Investigating the relation between social media activities and the mobility and activity types of users

K. Gkiotsalitis[a,1,*]

[a] *University of Twente, Center for Transport Studies, Horst - Ring Z-222, P.O. Box 217, 7500 AE Enschede, The Netherlands*

Abstract

In the past years, there has been an emerging number of studies on estimating the passenger demand in urban environments based on social media and cellular data. However, the study of the travel behavior at the individual level and the relation between social media activity and the activity/mobility patterns of users has received limited attention. To rectify this, this study examines Twitter data for unveiling the relations between geo-tagged Tweets and Twitter user sentiments, and the respective activity types performed in the real-world. In this work we try to find common patterns between users' Twitter activity and their actual mobility/activity patterns with the aim to provide some generalizations that can help to understand and model the travel behavior of users. This is achieved with the development of educated rules and probabilistic models that can predict the mobility transfers of users between different activities based solely on social media data. The validity of our generalizations is validated with the use of 4-month Twitter data from London. Only active Twitter users have been selected to study in deep the relations between social media activities/sentiments and the activity types performed in the real-world. Although our generalizations are case study-specific, they demonstrate that it is possible to extract the activity and mobility behavior of users with the use of social media and offer a first step in this direction.

Keywords: Twitter; sentiment analysis; travel behavior; mobility patterns; activity-based modeling.

1. Introduction

Since the 1970s the travel travel behavior of individuals is investigated using travel surveys (Maat et al. (2004), Stopher and Collins (2005)). Travel surveys have been extensively used for the development of Trip-based Models (TBM) and Activity-based Models (ABM) for travel demand prediction (Axhausen and Gärling, 1992; Dong et al., 2006; Rasouli and Timmermans, 2014). Over the years though, conducting a travel survey at a city level has become an expensive and labor intensive task. As a result, more and more surveys are conducted via telephone interviews and cover only a small portion of

*Corresponding author

Email addresses: k.gkiotsalitis@utwente.nl (K. Gkiotsalitis), francesco.alesiani@neclab.eu (K. Gkiotsalitis)

[1]Tel.: +31 534 891 870

travelers that is typically less than 3% of the city population (Stopher, 1996). In addition, the information from travel surveys is static and fails to capture the new travel trends because city-wide travel surveys are typically conducted every 2-10 years (Stopher and Greaves, 2007; Flyvbjerg, 2005; Flyvbjerg et al., 2006).

Apart from the small sample sizes and the static information, the information from travel surveys is only useful for understanding the mobility patterns within the area that they are conducted (Gkiotsalitis and Stathopoulos, 2015b). As a result, even if some cities with budget limitations attempt to borrow surveys from other cities with similar characteristics, there is no guarantee that the information from those surveys is portable and they cannot be used to build reliable mobility demand prediction (Gkiotsalitis and Stathopoulos, 2015a; Gkiotsalitis et al., 2018).

Considering the aforementioned inefficiencies of survey-based methods, the potential of understanding the travel (and activity) behavior of users with the use of social media data is examined. The specific focus of this study is on investigating whether social media can be used for identifying the travel behavior of individuals. Unlike the vast majority of past studies (i.e., Carrasco et al. (2008); Calabrese et al. (2011a)) that use social media or cellular data to estimate the passenger demand in one area based on the transmitted locations of individuals, our work has a more deep focus on investigating the potential of social media data to understand the travel behavior of individual users. That is to say, we do not use social media to estimate the current status of passenger demand but we try to investigate the link between the social media activity of a user and his/her travel behavior including the performed activities in the real-world.

In this pursue, we do not consider only the location information from the social media data. Instead, we explore qualitative information such as the sentiments of users when they Tweet, the frequency of posting messages, the changes of locations when posting subsequent messages, and the spatio-temporal variation of the posted messages of each single individual. This user-focused exploration aims to investigate the link between the social media activity and the individual mobility patterns. This can be beneficial for the development of travel behavior models that do not require qualitative data from labor-intensive, revealed preference travel surveys (Flyvbjerg et al., 2006).

In the remainder of this paper, we study the link between social media activity and travel behavior at the level of individual users. The paper is structured as follows: in section 2, the related literature and the motivation of this work are presented. In section 3, the notion of states that contain information about the individual's mobility and behavioral status is introduced. Moreover, the format of input data and the examined types of interaction are discussed. In section 4, the relation between the sentiments of individuals when posting messages on Twitter and their performed activities in the real-world is investigated. In section 5, we explore the potential of estimating the evolution of activity and travel states of individuals with the use of social media data. Furthermore, a Monte Carlo-based method for the simulation of the individuals daily travel schedules is provided. Additionally, a probabilistic state model which describes comprehensively the mobility patterns, the general behavior, and the high-level characteristics of one individual is presented. In section 6, we explore the accuracy of our techniques that try to identify the travel and activity behavior of individuals based on social media data. For this, we perform a validation with the use of four months of Twitter data from London that is kept for testing purposes and is not used in our initial exploration. Finally, in section 7 we provide some concluding remarks and propose future research directions.

2. Related Work

Mobility demand is estimated by using either trip-based, activity-based or agent-based models (Kitamura, 1988; Bowman and Ben-Akiva, 2001; Zhang and Levinson, 2004; Lin et al., 2008). Activity-based and agent-based models are oftentimes used instead of the traditional trip-based models and allow the modeling of soft policies (i.e., implementing congestion charging schemes or modifying the office hours in cities (Algers et al., 2005; Zheng et al., 2012)). Nevertheless, both activity-based and trip-based models are tailored towards survey-based data and they need to be adjusted for accommodating other sources of data.

In a comprehensive review on travel surveys, Stopher and Greaves (2007) concluded that, due to the small samples and the static data provided from travels surveys, it is beneficial to use new sources of data to enriching the available information. Several other works such as Van Der Zijpp (1997); White and Wells (2002); Pan et al. (2006); Zhou and Mahmassani (2006); Gonzàlez et al. (2008); Zhang et al. (2010); Baek et al. (2010); Calabrese et al. (2011b) and Shang et al. (2012) have also explored the use of continuously updated data, such as Floating Car Data (FCD), Automated Vehicle Identification (AVL) data, Inductive Loop Detector (ILD) data, and cellular data to predict the mobility of people. A special category is Smart Card (SC) data that can be used for estimating the spatio-temporal density of public transport passengers (Pelletier et al., 2011; Munizaga and Palma, 2012; Sun et al., 2012; Gkiotsalitis et al., 2019).

The aforementioned data types cannot substitute the travel surveys' detailed insights with regards to the travel behavior of individuals because they provide information only about the location of individuals and not about the actual reasons behind traveling. Hence, they are mostly used for predicting the passenger demand and the trajectories of individuals, and not the travel behavior. This research gap motivates our work. In more detail, the above-mentioned weakness related to data types that provide only information about the locations of individuals can be potentially addressed with the use of data that is both continuously updated and information rich. One data candidate is social media data that includes posted messages with additional information.

Regarding social media data, Twitter data has been used for a range of applications such as the validation of travel demand models (Lee et al., 2015), the understanding of human mobility (Jurdak et al., 2015), and the exploration of mobility patterns (Hawelka et al., 2014). Foursquare has been notably used for capturing the mobility and activity transitions of travelers (Wu et al., 2014) and for supplementing the travel demand survey data (Chaniotakis et al., 2016). However, there is little attention to the exploration of the link between social media data and the underlying reasons behind mobility and activity choices.

For this reason, this study explores the potential of using social media data to understand the travel behavior of individual travelers. More specifically, attention is given to the correlation between the individuals' Twitter posts (i.e., the interaction types, the time of tweeting, the recurrency and duration of interactions, the individuals' sentiments during the interaction) and their travel decisions. It is worth noting that, besides Twitter and Foursquare, user information can be received by city-released Apps and other message-posting platforms. The main requirement is that the individual does not merely share information about his/her location but also about his/her activity and mobility status, sentiments, and device type from which the message is posted.

As a final remark, we should note that past works have studied at a theoretical level the effects of social relations on mobility decisions (see Musolesi and Mascolo (2007); Arentze and Timmermans (2008); Carrasco et al. (2008) and Nguyen and Szymanski (2012)). Notwithstanding this, the effect of one individual's social relations to his/her travel decisions is not addressed in our work because, in most cases, the social media data of the close network of a user is not available given that not all individuals have an active Twitter account.

Summarizing, our work is one of the first that investigates the link between posted messages on social media and the travel behavior of users expressed in performed activities and travel choices. Its main contributions are the investigation of the relationship between the users' sentiments and their activity/travel behavior, the development of a transition model that tries to estimate the activity and mobility changes of an individual based on Twitter information, and the assessment of the ability of social media data to reveal information regarding the travel behavior of individuals.

3. Definition of States and Interaction Activities

3.1. Definition of States

The notion of states is introduced to describe the mobility patterns of an individual by associating mobility decisions with parameters such as location, activity types and time of day. The location/activity of an individual at any given time $\{t, k\}$, where t is the time of day and k is the date (i.e., day, month, year) is described by his/her state:

$$q_i^{t,k} = \{L_i^{t,k}, A_j^{t,k}\} \tag{1}$$

All possible visited locations of an individual are included to the set of his/her visited locations: $L_i^{t,k} \in \Lambda = \{L_1, L_2, ..., L_N\} \forall t, \forall k$. In addition, each activity $A_j^{t,k}$ is included in the set of all possible activities: $A_j^{t,k} \in A = \{A_1, A_2, ..., A_M\} \forall t, \forall k$. Finally, all states are included in the set of possible states: $q_i^{t,k} \in Q = \{q_1, q_2, ..., q_N\} \forall t, \forall k$.

We assume that an individual participates in one activity when visiting a predefined location. For facilitating the analysis the following set of activities is introduced:

$$A = \begin{cases} A_1 : \text{Home} \\ A_2 : \text{Work} \\ A_3 : \text{Fixed Activity} \\ A_4 : \text{Flexible Activity} \end{cases}$$

Furthermore, each location is described by the associated coordinates as they are retrieved from geo-tagged data. In order to reduce the total number of locations and minimize the effects of positioning errors, we group all locations which lay within a circle with a representative location as center and 300 meters distance as radius.

3.2. Input Data

The provision of public data from social networking websites and other sources grows fast and techniques for data analysis are required for exploiting the potential of using public data for mobility prediction. Since data from social networking websites can be highly heterogeneous, efforts have been made to minimize the data requirements for our techniques in order to enhance interoperability and facilitate the use of as many data-sources as possible. Therefore, only a five-parameter dataset is considered after omitting irrelevant information in order to allow the future integration of information from multiple sources (even if at the current state, this study uses only Twitter data as input). Potential future data sources can include other social networking websites (i.e, Foursquare, Facebook, Google+) and mobile phone Short Message Service (SMS) data.

The five-value dataset can be described in the form of a matrix $S_j^v[t,k]$, $j = 1, ..., 5$, where t is the exact time (sec., min., hour) and k the exact date (day, month, year) at which the individual generated content via social media. Furthermore, v, is the individual's ID (i.e., username) and j is the set of the five predetermined values that are provided at each interaction, j, where: $j = \{$1: latitude, 2: longitude, 3: source of interaction (i.e., mobile internet, web), 4: recipient of the interaction (if any), 5: purpose of interaction (i.e., simple text, photo upload, video upload)$\}$. in addition, individuals' socio-demographic information can be obtained manually from their public account profiles including their age category, gender, employment status, marital status and more.

Table .5 in the appendix presents the format of the Twitter data after it is mined using the Twitter API. The small sample in table .5 presents all related fields which are considered in the analysis and from which the above-mentioned five-value datasets is derived.

3.3. Interaction Activities

Three different interaction types are defined (the chatting activity, the continuous interacting activity and the reply activity), after the temporal analysis of data-exchange over time. "Chatting activity" at a location L_i at time t and date k describes the situation when the individual exchanges text messages continuously with another individual over a predefined time period. Therefore:

$$C_i^{t,k} = \begin{cases} 1 & \text{if a chatting activity was occured at location } L_i \text{ at time } t \text{ and} \\ & \text{day } k \\ 0 & \text{otherwise} \end{cases}$$

The "continuous interacting activity" refers to the occasion at which the individual composes messages, uploads photos or videos or shares information continuously over a predetermined time period. Continuous interacting describes the way of interacting from different locations.

$$O_i^{t,k} = \begin{cases} 1 & \text{if a continuous interacting was occured at location } L_i \text{ at time} \\ & t \text{ and day } k \\ 0 & \text{otherwise} \end{cases}$$

Finally, the "reply activity" refers to the occasion when the individual replies to a composed message sent by another individual and the disengages from the conversation ($R_i^{t,k} = 1$ if the individual replies to a message from location L_i, time t and day type k).

4. Educated Rules for Allocating Activities to Locations

As we discussed previously, four different activity types are introduced:

$$A = \begin{cases} A_1 : \text{Home} \\ A_2 : \text{Work} \\ A_3 : \text{Fixed Activity} \\ A_4 : \text{Flexible Activity} \end{cases}$$

For the assignment of representative locations to specific activities the embedded content of messages is used only for capturing the sentiments of individuals, without performing a comprehensive textual analysis on the embedded text. In addition, locations are not associated with activity types by using information regarding the main Points of Interest (POI) near to locations where a message was posted since we aim at developing techniques without external requirements that can comprehend the obtained data from a study area (i.e., city) and remain scalable. Thus, the developed educated rules are mainly based on continuous spatio-temporal and interaction analysis of the input data.

4.1. Associating Activities with User Sentiments

In contrast to cellular data that provides only information about the approximate position of an individual, the more detailed input from social media datasets can facilitate the study of individuals' sentiments when they participate in different activity types. In order to derive the sentiments of individuals, we extracted subjective information from embedded text messages by applying sentimental analysis. As we stated previously, we do not perform a comprehensive textual analysis but we are focusing on selected keywords that could reveal the sentiments of the individual. For determining sentiment a scaling system is adopted whereby words commonly associated with having a negative, neutral or positive sentiment are given an associated weight factor on a -10 to $+10$ scale. Nevertheless, regardless the used method, there are cases where individuals might comment on something allegorically or ironically and use words with negative meaning while they refer to something positive and vice versa. Therefore, there is a small chance that, in some cases, messages are assigned incorrectly to the wrong category.

After the analysis of a four-month sample of published data on social networks the observed sentiments of individuals while participating in different activity types are presented in table 1.

Table 1: Summary results of sentiment analysis for different activity types

Activity type	**Positive sentiment**	**Negative sentiment**
Home	26%	74%
Work	48%	52%
Fixed	54%	46%
Flexible	70%	30%

The observations from table 1 are indicative since they are based on individuals that reside in a specific metropolis; thus results might differ for different data samples from other cities.

4.2. Home Activity Type

In order to associate a location with the home activity type, let

$$x^{t,k} = \begin{cases} 1 & \text{if individual's location and activity at time } t \text{ and day } k \text{ is known} \\ 0 & \text{otherwise (i.e., no data)} \end{cases}$$

Let also $Y = \{y_1^{t,k}, y_2^{t,k}, ..., y_N^{t,k}\}$, where

$$y_i^{t,k} = \begin{cases} 1 & \text{if the examined individual is at location } L_i \text{ at time } t \text{ and day } k \\ 0 & \text{otherwise} \end{cases}$$

Moreover, let

$$w^{t,k} = \begin{cases} 1 & \text{if } k \text{ is a weekday} \\ 0 & \text{if } k \text{ is weekend day} \end{cases}$$

By examining individuals' data exchange we observed that each individual has a time period $\{Max_T \rightarrow Min_T\}$ at which he/she does not undertake any interaction activity (i.e., message composure, photo upload etc.). That period will be called from now on "inaction period" since the individual does not interact with social media over a significant time frame (i.e., one month) (see Fig. 1). During that time we assume that the individual is located at home as it is also proposed by other studies (Noulas et al. (2011)).

The lower bound Max_T is the latest time at which the examined individual interacted with social media and is derived after examining his/her interactions over at least one month's period. Furthermore, Min_T defines the upper bound and it is the earliest time in the morning at which an interaction is observed. After examining the temporal characteristics of individuals interactions we set 04:00 am as a threshold for dividing the night and morning periods; thus, interactions before 04:00 am are considered as night interactions and vice versa.

In a first attempt to associate a location with the home activity (rule I) it is assumed that home is the location with the more interactions over a significant time period (i.e., at least one month). Therefore, a location L_i is associated with the home activity type, A_1, if:

$$\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} w^{t,k} \geq \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_l^{t,k} w^{t,k} \quad \forall L_l \in \Lambda \tag{2}$$

In a second method (rule II) that was introduced in the work of Gkiotsalitis and Stathopoulos (2015b), the home activity might not be the location with the most interactions during a significant time period but the location which satisfies several criteria as they derived from the temporal analysis of the data and the analysis of individuals' sentiments while interacting. In summary, it is assumed that individuals interact more from home early in the morning and late at night and the possibility of being in a negative emotional state is higher while interacting from home - as it is observed from the

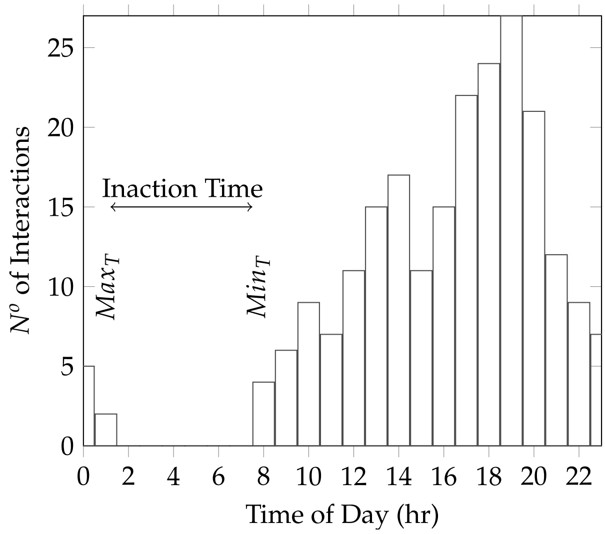

Figure 1: Temporal analysis of the number of tweets ("interactions") from a selected individual over a two-month period

emotion/activity type analysis presented in table 1. The following expressions formalize the three criteria of the empirical rule II.

$$\sum_{k=D_1}^{k=D_M} \sum_{t=Min_T}^{t=Min_T+2hr} y_i^{t,k} w^{t,k} + \sum_{k=D_1}^{k=D_M} \sum_{t=Max_T-3hr}^{t=Max_T} y_i^{t,k} w^{t,k} \geq$$
$$\sum_{k=D_1}^{k=D_M} \sum_{t=Min_T}^{t=Min_T+2hr} y_l^{t,k} w^{t,k} + \sum_{k=D_1}^{k=D_M} \sum_{t=Max_T-3hr}^{t=Max_T} y_l^{t,k} w^{t,k} \quad \forall L_l \in \Lambda \tag{3}$$

and

$$\sum_{k=D_1}^{k=D_M} \sum_{t=Min_T}^{t=Min_T+2hr} y_i^{t,k} w^{t,k} + \sum_{k=D_1}^{k=D_M} \sum_{t=Max_T-3hr}^{t=Max_T} y_i^{t,k} w^{t,k} \geq 0.3 \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_n} x^{t,k} \tag{4}$$

and

$$\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} l^{t,k} \leq 0 \tag{5}$$

where $l^{t,k}$ is a weight factor on a -10 to +10 scale that indicates the negative, neutral or positive affinity/mood of an individual after performing a sentiment analysis on the embedded text of the associated message. To calculate the distance of other locations from home we use the Haversine formula which returns the shortest distance between any two geographic coordinates. Therefore, the distance between a location L_i and the Home location, L_h, is set as F_i (in km) after calculating it with the Haversine formula.

4.3. Work Locations

Work activity is a special activity accumulating a number of specific attributes regarding its duration, starting time and day type. In a first empirical rule it is assumed that the location associated with the work activity is the one with the most interactions during day time (08:00 - 17:00). Nevertheless, after observing individuals' interactions we concluded that there are a number of criteria that should also be considered for improving the accuracy of the estimation. For instance, individuals interact from locations associated with the work activity type during daytime over weekdays and with a variance of more than six hours. In addition, interaction activities that reveal information about the duration of staying in a particular place such as chatting and continuous interacting are mainly observed at work location during day times.

Let $T_{L_i} = \{t_1, t_2, ..., t_\lambda\}$ be a set of day times at which the individual was at location L_i after considering the observed data from all dates. In a next step we sort the observed time set, T_{L_i}, in ascending order $T'_{L_i} = \{t'_1, t'_2, ..., t'_\lambda\}$, where, $t'_1 \leq t'_2 \leq ... \leq t'_\lambda$. We also define the 1^{st} and the 3^{rd} quartile, $Q_{25,i}$ and $Q_{75,i}$, where:

$$Q_{25,i} = t'_{\lfloor n_{25,i} \rfloor} + (n_{25,i} - \lfloor n_{25,i} \rfloor)(t'_{\lfloor n_{25,i} \rfloor + 1} - t'_{\lfloor n_{25,i} \rfloor}) \tag{6}$$

and

$$Q_{75,i} = t'_{\lfloor n_{75,i} \rfloor} + (n_{75,i} - \lfloor n_{75,i} \rfloor)(t'_{\lfloor n_{75,i} \rfloor + 1} - t'_{\lfloor n_{75,i} \rfloor}) \tag{7}$$

The values of $n_{25,i}$ and $n_{75,i}$ are calculated from the equations:

$$n_{25,i} = \frac{25}{100}(-1 + \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k}) + 1 \tag{8}$$

and

$$n_{75,i} = \frac{75}{100}(-1 + \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k}) + 1 \tag{9}$$

Then, a location L_i is a work location, A_2, if the three following criteria of the above-mentioned empirical rule are satisfied:

$$\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} w^{t,k} \geq 0.75(\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} - \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} w^{t,k}) \tag{10}$$

and

$$Q_{75,i} - Q_{25,i} \geq 360 \text{ minutes} \tag{11}$$

and

$$\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} (C_i^{t,k} w^{t,k} + O_i^{t,k} w^{t,k}) \geq 0.75(\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} (C_i^{t,k} w^{t,k} + O_i^{t,k} w^{t,k}) \quad) \tag{12}$$

4.4. Locations of Fixed Activities

A location L_i is the location of a fixed activity, A_3, if:

$$\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} \geq 0.05 \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} x^{t,k} \tag{13}$$

Fixed activities, A_3, is a general term that is used for describing several real-world activities such as visiting a shopping mall, a park, a restaurant, the school/university's premises and more. Fixed activities are more important than flexible activities since they are recurrent and in some cases is useful to know the exact type of a fixed activity.

To capture the actual types of fixed activities we introduce different categories which divide further the fixed activities into subgroups and link them to real-world activities. In the remaining of this section, we provide a rule-set for dividing further the fixed activities into six subcategories with respect to the input from the five-value dataset that is used in this study. Each category is linked to a number of real-world activities as they are derived after analyzing four-month data from published information on Twitter. For the calibration phase only, we adjusted the results of our rule-set to the real-world activities of individuals which were derived by using textual analysis and data from Geographic Information Systems (GIS) together with regional POIs.

One of the criteria for dividing fixed activities into different categories is their distance from the home location. After calculating the distance between the locations of fixed activities and the home location we introduce the following categories together with their associated real-world activities as presented in table 2.

The real-world activities are assigned to different categories according to their likelihood of occurrence as it is derived from the analysis of the four-month sample.

Table 2: Categories of Fixed Activities

Criteria	Category	Observed Visited Locations
$Q_{75,i} - Q_{25,i} < 4hr$ $F_i < 3.5km$	I: Local activities	Shopping Mall, Park, Sport Stadium, Restaurant/Bar, Gym, Language Class,University Lecture Theater, Children's School, Bus or Train Station
$Q_{75,i} - Q_{25,i} < 4hr$ $F_i \geq 3.5km$	II: Short-term, Long Distance Activities	Hotel, Business Trip Meeting Place, External Office, Home of a family member
$4hr \leq Q_{75,i} - Q_{25,i} < 4hr$ $0.05km \leq F_i < 15km$	III: Recurrent Non-Scheduled Activities	Hotel, Business Trip Meeting Place, External Office,Home of a family member
$4hr \leq Q_{75,i} - Q_{25,i} < 4hr$ $F_i \geq 15km$	IV: Medium Variance Activities, Long Distance	Hotel, Business Trip Meeting Place, External Office, Home of a family member
$Q_{75,i} - Q_{25,i} \geq 7hr$ $0.1km \leq F_i < 1km$	V: Recurrent Non-Scheduled Activities near Home	Working Place near Home, Local Shopping Center, Local Bus Stop
$Q_{75,i} - Q_{25,i} \geq 7hr$ $F_i > 1km$	VI: Large Variance, Long Distance Activities	Family's member house, Hotel, Vacation Place

5. Probabilistic Mobility Prediction Model

5.1. Educated Rules for Missing Data Imputation

Having allocated an activity to each representative location, we estimate the individual's travel schedule by establishing a set of rules. The rules are introduced to estimate the current states of individuals when data is not available and are particularly useful for datasets from social networking websites since, in most cases, individuals share information sparsely (see Fig. 2).

In general, the educated rules combine information from the pre-allocation of activi-

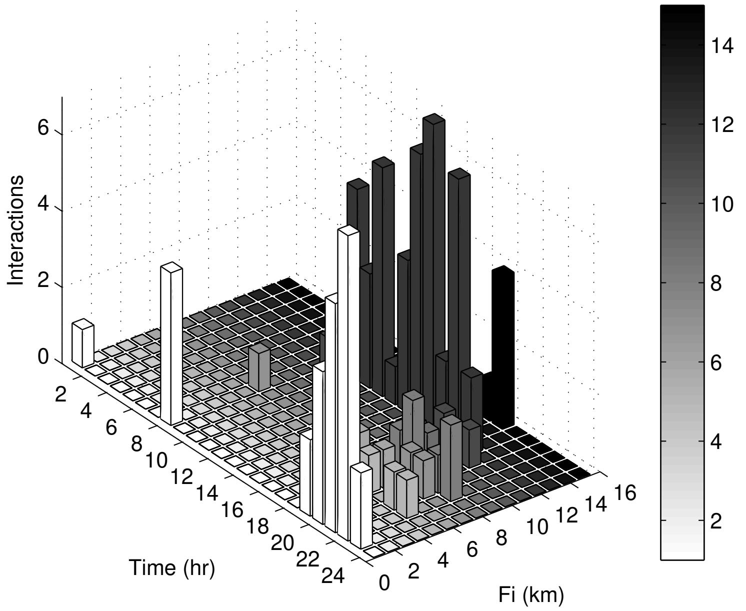

Figure 2: Spatio-temporal analysis of the interactions of a randomly selected individual over a two-month period

ties to representative locations with information about locations at which the individual generated content throughout a daily period to estimate his/her daily trajectory.

First, we assume that the individual stays at home during the inactive time period $\{Max_T \rightarrow Min_T\}$; thus, he/she is at the representative location which is associated with the home activity. When the user generates content from a different location than the location associated with the home activity for the first time $t_{1,l}$ at a day, D_l, we know that the user is at another representative location and participates in a fixed, A_3, a flexible, A_4, or work activity, A_2, at time $t_{1,l}$. The transfer time, $t_{s,l}$, when the user travels from A_1 to another representative location depends on the type of activity at the destination and is estimated as follows:

$$t_{s,l} = \begin{cases} \min\{Q_{15,i}; t_{1,l}\} & \text{if a fixed or work activity is performed at the destination location } L_i \\ t_{1,l} & \text{if a flexible activity is performed at the destination location } L_i \end{cases}$$

Since we have identified the duration and starting times of work and other fixed activities from the first processing of the individual's data, we adjust the transfer time from home to work or another fixed activity by considering not only the time at which the individual generated content from a different location but also the starting time and the duration of those activities. For performing such task, the 15th percentile of the activity duration is selected due to its better performance compared to the 25th percentile and the 5th percentile. For flexible activities we consider only their starting time since we do not collect information about their duration.

The location and activity values of the individual's state are changed only if he/she generates content from a different representative location than the one he/she was before.

If the user generates content at a location L_j at time $t_{h,l}$ and before he was at location $L_i \neq L_j$ at time $t_{h-1,l}$, then we assume that the user stayed at location L_i until time $t = t_{s,l}$ and changed locations at time $t_{s,l}$.

The transfer time depends on the interaction times $t_{h-1,l}$, $t_{h,l}$ and the type of the activities at locations L_i, L_j. For flexible activities we consider only the time at which the individual generated content. However, for a fixed or work activity we consider also the duration, the starting and ending times as we discussed above. In summary, after testing different scenarios the transfer times are set as:

$$t_{s,l} = \begin{cases} \max\{Q_{15,j}; t_{h-1,l}\} & \text{if the origin is a flexible activity and the destination is work or a fixed activity} \\ (t_{h-1,l} + t_{h,l})/2 & \text{if the origin is a flexible activity and the destination is home or a flexible activity} \\ (t_{h-1,l} + t_{h,l})/2 & \text{if the origin is home and the destination is a flexible activity} \\ \max\{Q_{15,j}; t_{h-1,l}\} & \text{if the origin is home and the destination is work or a fixed activity} \\ \min\{Q_{85,i}; t_{h,l}\} & \text{if the origin activity is work or a fixed activity and the destination home or a flexible activity} \\ (t_{h-1,l} + t_{h,l})/2 & \text{if the origin activity is work or a fixed activity and the destination work or a fixed activity} \end{cases}$$

In the case where the last daily interaction was at time $t_{y,l}$ from a different representative location than the one which is associated to home, we assume that the user returned to home at transfer time:

$$t_{s,l} = \begin{cases} \min\{Max_T; t_{y,l} + 2hr\} & \text{if the last interaction is from a flexible activity type location} \\ \max\{Q_{85,j}; t_{y,l}\} & \text{if the last interaction is from work or a fixed activity type location} \end{cases}$$

The educated rule-set for imputing the daily missing data returns the results of Fig. 3.

5.2. Modeling the choice of Activity Timing

In the previous section we estimated the timing, $t_{l,s}$ at which the individual transfers between consecutive activities on a deterministic way by using a set of educated rules. The estimated transfer times can be also adjusted with the use of theoretical models on top of our educated rules.

Models for modeling the timing of activities are based upon basic assumptions such as: individuals aim at optimizing the utility of their daily activity patterns (Meloni et al. (2004)), individuals derive a certain utility depending on the duration and the time of day they participate in activities and there is a disutility associated with the cost and the time spent while traveling (Akiva and Lerman (1985), Ettema et al. (2007)).

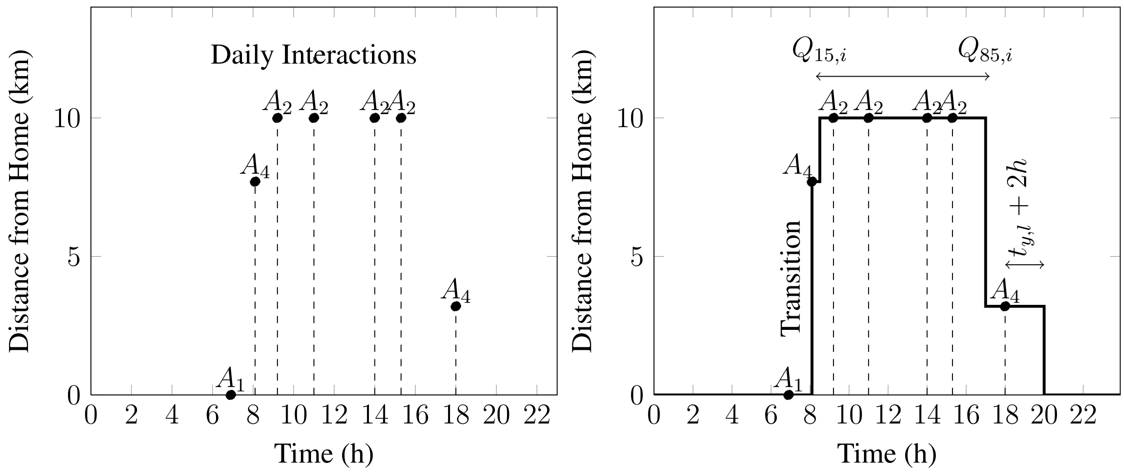

Figure 3: Estimating states evolution by imputing missing data

Assuming that $t_{l,s}$ is the ideal time when the next activity should start, we can use the notion of early and late schedule delay defined by Hess et al. (2007):

$$SDE_s = \max(0, (s - t_{l,s})) \tag{14}$$
$$SDL_s = \max(0, (t_{l,s} - s)) \tag{15}$$

where s is the actual arrival time at the next activity.

The arrival time s can be discretized to represent 5-minute wide intervals of arriving at the activity over a predetermined period (i.e., $s \in t_{l,s} - 30, t_{l,s} - 25, ..., t_{l,s} + 30$). Then, the utility of arriving at time s can be defined as:

$$U_s = \beta_1 SDE_s + \beta_2 SDL_s + \beta_3 + \varepsilon(s) \tag{16}$$

The parameters $\beta_1; \beta_2; \beta_3$ capture the trade-off between changes in the timing of arrival and scheduling costs. Furthermore, by assuming that the error term $\varepsilon(s)$ is independently and identically distributed according to a type 1 extreme value distribution, then the choice of arriving time is calculated from the following multinomial logit model:

$$P(s) = \frac{e^{\beta_1 SDE_s + \beta_2 SDL_s + \beta_3}}{\sum\limits_{l=t_{l,s}-30}^{l=t_{l,s}+30} e^{\beta_1 SDE_l + \beta_2 SDL_l + \beta_3}} \tag{17}$$

5.3. Simulating Individual's Daily Schedules

All dates k are divided into two main groups: w_1 which includes all weekdays and w_2 which contains all weekend days. The transition probability between two states can be defined as:

$$a_{i,j}^{t,k\in w_1} = \frac{\text{number of times being at state } q_j \text{ at time } t+1 \text{ while being at state } q_i \text{ at time } t \text{ and day type } w_1}{\text{number of times being at state } q_i \text{ at time } t \text{ and day type } w_1} \tag{18}$$

and

$$a_{i,j}^{t,k\in w_2} = \frac{\text{number of times being at state } q_j \text{ at time } t+1 \text{ while being at state } q_i \text{ at time } t \text{ and day type } w_2}{\text{number of times being at state } q_i \text{ at time } t \text{ and day type } w_2} \tag{19}$$

These probabilities can be used for estimating the daily schedule of an individual for a day type $k \in w_1$. The daily schedule is described by the evolution of different states during that day and is derived from the probability of being at state q_i at time $Min_T^{w_1}$:

$$P_{Min_T^{w_1}}(q_i) = P_{Min_T^{w_1}}(A_j|L_i)P_{Min_T^{w_1}}(L_i) = \frac{\sum_{k=D_1}^{k=D_M} y_i^{Min_T^{w_1}} w^{Min_T^{w_1},k}}{\sum_{k=D_1}^{k=D_M} x^{Min_T^{w_1}} w^{Min_T^{w_1},k}} \tag{20}$$

where $Min_T^{w_1}$ is the earliest time at which the individual posts a message on Twitter for day type $k \in w_1$.

The daily schedule of the individual is estimated via an iterative simulation by knowing the initial probability of being at a state q_i at time $Min_T^{w_1}$ and the transition probabilities $a_{i,j}^{t,k\in w_1}$. From the simulation, the probability of being at a state j at time t and day type $k \in w_1$ is given as:

$$\pi_j^{t,w_1} = \sum_i \pi_i^{t-1,w_1} a_{i,j}^{t-1,w_1} \tag{21}$$

The output is the estimation of the individual's daily schedule (Fig. 4) as a sequence of states: $\{q^{Min_T^{w_1}}, q^{Min_T^{w_1}+1}, ..., q^N\}$

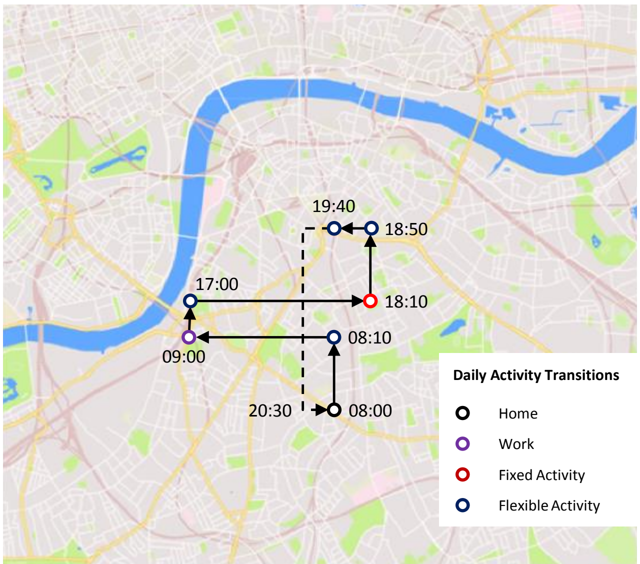

Figure 4: Simulating the daily evolution of states over a weekday [map source: Google maps]

6. Validating the Educated Rules for Allocating Activities to Locations

The rule-set for allocating activities to representative locations has been validated with the use of real data from Twitter (4-month sample from a metropolis in north Europe). The data includes information for the development of the five-value dataset. The validation phase examines the accuracy of the developed techniques on interpreting social media data and associating geo-tagged locations to activities. The results of allocating activities to locations (i.e., home activity, work activity, fixed activity, flexible activity) are compared with real data derived manually from GIS, data from POIs and the outcome of textual analysis on the content of users' messages.

By using the sample, 3572 activities were assigned to representative locations. To evaluate the performance of our techniques we introduce the following metrics:

$$E_h = \left(\sum_{z=1}^{z=S} \frac{\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} r_{i,z}^{t,k}}{\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_{i,z}^{t,k}} \right) / S \tag{22}$$

where E_h is the error of assigning the home activity to the wrong location, $r_{i,z}^{t,k} = 1$ if the corresponding location $L_{i,z}$ is assigned inappropriately to activity A_1 and $r_{i,z}^{t,k} = 0$ otherwise. In addition, S is the total number of individuals in the sample. The second metric is:

$$E_w = \left(\sum_{z=1}^{z=S} \frac{\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} R_{i,z}^{t,k}}{\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_{i,z}^{t,k}} \right) / S \tag{23}$$

Table 3: Errors in associating home and work-related activities of individuals to the right locations using the educated rules

Activity Type	**Sample Size**	**Errors**	
Home - using Rule I	1664	E_h=34.735%	F_h=29.931%
Home - using Rule II	1664	E_h=13.747%	F_h=11.765%
Work	552	E_w=39.1347%	F_w=317.432%

Table 4: Errors in associating fixed and flexible activities of individuals to the right locations

Activity Type	**Error**
Flexible (Failures: 138 — Sample Size: 1354)	9.845%
Fixed - Category 1 (Shopping Mall, Park, Sport Stadium, Restaurant/Bar, Gym, Language Class,University Lecture Theater, Children's School, Bus or Train Station)	1.312%
Fixed - Category 2 (Hotel, Business Trip Meeting Place, External Office, Home of a family member)	7.256%
Fixed - Category 3 (Hotel, Business Trip Meeting Place, External Office,Home of a family member)	12.526%
Fixed - Category 4 (Hotel, Business Trip Meeting Place, External Office, Home of a family member)	2.136%
Fixed - Category 5 (Working Place near Home, Local Shopping Center, Local Bus Stop)	9.254%
Fixed - Category 6 (Family's member house, Hotel, Vacation Place)	8.233%

where E_w is the error of assigning the work activity to the wrong location, $R^{t,k}_{i,z} = 1$ if the corresponding location $L_{i,z}$ is assigned inappropriately to activity A_2 and $R^{t,k}_{i,z} = 0$ otherwise. Similarly, E_f, is the error associated with assigning representative locations to flexible activities. In addition,

$$F_h = \frac{\sum\limits_{z=1}^{z=S} \vartheta_z}{S} \quad \text{and} \quad F_w = \frac{\sum\limits_{z=1}^{z=S} \varpi_z}{S} \tag{24}$$

are errors associated with the total failure of assigning home and work activities to representative locations where $\vartheta_z = 1$ and $\varpi_z = 1$ if the home and work activity of the z^{th} individual are not assigned to the appropriate representative location and $\vartheta_z = 0, \varpi_z = 0$ otherwise. The associated errors of the proposed techniques are presented in table 3 and table 4.

7. Concluding Remarks and Future Work

In this work we introduced educated rules that process information about the type, the timing, the duration and the recurrence of interactions together with the sentiments of individuals while interacting. The scope of this study was to associate representative locations to different activity types after processing Twitter data and develop a general framework that can be used for processing datasets from social media and SMS applications in order to automate the extraction of mobility-related information.

In addition, we introduced educated rules for imputing missing data which are particularly useful when social media are the only source of information due to the data sparsity. Finally, we proposed a Monte Carlo-based simulation method for estimating the evolution of daily states and transfers of the users over a weekday or a weekend day taking as input the outcome from the proposed techniques.

After validating our educated rules against a four-month sample the results were promising and the accuracy of allocating activities to locations was more than 85% with only exception the assignment of the work activity of users to the right location (82.67% accuracy). Thus, the study showed that it is possible to use social media as a standalone or supplementary method for surveying the population and -more importantly- for estimating the activities and the mobility behavior of individuals even if the data from social media is more sparse compared to the cellular data from mobile phones (Gkiotsalitis and Stathopoulos (2016)). Given the already promising results, we propose a detailed textual analysis in the content of the embedded text messages in order to leverage on the publicly shared content for improving the accuracy of activity allocations.

In future research, we envision the development of a mobility demand prediction model which uses the mobility patterns of individuals as they are estimated after applying our educated rules. Finally, we propose the integration of survey data with social media data in order to study potential accuracy improvements on predicting the mobility demand as another future research topic.

REFERENCES

Akiva, M. E. B., Lerman, S. R., 1985. Discrete Choice Analysis: Theory and Application to Predict Travel Demand. MIT Press.

Algers, S., Eliasson, J., Mattsson, L.-G., 2005. Is it time to use activity-based urban transport models? a discussion of planning needs and modelling possibilities. The Annals of Regional Science 39 (4), 767–789.

Arentze, T., Timmermans, H., 2008. Social networks, social interactions, and activity-travel behavior: a framework for microsimulation. Environment and Planning B: Planning and Design 35 (6), 1012–1027.
URL http://www.envplan.com/abstract.cgi?id=b3319t

Axhausen, K. W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transport reviews 12 (4), 323–341.

Baek, S., Lim, Y., Rhee, S., Choi, K., Mar. 2010. Method for estimating population OD matrix based on probe vehicles. KSCE Journal of Civil Engineering 14 (2), 231–235.
URL http://link.springer.com/article/10.1007/s12205-010-0231-4

Bowman, J. L., Ben-Akiva, M. E., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transportation research part a: policy and practice 35 (1), 1–28.

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011a. Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in boston metropolitan area. IEEE Pervasive Computing 99.

Calabrese, F., Lorenzo, G. D., Liu, L., Ratti, C., 2011b. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Computing 10 (4), 36–44.

Carrasco, J. A., Hogan, B., Wellman, B., Miller, E. J., 2008. Collecting social network data to study social activity-travel behavior: an egocentric approach. Environment and Planning B: Planning and Design 35 (6), 961 – 980.
URL http://www.envplan.com/abstract.cgi?id=b3317t

Chaniotakis, E., Antoniou, C., Grau, J. M. S., Dimitriou, L., 2016. Can social media data augment travel demand survey data? In: Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE, pp. 1642–1647.

Dong, X., Ben-Akiva, M. E., Bowman, J. L., Walker, J. L., 2006. Moving from trip-based to activity-based measures of accessibility. Transportation Research Part A: policy and practice 40 (2), 163–180.

Ettema, D., Bastin, F., Polak, J., Ashiru, O., Nov. 2007. Modelling the joint choice of activity timing and duration. Transportation Research Part A: Policy and Practice 41 (9), 827–841.
URL http://www.sciencedirect.com/science/article/pii/S0965856407000328

Flyvbjerg, B., Jul. 2005. Measuring inaccuracy in travel demand forecasting: methodological considerations regarding ramp up and sampling. Transportation Research Part A: Policy and Practice 39 (6), 522–530.
URL http://www.sciencedirect.com/science/article/pii/S0965856405000273

Flyvbjerg, B., Skamris Holm, M. K., Buhl, S. L., 2006. Inaccuracy in traffic forecasts. Transport Reviews 26 (1), 1–24.
URL http://www.tandfonline.com/doi/abs/10.1080/01441640500124779

Gkiotsalitis, K., Alesiani, F., Maslekar, N., Apr. 24 2018. Method and system for providing demand-responsive dispatching of a fleet of transportation vehicles, and a mobility-activity processing module for providing a mobility trace database. US Patent 9,953,539.

Gkiotsalitis, K., Stathopoulos, A., 2015a. Optimizing leisure travel: Is bigdata ready to improve the joint leisure activities efficiency? In: Engineering and Applied Sciences Optimization. Springer, pp. 53–71.

Gkiotsalitis, K., Stathopoulos, A., 2015b. A utility-maximization model for retrieving users willingness to travel for participating in activities from big-data. Transportation Research Part C: Emerging Technologies 58, 265–277.

Gkiotsalitis, K., Stathopoulos, A., 2016. Demand-responsive public transportation re-scheduling for adjusting to the joint leisure activity demand. International Journal of Transportation Science and Technology 5 (2), 68–82.

Gkiotsalitis, K., Wu, Z., Cats, O., 2019. A cost-minimization model for bus fleet allocation featuring the tactical generation of short-turning and interlining options. Transportation Research Part C: Emerging Technologies 98, 14–36.

Gonzàlez, M. C., Hidalgo, C. A., Barabàsi, A.-L., Jun. 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.
URL http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located twitter as proxy for global mobility patterns. Cartography and Geographic Information Science 41 (3), 260–271.

Hess, S., Polak, J. W., Daly, A., Hyman, G., Mar. 2007. Flexible substitution patterns in models of mode and time of day choice: new evidence from the UK and the netherlands. Transportation 34 (2), 213–238.
URL http://link.springer.com/article/10.1007/s11116-006-0011-7

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., 2015. Understanding human mobility from twitter. PloS one 10 (7), e0131469.

Kitamura, R., 1988. An evaluation of activity-based travel analysis. Transportation 15 (1-2), 9–34.

Lee, J. H., Gao, S., Goulias, K. G., 2015. Can twitter data be used to validate travel demand models. In: 14th International Conference on Travel Behaviour Research.

Lin, D.-Y., Eluru, N., Waller, S., Bhat, C., Dec. 2008. Integration of activity-based modeling and dynamic traffic assignment. Transportation Research Record: Journal of the Transportation Research Board 2076 (-1), 52–61.
URL http://dx.doi.org/10.3141/2076-06

Maat, K., Timmermans, H. J., Molin, E., 2004. A model of spatial structure, activity participation and travel behavior. In: 10th World Conference on Transport Research.

Meloni, I., Guala, L., Loddo, A., Feb. 2004. Time allocation to discretionary in-home, out-of-home activities and to trips. Transportation 31 (1), 69–96.
URL http://link.springer.com/article/10.1023/B%3APORT.0000007228.44861.ae

Munizaga, M. A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. Transportation Research Part C: Emerging Technologies 24, 9–18.

Musolesi, M., Mascolo, C., Jul. 2007. Designing mobility models based on social network theory. SIGMOBILE Mob. Comput. Commun. Rev. 11 (3), 59–70.
URL http://doi.acm.org/10.1145/1317425.1317433

Nguyen, T., Szymanski, B., 2012. Using location-based social networks to validate human mobility and relationships models. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 1215–1221.

Noulas, A., Scellato, S., Mascolo, C., Pontil, M., 2011. An empirical study of geographic user activity patterns in foursquare. ICwSM 11, 70–573.

Pan, C., Lu, J., Di, S., Ran, B., Jan. 2006. Cellular-based data-extracting method for trip distribution. Transportation Research Record: Journal of the Transportation Research Board 1945 (-1), 33–39.
URL http://dx.doi.org/10.3141/1945-04

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19 (4), 557–568.

Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. International Journal of Urban Sciences 18 (1), 31–60.

Shang, W. X., Zhang, Y., Zhou, J., Zhu, Y. F., Jul. 2012. Predicting dynamic transportation demand with mobility data. International Classification: G06Q10/04 U.S. Classification: H04W 4/02P6; G06Q 30/0202.

Stopher, P., Collins, A., 2005. Conducting a gps prompted recall survey over the internet. In: Transportation Research Board Annual Meeting, 84th, 2005, Washington, DC, USA.

Stopher, P. R., 1996. Methods for household travel surveys. Vol. 236. Transportation Research Board.

Stopher, P. R., Greaves, S. P., 2007. Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice 41 (5), 367–381.

Sun, L., Lee, D.-H., Erath, A., Huang, X., 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system. In: Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, pp. 142–148.

Van Der Zijpp, N., Jan. 1997. Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data. Transportation Research Record: Journal of the Transportation Research Board 1607 (-1), 87–94.
URL http://dx.doi.org/10.3141/1607-13

White, J., Wells, I., 2002. Extracting origin destination information from mobile phone data. In: Road Transport Information and Control, 2002. Eleventh International Conference on (Conf. Publ. No. 486). pp. 30–34.

Wu, L., Zhi, Y., Sui, Z., Liu, Y., 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. PloS one 9 (5), e97010.

Zhang, L., Levinson, D., 2004. Agent-based approach to travel demand modeling: Exploratory analysis. Transportation Research Record: Journal of the Transportation Research Board (1898), 28–36.

Zhang, Y., Qin, X., Dong, S., Ran, B., 2010. Daily o-d matrix estimation using cellular probe data.
URL http://trid.trb.org/view.aspx?id=910539

Zheng, N., Waraich, R. A., Axhausen, K. W., Geroliminis, N., 2012. A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. Transportation Research Part A: Policy and Practice 46 (8), 1291–1303.

Table .5: Indicative example of a small portion of Twitter data from one user using the Twitter API

Day	**Month**	**Date**	**Time**	**Lat.**	**Lon.**	**Recipients**	**Message**
Wed	Jan	29	07:54:33	51.52065563	-0.21149316	N/A	*
Thu	Jan	16	08:10:31	24.42988412	54.64383623	N/A	*
Thu	Jan	16	07:06:21	24.42994406	54.64364278	N/A	*

*Not presented here for user privacy reasons

Zhou, X., Mahmassani, H., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. IEEE Transactions on Intelligent Transportation Systems 7 (1), 105–114.

A small portion from a sample of Twitter Data from one user is presented in table .5. The text message of the tweets is considered at the sentiment analysis but is not presented in the table for privacy reasons.