T04

Building workers' travel demand models based on mobile phone data

Feng Liu[a], Davy Janssens[b], JianXun Cui[c], Geert Wets[b]

[a,b] Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6,
B-3590, Diepenbeek, Belgium

[c] Department of transport engineering, Harbin Institute of Technology (HIT), 1500, Harbin,
China

[a] Corresponding author: Tel: +32 0 11269125 fax: +32 0 11269199

E-mail addresses: feng.liu@uhasselt.be (F. Liu), davy.janssens@uhasselt.be (D. Janssens),
cuijianxun@hit.edu.cn (J.X. Cui), geert.wets@uhasselt.be (G. Wets)

Abstract

Daily activity-travel sequences of individuals have been estimated by activity-based transportation models. The sequences serve as a key input for travel demand analysis and forecasting in the region. However, the high cost along with other limitations inherent to traditional travel data collecting methods has hampered the models' further advancement and application, particularly in developing countries. With the wide deployment of mobile phone devices today, we explore the possibility of using mobile phone data to build such a travel demand model.

Our exploration consists of four major steps. First, home, work and other stop locations for each user are identified, based on their mobile phone records. All the obtained locations along with their particular orders on a day are then formed into stop-location-trajectories and classified into clusters. In each cluster, a Hidden Markov Model (HMM) is subsequently constructed, which characterizes the probabilistic distribution of activities and their related travel of the sequences. Finally, the derived models are used to simulate travel sequences across the entire employed population.

Using data collected from natural mobile phone usage of around 9 million users in Senegal over a period of one year, we evaluated our approach via a set of experiments. The average length of daily sequences drawn from the stop-location-trajectories and the simulated results is 4.55 and 4.72, respectively. Among all the 677 types of the stop-location-trajectories, 520 (e.g. 76.8%) are observed from the simulated sequences, and the correlation of sequence frequency distribution over all the types between these two sequence sets is 0.93. The experimental results demonstrate the potential and effectiveness of the proposed method in capturing the probabilistic distribution of activity locations and their sequential orders revealed by the mobile phone data, contributing towards the development of new, up-to-date and cost-effective travel demand modelling approaches.

Keywords activity-travel sequences, Hidden Markov Model, activity-based transportation models, travel surveys, mobile phone data.

1. Introduction

1.1. Activity-based transportation models

The main premise of *activity-based transportation models* is the treatment of travel behavior as a derived demand of activity participation. In this modeling paradigm, travel is analyzed through daily patterns of activity behavior related to and derived from the context of land-use and transportation network as well as personal characteristics such as social-economic background, lifestyles and needs of individuals (e.g. Bhat & Koppelman, 1999; Davidson et al., 2007; Wegener, 2013).

All the above information, complemented with a training set of household *travel surveys* which record the full daily activity-travel sequences of a small sample of individuals during one or a few days, is analyzed and translated into heuristic decision making rules, using machine learning techniques, e.g. decision trees (e.g. Arentze & Timmermans, 2004; Bellemans et al., 2010). These rules represent the scheduling process of activities and travel by the individuals. Once established, the activity-based models can be used as the probabilistic basis for a micro-simulation process using Monte Carlo methods, in which complete daily activity-travel sequences for each individual in the whole region are synthesized. The synthesized sequences are then aggregated into travel measures, e.g. the average number of trips or travel distances per day, or an *origin-destination (OD) matrix*. The OD matrix represents the number of trips between each pair of locations of the region, and it can be assigned to a road network through traffic assignment algorithms. The derived travel measures as well as the amount of travel assigned to specific roads can subsequently serve as essential input for travel analysis in the region, such as travel demand forecasting, emission estimates, and the evaluation of emerging effects caused by different transport policy scenarios. Fig. 1 illustrates the entire process of an activity-based transportation model.

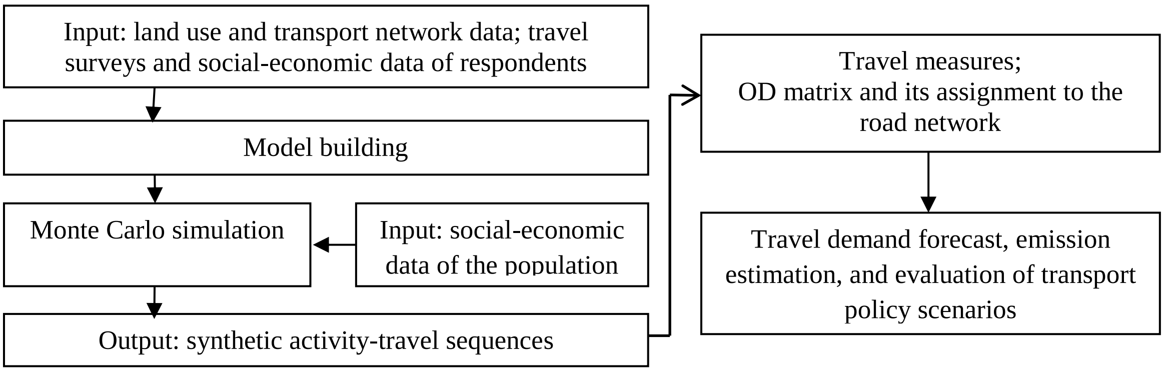
Input: land use and transport network data; travel surveys and social-economic data of respondents
Model building
Monte Carlo simulation
Input: social-economic data of the population
Output: synthetic activity-travel sequences
Travel measures; OD matrix and its assignment to the road network
Travel demand forecast, emission estimation, and evaluation of transport policy scenarios

Fig. 1. The entire process of an activity-based transportation model

1.2. Problem statement

Despite comprehension and advancement of activity-based transportation models, e.g. Albtross (Arentze & Timmermans, 2004), TASHA (Roorda et al., 2008), Feathers (Bellemans et al., 2010), the availability of household travel surveys has been a prerequisite condition for the model building, regardless of the following drawbacks of the data collection method (e.g. Asakura & Hato, 2006; Cools et al., 2009). (i) The entire survey is a lengthy process; from the initial data gathering to data cleaning and the exploitation of the first results, it could take months even years, causing a time lag between the data initially obtained and the results that are required for objective and up-to-date activity-travel behavior analysis. (ii) It imposes a significant burden on respondents, resulting in low response rates and under-reporting of short trips. (iii) Despite the above disadvantages, the data is very expensive to collect, leading to only a limited number of respondents and a (or a few) day(s) being involved in the surveys. Consequently, this tends to obfuscate the less frequent activities, such as sports or telecommuting activities which are often carried out once a week or once a month. Questions are also raised about the capability of such limited sample size in representing activity-travel behavior of a whole population.

Apart from travel surveys, travel information has also been gathered from sensors, e.g. loop detectors and video cameras, which are installed in a road network to monitor traffic flow. However, the sensors are usually set up on highways, as it is expensive to instrument a whole region with such static devices. Consequently, the collected data is only limited to the high-capacity roads, and sheds little light on the traffic flow in the rest of the area (e.g. Gühnemann et al., 2004).

Due to the data constraints, the existing methods on travel behavior analysis and travel demand modeling are restricted to only a (or a few) statistical average day(s) and a relatively small region as well as to a subset of the population, because of the lack of a large dataset that is spatially and temporally extensive as well as involves more individuals. Consequently, the results are difficult to be generalized to evaluate travel demand in various types of days (e.g. weekdays, weekend and holidays) and at a higher geographical scale (e.g. an entire city or a whole country). For a long time, data problems have been one of the essential challenges of the current research on travel demand modelling. The problems have seriously hampered further development and application of the existing techniques (e.g. Hartgen, 2013; Janssens et al., 2012). Having accurate, reliable, while affordable travel data for the estimation of travel demand and the subsequent analysis on transport network systems has thus been a major concern, particularly in developing countries.

1.3. Mobile phone data: a new data source for travel demand modelling

The wide deployment of mobile phones has created the opportunity to use the devices as a new data collection method to overcome the lack of reliable travel data (Jiang et al., 2013). Location data recorded from mobile phone devices reflects up-to-date travel patterns on a significantly large sample of a population, making the data a natural candidate for the analysis of mobility phenomena in the region (e.g. Do & Gatica-Pereza, 2013; Schneider et al., 2013). In addition, the data collection is a by-product of mobile phone companies for billing and operational purposes that generates neither extra expenses nor respondent burden.

The importance and added value of mobile phone data in the field of transportation research have been manifested by a variety of studies, ranging from the investigation of key dimensions of human travel, such as travel distances and time expenditure at different locations (e.g. González et al., 2008; Schneider et al., 2013; Song et al., 2010), to the discovery of typical mobility patterns (e.g. Bayir et al., 2009; Berlingerio et al., 2013; Calabrese et al., 2011), and to the examination of the status and efficiency of current transport network systems (e.g. Angelakis et al., 2013; Steenbruggen et al., 2013). Particularly, mobile phone data has been employed to explore the possibilities of building travel demand models, e.g. OD matrices (e.g. Becker et al., 2011; Calabrese et al., 2011; Shan et al., 2011). The research by (Shan et al., 2011) can represent the typical process of such exploration. The study utilizes mobile phone data of more than 0.3 million users collected in the metropolitan area of Lisbon, Portugal for an entire month. In this process, the two most frequent call cell towers for each of the users are first identified as the residential and employment locations, respectively. Using the two obtained locations, an OD matrix depicting home-to-work commuting trips in the morning is then built. Based on a census survey, this derived OD matrix is subsequently scaled up to account for the total employed population of 1.3 million in the study area. The adjusted matrix is ultimately used to compare against the travel demand during the same morning period forecasted by an integrated land use and transportation model developed in this region. The results show comparative performance of this OD matrix in estimating the morning travel demand in this region.

However, despite its advancement by incorporating mobile phone data into the modeling process, the OD-based method does not consider the sequential information which is imbedded in activity-travel patterns. A detailed analysis of the sequential dependencies of the daily activities from activity-travel behavior is thus ignored in the modeling process. It has been widely acknowledged that the choice of activities is dependent on the preceding activity engagement (e.g. Joh et al., 2008; Wilson, 2008), exemplified by the fact that, during one particular working day, it is highly probable that the combination of having breakfast, travel and working is observed together. On the contrary, if a sports activity is carried out in the morning, there is a small chance that it is performed again in the evening. The interdependencies of daily activities have been considered as a crucial factor in the activity-travel decision making process (e.g. Delafontaine et al., 2012; García-Díez et al., 2011). A modeling process, which takes into account the sequential information and generates activity-travel sequences that are consistent with the sequential constraints observed from real travel behavior, is thus important. The existing activity-based models have integrated the sequential information of daily activities into the modeling process. But as previously described, the activity-based models are constructed based on a small set of activity-travel sequences from travel surveys, thus subject to the shortcomings that are inherent to the traditional data collection methods. A model, which is based on massive mobile phone data while taking into account the sequential aspect of activity-travel behavior, has so far been lacking.

1.4. Research contributions

Extending the current studies on the application of mobile phone data to transportation research, and particularly addressing the above mentioned limitations in the development of travel demand models, our study proposes a new approach which is based on the phone data and considers the sequential information imbedded in activity-travel patterns. Specifically, this study is to build a workers' travel demand model based on mobile phone data using Hidden Markov Modeling (HMM) techniques. The derived model characterizes the probabilistic distribution of activities and their related travel on a day among workers. The models can be used to simulate new activity-travel sequences across the whole employed population. The synthesized sequences can be subsequently aggregated into certain travel measures which serves as important input for travel demand analysis in the region.

Compared to existing activity-based models, this approach offers the following advantages. (i) This method is built upon the observed current activity-travel behavior of a large proportion of population, thus providing a more representative and up-to-date modeling process. (ii) Through a long period of mobile phone data records, inter- and intra- personal variations of travel behavior as well as weekday, weekend and seasonal deviations are captured. (iii) The use of mobile phone data generates no extra financial cost in terms of data collection, making it a cost-effective approach. This is particularly practical in developing countries where, as stated before, the high cost of traditional travel data collection mechanisms combined with other disadvantages of the methods have deterred the much needed development of a new, effective and cheaply realized travel demand modelling technique. With the use of the large-scale mobile phone data, the proposed method can be regarded as a reality mining approach which places the realized trips of travellers in daily life directly at the centre of the analytical process. (iv) When this method is compared with the OD-based modeling approach, the OD-based method analyzes travel behavior in terms of the distribution of all individual trips over different pairs of origin-destination locations; it is an aggregated modeling process. While the approach developed in this study examines the entire activity-travel sequences and focuses on the sequential aspect of travel behavior. In this new approach, the locations which are accessed by an individual on the same day are viewed and tackled as a whole, rather than an isolated participation in activities. Both methods analyze activity-travel behavior from different perspectives, thus providing a complementary means of modeling travel demand based on mobile phone data. In addition, while the OD-based approach is just an end product of the observed behavior from the phone data, and reflects the current mobility phenomena; the model proposed in this study is able to predict travel demand in regions where no phone data is provided or in future scenarios, e.g. the displacement of residential areas or the establishment of new industrial sites.

The remainder of this paper is organized as follows. Section 2 introduces the mobile phone data and Section 3 details the proposed modeling approach. A case study is conducted in Section 4, and a comparison of the modeling results against the data in the validation set is carried out in Section 5. Finally, Section 6 ends this paper with major conclusions and discussions for future research.

2. Mobile phone data description

The mobile phone dataset consists of full mobile communication patterns of around 9 million users in Senegal between January 1, 2013 to December 31, 2013 (de Montjoye et al., 2014). The dataset contains the location and time when each user conducts a call activity, including initiating or receiving a voice call or text message, enabling us to reconstruct the user's time-resolved call location trajectories. The locations are represented with the identifications of base stations (cells) in a GSM network; the radius of each of the stations ranges from a few hundred meters in metropolitan to a few thousand in rural areas, controlling our uncertainty about the user's precise location. Despite the low accuracy of users' exact locations, the massive mobile phone data represents a significant percentage (i.e. 69%) of this country's total population, providing a valuable source and opportunity for the analysis on human travel behavior and for drawing relevant inferences that can be statistically sound and representative. In order to address privacy concerns, the original dataset has been split into consecutive two-week periods. In each period, users are randomly selected and assigned to anonymized identifiers. New random identifiers are chosen for re-sampled users in different time periods. The data process results in totally 25 randomly sampled datasets, each of which contains communication records of 300,000 users over two weeks. One of these datasets is selected for this study. Table 1 illustrates typical call records of an individual identified as *user20* on Thursday, January 24th, 2013.

Table 1. The typical call data of an individual

Time	11:57:00	13:40:00	16:59:00	17:43:00	21:28:00
Cell_id	751	749	177	751	751

3. Methodology

3.1. Overview of the approach

The method is composed of 4 major steps. (i) Home, work and other stop locations for each user are identified, based on their mobile phone records. (ii) The obtained location trajectories are clustered according to the travel features encoded in the sequences. (iii) In each cluster, a Hidden Markov Model is constructed, which characterizes the probabilistic distribution of the corresponding sequences. (vi) The obtained models are used to simulate activity-travel sequences across the whole employed population in the study region. The overall structure of the approach is shown in Fig. 2, and the detailed procedure is elaborated as follows.

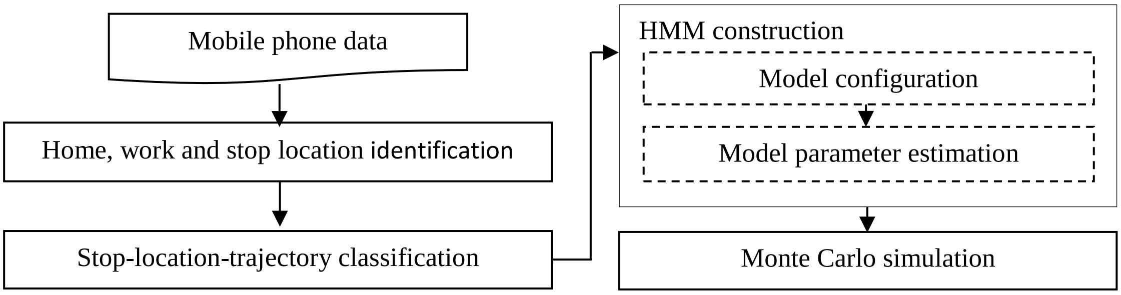
Mobile phone data
Home, work and stop location identification
Stop-location-trajectory classification
HMM construction
Model configuration
Model parameter estimation
Monte Carlo simulation

Fig. 2. The overall structure of the methodology

3.2. Home, work and other stop location identification

3.2.1. Mobile phone call location trajectories

A call location trajectory from a mobile phone user during a day, i.e. *call-location-trajectory*, is defined as a series of locations where the user makes calls when traveling or doing activities, as the day unfolds. It can be formulated as a sequence of l_1 -> l_2 -> ... -> l_n, where n is the *length* of the sequence, i.e. the total number of locations that the user has travelled to when making calls that day, and l_i $(1 \leq i \leq n)$ is the identification of the locations, e.g. cell IDs in this study. At each l_i, there could be multiple calls k_i $(k_i \geq 1)$, referred as *call-frequency*; the time for each of the calls is denoted as $T(l_i,1)$, $T(l_i,2)$, ..., $T(l_i,k_i)$, respectively. The time interval between the first and the last call time in the set of consecutive calls, i.e. $T(l_i,k_i) - T(l_i,1)$, is defined as *call-location-duration*. Accommodating the time signatures of the multiple calls, a call-location-trajectory can be represented as $l_1(T(l_1,1),T(l_1,2),\ldots,T(l_1,k_1))$ -> ... -> $l_n(T(l_n,1),T(l_n,2),\ldots,T(l_n,k_n))$, simplified as $l_1(T(1),T(2),\ldots T(k_1))$ -> ... -> $l_n(T(1),T(2),\ldots,T(k_n))$. Given the above call-location-trajectories constructed from the mobile phone data, the home and work locations are first predicted. This is followed by the identification of stop locations where activities are carried out.

3.2.2. Prediction of home and work locations

Various methods have been proposed to derive home and work locations from mobile phone data, mainly based on the visited frequency of a location during a particular time period (e.g. Becker et al., 2011; Calabrese et al., 2011). However, different time windows have been specified in these methods, depending on the context of the study area. In this study, a similar approach is adopted, but the time windows are empirically estimated from the mobile phone data as follows. The time period when call activities start to increase considerably in the morning during weekdays is chosen as the work start time, denoted as *work-start-time.* Similarly, the moment when the second peak of call activities start to appear in late afternoon is considered as the work end time, referred as *work-end-time*. Around this time, it is assumed that people start to communicate for off-work activity engagement.

Based on these two temporal points, a location is defined as the home location if it is the most frequent stop throughout the weekend period as well as during the night-time interval on weekdays between work-end-time and work-start-time. On the contrary, a location is considered as a work place if it satisfies the following criteria. (*i*) It is the most common place for call activities in the perceived work period between work-start-time and work-end-time on weekdays. (*ii*) It is not identical to the previously identified home location for the user. (*iii*) The calls at the location are not limited in only one day, they should occur at least 2 days a week.

With the above-defined identification criteria, we assume that people have only one home location and at most one work location. The additional locations, which are occasionally accessed for home or work activities, are regarded as a stop for non-mandatory activities. In addition, only individuals, who work in areas different from their home locations and who work at least two days per week, are included for the analysis of workers' travel behavior.

3.2.3. Identification of stop locations

After the identification of the distinct home and work locations for each user, the remaining locations in the call-location-trajectories are either *stop-locations* where people pursue non-mandatory activities or *non-stop-locations*. Each of these non-stop-locations can be further divided into either a *trip-location* where the user is traveling, or a *false-location* that is wrongly documented due to location update errors. The location update errors normally occur when call traffic is busy in the user's real location area, and consequently this location is shifted to less crowded cells for short time periods, causing location area updates, without the users' actual moving (e.g. Calabrese et al., 2011).

In addition, for the identified home or work locations, some occurrences of the locations could also be caused by non-stop reasons, e.g., people travelling in the same area as their home locations when making calls. Therefore, each location occurrence in the call-location-trajectories will be classified into stop-locations and non-stop ones, regardless its activity type.

The scenarios, where the two types of non-stop-locations could occur, can be illustrated with the call records of two typical users. The trajectory from the first user, identified as *user265,* is $l_1(17{:}06,17{:}43)$ -> $l_2(17{:}51)$ -> $l_3(17{:}56,19{:}41)$ -> $l_4(21{:}55)$, where 4 locations are observed, with the call-location-duration as 37, 0, 105 and 0 min respectively. Each of these locations needs to be identified as either a stop visit or just a passing-by place. The trajectory of the second user, i.e. *user72,* is $l_1(13{:}21,20{:}11)$ -> $l_2(22{:}00)$ -> $l_3(22{:}02)$ -> $l_4(22{:}05)$ -> $l_2(22{:}07,23{:}12)$. This user has 5 location updates, with the call-location-duration as 410, 0, 0, 0 and 65 min respectively. It should be noted that the time interval between the first and second visit to location l_2 is only 7 min. Although there is a possibility that this user may have travelled at a high speed during this period, the temporary interruption of l_2 by the extra locations l_3 and l_4 in such a short interval is most likely resulted from the location update errors. Consequently, locations l_3 and l_4 are falsely connected to the user's mobile phone at 22:02 pm and 22:05 pm although he/she had been actually remaining at location l_2 during this period.

In order to identify the stop-locations, the approach proposed in the study (Liu et al., 2014) is used, which consists of the following steps. (*i*) For each location l_i, the call-location-duration is first examined. If it is longer than a certain time limit, denoted as $T_{call\text{-}location\text{-}duration}$, this location is considered as a stop-location. (*ii*) Otherwise, if the condition does not hold (e.g. only a single call made at l_i), and if the location appears in the middle of a daily sequence of n, i.e. $1 < i < n$, a second parameter, namely *maximum-time-boundary*, defined as the time interval between the last call time at l_i's previous location and the first call time of its next location, i.e. $T(l_{i+1},1) - T(l_{i-1},k_{i-1})$, is computed. If this time period is longer than a threshold value, defined as $T_{maximum\text{-}time\text{-}boundary}$, l_i is perceived as a stop visit. (*iii*) When l_i is in the first or last position of a trajectory and the call-location-duration is shorter than $T_{call\text{-}location\text{-}duration}$, there is no sufficient information to estimate maximum-time-boundary for this visit. Thus, all the distinct locations, where the user has stayed at least once for conducting an activity over the entire survey period, are collected. These locations are considered as potential stop locations that are on the user's daily activity agenda and that are visited either routinely or once in a while. If l_i is one of these locations, it is assumed to be a stop for activity purposes. In contrast, if l_i is the place where the individual has not been observed doing activities, it is then considered as a passing-by place or being recorded as a localization error and therefore removed.

After the removal of locations that are either trips or stem from localization errors, all the remaining locations from a call-location-trajectory are regarded as stops and formed a *stop-location-trajectory*. Based on the above described identification process, if a duration of 30 and 60 min are used for $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ respectively, as set up in our experiment described in Section 4, the obtained stop-location-trajectories for *user265* and *user72* are l_1 -> l_3 -> l_4 and l_1 -> l_2 respectively.

3.3. Stop-location-trajectory classification

Each location l_i in the previously obtained stop-location-trajectories is complemented with its function, denoted as *activity*(l_i), categorized into home, work and non-mandatory activities, represented as 'H', 'W' and 'O', respectively. While H and W encapsulate all activities performed at home and work (including school) places respectively; O refers to all activities undertaken outside home and work places, differentiated between maintenance activities (e.g. shopping, banking or visiting doctors) and discretionary activities (e.g. social visits, sports or going to restaurants) (e.g. Arentze & Timmermans, 2004). Travel is implicit in between each two consecutive locations of the sequences.

Various methods have been used to classify activity sequences, mainly based on either a priori scheme or a numerical distance measure. A priori scheme aims to cluster the sequences according to predefined variables, e.g. socio-demographic factors of respondents or activity-travel features of the sequences. For example, researches (Spissu et al., 2009) first extract activity sequences of all employed people and then divide the sequences into HWH, HOH, HOWH, HWOH and HWOWH, depending on whether non-mandatory activities are involved, and if so, on when these non-mandatory activities are conducted. This classification method provides a simple way to build the clusters and to analyze the correlation between the behavior of each cluster and the socio-demographic characteristics of the corresponding individuals. Numerical distance measure methods, on the other hand, classify activity-travel sequences based on some measures of distances between the sequences, such as the number of identical activities (e.g. Roorda & Miller, 2008) or the similarities of the activities and their sequential order derived using sequence alignment methods (SAM) (e.g. Joh et al., 2008; Saneinejad & Roorda, 2009).

In this study, the stop-location-trajectories are classified based on the travel features of the sequences, i.e. the number of home based tours on the days. Two types *of home-based tours*, including *home-based-work-tour* and *home-based-non-work-tour*, are defined as a chain of locations (trips) that starts and ends at home and accommodates at least one work or one non-mandatory location visit, respectively. Based on this definition, a stop-location-trajectory for a working day can be classified into 1-home-based-work-tour (e.g. HWH), 2-home-based-work-tours (e.g. HWHWH), or 3 (or more)-home-based-work-tours (e.g. HWHWHWH), referred as 1_HBWT, 2_HBWT or 3_HBWT, respectively. While for a non-working day, the trajectory can be assigned into 1-home-based-non-work-tour (e.g. HOH), 2-home-based-non-work-tour (e.g. HOHOH), or 3 (or more)-home-based-non-work-tour (e.g. HOHOHOH), namely 1_HBNT, 2_HBNT or 3_HBNT, respectively. Apart from the above 6 classes, the weekday days when an individual does not make any trips are characterized into an additional class, represented as the single letter of H.

Given a group of users along with the distances between the home and work locations of the individuals, referred as *d*, their stop-location-trajectories can be attributed to the above corresponding classes. The relative frequencies of the trajectories in each of the 7 clusters over the total number of the sequences, in each particular range of distance *d*, is referred as *distance-based-tour-class-distribution*, which characterizes the observed probabilities of the sequences in each tour class with respect to the home-work distances.

3.4. Hidden Markov Model construction

3.4.1. Model configuration

A pHMM is a probabilistic representation that can capture statistical relevant information implicit in a group of related sequences. It was introduced into bio-informatics in the 1990s (Krogh et al., 1994) and has since been widely used for large-scale protein sequence analysis (e.g. Finn et al., 2014). The information extracted from a group of sequence includes: (i) a sequence of positions, each with its own distribution overall all possible letters; (ii) the possibility for either skipping a position or inserting extra letters between consecutive positions.

In this study, the HMM building process for the two classes, including 1_HBWT and 1_HBNT, are described. The similar process applies to the remaining tour classes including 2_HBWT, 3_HBWT, 2_HBNT and 3_HBNT.

A HMM for the 1_HBWT class is designed as follows (*see* Fig. 3). It divides a sequence into four different parts, including: (i) before-going-to-work sub-sequences which represent the activities and travel undertaken before leaving home to work, e.g. HOH; (ii) commute sub-sequences which account for the activities and travel pursued during the home-to-work and work-to-home commutes respectively, e.g. HOW or WOH; (iii) work-based sub-sequences which accommodate all activities and travel conducted from work, e.g. WOW; (iv) after-work sub-sequences which comprises the activities and travel engaged after arriving home from work, e.g. HOH.

Based on the above segmentation of the sequences, a total of 8 *states* is defined, including the start home, work and end home locations, defined as m_1, m_2 and m_3 respectively, and the other stop locations corresponding to each part of the sequences, defined as $m_{1,1}$, $m_{1,2}$, $m_{2,1}$, $m_{2,2}$ and $m_{3,1}$, respectively. Each of these states can emit an letter, i.e. x, from all possible types of x governed by a distinct *emission probability distribution*, defined as $p_{emit}(x|state)$.

At each of the states, maximum 3 possible transition probabilities πs are assigned to describe the likelihood of movement between each two connected states as follows. (i) Transitions linking state m_k $(k=1, 2)$ to the other 3 possible states, including: to state $m_{k,1}$, i.e. $\pi(m_{k,1}|m_k)$, when a trip is made in the morning before going to work ($k=1$) or at noon during work period ($k=2$); to state $m_{k,2}$, i.e. $\pi(m_{k,2}|m_k)$, when an activity is conducted during the commuting way from home to work ($k=1$) or from work to home ($k=2$); to state m_{k+1}, i.e. $\pi(m_{k+1}|m_k)$, when no stops occur on the commuting ways from home to work ($k=1$) or from work to home ($k=2$). (ii) Transitions from state m_3 to only a state $m_{3,1}$, i.e. $\pi(m_{3,1}|m_3)$, when a trip is made in the evening after coming back from work. (iii) Transitions from state $m_{k,1}$($k=1, 2, 3$) to state m_k, i.e. $\pi(m_k|m_{k,1})$, when the person returns back home after finishing all activities outside in the morning or in the evening ($k=1$ *or* 3), or when the person returns to work after finishing activities outside at noon ($k=2$); or to itself, i.e. $\pi(m_{k,1}|m_{k,1})$, when an extension of multiple activities is done in the respective periods. (iv) Transitions from state $m_{k,2}$ $(k=1, 2)$ to state m_{k+1}, i.e. $\pi(m_{k+1}|m_{k,2})$, when all the activities are finished on the commuting way from home to work ($k=1$) or from work to home ($k=2$); or to itself, i.e. $\pi(m_{k,2}|m_{k,2})$ when an extension of multiple activities is done on the commute trips.

Apart from the above 8 states for stop locations, an additional *End* state is added to the end of the model, allowing transitions from m_3 to the end of the sequence; the corresponding transition probability is defined as $\pi(End \mid m_3)$.

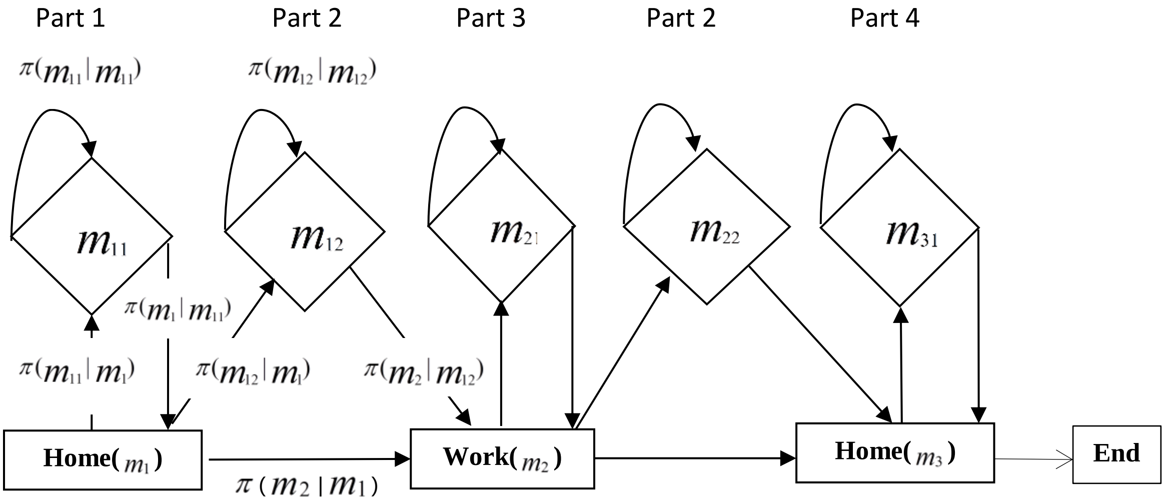

Fig. 3. The HMM for a home-based-work-tour

The above-defined model configuration thus turns the home-based-work-tours into a network system of a set of states. States m_k *(k=1, 2, 3)* underline the basic structure of the sequences, i.e. the home and work locations, while the introduction of the remaining states accommodates the situation where activities are conducted at different periods that are formed based on the home and work places. The transition probabilities πs reveal the intensity of the conversion between different states (situations).

Alongside the transition probabilities, the model also accommodates the emission probability of letter *x* at each state, i.e. $p_{emit}(x|state)$. In the current study, variable *x* represents the type of different activities; however, it can also be used to characterize other dimensions of the sequences, e.g. travel start time, distances and travel modes, thus capable of modeling multiple aspects of activity-travel behavior.

Fig. 4 illustrates the HMM for the 1_HBNT class. It has only 3 states, including the states for start and end home locations, i.e. m_1 and m_2, respectively, and the third one, i.e. $m_{1,2}$, representing locations for non-work activities conducted during the home-based tour.

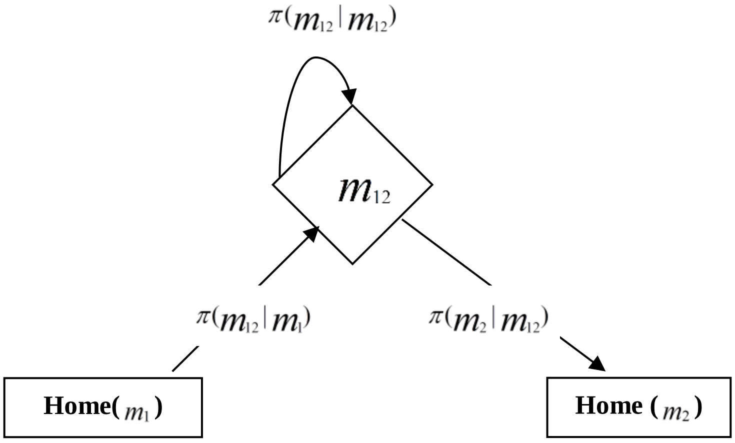

Fig. 4. The HMM for a home-based-non-work-tour

3.4.2. Model parameter estimation

After the model structure is defined, the next step involves the estimation of the specific parameters including the transition probabilities and emission probabilities. The probabilities πs and $p_{emit}(x|state)$ can be obtained by the observed frequencies of the letters at the corresponding periods of the sequences (e.g. Durbin et al., 1998). Let $A(r|q)$ as the frequencies of the transitions from a state, denoted as q, to another state, denoted as r, and $E(x|state)$ as the frequencies of letter x at state $state$, respectively. The estimators for the parameters are given by the following formula.

$$\pi(r\,|\,\mathrm{q}) = \frac{A(r\,|\,\mathrm{q})}{\sum_r A(r\,|\,\mathrm{q})},\ p_{emit}(x\,|\,\mathrm{state}) = \frac{E(x\,|\,\mathrm{state})}{\sum_{x'} E(x'\,|\,\mathrm{state})}$$

Where, *x, x'*$\in$*{set of all letter types at the state}.*

In the parameter estimation process, a *pseudocount* is set, which is a small value added to $A(r|q)$ or $E(x|state)$ if the instances of the corresponding observed cases are zero. This is to adjust the probability of rare but not impossible events so that the events are not completely excluded. The relative values of pseudocounts represent the prior knowledge on the expected probabilities of the corresponding events.

3.5. Monte Carlo simulation

3.5.1. The whole process of the simulation

Using the constructed HMMs and the distance between the home and work locations of an individual, the Monte Carlo method can be used to generate a new sequence. Monte Carlo simulation is a process that approximates solutions to quantitative problems, e.g. determining the properties of some phenomenon or behavior, through repeated statistical sampling. In this process, the investigated system is simulated a large number of times; for each simulation, all of the uncertain parameters in the system are sampled according to their respective probabilistic distribution. The simulation results are a large number of separate and independent realizations, each representing a possible "future" for the system. The results can be used for subsequent statistical analysis on the properties of the system.

In the simulation process, we first generate a tour class according to the probabilistic distribution characterized in the distance-based-tour-class-distribution. From this selected class, an entire daily sequence for this individual is then simulated based on the HMM derived from the specific class. The detailed simulation procedure based on the HMM for 1_HBWT class is described in the following section; a similar process can be applied to other classes using the respective models.

3.5.2. HMM simulation

Given distance d and the HMM as demonstrated in Fig. 3, the new sequence, i.e. s, is generated as follows. (1) Sequence s is initiated by the start home activity at state m_1 (i.e. s=H). (2) The next state is decided among the three states of m_{11}, m_{12} and m_2, according to the corresponding transition probabilities of $\pi(m_{11}|m_1)$, $\pi(m_{12}|m_1)$ and $\pi(m_2|m_1)$. (3) If m_{11} is chosen, activity x emitted from probability distribution $p_{emit}(x|\ m_{11})$ is added to the sequence (i.e. s=Hx). At this state, a next transition needs to be chosen between going back to m_1 (i.e. s=HxH) or continuing on this state (i.e. s=Hxx), based on $\pi(m_1|m_{11})$ and $\pi(m_{11}|m_{11})$ respectively. If the latter situation is selected, the loop at m_{11} continues until a transition to the home location at m_1 occurs (i.e. s=*Hxx..xH*). (4) If m_{12} is selected, x is added to the sequence (i.e. s=*Hx*). At m_{12}, a new transition is decided to either move to m_2 (i.e. s=*HxW*) or remain on this state (i.e. s=*Hxx*), governed by probabilities $\pi(m_2 | m_{12})$ and $\pi(m_{12} | m_{12})$ respectively. The remaining on this state continues until a transition to m_2 is chosen (i.e. s=*Hx..xW*). (5) If m_2 is selected, activity W is added to the sequence (i.e. s=HW). (6) The similar procedure described in steps 2-5 is repeated for next states including m_2 and m_3, using the corresponding transition probabilities. The simulation process finally stops when the transition from m_3 to the *End* state of the model is realized based on $\pi(End | m_3)$.

4. **Case study**

In this section, adopting the proposed approach and using the mobile phone dataset described in Section 2, we carry out a case study. In this process, a set of stop-location-trajectories for workers are first identified. The corresponding individuals are then randomly divided into two parts with the ratio as 4 to 1, for model training and validation, respectively. From the training set, the stop-location-trajectories are classified; in each cluster, a HMM is constructed. Based on the derived HMMs, new activity-travel sequences for individuals in the validation set are simulated.

4.1. Stop-location-trajectory construction

4.1.1. Work-start-time and work-end-time

Fig. 5 describes the distribution of the frequencies of calls made in each hour of the weekdays, showing that from 8am in the morning, calls start to increase considerably and reach their peak at noon; while at 20pm in the evening, a second climax of call activities starts to occur. These two morning and evening temporal points are chosen as the work-start-time and work-end-time, respectively.

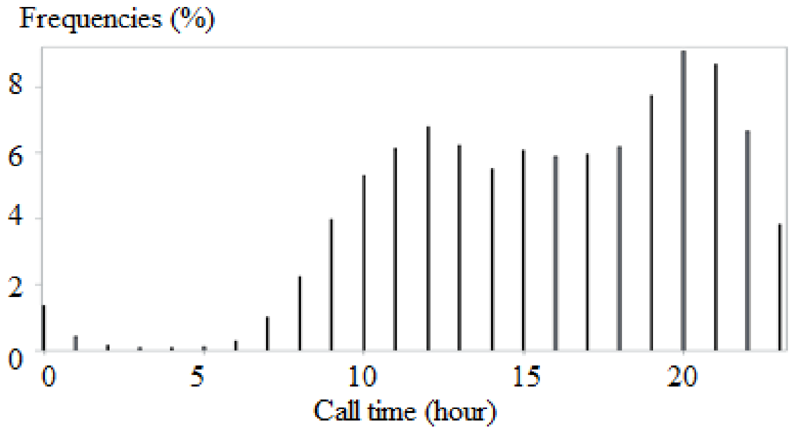

Fig. 5. The distribution of the time of calls

Based on the pre-defined criteria for home and work location identification, 319,492 users (i.e. 99.9% of the total users in the mobile phone dataset) have their home locations discovered. The remaining 0.1% are those who made no calls at weekend or in the night period from 20pm to 8am across the two surveyed weeks. As a result, their homes cannot be spotted by these rules. Meanwhile, 89,643 users are screened out as employed people, if they work between 8am and 20pm at least two weekdays per week. By contrast, those who work in the same location as their homes, who work at night shifts or at weekends, who work less than two days a week, or who make few calls at work, are left out. Although the final obtained workers account for only 28.1% of the total users in the selected dataset, they represent the part of population who regularly travels to work during the day time period among weekdays, thus they are an important target group for travel behavior analysis and transport network management. All the 7,897,854 call records of these individuals during weekdays are extracted, and the consecutive calls made at a same location are aggregated. This reduces the records to 3,479,532 locations. The locations for a same user on a same day are linked according to the temporal order, resulting in total 781,817 call-location-trajectories that will be used for further analysis.

4.1.2. $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$

For each location in the call-location-trajectories, a distinction must be made between stop-locations and non-stop ones which include trip- and false-locations. Two parameters characterize this identification process. The first one $T_{call\text{-}location\text{-}duration}$ defines the minimum time interval at a location, above which the location is considered as a possible stop. The other parameter $T_{maximum\text{-}time\text{-}boundary}$ estimates the total time that is required to travel from the previous cell to the current one and from the current one to the next cell. In addition, it should also be able to detect location update errors which usually occur in a short time interval.

In this experiment, $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ are set as 30 min and 60 min respectively. Under these thresholds, 33.3% of all the locations from the call-location-trajectories are removed; the remaining locations in these sequences form the set of stop-location-trajectories. The average length of these trajectories is 2.97. Based on the assumption that a user starts and ends a day at home, the stop-location-trajectories are added with a home activity at the beginning and/or end of the sequences if the home activity is absent from these two positions. All the obtained stop-location-trajectories are divided into training and validation sets.

4.2. Stop-location-trajectory classification

The obtained stop-location-trajectories from the training set are classified according to the number of home-based-work-tours and home-based-non-work-tours accommodated in the sequences. The average frequencies of sequences in each class relative to the total number of the sequences are 63.05%, 5.29%, 0.84%, 22.31%, 1.86%, 0.26% and 6.39% for classes 1_HBWT, 2_HBWT, 3_HBWT, 1_HBNT, 2_HBNT, 3_HBNT and H, respectively. The sequences in each class are further split based on distance *d* of the corresponding users. Fig. 6 shows the distribution of the sequence frequencies in each class, across each kilometer of *d*. In this figure, each curve represents a particular class. It is noted that, as d increases, most of the curves do not remain constant; variation in the distribution of the frequencies within each of the classes is observed. For instance, for the top curve representing the most typical class 1_HBWT, the frequencies increase as *d* gets larger but starts to decrease when d reaches a certain distance, e.g. 11km. While for the second top curve featuring class 1_HBWT, the frequencies show a stable rising trend as d increases. It suggests that, given a certain distance *d,* the observed sequence probabilities of each tour class slightly differ from the average frequency over all distance values in the class.

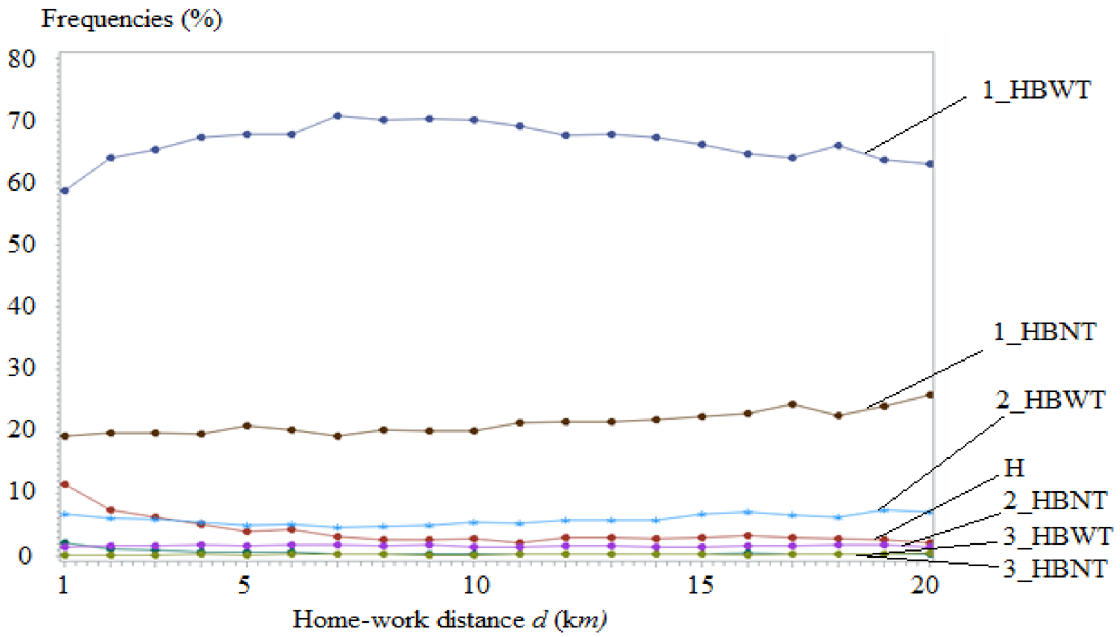

Fig. 6. The distribution of sequence frequencies in each class over home-work distances

Based on the observation from Fig. 6, we thus divide *d* into 4 intervals including ≤*2km*, *2-6km*, *6-11km*, and *>11km*. The frequencies of each class in each of these intervals characterize the distance-based-tour-class-distribution. Table 2 lists the obtained results; the average over all distance values are also presented as a comparison. This table further demonstrates the variations among different distance intervals. For instance, for class 1_HBWT, when *d* increases, the frequencies become higher, implying that more people conduct one home-based-work-tour for more days. However, when *d* is larger than a certain value, e.g. 11km, people start to perform less home-based-work-tours. Instead, they tend to stay at home or only conduct 1 tour for non-work purposes, as reflected from the frequencies of 28.62% and 7.87% in the interval of *d>11km* for classes 1_HBNT and H which are the highest probabilities over all distance intervals in these two classes.

A further test on this table obtains a statistics of 30569.65 with a significant p-value (i.e. <0.0001), signaling considerable differences in the frequencies across various distance intervals.

Table 2. The sequence frequencies of each class in each of the distance intervals (%)

Distance(d)	1_HBWT	2_HBWT	3_HBWT	1_HBNT	2_HBNT	3_HBNT	H	Total
≤2	61.00	9.76	1.70	19.42	1.57	0.15	6.41	100
2-6	66.69	5.14	0.77	20.05	1.68	0.21	5.47	100
6-11	70.05	2.69	0.31	20.15	1.60	0.25	4.95	100
>11	58.80	1.61	0.20	28.62	2.46	0.45	7.87	100
Average	63.05	5.29	0.84	22.31	1.86	0.26	6.39	100

4.3. Hidden Markov Model construction

From all the trajectories in each cluster, a HMM is constructed and the corresponding parameters are estimated. Table 3 presents the transition probabilities for the model derived from the 1_HBWT cluster, with parameter *Pesucount* being tuned as 0.02. Based on the structure of the model defined in Fig. 3, at the *End* state m_3, transitions including $\pi(m_{k,2}|m_k)$, $\pi(m_{k,2}|m_{k,2})$ and $\pi(m_{k+1}|m_{k,2})$ are not expected, they are thus represented with '*Null*'.

Table 3. Transition probabilities of the HMM derived from the 1_HBWT cluster

Locations	$\pi(m_{k,1} \mid m_k)$	$\pi(m_{k,2} \mid m_k)$	$\pi(m_{k+1} \mid m_k)$	$\pi(m_{k,1} \mid m_{k,1})$	$\pi(m_k \mid m_{k,1})$	$\pi(m_{k,2} \mid m_{k,2})$	$\pi(m_{k+1} \mid m_{k,2})$
Start home (m_1)	0.02	0.29	0.72	0.02	0.02	0.38	0.62
Work (m_2)	0.18	0.31	0.51	0.24	0.76	0.41	0.59
End home (m_3)	0.05	Null	0.95	0.24	0.76	Null	Null

Regarding the emission probabilities $p_{emit}(x|state)$, in this study, as all activities at the other stop locations except the home and work places, are classified into a single type O, thus $x=$ 'O' and $p_{emit}(x|state)=1$ for all activities generated at these locations.

4.4. Monte Carlo simulation

Based on the derived distance-based-tour-class-distribution and HMMs, new sequences for users from the validation set who consist of different workers from those included in the training set, are simulated. In this process, the home-work distance *d* is first derived from each of the users, and a tour class is chosen based on the probabilities described in the distance-based-tour-class-distribution. In this case study, only when the 1_HBWT class is selected, an entire sequence for the particular user is then further generated according to the HMM derived from the corresponding cluster.

5. Comparison of the simulation results with the validation set

To examine the performance of the proposed modelling approach, we compare the sequences simulated from the models with the original stop-location-trajectories drawn from the validation set. The comparison is carried out in two aspects, including the aspect of individual locations, e.g. the average number of locations visited each day, and the sequential aspect of the locations.

5.1. The average number of locations each day

Among all 156374 stop-location-trajectories observed from 18284 users in the validation set, 61.91% of them fall into the 1_HBWT cluster. The average length of the sequences from the considered cluster is 2.79, and it increases to 4.55 after H is added to the two ends of the sequences.

For all the 18284 users, the tour class is first simulated based on their home-work distances. This results in 62.92% of the users falling into the 1_HBWT cluster. For the obtained users, the entire sequences are generated according to the HMM built from this cluster; the average length of the simulated sequences is 4.72, a close match to the average length of the sequences in the validation set.

5.2. The sequential aspect of the locations

From all the validation sequences in the 1_HBWT cluster, 677 types which are formed by the various combinations of activity locations in particular orders, are found. While for the simulated sequences, 948 types are generated; 520 of them are also observed among the validation sequences. Table 4 lists the sequence frequencies for the 13 most prevalent types, each of which accounts for more than 1% of the total number of sequences in the corresponding sets.

The relationship of the sequence frequencies over all the types between the two data sets is shown in Fig. 7, with the coefficient R as 0.93. The high value of R suggests that the derived HMM model is able to capture the probabilistic distribution of the activity locations and their temporal sequencing revealed by the mobile phone data, and can properly represent workers' travel behavior in a study area. As a result, the sequences generated from the derived models can accurately reflect the travel demand in the region.

Table 4. The sequence frequencies for the 13 most prevalent types in each set (%)

Types	HWH	HWOH	HOWH	HWOWH	HWOOH	HOWOH	HOOWH
Validation	41.38	11.38	7.97	4.60	3.53	3.26	2.21
Simulated	35.09	12.29	8.39	4.79	4.92	2.99	3.13
Types	HWOWOH	HWOOOH	HWHOH	HOWOOH	HOWOWH	HOOWOH	
Validation	1.83	1.34	1.33	1.21	1.12	1.01	
Simulated	1.75	2.01	1.42	1.24	1.12	1.18	

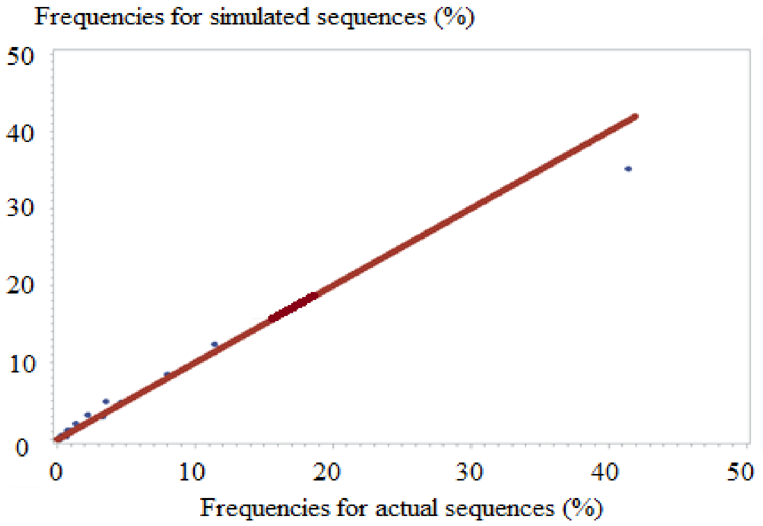

Fig. 7. The correlation of sequence frequencies for each type between actual phone location sequences and simulated ones

6. Discussions and conclusions

In this paper, we have developed a new method of modelling workers' travel demand based on mobile phone data. The advantage of this approach is that it does not depend on conventional travel data survey methods. The data requirement is fairly simple and its collection cost is low. In addition, the massive mobile phone data monitors current travel behavior in a large proportion of the population over a long time period. The models derived from the data are thus capable of providing a more general and objective representation of current mobility demand. Apart from the benefits that are realized by the use of the mobile phone data, this approach also provides added value in taking into account the sequential constraints of activity-travel patterns into the modelling process.

Once the models are developed in a region, they can be used to simulate activity-travel sequences for each of the employed people in the whole area, given the home and work locations of the individuals. The generated sequences can then be aggregated and subsequently be employed for travel demand analysis, e.g. the average number of trips made in the morning before going to work, on the commuting way, or in the evening after arriving at home. The models can also be utilized to forecast travel demand for future scenarios, e.g.

the displacement of residential areas or the establishment of new industrial sites, which could cause changes in the home-work distances of the workers. Furthermore, travel sequences in a new region, where no phone data is available, can also be predicted by the models, under the assumption that these two regions share similar activity-travel patterns of individuals, e.g. regions from a same country.

With respect to the performance of the approach, data collected from people's natural mobile phone usage in Senegal in the whole year of 2013 has been used, and the test results show the following major strengths of the proposed method. (i) While the average length of daily sequences from the 1_HBWT cluster in the validation set is 4.55, a close average value of 4.72 is achieved for the simulated sequences. (ii) Among all the 677 different types of the validation sequences, 520 (e.g. 76.8%) are also observed from the simulated sequence set. Particularly, the distribution of sequence frequencies over the 13 most prevalent types shares a high level of similarity between these two sequence sets. (iii) An overall comparison on the frequency distribution over all the 677 sequence types between these sets reveals a correlation of 0.93. All the above results suggest that the derived HMM model is able to capture the probabilistic distribution of activity locations and their sequential orders revealed by the mobile phone data. As a result, the sequences generated from the models can properly represent workers' travel behavior and lead to an accurate travel demand estimation in the region.

Despite the promising experimental results, the method could be enhanced and extended in the future research in terms of data processing, sequence clustering and model building. Concerning data processing, firstly, by using a fixed work period (e.g. 8am-20pm on weekdays in this experiment), individuals who work during night shifts are ignored. The prediction accuracy of home and work locations could be improved by taking into account the detailed information on individuals' work regime. Secondly, in the process of stop location identification, two parameters, namely $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ are used. $T_{call\text{-}location\text{-}duration}$ defines the maximum time duration needed to traverse a single cell area; while $T_{maximum\text{-}time\text{-}boundary}$ estimates the total time required for the travel from a previous cell to the current one and from the current one to the next cell. Instead of using overall threshold values of 30 min and 60 min for these two parameters respectively, the settings could be tailored to each particular individual and cells, through the use of the individual's travel speed and the size of the cell areas.

In terms of sequence clustering, the number of home-based tours encoded in the sequences as well as the home-work distances of the corresponding individuals are used as the classifiers. However, travel behavior is shaped by a range of multiple factors including the conditions of land use and transportation network as well as the social-economic characteristics of individuals. The social-economic information of the phone users could be inferred based on the mobile phone data, and the information could be integrated into the clustering process.

As to model building, improvement can also be made in terms of the following aspects. Firstly, in the designing of the HMM (*see* in Fig. 3), locations among different parts of the sequences are modelled independently, the correlation between these parts is thus unaccounted for. The interdependencies of activities performed on a day should be integrated in the modeling process, e.g. through conditional probabilities. Secondly, instead of considering only one-dimensional location sequences consisting of home, work and other stop locations, more dimensions of activity-travel patterns could be characterized using the emission probabilities $p_{emit}(x|state)$ at each state of the HMM, thus modelling the multiple aspects of travel behavior. For instance, the locations for other activities O can be distinguished among detailed activity categories. A number of research has been dedicated to annotating activity purposes on the mobile phone locations (e.g. Liu et al., 2013). Similar to activity types, other dimensions, e.g. travel start time and travel distances, can also be incorporated into the models. In particular, the travel distance at a stop location should be measured relative to the home or work place, and the distribution of the travel distances at this stop is characterized with the emission probability, i.e. $p_{emit}(x|state)$. Once the model is built and a new sequence is simulated for an individual, the specific geographic position of this stop location can be derived based on the obtained distance value, the home or work position of the corresponding individual, as well as the land use data describing the distribution of activity locations surrounding the home or work place.

When being faced with the challenge of acquiring both mobile phone data and real travel survey data from a same or similar study region, in this study the modeling results are tested using mobile phone data of users who are different from those involved in the model training process. However, due to the event-driven nature of the data collection, mobile phone data only reviews the presence of a user at a certain location and time point when his/her phone device makes GSM network connections. The places, where the individual has stayed but no calls were made, are missed. Thus, in the future research, the proposed method must be compared against a real travel survey from the study region or from a region with a similar context. The discrepancies between the simulated sequences and the actual travel sequences could be examined and handled e.g. through an overall scaling factor used by the research (Shan et al., 2011) described in Section 1. Alternatively, the technique developed in the study (Liu et al., 2014), which transforms each of the stop-location-trajectories into actual travel sequences, could be adopted. The obtained actual travel sequences can subsequently be used for the construction of the HMMs.

With the rapid development of mobile phone based services in the future (e.g. Liu & Chen, 2013; Monares et al., 2013), the amount of location data, which is recorded not only when people make calls but when they use the application services on their phones, will continuously grow. The data will reveal more activity locations and travel episodes, thus providing another prospect of improving the model performance and leading to an even better travel demand estimation.

Acknowledgements

The authors would like to acknowledge the support of the European Union (EU) through the project grant of Data Science for Simulating the Era of Electric Vehicles (DataSim). We also want to thank the Second Mobile Phone Data for Development Challenge (D4D-Senegal) committee for the provision of mobile phone data.

References

Angelakis, V., Gundlegård, D., Rajna, B., Rydergren, C., Vrotsou, K., Carlsson, R., Forgeat, J., Hu, T. H., Liu, E. L., Moritz, S., Zhao, S., & Zheng, Y. T. (2013). Mobility Modeling for Transport Efficiency - Analysis of Travel Characteristics Based on Mobile Phone Data. Third International Conference on the Analysis of Mobile Phone Datasets. NetMob, Special session on the D4D challenge, MIT, May 1-3, 2013.

Arentze, T. A., & Timmermans, H. J. P. (2004). A learning-based transportation oriented simulation system. Transportation Research Part B: Methodological, 38(7), 613-633.

Asakura, Y., & Hato, E. (2006). Tracking individual travel behavior using mobile phones: recent technological development. Paper presented at 11th International Conference on Travel Behaviour Research, Kyoto.

Bayir, M. A., Demirbas, M., & Eagle, N. (2009). Discovering spatiotemporal mobility profiles of cellphone users. World of Wireless, Mobile and Multimedia Networks & Workshops, WOWMOM, IEEE, 1-9.

Becker, R., Cáceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Computing, 10(4), 18–26.

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., & Timmermans, H. J. P. (2010). Implementation Framework and Development Trajectory of Feathers Activity-Based Simulation Platform. Transportation Research Board: Journal of the Transportation Research Board, 2175, 111-119.

Berlingerio, M., Calabrese, F., Lorenzo, G. D., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. Third International Conference on the Analysis of Mobile Phone Datasets. NetMob, Special session on the D4D challenge, MIT, May 1-3, 2013.

Bhat, C. R., & Koppelman, F. S. (1999). A Retrospective and Prospective Survey of Time-Use Research. Transportation, 26(2), 119-139.

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 10(4), 36-44.

Cools, M., Moons, E., Bellemans, T., Janssens, D., & Wets, G. (2009). Surveying activity-travel behavior in Flanders: Assessing the impact of the survey design. Proceedings of the BIVEC-GIBET Transport Research Day, Part II, VUBPress, Brussels, 370, 727-741.

Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J., & Picado, R. (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling. Transportation Research Part A: Policy and Practice, 41(5), 464–488.

Delafontaine, M., Versichele, M., Neutens, T., & Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. Applied Geography, 34, 659–668.

de Montjoye, Y., Smoreda, Z., Trinquart, R., Ziemlicki, C., & Blondel, V. D. (2014). D4D-Senegal: The Second Mobile Phone Data for Development Challenge.

Do, T. M. T., & Gatica-Pereza, D. (2013). Where and what: Using smartphones to predict next locations and applications in daily life. Pervasive and Mobile Computing. http://dx.doi.org/10.1016/j.pmcj.2013.03.006

Durbin, R., Eddy, S.R., Krogh, A., & Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. Nucleic Acids Res. 42(D1): D222–D230.

García-Díez, S., Fouss, F., Shimbo, M., & Saerens, M. (2011). A sum-over-paths extension of edit distances accounting for all sequence alignments. Pattern Recognition, 44(6), 1172–1182.

González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. Nature, 453, 779-782.

Gühnemann, A., Schäfer, R. P., &Thiessenhusen, K. U. (2004). "Monitoring Traffic and Emissions by Floating Car Data." Institute of Transport Studies, Working Paper, Issue ITS-WP-04-07.

Hartgen, D. T. (2013). Hubris or humility? Accuracy issues for the next 50 years of travel demand modeling. Transportation, 40(6), 1133-1157.

Janssens, D., Giannotti, F., Nanni, M., Pedreschi, D., & Rinzivillo, S., (2012). Data Science for Simulating the Era of Electric Vehicles. KI - Künstliche Intelligenz, 26(3), 275-278.

Joh, C. H., Ettema, D., & Timmermans, H. J. P. (2008). Improved Motif Identification of Activity Sequences: Application to Interactive Computer Experiment Data. Transportation research record: Journal of the Transportation Research Board, 2054, 93-101.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235(5):1501-1531.

Liu, C. C., & Chen, J. C. H. (2013). Using Q methodology to explore user's value types on mobile phone service websites. Expert Systems with Applications, 40(13), 5276–5283.

Liu, F., Janssens, D, Cui, J.X. , Wang, Y.P., Wets, G, & Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone da. Expert Systems with Applications. 41(14), 6174–6189.

Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Applications, 40(8), 3299–3311.

Monares, A., Ochoa, S. F., Pino, J. A., Herskovic, V., Rodriguez-Covili, J., & Neyem, A. (2013). Mobile computing in urban emergency situations: Improving the support to firefighters in the field. Expert Systems with Applications, 38(2), 1255–1267.

Roorda, M. J., Miller, E. J. & Habib, K. M. N. (2008). Validation of TASHA: A 24-H Activity Scheduling Microsimulation Model. Transportation Research Part A: Policy and Practice, 42(2), 360-375.

Saneinejad, S., & Roorda, M. J. (2009). Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. Journal Transportation Letters: The International Journal of Transportation Research, 1(3), 197-211.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling Daily Human Mobility Motifs. Journal of The Royal Society Interface, 10(84),

Shan, J., Viña-Arias, L., Ferreira, J., Zegras, C., & González, M. C. (2011). Calling for Validation, Demonstrating the use of mobile phone data to validate integrated land use transportation models. In Proceedings 7VCT 2011.

Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. Science, 327(5968), 1018-1021.

Spissu, E., Pinjari, A. R., Bhat, C. R., Pendyala, R. M., & Axhausen, K. W. (2009). An analysis of weekly out-of-home discretionary activity participation and time-use behavior. Transportation, 36(5), 483-510.

Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., & Scholten, H. (2013). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal, 78(2), 223-243.

Wegener, M. (2013). The Future of Mobility in Cities: Challenges for Urban Modelling. Transport Policy, 29, 275–282.

Wilson, C. (2008). Activity patterns in space and time: calculating representative Hagerstrand trajectories. Transportation, 35(4), 485-499.