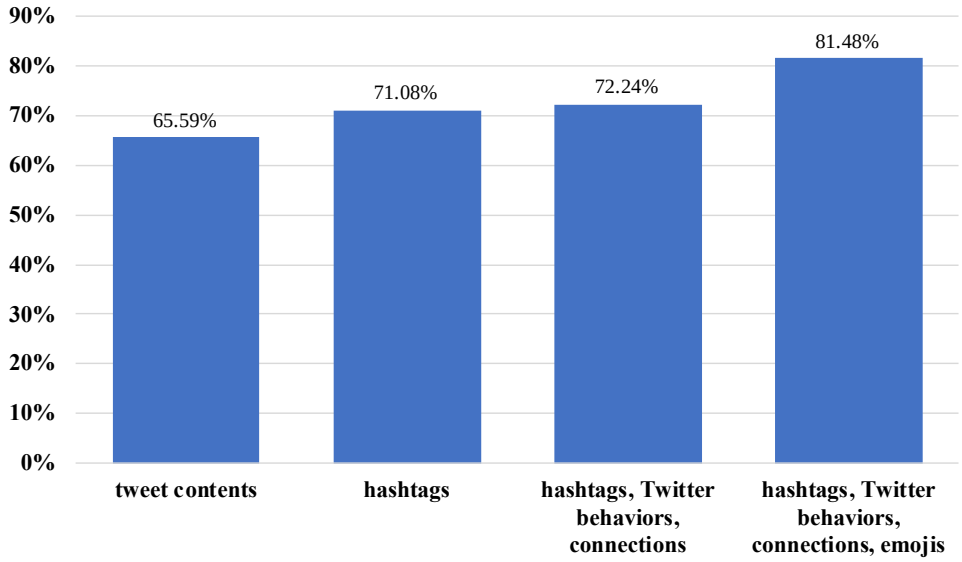


Figure 13 Accuracy of models with different features for predicting the gender

Surprisingly, it is found that emojis are the most important features to indicate gender, and from Figure 5, the accuracy is improved significantly. This accuracy is comparable with the models in the literature review section. There are total 1644 emojis, and we separated them into 67 small categories. These 67 small categories belong to faces, people, sports, bodies, animals, plants, foods, places, transports, times, weather, activities, phones and computers, symbols, and flags. For people-related emojis, there are emojis about roles and people doing sports that can distinguish genders. For example, there are emojis for woman, man, grandmother, grandfather, female painter, male painter, man biking, and woman biking, etc. Table 5 shows the usage of emojis between males and females. And it is revealed that males and females are prone to use emojis with the gender the same as their own gender. Also, the number of emojis used per female is almost twice as many times as per male.

Table 5 Number of gender related emojis posted by different genders

	Gender	Male users	Female users
Emoji	Male-related	5708	2223
	Female-related	3642	6109
	Total	9350	8332
	Per person	15.87	28.24

It is worth noting that past models with high prediction accuracy typically used name-related features, especially first names. In these studies, name-related features are obtained from social media platforms such as Facebook before the limitation of user profile retrieval, or from manual labeling by employing Amazon Mechanical Turk, or inferred from user screen names. Each method has limitations. First, mining from Facebook is not working any longer. Second, manual labeling is slow, and buying from merchants is costly. Third, inferring from Twitter users' screen names is not accurate. It is because some users do not use their real names as screen names, And some usernames only contain part of real names; Moreover, some usernames are a combination of first names and last names. Therefore, in this study, we do not any include name-related features in models. And this can avoid some issues from inaccurate name inference.

3.2.1 Twitter Demographics - Age

Age is an important characteristic of a traveler because most people follow different travel patterns according to their ages. For example, if people are under 22, mostly they are students and go to school, and if people are around 40 years old, probably they go to work and perform more family-related trips. It has been discovered that gender is difficult to predict by using language features since people do not use stereotypical language associated with their gender (Nguyen et al., 2014). However, unlike gender, there are certain language patterns corresponding to the certain life stages and ages. As aforementioned, young people like using slang words, and older people's posts are well organized. Moreover, people in different ages post about different topics and different life focuses.

In this study, we divided ages into 3 categories, younger than 30 years old, ages between 30 to 45 years old, and older than 45 years old. The reason to divide ages like this is to make data entries in each category more balanced. The population of Twitter users is younger than the population of the real world. Therefore, we set the older group for users older than 45 years old. The following bullet points illustrate the top 9 meaningful words for each category, which show people's focus in each age group.

- Age<30: work, game, friends, aquarium, school, guys, party, movie, dude
- Age 30-45: work, video, park, airport, family, bar, trump, business, Disneyland
- Age>45: trump, artwork, family, life, baby, dad, kids, children, mom

For people under 30 years old, they focus on work, game, and friends. Since they graduated from school not too long ago, they still want to hang out with their friends and use slang words. In addition, some of them are still at school, so they also post things about their school life. People between 30 to 44 years old also tend to focus on their jobs, so there are words like work, business, and airport for their business travels. Most of them would like to discuss activities about their families (e.g. trips to Disneyland). Moreover, they also follow political affairs and try

to relax at bars after work. For people older than 45 years old, they are more focused on their own lives and their kids. They like posting about their children and referring to themselves as dad or mom. And most of them tend to follow political topics. In this section, we first applied several traditional machine learning methods and found out that they did not work very well in this task. Further, we implemented deep learning (Cui et al., 2018a) which is a strong natural language processing tool. We constructed a neural network with one input layer, three hidden layers, and one output layer. Between each layer, the activity function is Rectified Linear Unit (ReLU) function, and the cost function is softmax cross entropy with logits. The number of hidden nodes for each hidden layer is 4096, 1024 and 256, respectively. The model performance is shown in Table 6. Most precisions and recalls are larger than 90%. Therefore, deep learning is an appropriate method for age prediction using language features.

Table 6 Model performances of age bin (deep learning), ethnicity (random forest) and education level (random forest)

Age Bin	**Precision**	**Recall**	**F1 Score**
Age<30	95.31%	96.06%	0.957
Age30-44	93.02%	88.89%	0.909
Age>=45	91.18%	93.94%	0.925
Average	93.17%	92.96%	0.931
Ethnicity	**Precision**	**Recall**	**F1 Score**
Asian	69.29%	78.22%	0.735
Black	88.95%	85.96%	0.874
White and Others	76.55%	67.41%	0.717
Average	78.26%	77.20%	0.775
Education level	**Precision**	**Recall**	**F1 Score**
Below Bachelor	83.90%	95.09%	0.891
Bachelor	95.00%	77.64%	0.854
Graduate	94.39%	98.68%	0.965
Average	91.10%	90.47%	0.908

3.2.3 Twitter Demographics - Ethnicity

Different ethnicities have different cultures, which may result in different travel behaviors and frequency of visiting places. People also pay more attention to the news about their ethnic group. For example, Asians like shopping in supermarkets which sell their ethnic group's traditional food. Also, people of each ethnicity likely prefer to visit restaurants with their special cuisines.

Image recognition is an easy way to identify the ethnicity and gender of a Twitter user as well. However, this method contains several shortages. First, Twitter users may not use their real picture as their profile photos. Some users use photos of their favorite celebrities, and some users use their pets, landscapes, and illustrations. Moreover, numbers of users' profile photos contain more than one person, which makes it difficult to identify who is the actual user. Second, the image recognition technique is sometimes not very reliable. Therefore, this study does not utilize

any photo features. However, for ethnicity identification, there are still several useful features including emojis and language features. Emojis include ethnicity-related information, for example, Asian foods, mosques, and national flags. Moreover, Twitter can recognize language used for each post and this property can extract by using Twitter API. We assume individuals belonging to each ethnicity can speak and post tweets in the corresponding languages.

In the ethnicity prediction task, we divided Twitter users into 3 categories, which are “Asian” (213 people), “Black” (95 people) and “White and Others” (512 people). We manually labeled “Asian” and “Black” from their profile photos. For the rest of users, due to lack of ground truth, we could not identify more specific races. Therefore, we grouped them as “White and Others”. We found that Random Forest is the best choice. We also balanced the dataset and employed 10-fold cross-validation. There are total 2939 features including Twitter behavior features, tweet text features, and connection features. The plot for the number of variables in models and errors are shown in Figure 14. We chose 184 features as the number of features and selected the top 184 important features to run the model again. There are 46 language features in total, and 25 of them are within the top 184 important features. Others that are not included in the top important features are some languages which are either used by too many users, such as English, or seldom used, for example, Ukrainian and Tamil.

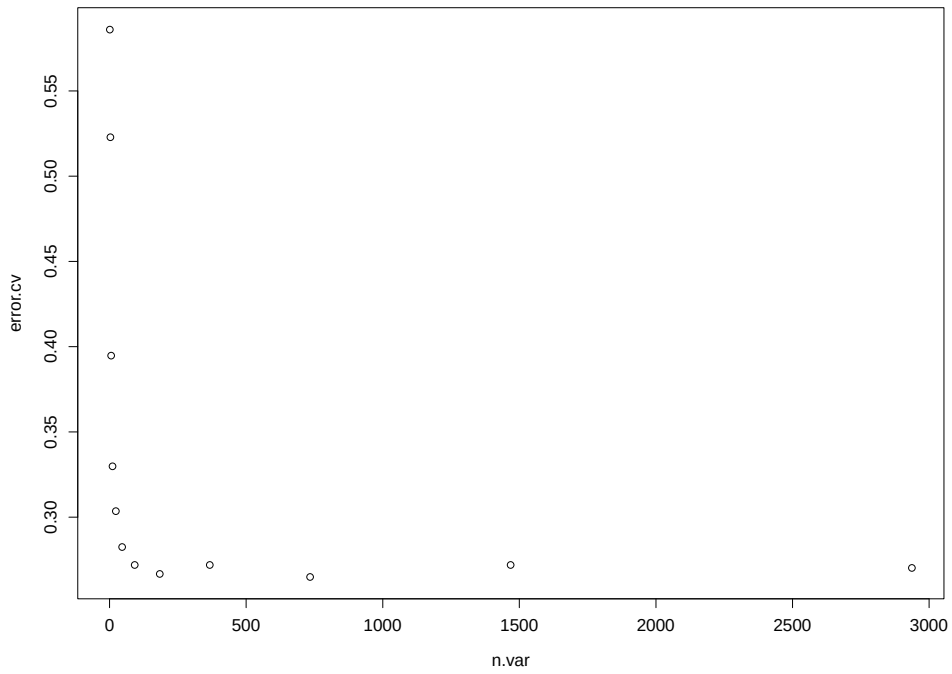


Figure 14 Number of variables vs. error for Random Forest

The final classification results are shown in Table 6. It shows that “Black” prediction achieves more than 85%, which outperforms the others. And the accuracies of “Asian” and “White and Others” are around 70%. This may be because, first, “White and Others” category contains multiple ethnicities which are very complex. Second, Asians may speak one or multiple

languages since some of them are immigrants in the Bay Area. It makes their tweet language complicated to be identified. It is worth noting that past studies typically employed last names to predict ethnicity and gender. However, for the same reason as gender, we do not employ the name feature for ethnicity in this study. The performance of these features (tweet post features, tweet behavior features, emojis, and languages) with the Random Forest model is reasonable.

3.2.4 Twitter Demographics - Education level

Education level is an important characteristic for individuals. First, education levels are associated with age and vice versa. For example, a 20-year-old person who is a college student has a high probability that his/her highest education level is high school. Second, education levels are highly related to income levels, positions, interests, living circle and so on. These differences influence people's living patterns, travel behaviors, and frequency of visiting places, etc.

This study divides education level into 3 categories, "below bachelor", "bachelor", and "graduate". Note that the education level here represents the highest education level. We first explored tweet text features by using a deep learning approach. However, it does not work well. Then we replaced tweet text features with hashtags, which are considered as concentrated posts. Beside hashtags, we also included Twitter behavior features and connection features. Moreover, the model also takes into account the number of languages the user used to post. We assumed that individuals with a higher education level can master more languages.

The performance results of the Random Forest model, which is the best among different methods, are shown in Table 6. As one can see, the accuracy of predicting "below bachelor" is lower than others. Maybe this is because the population in the "below bachelor" category is more complex. This category includes users that are college students but not graduated yet and people that have only completed degrees under the bachelor level. Therefore, the ages of these users could range from the young to the senior who did not go to college when they were young.

The relationship between education level and the number of languages mastered has also been analyzed, and the results are shown in Figure 15. Individuals who only master one language are almost even in the three categories. More people with a higher education degree could utilize more than 3 languages. Surprisingly, the descending order of proportion of two languages used is "below bachelor", "bachelor", and "graduate". This indicates that even though the education level is not very high, there are still a lot of bilingual people. One possible reason could be that the Bay Area has a large number of immigrations.

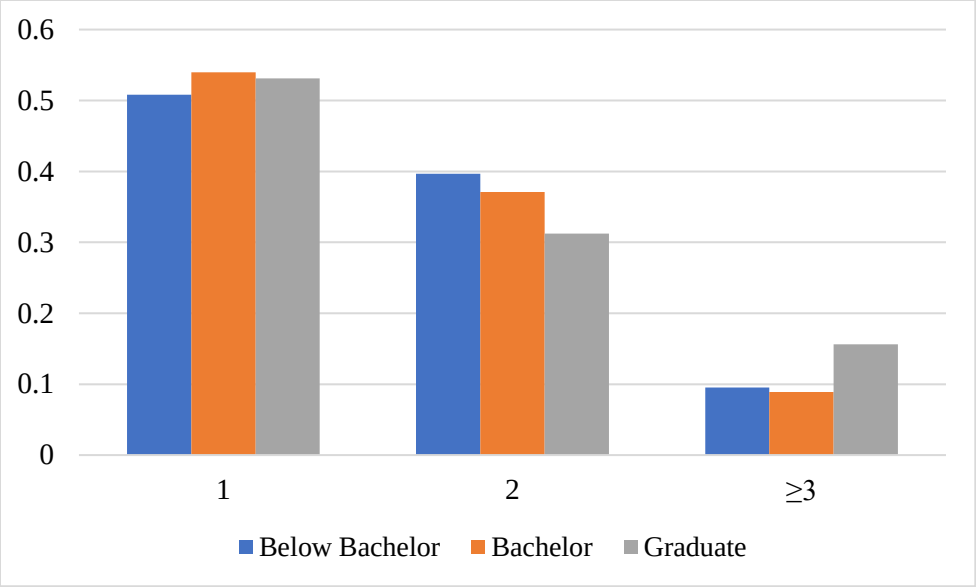


Figure 15 Plot of Education level vs. Number of Languages Mastered

3.2.5 Twitter Demographics - Resampling and comparing with CHTS data

All the efforts and models stem from the sake of resampling data from Twitter and ensuring the demography distributions of Twitter users are the same as the one of the real population. For example, in reality, the total number of males is slightly less than females. However, Twitter users in the United States include slightly more men than women. In addition, the gender distribution in the data we collected from Twitter is also not the same as for the whole Twitter population. It appears highly unbalance due to privacy and safety concerns. The same phenomenon is also appearing in other socio-demographic attributes, such as age, ethnicity, and education levels.

In the previous section, there are 4 classification tasks, which are gender ("male", "female"), age ("younger than 30", "aging between 30 and 45 years old", and "older than 45"), ethnicity ("Asian", "Black", "White and Others") and education ("below bachelor", "bachelor" and "graduate"). Therefore, there is a total of 54 (2x3x3x3) types of socio-demographics. The CHTS survey data only contains 52 types, and the missing scenarios are "black male with a graduate degree and younger than 30 years old", and "black female with a graduate degree and younger than 30 years old".

In this study, we resample Twitter data according to the socio-demographic distribution of the CHTS survey data. The equation of resampling is shown below.

$$\frac{x_{i,j,k,l}^{soical}}{\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} x_{i,j,k,l}^{soical}=p_{i,j,k,l}^{survey}} = p_{i,j,k,l}^{survey} \tag{5-1}$$

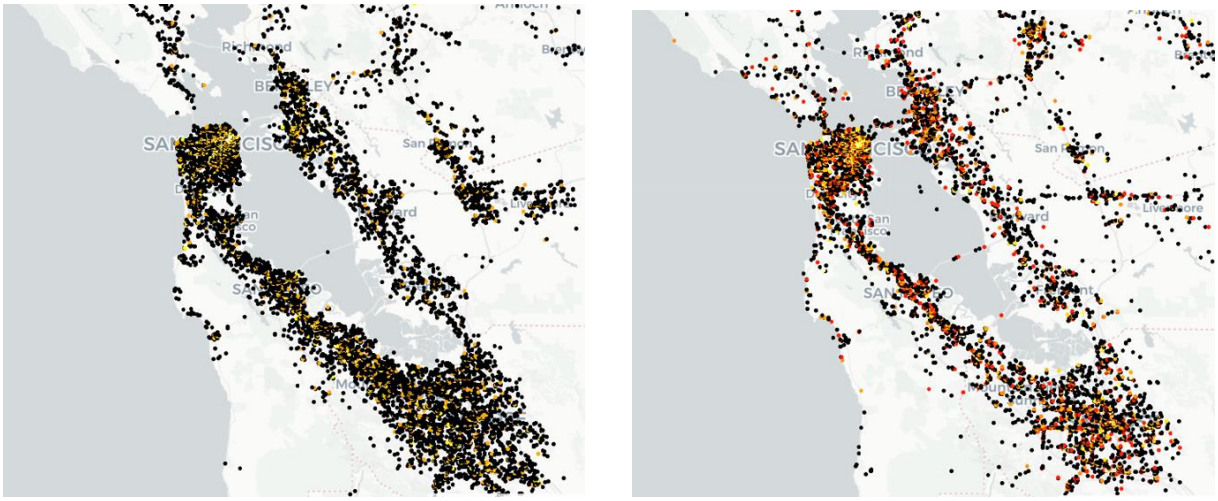
where $i \in \{male, female\}$, $j \in \{age < 30, 30 \leq age \leq 45,\ age > 45\}$, $k \in \{Asian, Black, White\ and\ Others\}$ and $l \in \{below\ bachelor, bachelor, gradute\}$, $x_{i,j,k,l}^{social}$

represents the number of social media users with socio-demographics of gender i, age j, ethnicity k and education level l. $p_{i,j,k,l}^{survey}$ represents the proportion of respondents with corresponding socio-demographics in survey data, and it is calculated by Equation 2 below.

$$p_{i,j,k,l}^{survey} = \frac{x_{i,j,k,l}^{survey}}{\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} x_{i,j,k,l}^{survey}} \quad (5\text{-}2)$$

Table 7 Distribution comparisons of each socio-demographic of social media data before and after resampling

Socio-demographics		Social Media Data		Survey Data
		Before Resampling	After Resampling	
Gender	Male	66.63%	51.03%	50.33%
	Female	33.37%	48.97%	49.67%
Age	<30	17.27%	22.93%	24.92%
	30-45	80.30%	46.32%	42.59%
	>45	2.43%	30.75%	32.48%
Ethnicity	Asian	23.94%	10.58%	12.74%
	Black	12.12%	7.61%	2.81%
	White and Others	63.94%	81.81%	83.45%
Education Level	Below Bachelor	12.42%	33.21%	29.66%
	Bachelor	79.09%	46.74%	43.96%
	Graduate	8.49%	20.05%	26.38%



(a) (b)

Figure 16 Visualized survey data (a) and resampled social media data (b)

Table 7 shows the distribution comparisons of each socio-demographic of social media before and after resampling. According to the table, some proportion of socio-demographics are changed dramatically, for example, Male, Female, Age 30-45, Age >45, and Bachelor. After resampling from Twitter data according to the socio-demographic distribution of 2009 California Household Travel Survey (CHTS), we geographically plotted the survey data and Twitter data, separately as shown in Figure 16. Each dot in Figure 16(a) represents a trip end location or an activity location in CHTS data, and each dot in Figure 16(a) represents a geo-tagged tweet. Colors of dots range from black to red which represents the visiting frequency from low to high. There are 7330 participants in survey data, whereas only 330 users[1] in social media data. Nevertheless, social media data is comparable with survey data because it contains much more data entries per user. The visiting patterns of both two plots overlay each other, indicating that resampled social media data captures most trip end locations in survey data. Most dots in Figure 16(a) are black, indicating that the visiting frequency of these places is very low for CHTS data. It is anticipated since the CHTS data only contains one-day survey records so that the visiting places are sparse. On the contrary, there are a lot of yellow, orange and red dots in Figure 5-6(b) for social media data. For each Twitter user, we collected up to 3000 tweets from him/her timeline. Therefore, since users can post numbers of tweets at one location, there are more frequently visited locations than survey data.

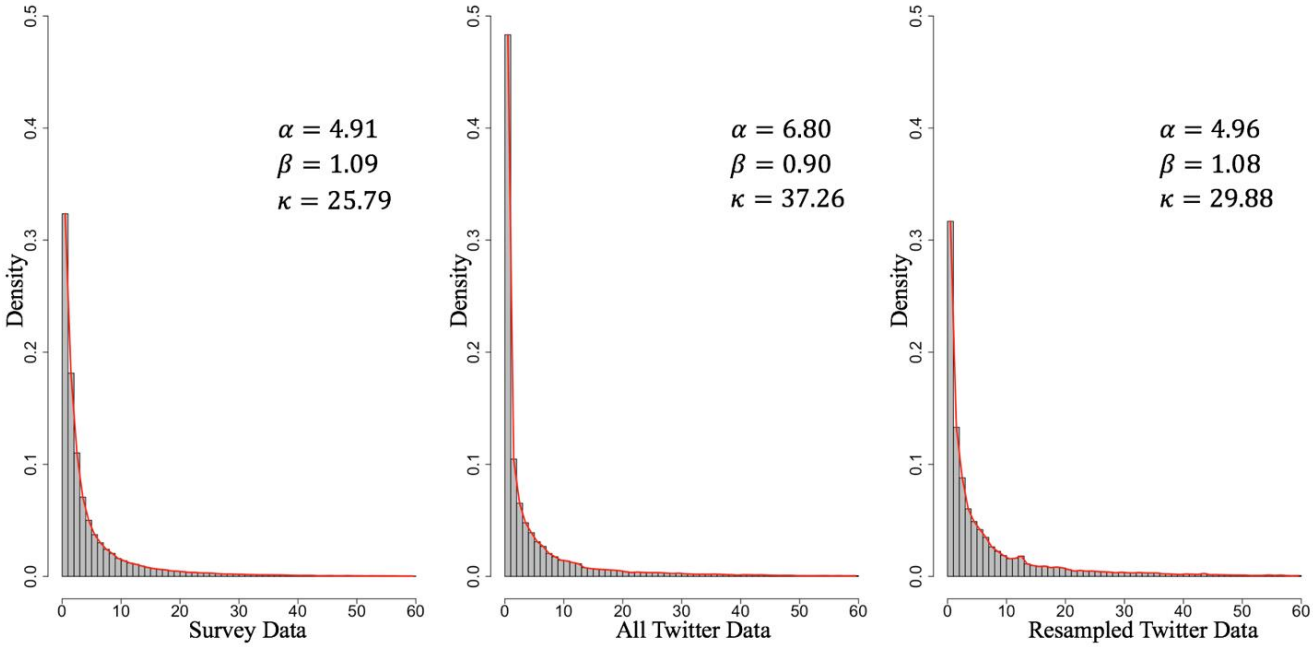


Figure 17 The comparison of trip length among survey data, all Twitter data and resampled twitter data.
The x-axis is the trip distance and the unit is one mile; y-axis is the probability density.

(Gonzalez et al., 2008) found that the distribution of displacements of travelers is can be well represented by a truncated power-law, and the equation of probability density function shows below,

[1] We have 832 unique active Twitter users with geo-tagged tweets in total. However, we need Twitter users who have valid information in gender, age, ethnicity and education. Therefore, the number of Twitter users are screened down to 330 after resampling.

$$P(x) = (x + \alpha)^{-\beta} \exp(-\frac{x}{\kappa}) \quad (5\text{-}3)$$

The comparison of the trip length among survey data, all Twitter data and resampled Twitter data is shown in Figure 17. The x-axis is the trip distance and the unit is one mile; y-axis is the probability density. All the estimated parameters for truncated power-law are shown on the plot. The parameters indicate that resampled Twitter data is more similar to survey data. Moreover, three datasets show heavily right-skewed, and most individuals conducted trips within 30 miles. In addition, Twitter data captures more short distance trips than survey data. Even though there are more short distance trips in Twitter data, the value of β are almost the same. Similar βs indicate similar travel behaviors between survey data and Twitter data. This is also verified in a previous study (Zhang et al., 2017).

Table 8 shows the comparison of parameters of truncated power-law for 4 categories with most people of survey data and resample Twitter data. The similarity of each pair of truncated power-law distribution is calculated by using Bhattacharyya distance. Bhattacharyya distance measures the similarity between two probability distributions, and the value of Bhattacharyya distance represents the similarity of these two probability distributions. This value ranges from 0 to 1. 0 indicates that these two probability distributions are not similar at all, and 1 means that they are exactly the same (Cui, 2016). From this table, we find out the similarity of each pair of truncated power-law distributions are very high, which means the travel distance distribution of survey data and resampled Twitter data are similar to each other. Therefore, the resampled Twitter data are close to the survey data.

Table 8 Comparison of parameters of truncated power-law for the top 4 categories of survey data and resampled Twitter data

Socio-Demographic	Survey Data			Resampled Twitter Data			Similarity
	α	β	κ	α	β	κ	
2-2-2-1	4.38	1.12	22.8	4.93	1.08	27.8	94.92%
1-2-2-1	4.83	1.15	28.6	5.09	1.32	29.75	90.72%
2-1-2-1	4.98	1.21	26.12	5.18	1.14	28.72	90.13%
1-1-2-1	5.18	1.01	29.12	5.22	1.27	29.65	94.70%

* 2-2-2-1 means female-age30-45-bachelor-white and others, 1-2-2-1 means male-age30-45-bachelor- white and others, 2-1-2-1 means female-age<30-bachelor-white and others, 1-1-2-1 means male-age<30-bachelor-white and others

The resampled social media data removed bias from oversample and undersample of original social media data to a great extent. Therefore, using resampled social media data is more reasonable in travel behavior analysis. Moreover, abundant social media data also supply and extend survey data. This kind of data can support survey data in uncovering travel patterns of the real population. It also helps reveal the frequent visiting places for people with different socio-demographics.

3.2.6 Activity Sequence Simulator - Individual Level Validation

In this study, each participant has more than 20 weeks data, so we can do individual level validation. Aforementioned, people travel patterns and frequently visited locations are evolved over time, we cannot use earlier data as training data and later data as validation data. In order to decrease this bias and error, we treat every four weeks as a time period, and select first three weeks' data as training data and last one-week data as validating data.

In this session, we introduce validation results in individual level. According to information of validation data (previous visited place and current time), predict current place and staying duration of current place using proposed methodology. If the predicted current place is same as validation data, it is a right prediction and correctness labeled as 1, otherwise, labeled as 0. The final prediction accuracy is the average of correctness. Mean absolute percentage error (MAPE) also employed to measure accuracy of predicted staying duration. The individual level validation results is shown in Table 9 and Figure 19.

Table 9 Individual Level Validation Results

	Accuracy of Activity-location choice	MAPE of Duration
Level 1	70.83%	24.97%
Level 2	64.18%	23.29%
Level 3	63.19%	22.61%

After we get accuracies for each simulation entries, the final accuracy is the average of these accuracies. For Level 1, Level 2 and Level 3 activity sequence simulators, accuracies are 70.83%, 64.18% and 63.19%, respectively. We cannot simulate people's long-term travel behaviors in one-time simulation, and individuals travel behaviors cannot be captured just by one activity sequence. Validation data has activity sequence never appeared in training data. Moreover, people cannot act exactly as a typical day every day, activities vary day of time and time of day, the accuracy is not expected to be 100%. However, from this result, most activities captured by the simulator, such as fixed and routine activities, including "Home" and "Work". Figure 6-12 shows hourly accuracy, unsurprisingly, high accuracy will be obtained in during the night and accuracy of daytime is relatively lower. Because, activities do not vary too much during night, however, activities may vary a lot during the daytime.

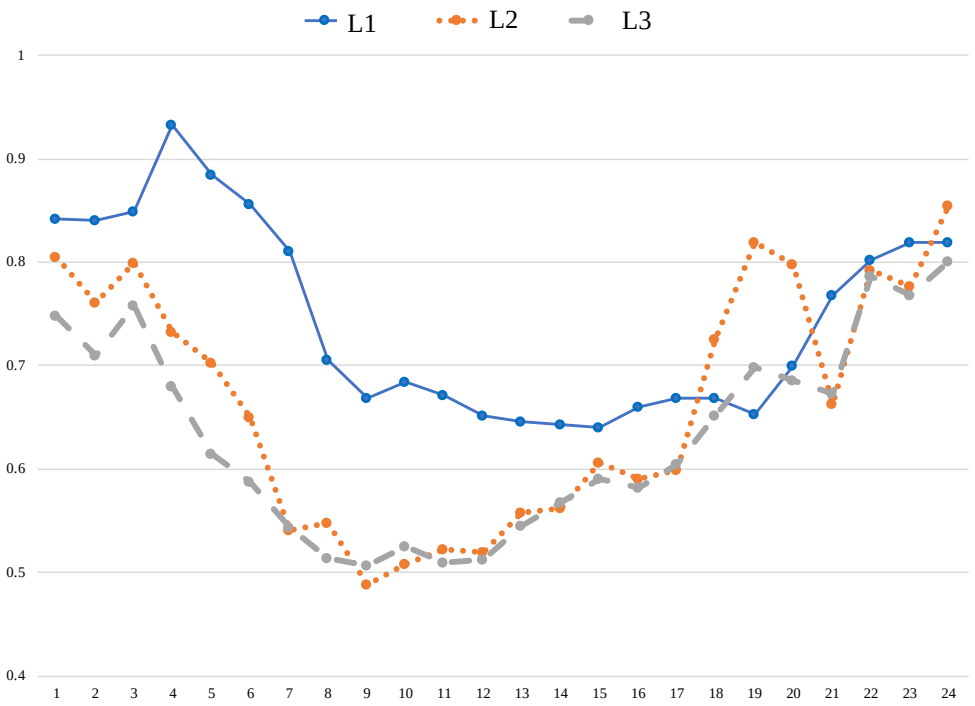


Figure 18 Hourly Accuracy

3.2.7 Activity Sequence Simulator - Aggregated Accuracy

The simulation result also can be validated in an aggregated way. In this session, 300 times monte carol simulation are conducted for one participant. Table 10 shows the distribution of activity type for survey data and simulation result. From this table, difference between the percentage of each activity type is not big. The largest one is "Work" that it is underestimated for 3%. Moreover, Figure 20 displays the distributions of activity length for survey and simulation result. These two lines are also very close to each other with indicate small difference between each other. The percentage of activity duration decrease when the activity duration increase, however, there are a lot of duration of activity are longer than 4 hours. These activities may include "Home", "Work", or visit relatives' and friends' house which belongs to "Recreation". Comparing with the result of the paper of Kitamura et al. (Kitamura et al., 2000), our results improve a lot. Our data is a long-term survey data (20 weeks), this can reduce the imperfection of data. It performs better than the one- or two- day survey data. Table 11 shows the comparison result of staying duration for different activity types. And the table shows simulation model tends to underestimate the duration for fixed activities, including "Home", "Work" and "School" activities. This finding is also found in (Kitamura et al., 2000)

Table 10 Distributions of Activity Type for Survey data and Simulation Result

Distribution of Activity Type	Survey		Simulation		Error
	Frequency	Percentage	Frequency	Percentage	Percentage

Home	369687	51.31%	1698507	52.59%	-1.28%
Work	83169	11.54%	282297	8.74%	-2.80%
School	19174	2.66%	71566	2.22%	-0.44%
Shopping	52153	7.24%	329176	10.19%	2.95%
Recreation	145921	20.25%	639505	19.80%	-0.45%
Personal Business	36954	5.13%	149646	4.63%	-0.50%
Transportation	1373	0.19%	8830	0.27%	-0.08%
Other	12025	1.67%	50299	1.56%	0.11%
Total	720456	100%	3229826	100%	---

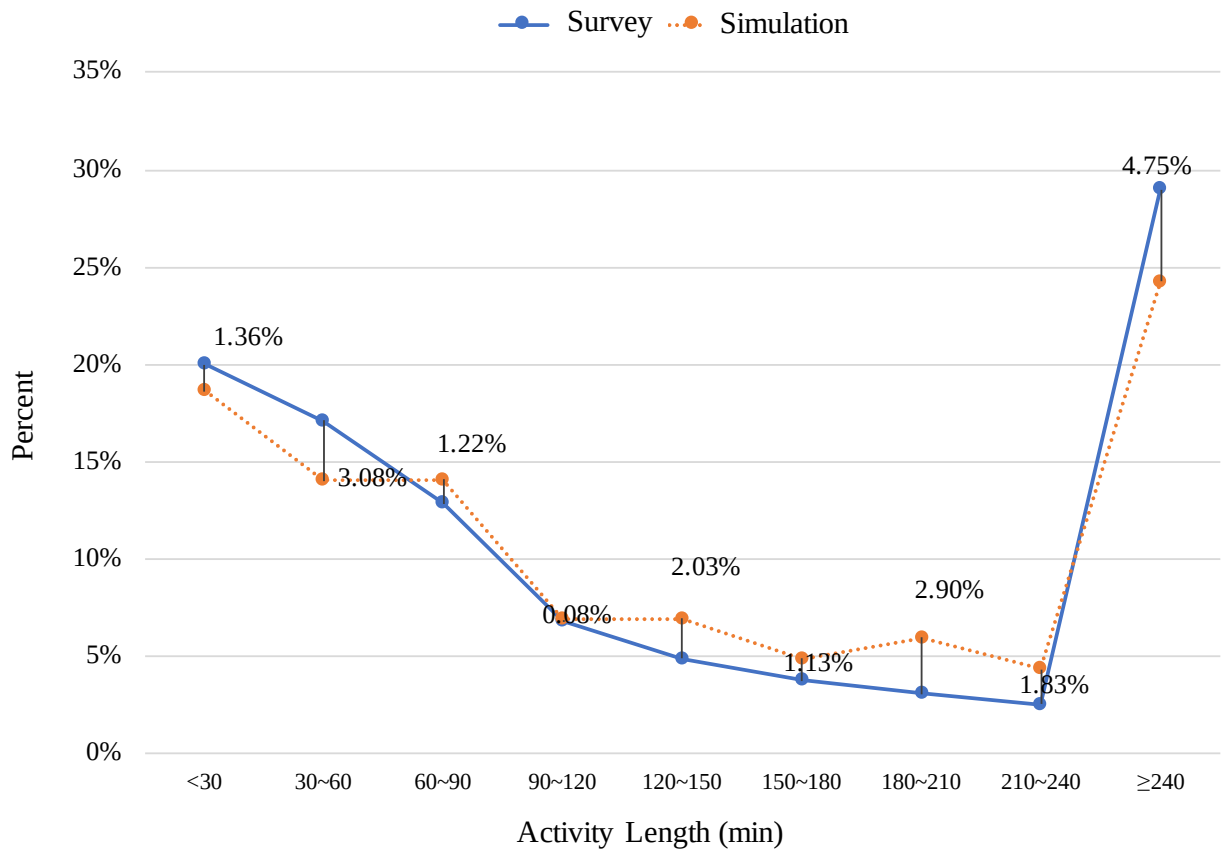


Figure 19 Distributions of Activity Length for Survey data and Simulation Result

Table 11 Comparison of Duration for Different Activity Types for Survey and Simulation Results

Activity Type	Survey (hour)	Simulation (hour)	MAPE
Home	6.61	5.63	14.86%

Work	4.10	3.86	5.83%
School	2.60	2.02	22.20%
Shopping	1.17	1.72	46.71%
Personal	2.56	2.27	11.09%
Recreation	2.19	2.27	3.61%
Trans	2.00	1.61	19.36%
Other	2.36	2.40	1.53%

4. Conclusions

4.1 Conclusions of Twitter Demographic Analysis

This study presents a study on the socio-demographics classification for correcting sampling bias with social media data. In the classification task, we classified 4 types of socio-demographics, which are gender, age, ethnicity, and education level. We employed different features comparing to previous research. We included neither the name-related features in this study nor the image recognition since these features contain bias and errors if used. However, besides traditional Twitter features, such as tweet text, number of followers, number of followings, we also took into account several other features mined by using Twitter API. These features are emojis, languages, and the number of languages used.

For each classification task, there were interesting findings. Gender-related emojis are important in gender prediction. Even though the contents of posts are not effective in the gender classification, it performs well in age prediction. Languages used in posts show a significant effect on ethnicity prediction, as well as emojis with country characteristics. In addition, the number of languages used also acts as an important role in education level prediction.

All the efforts are dedicated to resampling according to socio-demographics. The resampled social media data can capture similar information as survey data, and it also shows several advantages. First, social media data is a type of long-term data, therefore, it is useful in inferring longitudinal travel behaviors and demand. Second, social media data works well in analyzing prevalent places for different groups of people because it is much denser than survey data. Last but not least, social media data also can be utilized as a proxy in CO_2 emission and carbon footprint research.

4.2 Conclusions of Activity Sequence Simulator

In this study, a three-level probabilistic activity sequence simulator is developed with a sustainable and imperfect GPS-based survey. This three-level activity sequence simulator can handle unknown information in different levels: non-unknown places (level 1), all unknown places as one new category (level 2), and individual unknown places (level 3). In order to take unknown places into detailed consideration, DBSCAN is involved to identify the activity

locations. This simulator can capture most activities in a synthetic day. In addition, running this probabilistic simulator multiple times can generate more infrequent activity-travel patterns. The test accuracy in individual level is high during night and relatively lower during daytime due to activity variability at different times. The average individual-level accuracies are 70.83%, 64.18% and 63.19% for level 1, level 2 and level 3 simulation, respectively. From the aggregated validation, the simulation results are similar to survey results in terms of activity type distribution and activity duration. Therefore, the simulation results are close to the real activity sequences, and it proves that our method is appropriate in the activity sequence generation, and it can handle imperfect data.

5. Recommendations

5.1 Recommendations of Twitter Demographic Analysis

In travel behavior analysis and travel demand forecasting study, if social media data are employed in these study, a resampling procedure need to be conducted before research. This will reduces the error and bias introduces from social media data, since the users' population is different from the real population.

In the future, more social media can be collected to develop a more powerful the socio-demographics prediction model. In addition, one could examine more resampling techniques and compare more travel behavior characteristics between resampled social media data and travel survey data.

5.2 Recommendations and Discussions of Activity Sequence Simulator

Our survey data has both similar and different characteristics compared to traditional survey data. Our data contains similar socio-demographic and individual and household information. The first different aspect is participant sampling. Participants of the traditional survey are selected according to the real population, while for our experiment, the participants are randomly recruited. However, we still can resample our data according to match household travel survey. The second difference is that instead of one- or multi-day travel diaries, our data consists weekly reported five frequently visited places. Trip purpose and travel mode are not including in weekly reported frequently visited places (20 weeks), and places never reported are missing in our dataset. However, this kind of imperfect information can be complemented by the long data collection period. The last difference is that the traditional travel surveys assisted by GPS devices consist high frequent GPS points. However, GPS points in our data are low frequently sampled to retain the battery life. Therefore, the traditional travel surveys contain valuable information which is irreplaceable. Our dataset augments traditional travel survey by extending the survey period. Long-term data can capture more information and variations in travel behaviors and patterns.

There are two travel behavior characteristics that are not included in our dataset. One is travel mode, and the other one is trip or activity purpose. Travel mode can be hardly obtained given by low frequent and low precision GPS points. Travel mode is mostly detected by speed-based methods, which determine travel mode according to speed and time (Bohte and Maat, 2009,

Yang et al., 2016). (Jiang et al., 2017) found that the displacement of different travel modes follows different distributions. Subway trips follow the gamma distribution, and exponential distributions are used to fit the displacement bus and taxi trips with different parameters. Machine learning methods are also employed in travel model detection (Wang et al., 2017). These methods mainly depend on the frequency of GPS points. The frequency of GPS data points is at least 1Hz. However, in our dataset, the smallest time interval between two adjacent points is around 5 minutes. Trip purpose is more difficult to analyze than travel mode. It is because trip purpose identification needs to cooperate with both GPS trajectory data and other sources of data, including land use information, temporal information, and socio-demographics. Methodologies for trip purpose prediction can be mainly divided into two categories, rule-based methods (Wolf et al., 2004, Chen et al., 2010), probabilistic-based methods. Recently, social media data also was involved in trip purpose studies (Meng et al., 2017, Cui et al., 2018b). These methods rely on the accuracy of GPS points. If the precision of GPS is too low, the real activity locations or trip ends are far from the recorded GPS points. Thus, there will involve a large uncertain activity location and many nearby POIs which make the inference process difficult. In this study, if there is no position change for 500 meters or more, IOS devices stops recording GPS points. However, there are plenty of POIs within a 500-meter radius circle. In order to obtain accuracy travel mode and trip purse, it is better to increase the frequency to be at least 0.1Hz) and precision of GPS points to be at most 50 meters in haccuracy.

Our model can generate activity sequences for participants who have reported their historical travel data. With regard to unknown travelers, there still two ways to simulate their activity sequences. The first method is to generate travel information according to other datasets collected within the same area (e.g. travel survey data, smart card data, social media check-in data, ridesharing data, etc.). The second way is to associate socio-demographic data with travel behavior to generate activity sequences by employing discrete choice models or machine learning models.

Further, joint activity data plays an important role in future travel survey and travel behavior analysis. For example, one can explore how two household heads plan their trips jointly and analyze how travelers show up at the same locations within similar time spans. For survey data, members of the same household and co-workers or schoolmates are easily identified. Moreover, social media data can be collected for this research. Friendships in social media are easy to be detected, as well same locations they visit together. Combining such two pieces of information, one can construct a detailed social network with information about household members, co-workers, and friends. However, how to link social media data with travel survey data is still a challenging task as well.

List of Figure

List of Table

Reference

APPLE. 2018. *Using the Significant-Change Location Service* [Online]. Available: https://developer.apple.com/documentation/corelocation/getting_the_user_s_location/using_the_significant-change_location_service [Accessed].

BIERLAIRE, M., CHEN, J. & NEWMAN, J. 2013. A probabilistic map matching method for smartphone GPS data. *Transportation Research Part C: Emerging Technologies,* 26**,** 78-98.

BOHTE, W. & MAAT, K. 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies,* 17**,** 285-297.

CHEN, C., GONG, H., LAWSON, C. & BIALOSTOZKY, E. 2010. 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,* 44**,** 830-840.

CUI, Y. 2016. *Behavior-based Traveler Classification Using High-Resolution Connected Vehicles Trajectories and Land Use Data.* University at Buffalo, SUNY, USA.

CUI, Y., HE, Q. & KHANI, A. 2018a. Travel Behavior Classification: An Approach with Social Network and Deep Learning. *Transportation Research Record***,** 0361198118772723.

CUI, Y., MENG, C., HE, Q. & GAO, J. 2018b. Forecasting current and next trip purpose with social media data and Google Places. *Transportation Research Part C: Emerging Technologies,* 97**,** 159-174.

ESTER, M., KRIEGEL, H.-P., SANDER, J. & XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 1996. 226-231.

FINK, C., KOPECKY, J. & MORAWSKI, M. Inferring Gender from the Content of Tweets: A Region Specific Example. ICWSM, 2012.

GONG, H., CHEN, C., BIALOSTOZKY, E. & LAWSON, C. T. 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems,* 36**,** 131-139.

GONZALEZ, M. C., HIDALGO, C. A. & BARABASI, A.-L. 2008. Understanding individual human mobility patterns. *nature,* 453**,** 779.

HABIB, K. M. & MILLER, E. J. 2008. Modelling daily activity program generation considering within-day and day-to-day dynamics in activity-travel behaviour. *Transportation,* 35**,** 467.

HARIHARAN, R. & TOYAMA, K. Project Lachesis: parsing and modeling location histories. International Conference on Geographic Information Science, 2004. Springer, 106-124.

HUDSON, J. G., DUTHIE, J. C., RATHOD, Y. K., LARSEN, K. A. & MEYER, J. L. 2012. Using smartphones to collect bicycle travel data in Texas. Texas Transportation Institute. University Transportation Center for Mobility.

JIANG, S., GUAN, W., ZHANG, W., CHEN, X. & YANG, L. 2017. Human mobility in space from three modes of public transportation. *Physica A: Statistical Mechanics and its Applications,* 483**,** 227-238.

KICKFACTORY. 2016. *The Average Twitter User Now has 707 Followers* [Online]. Available: https://kickfactory.com/blog/average-twitter-followers-updated-2016/ [Accessed June 23, 2016 2016].

KITAMURA, R., CHEN, C., PENDYALA, R. M. & NARAYANAN, R. 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation,* 27**,** 25-51.

KONG, X., LI, M., MA, K., TIAN, K., WANG, M., NING, Z. & XIA, F. 2018. Big Trajectory Data: A Survey of Applications and Services. *IEEE Access,* 6**,** 58295-58306.

LEE, J. H., DAVIS, A. W., YOON, S. Y. & GOULIAS, K. G. 2016. Activity space estimation with longitudinal observations of social media data. *Transportation,* 43**,** 955-977.

LIAO, C.-F., CHEN, C. & FAN, Y. A review on the state-of-the-art smartphone apps for travel data collection and energy efficient strategies. Transportation Research Board 2017 Annual Meeting Compendium of Papers, 2017. 17-00436.

LIAO, Z.-X., LEI, P.-R., SHEN, T.-J., LI, S.-C. & PENG, W.-C. Mining temporal profiles of mobile applications for usage prediction. Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, 2012. IEEE, 890-893.

LIAO, Z.-X., PAN, Y.-C., PENG, W.-C. & LEI, P.-R. On mining mobile apps usage behavior for predicting apps usage in smartphones. Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013. ACM, 609-618.

LIN, L., NI, M., HE, Q., GAO, J. & SADEK, A. W. 2015. Modeling the impacts of inclement weather on freeway traffic speed: exploratory study with social media data. *Transportation Research Record: Journal of the Transportation Research Board***,** 82-89.

MENG, C., CUI, Y., HE, Q., SU, L. & GAO, J. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. Big Data (Big Data), 2017 IEEE International Conference on, 2017. IEEE, 1319-1324.

MISLOVE, A., LEHMANN, S., AHN, Y.-Y., ONNELA, J.-P. & ROSENQUIST, J. N. 2011. Understanding the Demographics of Twitter Users. *ICWSM,* 11**,** 25.

MONTINI, L., PROST, S., SCHRAMMEL, J., RIESER-SCHÜSSLER, N. & AXHAUSEN, K. W. 2015. Comparison of travel diaries generated from smartphone data and dedicated GPS devices. *Transportation Research Procedia,* 11**,** 227-241.

NGUYEN, D., TRIESCHNIGG, D., DOĞRUÖZ, A. S., GRAVEL, R., THEUNE, M., MEDER, T. & DE JONG, F. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014. 1950-1961.

NHTS, N. H. T. S. 2011. Uses of National Household Travel Survey Data in Transportation. Using National Household Travel Survey Data for Transporation Decision Making A Workshop.

NI, M., HE, Q. & GAO, J. 2017. Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems,* 18**,** 1623-1632.

OUIMET, M. C., SIMONS-MORTON, B. G., ZADOR, P. L., LERNER, N. D., FREEDMAN, M., DUNCAN, G. D. & WANG, J. 2010. Using the US National Household Travel Survey to estimate the impact of passenger characteristics on young drivers' relative risk of fatal crash involvement. *Accident Analysis & Prevention,* 42**,** 689-694.

PALMA, A. T., BOGORNY, V., KUIJPERS, B. & ALVARES, L. O. A clustering-based approach for discovering interesting places in trajectories. Proceedings of the 2008 ACM symposium on Applied computing, 2008. ACM, 863-868.

PATTERSON, Z. & FITZSIMMONS, K. 2016. DataMobile: Smartphone travel survey experiment. *Transportation Research Record: Journal of the Transportation Research Board***,** 35-43.
PICORNELL, M., RUIZ, T., LENORMAND, M., RAMASCO, J. J., DUBERNET, T. & FRÍAS-MARTÍNEZ, E. 2015. Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation,* 42**,** 647-668.
POLZIN, S. E., CHU, X. & RAMAN, V. S. 2008. Exploration of a shift in household transportation spending from vehicles to public transportation.
RASHIDI, T. H., ABBASI, A., MAGHREBI, M., HASAN, S. & WALLER, T. S. 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies,* 75**,** 197-211.
REDDY, S., MUN, M., BURKE, J., ESTRIN, D., HANSEN, M. & SRIVASTAVA, M. 2010. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN),* 6**,** 13.
SHEN, L. & STOPHER, P. R. 2014. Review of GPS travel survey and GPS data-processing methods. *Transport Reviews,* 34**,** 316-334.
STATISTA. 2018. *Distribution of Twitter users in the United States as of January 2017, by gender* [Online]. Available: https://www.statista.com/statistics/678794/united-states-twitter-gender-distribution/ [Accessed 2018].
TANG, J. & MENG, L. Learning significant locations from GPS data with time window. Geoinformatics 2006: GNSS and Integrated Geospatial Applications, 2006. International Society for Optics and Photonics, 64180J.
THIERRY, B., CHAIX, B. & KESTENS, Y. 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International journal of health geographics,* 12**,** 14.
VLACHOS, M., YU, P. & CASTELLI, V. On periodicity detection and structural periodic similarity. Proceedings of the 2005 SIAM international conference on data mining, 2005. SIAM, 449-460.
WANG, L., MA, W., FAN, Y. & ZUO, Z. 2017. Trip chain extraction using smartphone-collected trajectory data. *Transportmetrica B: Transport Dynamics***,** 1-20.
WOLF, J., BRICKA, S., ASHBY, T. & GORUGANTUA, C. Advances in the application of GPS to household travel surveys. National Household Travel Survey Conference, Washington DC, 2004.
XIAO, Y., LOW, D., BANDARA, T., PATHAK, P., LIM, H. B., GOYAL, D., SANTOS, J., COTTRILL, C., PEREIRA, F. & ZEGRAS, C. Transportation activity analysis using smartphones. Consumer Communications and Networking Conference (CCNC), 2012 IEEE, 2012. IEEE, 60-61.
YANG, F., YAO, Z., CHENG, Y., RAN, B. & YANG, D. 2016. Multimode trip information detection using personal trajectory data. *Journal of Intelligent Transportation Systems,* 20**,** 449-460.
YE, Y., ZHENG, Y., CHEN, Y., FENG, J. & XIE, X. Mining individual life pattern based on location history. Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on, 2009. IEEE, 1-10.
ZHANG, Z. & HE, Q. 2019. Social Media in Transportation Research and Promising Applications. *Transportation Analytics in the Era of Big Data.* Springer.

ZHANG, Z., HE, Q., GAO, J. & NI, M. 2018. A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C: Emerging Technologies,* 86**,** 580-596.
ZHANG, Z., HE, Q. & ZHU, S. 2017. Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. *Transportation Research Part C: Emerging Technologies,* 85**,** 396-414.
ZHOU, C., JIA, H., JUAN, Z., FU, X. & XIAO, G. 2017. A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data. *IEEE Transactions on Intelligent Transportation Systems,* 18**,** 2096-2110.

Appendices

Table of Rules for Dividing Places Types Obtained from Google Places API to Place Categories

Place Category	Place Types Obtained from Google Places API
School	university, school, library
Shopping	clothing_store, store, liquor_store, supermarket, department_store, grocery_or_supermarket, bakery, shoe_store, pet_store, shopping_mall, convenience_store, home_goods_store, car_dealer, book_store, furniture_store, hardware_store, pharmacy, meal_delivery, electronics_store, meal_takeaway, veterinary_care, florist, brewery, brewing, bicycle_store, jewelry_store
Recreation	lodging, park, food, restaurant, cafe, bar, club, historical_landmark, residence, museum, stadium, gym, natural_feature, bowling_alley, zoo,movie_theater, orchestra, night_club, casino, farm, movie_rental, art_gallery, amusement_park, recreation, neighborhood, theater, theatre, campground, rv_park, staduim
Personal Business	wedding_hall, convention_center, banquet_hall, funeral_home, office, laundry, city_hall, post_office, church, car_repair, doctor, lawyer, real_estate_agency, gas_station, bank, plumber, local_government_office, health, animal_shelter, oragnization, beauty_salon, travel_agency, car_wash, wedding_venue, hair_care, car, skin_care, logistics, finance, physiotherapist, insurance_agency, hospital, dentist, spa, moving_company, general_contractor, police, courthouse, offic, office, cemetery, accounting, storage, agency, place_of_worship, electrician, atm, car_rental, hindu_temple, fincance, roofing_contractor, Dig Coworking Space, fire_station, hospical, army, wedding_vener, organization
Transportation	airport, parking, train_station, bus_station, trainsit_station, transit_staion
Other	point_of_interest, street_address, route, intersection, locality, administrative_area_level_3, political, postal_code