Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method

Zhenhua Zhang\textsuperscript{a}, Qing He\textsuperscript{b,\textsuperscript{*}}, Shanjiang Zhu\textsuperscript{c}

\textsuperscript{a} Here Global B.V., 425 W Randolph St, Chicago, IL 60606, United States
\textsuperscript{b} Department of Civil, Structural and Environmental Engineering, Department of Industrial and Systems Engineering, The State University of New York, Buffalo, NY 14260, United States
\textsuperscript{c} Sid and Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, VA 22030, United States

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Longitudinal travel behavior
Social media
Travel survey
Clustering

\textbf{ABSTRACT}

This study explores the possibility of employing social media data to infer the longitudinal travel behavior. The geo-tagged social media data show some unique features including location-aggregated features, distance-separated features, and Gaussian distributed features. Compared to conventional household travel survey, social media data is less expensive, easier to obtain and the most importantly can monitor the individual’s longitudinal travel behavior features over a much longer observation period. This paper proposes a sequential model-based clustering method to group the high-resolution Twitter locations and extract the Twitter displacements. Further, this study details the unique features of displacements extracted from Twitter including the demographics of Twitter user, as well as the advantages and limitations. The results are even compared with those from traditional household travel survey, showing promises in using displacement distribution, length, duration and start time to infer individual’s travel behavior. On this basis, one can also see the potential of employing social media to infer longitudinal travel behavior, as well as a large quantity of short-distance Twitter displacements. The results will supplement the traditional travel survey and support travel behavior modeling in a metropolitan area.

\section{Introduction}

Traditional household travel survey can directly provide trip features and unveil regional traveler demographics. The travel survey is of crucial importance for transportation planners and policy makers because the detailed individual data in the travel survey cannot only explain current travel behavior but also to forecast future travel demand. Current survey methodologies vary in the form of telephone interview, email or mail, home visit, GPS device tracking, etc. and the process is usually well designed to guarantee accuracy. The respondents are also drawn scientifically from a cross-section of age, gender, household income, educational background, etc.

However, even a well-designed travel survey suffers from several well-known issues. First, the sample size per person is usually small, and the response rate is low. For instance, due to high costs for these methods, 2009 National Household Travel Survey\cite{Santos et al., 2011} only covers about 150,000 people across the U.S. with both the national samples and the States and MPO add-on samples combined. Second, NHTS only provides one-day snapshot on the travel behavior of the sampled household and current household travel survey studies usually have some well-documented quality problems. Not all respondents are able to provide accurate trip
histories. For example, the respondents may not accurately recall their trip details, especially for non-recurrent and occasional short-distance trips. What’s more, the travel survey is always costly, time-consuming, and labor-consuming and almost no organizations or institutes can afford an annual travel survey. For instance, the National Household Travel Survey after millennia is conducted only in 1969, 1977, 1983, 1990, 1995, 2001, 2009 (Santos et al., 2011) and 2016. Therefore, the conventional survey methods cannot provide up-to-date, continuous and longitudinal observations.

To mitigate above shortcomings and broaden the horizons of travel behavior studies, data crowdsourcing is a good option in complementing the traditional methods. Similar attempts can be seen in (Chen et al., 2016; Lin et al., 2015; Wang et al., 2016; Zheng et al., 2016). In this study, we introduce social media as a viable tool to infer the longitudinal travel behavior and activity patterns. The study area is the metropolitan area in Northern Virginia (NOVA) which has long been known for its heavy congestion. Its fluctuated traffic conditions and complex road networks make it a good testbed for studies such as traffic flow patterns (Zhang et al., 2016b), and travel time reliability (Zhang et al., 2016a). Our main contributions lie in the following three aspects. First, we unveil the characteristics of individual travel behavior especially the location aggregation, separation distances between locations and the clustering features of certain groups of locations. Second, based on these features, we propose a sequential model-based clustering method to capture the clustering features of traveler’s hourly locations and extract directed travel between primary locations. Third, the results of travel behavior obtained from Twitter are compared with those from the household travel survey and advantages and disadvantages of Twitter travel behavior are discussed.

Our findings and modeling results are fully detailed in Fig. 1: Section 2 reviews the visions in the traditional travel survey and current efforts of social media in travel behavior studies. Based on the review, Section 3 examines the Twitter location data and finds three features; In Section 4, an algorithm of the sequential model-based clustering method is proposed which fully considered the data features of the Twitter locations and this algorithm can cluster the locations and eliminate the effects of undirected travel. Section 5 unveils the travel behavior features of the Twitter displacements, compared and validated by the household travel survey. Based on these Twitter displacement features, Section 6 identifies the potential of Twitter to conduct the longitudinal travel behavior studies. The paper ends with an elaborate conclusion and discussions in Section 7.

2. Literature review

2.1. Review of traditional travel behavior studies

Traditional travel behavior studies based on the travel survey or other methods are good references for our study. Most of the works are well founded after years of dedications, and their conclusions are insightful. They can not only unravel the interesting properties of the underlying mobility patterns (Schneider et al., 2013) but also have some practical meanings such as features of epidemics spreading (Anderson et al., 1992), population diffusion (Petrovskii and Morozov, 2009), and social networking (Centola, 2010).

One of the key elements in travel behavior studies is the movement. Past studies tend to clarify the movements by predefining the scales of zones (Balcan et al., 2009; Lenormand et al., 2014). The zone scale in each trip may vary from hundreds of meters to a few miles. Usually, movement between the zones is taken as directed travel while those within the zones is undirected travel (Schneider et al., 2013). The directed travel is the primal trip activity which holds the common view that human mobility is a motif-driven activity. These motifs are general human mobility characteristics that can account for the activities mostly. The primal trip activity is commonly seen and driven by purposes, such as traveling to a special event (Ni et al., 2014), commuting to the workplace, going to school, etc. In contrast, undirected travel represents the secondary trip activity (as this activity usually cannot be taken as a trip, we
call secondary activity in the following) which comes from the idea that the destination is sometimes ancillary to the travel rather than the converse which is usually assumed (Mokhtarian and Salomon, 2001). A study showed that the share of total travel that is completely undirected is presumed relatively small but this travel is not a byproduct of the activity, but itself constitutes the travel activity (Mokhtarian and Salomon, 2001). The secondary travel activity is also common, such as tourists wandering around, joy-riding, strolling in the park, etc. in which travelers' attitudes and personality are more important determinants than objective travel (Ory and Mokhtarian, 2005).

The destination locations and the corresponding motifs of secondary activities are usually difficult to predict because under most circumstances they are controlled by undetermined travelers' minds. As not driven by the special needs and purposes, secondary activity is expected to be shorter than the primal trip, and the spatial scale of secondary activities should be possibly limited so that it does not influence the primal trips driven by daily motifs. The underlying causes of these location deviations are usually complex due to the fact that there are many potential influential factors: residents of neighborhoods with higher levels of density, land-use mix, transit accessibility, and pedestrian friendliness drive less than residents of neighborhoods with lower levels of these characteristics (Handy et al., 2005). In contrast, some studies showed the causality between land use and peoples' preferences can be just reverse: the prevalence of walking and transit use may be caused by self-selection; that is, people who prefer walking or transit may choose neighborhoods that support their predilections (as opposed to neighborhood designs strictly influencing choices) (Ewing and Cervero, 2001). The GPS errors should not be ignored either.

Despite the ambiguous cause-and-effect relationships between travelers' attitude and the external influential factors, the effects of secondary activity should be taken seriously when using high-resolution Twitter data. The problem is, on the one hand, GPS location information recorded by Twitter can potentially unveil more detailed information than traditional travel survey; while on the other hand, the data may suffer from the effects of secondary activity which is usually not a problem in the travel survey. To counter the problems, previous works relied on different heuristic clustering methods such as K-means, hierarchical clustering, etc. and extract the useful displacement information. In this study, we propose a sequential model-based clustering method which not only inherits the merits of traditional model-based clustering but also adapts well in using empirical features found in geo-tagged Twitter locations. Model-based clustering methods assume that there exist mixture models in the multi-dimensional data (Everitt, 1981; Fraley and Raftery, 2002; Wolfe, 1970). In the mixture model, each mixture distribution corresponds to a cluster. The selection of a proper distribution model remains a model choice problem for a specific research purpose. The parameters of the distribution need to be estimated which is an optimization problem. The most widely-used mixture model is the Gaussian Mixture Model (GMM) (Banfield and Raftery, 1993) and the well-recognized method to estimate the parameters is the Expectation-Maximization (EM) (Dempster et al., 1977). This paper modifies the classical model-based clustering method to study the Twitter locations, and the procedure is detailed in the following sections.

2.2. Review of social media in travel behavior studies

Recently, social media has been gradually accepted as a user-contributed data source in event detection and summarization. It has been proved a viable data source for applications such as transport information retrieval (Cottrill et al., 2017; Kuflik et al., 2017), activity pattern classification (Hasan and Ukkusuri, 2014), incident detection (Gu et al., 2016; Zhang and He, 2016), etc. Social media tools such as Twitter, Foursquare, Facebook, and Sina Weibo can provide the public available data including the time, posts, and the high-resolution location information. For example, Twitter can serve as a useful proxy for tracking and predicting human movement (Jurdak et al., 2015). Based on the individual Twitter check-in data, one can directly access the individual activity information which can infer the personal trip purpose (Hasan and Ukkusuri, 2014; Pianese et al., 2013), travel activity pattern (Cao et al., 2014; Hasan and Ukkusuri, 2014), drivers' routing behavior (Pan et al., 2013), etc. Recent studies even found that location clusters with more employment opportunities and more types of employment is associated with more social media check-ins (Huang et al., 2017). Compared with traditional travel survey, social media can access the individual mobility information in a much longer term. Thus, data crowdsourced from social media can be regarded as a valuable source for long-term activity space research which addresses the problems caused by short-term data (travel survey) (Lee et al., 2016).

Travel behavior studies based on social media is emerging. The aggregated findings coincide with consensus by traditional travel survey. For instance, Cheng et al. (2011) found that the footprints (check-in) left by the social media users follow a Levy Flight mobility pattern. Also, other studies showed that social media is a meaningful complement to traditional travel behavior studies. For instance, Rashidi et al. (2017) reviewed the current state-of-the-art methods of social media studies and concluded that it had an enormous potential of improving our knowledge in activity participation behavior; Zhu et al. (2014) investigated the location-based social networks and achieved over 75% accuracy in predicting trip purposes combining with the traditional travel survey; Zheng et al. (2016) even combined the social media with floating sensors and incident report to predict the human mobility and controlled the traffic in both the physical and cyber spaces.

Before social media is introduced, there are other attempts that use big datasets, including bank notes and cellphone records. Compared to these methods, social media has several advantages: First, the geo-tagged tweets can provide real-time locations of the Twitter user with a relatively high accuracy. Such location information can be further converted into the movements in time and space, and further unveil the trip patterns in a region. In comparison, the bank notes (Brockmann et al., 2006) used to trace the interstate or inter-city travel can usually only survey locations at a time interval of several days. Cellular data (Song et al., 2010) can generate more frequent mobility information but the geo-precision is in kilometers because the average service area of each mobile tower is approximately 3 km² (Gonzalez et al., 2008). Second, other methods using GPS tracer (Rhee et al., 2011) suffered from the issues of small sample size and the lack of generality. Same problems also exist in the newly emerged data source of Flicker (Beiró
et al., 2016) because of its specific design for photo sharing. In contrast, Twitter data provides a wide geospatial coverage and high sampling rate.

Our study employs Twitter data to explore several major variables of travel behavior studies: travel distance, travel time and departure time. We also examine the location features that are usually overlooked by traditional travel survey. In addition, we consider not only the features of Twitter data but also the knowledge of human mobility patterns generalized in previous studies.

3. Empirical findings

3.1. Data description

Tweets are public domain data that can be accessed through Streaming API with geolocation filter (2016). All tweets are time-stamped and associated with a unique user ID. A subset of them is also geo-tagged, containing four pieces of location information: longitude, latitude, names of county and state. The theoretical precision of latitude and longitude can be as high as $10^{-8}$ degree. These geo-tagged tweets represent a sample of human movements over time and space. The study area is located in the Northern Virginia (NOVA) and District of Columbia, and we collect nearly 6,000,000 geo-tagged tweets during the year of 2014.

The sampling rate of Twitter is much lower than the conventional GPS-based studies, and thus could not provide a complete picture of the travel trajectory of a particular Twitter user on a particular day. However, when data from a huge number of Twitter users over a long time are aggregated, they can provide a repeated sample of the human mobility space and reveal patterns and trends in human mobility for a large area. Therefore, a minimum number of samples are required to reveal enough details of the spatial-temporal space of human mobility. This minimum number depends heavily on the stability of human mobility patterns, which will be explored in this paper. The tweeting frequency varies greatly among users, and an active user can tweet more than 50 times per day.

In this paper, we focus on the Twitter user who tweet more than 100 times per year. To avoid creating a biased dataset, for those users that tweet more than once during an hour, we take the most frequently-visited location during that hour period as the representative of the hourly location.

3.2. Location-aggregated features

For people who follow routine all the time, their whereabouts over time are highly predictable, and a relatively low sample rate would be sufficient to capture their activity and mobility patterns. In contrast, a much higher sample rate may be required for people whose lifestyle is more liberal. To address this issue, we first analyze the patterns in frequently-visited locations. We round the precision of the latitudes and longitudes into $10^{-3}$ to reduce noises in location precision. Fig. 2 (a) shows that the frequency of which five randomly selected Twitter user (in different colors) stayed at the most frequently-visited locations during each hour of the day. For example, the whereabouts of the user represented in blue are highly predictable between 20:00 pm and 6:00 am, but less so for the rest of the day (at 12:00 p.m., this user only stayed at the most frequently-visited location for that period for one-third of the time). In contrast, the activity pattern of the user in green is less predictable over time. Even at night, this user only stayed at the most frequently-visited location for that period one-third of the time. Although people show some diversity in their activity patterns, their whereabouts are predictable to a large extent. Fig. 2(b) shows the visiting frequency of the top 10 locations for the same 5 individuals.

![Fig. 2.](image-url)

Fig. 2. (a) The ratio of the most frequently-visited locations over all locations for 5 individuals over different hour periods; (b) The ratio of the top 10 most frequently-visited locations over all locations for the same 5 individuals at 18:00 p.m. in different days. Each color of points represent a Twitter user.
at 18:00 p.m. on different days. All of them stayed in one of the top two most visited locations for that period over 80% of the time. We examine all the Twitter users’ locations and find that for each tweet user, if we rank descendingly the Twitter locations by their visiting frequency, the top 5% locations account for an average 74.4% of the whole record.

The spatial-temporal space of human mobility is usually anchored around primary activities (e.g., staying at home, commuting to work, and going to school), most of which would co-locate with some frequently-visited locations. Surrounding these most frequently-visited locations, people may also get engaged in some secondary activities and make a few short-distance movements. Therefore, if we sample human locations randomly over time, we should find a cloud surrounding a few primary activity locations that would help to reveal human mobility. Fig. 3 gives one example. Previous studies argued about the existence of a few “well-defined area” that would help to anchor activity space, but fall short on how such space should be identified (Gonzalez et al., 2008). As shown in Fig. 3, the spatial scales of these location clusters vary from different Twitter users, and for each user, there may exist more than one cluster. A robust method is needed to identify these primary activity locations.

Above all, we define the features of the locations as the location-aggregated features, and they are very important references for proposing methods to extract the primal trips. These features can be concluded as follows:

- The locations of a Twitter user are more likely to appear in several clusters. The scales of the clusters for each tweet user are unique and should be further identified.
- The locations within the same cluster should be taken as those driven by the same travel motif. Displacements of locations within the same cluster are considered as secondary activities without further information.
- The most frequently-visited locations are the most important in the clustering process. We need to place all the locations in sequence according to their visiting frequencies.

3.3. Distance-separated features

Besides the location characteristics discussed above, one can also trace the movement of a Twitter user over time based on location information. Although these movements in human mobility space revealed by Twitter data are not equivalent to trips (these individuals may have visited more places between posting two consecutive geotagged tweets), their characteristics may also reveal Twitter users’ travel behavior. For example, human mobility space may depend on both individual characteristics and the built environment according to Heath et al. (2006). Fig. 4 gives a close look at the subareas of a traveler’s locations and the movement between locations. The locations in Fig. 4 can be divided into 3 distinct clusters. The scale of different clusters, defined by the Euclidean distances between locations, are different. The radius of each cluster reveals the geographic span of secondary activities. The number of sampled locations (may be a surrogate of duration and frequency of visits) are also different for each cluster.

In order to develop an algorithm that could systematically detect the clusters and extra characteristics of individual mobility patterns, we randomly pick two Twitter users from Section 3.2 and plot the percentiles of Euclidean distances among all locations for...
them. The scales of these distances, as shown in Fig. 5(a) and (b), are different, and the percentile values increase following a stair-like shape. A stair-step indicates a scale of displacements that are frequently found for this Twitter user, and we define the stair step as the separation distance. The long separation distance usually represents the scale of primary trips and could be for travel motifs such as the daily commuting or other regular activities discussed in Section 2. The short separation distance is likely to be associated with secondary activities surrounding the primary activities.

We further investigate properties of the small separation distances for the two tweet users as shown in Fig. 5(c) and (d). These figures show that the distances between points associated with secondary activities are also different between individuals. Therefore, the scale of the location cluster is also quite different among users, which is proved in Fig. 3. Through our empirical findings, the distance-separated features of Twitter users can be generalized as follows:

- First, each Twitter user has a unique set of scalars that would separate distances between tweet-revealed locations into clusters. They can be derived from a data-driven method, then the scale of the location clusters can be identified.
- Second, even though the distance between two locations is smaller than a separation distance, they are not necessarily driven by the same travel motif. This is true when two frequently-visited locations are close to each other.
- Third, our empirical results show that all travelers have a separation distance equal to 0.001 (equivalent to 100 m) with highest probability and trips for secondary activities are usually to shorter than 0.01 (equivalent to 1000 m).

We use the Euclidean distance instead of the actual distance because the calculating the former is much more efficient than the later in ArcGIS, which is critical for processing large dataset.

3.4. Approximately Gaussian-distributed features

As the locations in the same cluster are assumed anchored around the same primary activity location and driven by the same primary travel motif, these locations may follow a specific geo-distribution. Gonzalez et al. (2008) modeled this kind of geo-distribution as $\Phi_a(x,y)$, which is the probability to find an individual $a$ in a given position $(x,y)$. This distribution can be approximated by a Gaussian distribution and we also validate similar results with our high-resolution Twitter locations.

From a spatial view, in each cluster, the visiting frequency of locations should follow a two-dimensional Mixture Gaussian distribution as shown in Fig. 6. In this distribution, there is a high peak representing the most frequently-visited location as the barycenter, and all other locations within the same cluster are distributed around it. The cross-section of a distribution shown in Fig. 6(a) depicts the probability distribution within a cluster where the visiting frequencies are converted proportionally into the density probability of the Gaussian distribution. Thus, we can assume that all the locations of a Twitter user can be approximately represented by a set of Gaussian distributions with unknown parameters, together with a set of isolated locations. The Gaussian-distributed features should have two important components:

- Gaussian-distributed features refer to the relationship between the frequently-visited location (barycenter) to their surrounding locations.
- For each barycenter, its surrounding locations may not necessarily follow the same Gaussian distribution because their separation
distances to the barycenter are different.

One can see that the location distribution is approximated using Gaussian distribution. This is one of the significant assumptions for the following clustering model.

4. Sequential model-based clustering method

The objective of this section is to group the locations identified by geo-tagged tweets in clusters and ensure that the locations driven by the same primary travel motif are placed into the same cluster. Section 3 unveils the features of human travel locations including the location-aggregated, distance-separated and Gaussian-distributed features. According to these features, we propose a data-driven clustering method called the sequential model-based clustering method, a variant of classical model-based clustering methods. This method proceeds sequentially from the smallest separation-distances to the largest and, in each loop employs the probability of Gaussian distribution model as the criteria to cluster the locations.

The classical model-based clustering method assumes that a population of interest consists of several different sub-populations (Banfield and Raftery, 1993). In each sub-population or cluster, a component is described by a probability density function. Each component is multi-dimensional (2 dimensional for Twitter locations) and has an associated weight (the visiting frequency for Twitter locations). For a Twitter user, we can assume there are $J$ different Twitter locations and $I$ different barycenters (in the

Fig. 5. (a and b) The distribution of percentile values of the separation distances (measured by the differences of longitudes and latitudes) among all locations for two travelers. (c and d) The frequency of the separation distances (measured by the differences of longitudes and latitudes) smaller than 0.01, which is equivalent to 1000 meters, among all locations of two Twitter users.
following section, we will keep using upper-case letters to denote vectors and lower-case letters to singular items). The Twitter locations here refer to the hourly Twitter locations and they are extracted by taking the most frequently-visited locations during each hour period. Assuming in the $i^{th}$ cluster, the barycenter is $\mu_i = (\mu_{i,\text{lon}}, \mu_{i,\text{lat}})$, where $\mu_{i,\text{lon}}$ and $\mu_{i,\text{lat}}$ are the longitude and latitude and variance of Gaussian distribution is $\sigma^2$. Then, the probability density of the $j^{th}$ location $x_j$ can be derived from Gaussian distribution which has been demonstrated in Section 3.4:

$$\phi(x_j | \mu_i, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x_j - \mu_i)^2}{2\sigma^2} \right)$$

(1)

where $|x_j - \mu_i|$ denotes the Euclidean distance between the two locations. The probability of $x_j$ given all barycenters can be further written as:

$$\varphi(x_j | \mu_i, \sigma) = \sum_i f_i \phi(x_j | \mu_i, \sigma) = \sum_i \frac{f_i}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x_j - \mu_i)^2}{2\sigma^2} \right)$$

(2)

where $f_j$ is the visiting frequency of the $j^{th}$ location $x_j$. The likelihood function of this Twitter user can be written as:

$$L(\mu_1, ..., \mu_n, \sigma) = \prod_j \varphi(x_j | \mu_i, \sigma)$$

(3)

Both $\mu_i$ and $\sigma$ are unknown parameters and can be estimated by maximizing the likelihood function. We can employ Expectation-Maximization (EM) algorithm, an iterative method for finding maximum likelihood estimates of parameters in statistical models. Usually, the variance $\sigma$ can also be different in different clusters, while sometimes the density function $f_j$ can be the same in different clusters. This data-driven clustering method relies heavily on the distributed features in the data and the likelihood function may not work for our Twitter locations for the following reasons:

- First, Eq. (3) does not provide criteria on whether to include a location $x_j$ into a given cluster.
- Second, the estimation of Gaussian distribution $\phi(x_j | \mu_i, \sigma)$ in most model-based clustering method is purely data-driven and may sometimes yield unrealistic clustering patterns (e.g. a cluster with a radius too large for secondary activity).
- Third, the barycenters of the distributions $\mu_i$ are better to be set as the real locations, such as home, workplace, etc. While in classical model-based clustering, they are sometimes virtually created and not real Twitter locations.

Thus, we propose the following improvements to adapt the data-driven method for our applications:
4.1. Modification 1: selecting criteria

We should consider the typical patterns for secondary activities: the separation distance and the variance should reflect the Twitter users’ personal activity patterns and preferences. For different Twitter users, the scalars separating different clusters should be quite different and can be extracted by clustering each user’s hourly locations. Here, in order to find these scalars, we employ the X-means clustering method (Pelleg and Moore, 2000) to cluster these distance data. The X-means method is built on the K-means clustering method and has certain advantages. The X-means method does not require to predefine the number of clusters and cluster centers. Instead, it can efficiently search the space of cluster locations and a number of clusters in an iterative process. The good clustering result must conform with Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) measure (Pelleg and Moore, 2000). The Algorithm should be applied on each Twitter user, and it includes three steps:

- Step 1. From all the Twitter locations of each user, extract the most frequent visited coordinates during each hour period in each day as the hourly locations: $x_i$.
- Step 2. Calculate the Euclidean distance between all $x_i$: $L = \text{dist}(x_i)$ and extract the distances that are lower than a threshold.
- Step 3. Implement X-means on $L$ and obtain the separation distances $D = \{d_i\}$ (this paper uses $[\ ]$ to denote a set of data).

In our study, we set the distance threshold as 0.01 in longitude and latitude, and one may set a higher value if needed. As the scale of each cluster is finite, for each Twitter location, we do not need to calculate the probabilities of all barycenters but to select the proper barycenter which is within the threshold distance. As discussed in Sections 3.2 and 3.4, the Twitter locations have aggregated features and locations within the same cluster are Gaussian-distributed around the most frequently-visited location which are taken as the barycenters. For a barycenter location $\mu_i$, one can estimate the density value of the location $x_i$ using Eq. (1). If a location should be included into a cluster, the distribution of the frequency of this location and the frequency of its corresponding barycenter location should be approximated by a normal distribution as discussed in Section 3.4, which can be written as:

$$f_i = \frac{\Phi(x_i|\mu_i, \sigma)}{\Phi(\mu_i, \mu_i, \sigma)} = \exp\left(-\frac{|x_i - \mu_i|^2}{2\sigma^2}\right)$$  \hspace{1cm} (4)

where $f_i$ is the visiting frequency of barycenter location $\mu_i$; $f_i^*$ is the threshold frequency of location $x_i$ and $f_i^*$ should be larger than the actual visiting frequency $f_i$ if $x_i$ belongs to the cluster of $\mu_i$.

$$C(x_i) = C(\mu_i), \text{ if } f_i < f_i^*$$  \hspace{1cm} (5)

where $C(x_i)$ denotes the cluster label created for location $x_i$.

4.2. Modification 2: separation-distance effect

A barycenter location and a separation distance can together determine a cluster’s location and the scale of this cluster anchoring at this primary activity location. A location which is within a separation distance from a barycenter location should be tested whether or not the location belongs to the cluster. The confidence level for this location $x_i$: $|x_i - \mu_i| = d_k$ to be included into the cluster $\mu_i$ can be assumed to be $\alpha$ where:

$$\Phi(x_i|\mu_i, \sigma) = \frac{1}{2} \left(1 + \text{erf}\left(\frac{d_k}{\sqrt{2}\sigma}\right)\right) = 1 - \alpha$$  \hspace{1cm} (6)

where $\Phi$ is the cumulative function of Gaussian distribution; $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}dt$ is the error function; $d_k$ is the $k$th separation distance; $\alpha$ is the confidence level set to be 0.05 in this paper. One can further derive the variance:

$$\sigma^2 = \left(\frac{d_k}{\sqrt{2\text{erf}^{-1}(1-2\alpha)}}\right)^2$$  \hspace{1cm} (7)

4.3. Modification 3: location-aggregated effect

According to the location-aggregated features, the clustering algorithm iterates from the barycenter locations with the highest visiting frequency. Thus, for each location of the Twitter user, if there are more than one qualified barycenter locations, this location should belong to the cluster which minimizes $\frac{f_i}{f_i^*}$:

$$C(x_i) = \left[C^i(x_i)|\arg\min_i\left(\frac{f_i}{f_i^*}\right)\right] C(x_i) = \left[C^i(x_i)|\arg\min_i\left(\exp\left(-\frac{|x_i - \mu_i|^2}{2\sigma^2}\right)\right)\right]$$  \hspace{1cm} (8)

where $C()$ is the notation of cluster label. The algorithm for each Twitter user can be detailed as follows:
Algorithm: Sequential model-based location clustering

Input: separation distance: $D = [d_k]$;
hourly locations for each traveler: $X = [x_j]$ and
their visiting frequency $F = [f_j]$;
confidence level: $\alpha$;
Output: cluster label for $x_j$: $C(x_j)$.

Assign unclustered locations: $U = X$;

Let $N = 1$;
Sort $D$ from the smallest to the largest;
For all $d_k$ in $D$ begin
Select from $U$ a barycenter location $u^c$ whose frequency $f^c$ is the max;
Calculate the distance between $u^c$ and $x_j$;
Calculate the threshold frequency for $x_j$:
$$ f^T_j = f^c \cdot \exp \left( -\frac{(|x_j - u^c| - \mu_i \cdot \text{erf}^{-1}(1 - 2\alpha))^2}{d_k} \right), $$ where $\text{erf}()$ is error function;
Select all locations $T = \{x_j | f_j \leq f^T_j, |x_j - u^c| \leq d_k\}$;
if sizeof $(T) < 0$ begin
label $T$ and $u^c$ with Cluster ID: $C^k(T) = C^k(u^c) = Num$;
$Num = Num + 1$;
remove $T$ and $u^c$ from $U$;
else
label $l$th $u_i$ in $U$ with Cluster ID: $C^k(u_i) = Num + l$;
remove all locations from $U^0$ and $U^0 = ()$;
End
End
$U^0 = U$;
End

For $x_j$ with more than one Cluster ID: $C^k$;
Select $C(x_j) = \left( C^k(x_j) | \arg \min_{i,k} \left( \exp \left( -\frac{(|x_j - u_i| - \mu_i \cdot \text{erf}^{-1}(1 - 2\alpha))^2}{d_k} \right) \right) \right)$.

The sequential model-based clustering is derived based on the three major location features found in Section 3 and can properly group the Twitter locations with approximately the same travel motif. Note that we use a general term: same travel motif instead of same trip purpose to describe the locations in the same cluster. This is because, in this paper, travel by the same motif refers to going to the same geographic area and the method can automatically decide the scale of this area. This approach will eliminate the effects of secondary activity and after that we can further analyze the directed travel features in Section 5.

5. Travel behavior features from Twitter

5.1. Twitter displacement and its representativeness

According to the clustering results in Section 4, we can extract displacements between every two consecutive hourly locations. Displacements between different clusters correspond to primal trips while those within the same cluster should be secondary. By the clustering results, one can easily extract the individual displacements. Fig. 7 gives an example of a Twitter user’s locations from 18:00 p.m. to 19:00 p.m. across the year of 2014. Fig. 7(a) shows the most frequently-visited location during 18:00 p.m. and Fig. 7(b) depicts three corresponding destinations during 19:00 p.m. on different days. For display purposes, the clusters of destination locations is set in different colors and shapes.

As one can see from Fig. 7, Twitter displacements can track individual trajectories and unveil their travel behavior features. What’s more, by aggregating data from a long period, researchers can infer the unique travel behavior features (e.g. trip purpose, trip length, etc.) of Twitter users. Despite the heterogeneities among individuals, the total population of Twitter users is so large that Twitter users may reveal travel behavior of different subgroups of the entire population to some extent. There are several unique features of Twitter users which may make the travel behavior results quite different from the travel survey. Such comparison provides
Fig. 7. Displacements that originate from (a) 18:00 p.m. and end in (b) 19:00 p.m. and the corresponding tweets for location: (1) “i hate this haircut”; (2) “ive never seen this movie that they playing on bet”; (3) “xisthatnigga marchmadness this is marchmadness this happening to duke makes this so great”.

Table 1
the basic statistics of Twitter and household (HH) travel survey.

<table>
<thead>
<tr>
<th></th>
<th>Displacement/trip size</th>
<th>Number of participants</th>
<th>Displacements/trips per person</th>
<th>Observation time</th>
</tr>
</thead>
</table>

Fig. 8. Comparisons of (a) age groups, (b) household income and (c) education level between Twitter and household travel survey.
us a basis for re-sampling the household travel survey data to be consistent with the demographics of Twitter users for comparing the travel behavior patterns extracted from the two data sources.

First, in our study, there are more than 9000 Twitter users involved, and an average of 43.98 Twitter displacements is recorded for each user. The basic statistics in Table 1 show the differences between two data sources. Twitter has relatively less number of participants but a larger sample size per person. Also, due to the low marginal costs and continuity of crowdsourced data source: Twitter provides an effective and efficient way to monitor the travel behaviors in the long term. The large sample size per person and longitudinal travel behavior monitoring are key features of travel behavior studies based on Twitter.

Second, “Twitter survey” represents particular groups and such groups of people may have different distribution features in income, education, etc. Fig. 8 shows the demographics of Twitter users which are quite different from those of travel survey. The user demographics were collected from comScore (Adam and Andrew, 2016), Pew Research Center (Duggan and Brenner, 2013) as well as population demographics of Fairfax County in Northern Virginia (Fatima et al., 2016). Compared with the population of survey respondents, Twitter users are much younger, and their incomes are much lower, while their overall education level is a little higher. Therefore, their travel behavior features are expected to be different from that of the traditional survey. It is worth mentioning that due to the limits of the open data for Twitter demographics, there is a discrepancy in the comparison. The household travel survey is based on that the entire State of Virginia while Twitter data are only for Northern Virginia. The comparison between two data sources can be better matched up in future studies when there is more detailed Twitter demographics.

5.2 Travel behavior feature exploration and validation

In this sub-section, we unveil the travel behavior features from Twitter and demonstrate the potential of Twitter displacements in inferring the related travel behavior such as the trip distance, duration, departure time, etc. To both compare and validate the results, we conduct a stratified sampling on the records in household travel survey and compare its results with the Twitter’s. The attributes of the samples in the stratified survey are consistent with those of Twitter users including gender, age, household income, and education. In this process, we first divide the total records into subsets with different combinations of attribute levels. There are 4 different attributes and totally 150 different subsets of combinations. We group them by random sampling with no duplicates, and the size of each subset is determined according to the distributions among Twitter users.

Also, Twitter is designed as a social network tool, and Twitter displacements are consequently a byproduct of retrieving passive crowdsourced social information making it different from the traditional survey. Thus, certain uniqueness should be clarified:

- First, the smallest trip length recorded by Survey is 1/9 mile (178 m) while Twitter displacements can be much smaller. In this subsection, we only keep Twitter displacements higher than 1/9 mile, and the smaller displacements will be discussed in a later section.
- Second, our examinations show that 96% of the trips in travel survey of Virginia are made within 1 h. Long-distance travel is rare for most of the travelers and will be the focus of further studies. In this study, we focus on Twitter displacements made within 1 h.
- Third, for similar reasons, we only focus on the travel within a metropolitan area and do not include the inter-city or inter-state trips.

Fig. 9 shows power law distribution of the trip length of travel for Twitter displacements and compares it with the stratified and original survey trips. We employ the power law function to approximate the distribution of both the displacements and trips and the statistical properties can capture the fundamental mechanism of driving human mobility patterns (Gonzalez et al., 2008).

\[ P(x) = (x + \gamma)^{-\beta} \cdot \exp\left(-x/\alpha\right) \]  

(9)

![Fig. 9. Comparisons of power law distributions among Twitter displacements, stratified survey and original survey trips. The x-axis is the discretized distance and the unit is one mile; y-axis is the ratio of each distance discretized group.](image-url)
where \( P() \) refers to the probability density function. From Fig. 9, one can see that the parameters \( \alpha, \beta \) and \( \gamma \) in Twitter displacements resemble stratified survey more than the original one. Besides this, one can also see that both datasets are heavily right-skewed and the median displacement/trip length can be a better representative of regional travel behavior features.

Although there are much more short-distance trips in Twitter than that in Survey, the similarity of \( \beta \) between Twitter and stratified travel survey shows that they have almost the same mechanism of travel, in which the scale of two data may not be the same, but their distributed features resemble each other. We further compare the displacement distance/trip length of the three datasets in Fig. 10 and the start time distribution of displacements/trips in Fig. 11. Between Twitter displacements and stratified household travel survey, Fig. 10 demonstrates, to some degree, many similarities:

- Fig. 10(a) truncates both Twitter displacement and trip length that are less than 1 mile. Tweet data that contains lots of small displacements underestimates the actual trip length in general. This is not surprising since people can post tweets while walking...
on the street, going from garage to the building, or moving between different rooms within a building. Many of these movements are ignored by the conventional travel survey, but recorded by Twitter data.

- Fig. 10(b) and (c) shows the median displacement/trip length longer than 4 miles and 7 miles from the Twitter data and the travel survey, respectively. The impact of short displacements in Twitter data is minimized in these two cases, and thus the medians are much closer in figure (b) and (c). In all three cases, the median of trip length is much closer to the median of displacement (in red) after social-demographic factors are considered through additional stratifications (blue line vs. the green line).
- Twitter displacements show a larger AM peak hour effect than that of survey but almost no PM peak hour effect. The absence of PM peak hour effect may be accounted by many reasons. We can also find similar results, such as the trip start time comparison in Fig. 11.
- Twitter displacements are smaller than that of the survey at night 19:00–24:00 and 1:00–5:00. This may be due to the small sample sizes in travel survey and Twitter as shown in Fig. 11. Late night Twitter displacements also show unique features and will be further discussed in later sections.

Fig. 11 compares the percentage of trips starting from the different time of day among all trips derived from survey data (blue and green lines for stratified and unstratified survey data, respectively) with the same results derived from Twitter data. The categorization of the start time in Fig. 11 refers to that in (Santos et al., 2011). The patterns are very consistent through a day except for the period between 4 pm and 7 pm. This means that during PM peak, we usually find fewer displacements from Twitter users. Similar results can also be found in the comparison of displacement/trip median in Fig. 10. The AM peak comparison is much better and the comparison during the late night also shows that Twitter captures more displacements than the traditional survey. More studies are needed to explain this deviation.

Fig. 12 compares the duration of Twitter displacement with the travel time of the survey. As one can see, for displacements and trips larger than 4 miles, the median of trip duration derived from Twitter data (a) is slightly longer (1–3 min longer during the daytime) than those derived from the survey data in Fig. 12(b) and non-stratified in Fig. 12(c). The comparison between medians are meaningful than means because of the right-skewed features of Twitter displacements and the differences between these two data sources, especially during the daytime, are acceptable. Also, Twitter duration even has a larger variance during different hour periods as compared to the survey which may be due to the larger sample size, especially during the late night. It is worth mentioning that the trip duration records in the survey are discretized recall data and thus do not have the same resolution as that of Twitter. This is why we can see a relatively larger fluctuation of the median in Twitter but a relatively flat and even unchanged median in the survey.

By comparing the Twitter displacements with the traditional household survey, we can see the validity of Twitter in studying the travel behavior of certain groups of people. Besides, one can see the importance of Twitter in complementing the traditional travel behavior studies, by comparing the hourly displacement/trip length distribution, start time distribution and the duration/travel time distribution between the two data sets.

- First, our comparison demonstrates that one can convincingly derive aggregated travel behavior information from Twitter users. These Twitter users are from people with designated gender, age, income and education distributions. The online open Twitter data are massive, low-cost, and real-time. The future applications using social media to infer certain features of travel demographics will become much easier than the traditional survey.
- Second, it shows great advantages of data crowdsourcing: the high precision GPS locations have certain advantages over the survey answers from the survey respondents. Travel behavior studies using social media data solve, to some extent, the “memory” problem (imprecise recall) of most respondents in some previous studies such as (Mokhtarian and Cao, 2008).
Third, the new findings from the derived travel behavior are also very important: one can see clearly from Fig. 10 that Twitter records more overnight displacements; and Fig. 11 shows that Twitter gives more travel histories; besides, we can see a large quantity of short-distance displacements (those shorter than 1/9 mile and not covered by survey), which benefits the longitudinal travel behavior monitored by Twitter.

6. Longitudinal travel behavior monitoring

6.1. Potentials

For a Twitter user, the observation of Twitter data is continuous and longitudinal which is the major advantage of social media. Thus, the process of information retrieval from Twitter can potentially contribute to next-generation passive and continuous travel behavior monitoring. This can enhance longitudinal travel behavior monitoring in several ways.

First, Twitter data can capture a lot of short-distance trips (especially through walking), which may have been under-reported in the traditional survey. Discussion in Section 5.2 shows that Twitter changes a great quantity of short-distance displacements. These displacements under 1 mile indicate short trips going a few blocks to neighborhoods surrounding the primary activity locations. It is worth mentioning that these short-distance displacements are still primal trips and the secondary activities have been excluded. Due to the large number of Twitter users, the short-distance displacements capture unprecedented details of human travel even though each of them is short and without detailed trajectory information. If we aggregate all these short displacements, they will unveil an ever-elaborate depicts of the Northern Virginia networks as shown in Fig. 13(a). By comparing with the authentic road information, we can prove the validity of these short-distance travels as shown in figure (b) and (c). These maps of aggregated short-distance travel analyzed the visiting patterns of the users to different places in a city like most of the studies discovering mobility patterns (Rashidi et al., 2017) and thus is unique in analyzing the aggregated mobility behavior. It can reflect the scale of economic activities, and the social-economic connections between different blocks cannot be detected by traditional household travel because the sample size per person in the survey is much smaller than that in Twitter. Therefore, the continuous travel behavior monitoring of Twitter captures the local travel patterns of participants in large quantities.

Second, Twitter displacements can monitor the travel behavior in real-time and during a much longer period. To realize the longitudinal travel behavior monitoring, more Twitter location data should be incorporated into the sequential model-based clustering algorithm, and the results potentially enrich the current attempts in the travel behavior studies:

- It tackles the limitations in studies such as (Papagiannakis et al., 2017), the arguably much intrapersonal variations due to the limited snapshot of travel for survey respondents. The “consistency” feature, which is a major concern by Mokhtarian and Cao (2008), of the continuous monitoring will diminish the variations when sampling the personal daily activities.
- The recent feature (Kah et al., 2016; Mokhtarian and Cao, 2008) found by the continuous Twitter travel behavior gives the latest information for scholar studies. It potentially brings positive impact to the studies related to social influence. These studies may include the travel behavior changes during the economic crisis, climate shift (Aamaas et al., 2013), urban construction (Zhang et al., 2016b), social events (Ni et al., 2017; Zhang et al., 2016c), etc.

6.2. Deficiencies

Our study also shows some deficiencies of Twitter data in travel behavior analysis, which requires more studies in future. The first major problem is that there are still some inherent problems in Twitter. We list four inevitable scenarios which will make the results inaccurate:

- Inaccurate estimation of travel time: If the Twitter user only tweets twice during the trip, and the tweet time is not the starting or ending time, the Twitter duration may be either longer or shorter than the actual travel time.
- Inaccurate estimation of trip length: If the Twitter user only tweets twice during the trip, and the Twitter location is not the starting or ending location, the Twitter displacements may be either larger or smaller than the actual trip length.
- Inaccurate estimation of travel time and trip length: If the Twitter user tweets more than twice during the trip, both the Twitter displacements and duration only reflect the fragment of the actual trips.
- Inaccurate sampling with tweet data. As Twitter users have different tweeting behavior, the size of their tweets in a year varies. Consequently, the Twitter displacement may be biased and oversampling for certain groups of people. Also, this study used the geo-tagged tweets which may also be different from non-geo-tagged tweets.

These errors may exist in all trip inference studies based on GPS information. One possible way to counter the problem is to conduct a field study, collect the ground truth of trip ends and further build models to calibrate the displacement results based on social media. Compared with both Twitter and travel survey, researchers can focus on smaller samples due to the inevitable costs mentioned in Section 1.

Another deficiency is that if the data sample size is not large, the travel behavior results may not be accurate and also have a large variance. In Section 5.2, the travel behavior results of Twitter locations when compared with the travel survey are promising because we conduct the sequential model-based clustering method on the whole-year Twitter location data. Large datasets guarantee the accuracy to some extent. However, our method may not lead to the same conclusions with one-week or even one-month Twitter data.
in which the sample size is relatively small. Table 2 compares the monthly travel survey results and Twitter displacements when we apply our method on Twitter location data on February, May, August and December in 2014. The Twitter displacements are relatively smaller than the survey and have a larger fluctuation over the month. One can see that, compared with displacements extracted from

Fig. 13. (a) Geographic distribution of short-distance displacements in Northern Virginia (the square noted area is Dale City); (b) Short-distance displacements in Dale City area; (c) Road networks in Dale City area.
the whole-year data, using one-month Twitter data may not produce reliable travel behavior results. However, given the undesirable results of this test, we still believe that the proposed method still has unique contributions since massive volume is the nature of social media data.

Finally, we need to say that Twitter cannot provide social-demographics of users. However, by analyzing the longitudinal patterns of tweets posted by any users, researchers may infer additional travel behavior-related information such as home-based or work-based travel, travel mode, and attitudes towards different modes all of which will be addressed in further investigation. Additional insights may also be generated by integrating Twitter data with other data sources such as the Connected Vehicles, land use, etc.

7. Conclusions

This study proposes a sequential model-based clustering method to study the social media (Twitter) based displacements and investigates the potential of social media to realize the longitudinal household survey.

First, we draw several important empirical findings for social media locations:

- Geo-tagged tweets provide a sample of human activity space through a cloud of locations. These locations may be aggregated in clusters of different scales, showing individual activity patterns.
- Distances between locations show unique clustering features, which can be used to separate primary activity locations with all secondary activities surrounding them.
- In each location cluster, the visiting times of locations are assumed to follow a multivariate Gaussian distribution across the geographic span.

Second, a sequential model-based clustering method is proposed to group the Twitter locations into clusters driven by the same travel motif. Displacements between clusters show similar features in distance, duration, distributions and start time to that of the national household survey in the same geographic area. Further stratification using the social-demographics of Twitter users helps to close the gap between the results derived from the Twitter data and the survey data, showing the importance of controlling demographics when using the Twitter data for travel behavior studies. Due to the unique representativeness of Twitter users, the results are useful for scholars and professionals for future research.

Third, this study also discusses and verifies the promises of Twitter for longitudinal travel behavior studies in two important respects: (1) Twitter provides a vast amount of short-distance displacements which can reconstitute the travel preferences of Twitter users in the road networks and social economic connections within the micro-structure of a city; (2) Twitter is a low-cost and real-time method to capture the longitudinal travel behaviors.

Future studies shall conduct the supervised survey and experiments to counter the limitation of Twitter displacements. For now, the travel behavior features from Twitter may be biased and the representativeness of the Twitter users are not fully studied. Studies can gradually narrow the research scope into several important respects such as the automatic detection of home and workplaces, commuting behaviors in urban road networks. According to the historical locations of an individual Twitter user, one may also predict his future travel behaviors. The prediction of individual travel behavior will surely enlighten the applications of social media. In addition, researchers can also move one step further by increasing the geographic span to study the inter-city or even inter-state travels. One may also see that this study also shows the potential power of tweet contents in the travel behavior study, as shown in Fig. 7. Additional semantic analysis of tweets will surely give more detailed travel behavior information. With increasing coverage of social media, the social media-based travel behavior studies will become more representative and convincing, and the results will become more applicable.

Acknowledgement

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

API. 2016. Twitter Streaming APIs, Twitter Developer Documentation. Twitter Inc.


Heath, G.W., Brownson, R.C., Kruger, J., Miles, R., Powell, K.E., Ramsey, L.T., 2006. The e...

