Building a validation measure for activity-based transportation models based on mobile phone data

Feng Liu[a], Davy Janssens[b], JianXun Cui[c], YunPeng Wang[d], Geert Wets[b], Mario Cools[e]

[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

[d] School of Transportation Science and Engineering, Beihang University, Beijing 100191, China

[e] LEMA, University of Liège, Chemin des Chevreuils 1, Bât B.52/3, 4000 Liège, Belgium

E-mail addresses: feng.liu@uhasselt.be (F. Liu), davy.janssens@uhasselt.be (D. Janssens), cuijianxun@hit.edu.cn (J.X. Cui), ypwang@buaa.edu.cn (Y.P. Wang), geert.wets@uhasselt.be (G. Wets), mario.cools@ulg.ac.be (M. Cools)

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

Abstract

Activity-based micro-simulation transportation models typically predict 24-hour activity-travel sequences for each individual in a study area. These sequences serve as a key input for travel demand analysis and forecasting in the region. However, despite their importance, the lack of a reliable benchmark to evaluate the generated sequences has hampered further development and application of the models. With the wide deployment of mobile phone devices today, we explore the possibility of using the travel behavioral information derived from mobile phone data to build such a validation measure.

Our investigation consists of three steps. First, the daily trajectory of locations, where a user performed activities, is constructed from the mobile phone records. To account for the discrepancy between the stops revealed by the call data and the real location traces that the user has made, the daily trajectories are then transformed into actual travel sequences. Finally, all the derived sequences are classified into typical activity-travel patterns which, in combination with their relative frequencies, define an activity-travel profile. The established profile characterizes the current activity-travel behavior in the study area, and can thus be used as a benchmark for the assessment of the activity-based transportation models.

By comparing the activity-travel profiles derived from the call data with statistics that stem from traditional activity-travel surveys, the validation potential is demonstrated. In addition, a sensitivity analysis is carried out to assess how the results are affected by the different parameter settings defined in the profiling process.

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

1. Introduction

1.1. Activity-based transportation models

The main premise of activity-based micro-simulation 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; Fan & Khattak, 2012; Lemp et al., 2007; Wegener, 2013).

All the above information, complemented with a training set of household travel surveys which document 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 strategy rules. These rules represent the scheduling process of activities and travel by the individuals (e.g. Arentze & Timmermans, 2004; Bellemans et al., 2010). Once established they can be used as the probabilistic basis for a micro-simulation process, in which complete daily activity-travel sequences for each individual in the whole region are synthesized, using Monte Carlo simulation methods. The synthesized individual activity-travel sequences are then aggregated into origin-destination (OD) matrix, with each matrix element representing the number of trips between each pair of locations of the region. This matrix, after being assigned to a road network through traffic assignment algorithms, 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 transportation policy scenarios. Fig. 1 illustrates the entire process of a micro-simulation model.

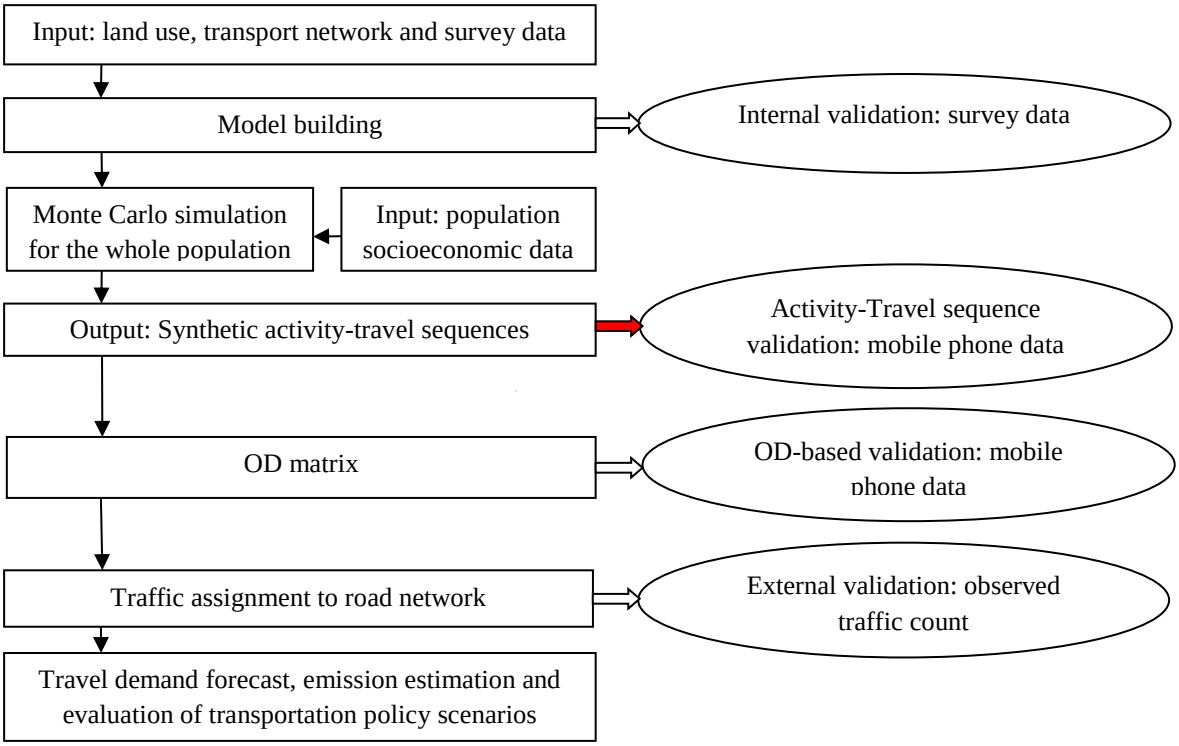

Fig. 1. The entire process of an activity-based transportation model

1.2. Problem statement

Despite comprehension and advancement of the activity-based transportation modeling system, the lack of reliable data in sufficient size does not enable one to have a decent benchmark and evaluation criterion for the model output (e.g. Cools et al., 2010a; Cools et al., 2010b). Typically, for this purpose, one examines the results of the model both internally and externally at different stages of the simulation process, as indicated in Fig. 1 (e.g. Bellemans et al., 2010; Rasouli & Timmermans, 2013; Yagi & Mohammadian, 2010). The internal validation involves the comparison of the estimated results with expanded travel survey data which is not used in the training phase of the model but usually collected in the same survey period. In this validation, certain aggregated measures, e.g. the average travel distance and travel duration, derived from both the predicted sequences and the observed ones, are examined (e.g. Cherchi & Cirillo, 2010; Roorda et al., 2008). The sequence alignment method (SAM), which compares two sequences based on the composition and temporal ordering of the daily activities (Abbott & Forrest, 1986; Wilson, 1998), is also employed to assess the similarities between each of the observed sequences and its predicted counterpart (e.g. Sammour et al., 2012). However, the process involved in the development of the model, from initial data gathering to exploitation and validation of the first results, is lengthy and may take years, imposing a time lag between the data initially obtained and the data that is required for an objective and up-to-date validation measure. In addition to this time limitation, the high cost related to the surveys, makes it a challenge to collect samples in sufficient size, capable of providing a good representation of activity-travel behavior of a population. Moreover, travel surveys usually query information of only one or two days, in order to limit the negative effects associated with respondent burden. Consequently, this tends to obfuscate the less frequent activities, such as sports or telecommuting activities which are often carried out only once a week or once a month. These shortcomings have been well reported in the literature (e.g. Asakura & Hato, 2006; Cools et al., 2009).

In contrast to the internal validation, the external validation consists of an indirect evaluation of the model output at a later phase, i.e. traffic assignment stage (see Fig. 1). The estimated

traffic volumes at a number of predefined road segments are compared against information from external sources, such as traffic counts collected by inductive loop detectors which are deployed on the road segments.
However, the external validation process encompasses an aggregation step to compose the OD matrix as well as an assignment step to allocate the travel demand matrix to the road network. Valuable information may be lost in these two steps. Consequently, positive outcomes of the compared results might be artifacts of the validation process itself, and thus provide no real guarantee of the accuracy of the model. Moreover, when mismatches are found, there exists no clear procedure to trace back the causes, thus limiting the discovery of remedies to improve model construction. Nevertheless, despite such limitations, at the present, the indirect external evaluation is essentially the only option for model quality assessment in practice, as no well-established methods are found for operating closer to the model itself (e.g. Janssens et al., 2012). This is a problem that seriously hampers further model development and model application (e.g. Hartgen, 2013). Having useful and reliable benchmark and evaluation criteria for activity-based micro-simulation models has thus been a major concern.

1.3. Mobile phone data: a new data source for transportation modelling and validating

The wide deployment of mobile phone devices has created the opportunity to use the devices as a new data collection method to overcome the lack of reliable benchmark data (Jiang et al., 2013). Location data recorded from the mobile phone devices reflects up-to-date travel patterns on a significantly large sample of the population, making the data a natural candidate for the analysis of mobility phenomena (e.g. Do & Gatica-Pereza, 2013; Schneider et al., 2013). In addition, the data collection is a by-product of the 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 study of travel behavior and transportation modeling have been manifested by a variety of research efforts, ranging from the investigation of key dimensions of human travel, such as travel distance and time expenditure at different locations (e.g. González et al., 2008; Schneider et al., 2013; Song et al., 2010), to the discovery of mobility patterns and the construction of OD matrices (e.g. Bayir et al., 2009; Berlingerio et al., 2013; Calabrese et al., 2011; Huang et al., 2009), and to the examination of the status and efficiency of current transport systems (e.g. Angelakis et al., 2013; Hansapalangkul et al., 2007; Steenbruggen et al., 2013). Alongside these studies, mobile phone data has also been investigated to explore the possibilities of building model evaluation measures. Two recent research efforts can represent the state-of-art of such exploration. The first one (Shan et al., 2011) involves the use of mobile phone data of more than 0.3 million users collected in the metropolitan area of Lisbon, Portugal over a time period of an entire month. In their study, the two most frequent call cell towers for each of the users are first identified as the residential and employment locations. An OD matrix depicting the home-to-work commuting trips in the morning is then built, using the identified residential and employment locations as well as the call data. Based on a census survey, this 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 predicted by an integrated land use and transportation model developed in this region. In the entire activity-based modeling process, the above-described OD-based validation method can be positioned at the stage of OD matrix generation, as shown in Fig. 1.
Instead of an OD matrix, in the second research (Kopp et al., 2013), other mobility parameters, such as the number of frequently visited locations and the spatial extent of a person's daily mobility, are derived from mobile phone data collected in two separate regions,

namely a central region of Italy and a large area of Lausanne, Switzerland. As opposed to directly using the derived travel parameters to examine the simulation results from a transportation model, the research compares the parameter values derived from the mobile phone data with the results inferred from GPS data that is recorded in the same regions during the same periods and therefore portrays the same mobility phenomenon. The comparison results demonstrate the capability of the mobile phone data in reflecting real travel patterns of the population, thus suggesting its potentials of building an objective and more detailed evaluation standard for transportation models.

The two studies have provided deep insights into the characteristics of mobile phone data and illustrated its potentials for constructing an improved validation measure. In particular, the first research (Shan et al., 2011) has proposed a specific OD-based validation method and used a real transportation model to test this method's applicability and viability. However, despite its advancement by incorporating mobile phone data into the model assessment, the OD-based method does not consider the sequential information which is imbedded in the activity-travel patterns. A detailed examination of the sequential dependencies of the daily activities from the simulated travel sequences is thus ignored in the validation process. It has been widely acknowledged that the choice of activities is dependent on the preceding activity engagement (e.g. Joh et al., 2007; 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; Saneinejad & Roorda, 2009; Shoval & Isaacson, 2007). The examination of how the predicted activity-travel sequences are consistent with the sequential constraints that are observed from the real travel patterns is thus important. In the existing validation methods for activity-based models, SAM has been employed to assess the similarities between each observed sequence and its predicted counterpart (e.g. Sammour et al., 2012), as previously described. But this evaluation is carried out against a small set of activity-travel sequences from travel surveys, thus subject to the shortcomings that are inherent to the traditional data collection method. A validation measure, 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 research on the application of mobile phone data to travel behavior analysis and transportation modeling, and particularly addressing the above mentioned limitations in the development of reliable validation measures for activity-based models, our study proposes a new approach which is based on the phone data and which considers the sequential information hidden in activity-travel patterns. Specifically, the goal of this approach is to build a profile of workers' activity-travel behavior, i.e. the relative frequency of each *typical pattern* which represents a certain class of activity-travel sequences, based on the mobile phone data. This profile can then be used to directly evaluate the sequences yielded from the simulation models, by comparing it against the frequencies of the corresponding pattern classes which are obtained from the simulated sequences (see Fig. 1). This comparison is carried out at the level of the generated activity-travel sequences, which enables the capability of detecting problems that are directly caused by the model itself and providing immediate feedback for the enhancement of the model.

Compared to existing validation measures, 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 validation

measure. (*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 validation measure. (*iv*) This evaluation method directly examines the simulated travel sequences, and can thus offer immediate solutions to problems which are linked to the model system itself. (*v*) When this approach is compared with the recently developed OD-based validation method, the OD-based method examines the simulated sequences in terms of the distribution of the trips over different pairs of origin-destination locations; while the approach developed in this study focuses on the sequential aspect of the simulated sequences, and evaluates the distribution of the sequences over various classes of typical activity-travel patterns. 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 measures assess the simulated sequences from different angles, thus providing a complementary means of benchmarking activity-based transportation models based on the mobile phone data.

The remainder of this paper is organized as follows. Section 2 describes the *typical patterns* which characterize workers' activity-travel sequences. Section 3 introduces the mobile phone data and Section 4 details the construction process of location trajectories based on the data. The call location trajectories are then transformed into complete travel sequences by a method proposed in Section 5. Section 6 classifies both the call location trajectories and the travel sequences into the typical patterns which have been established in Section 2, and the profiles which describe the relative frequency of each pattern class are drawn. A case study is subsequently conducted in Section 7, and a comparison of the results against the outcomes of real travel surveys is carried out in Section 8. An in-depth analysis on the sensitivity of this approach is further performed in Section 9. Finally, Section 10 ends this paper with major conclusions and discussions for future research.

2. Activity-travel sequence classification

Individuals make choices about the different activities being pursued, and travel may be required to participate in these activities. Traditionally, all activities performed at home are considered as *home* activities; while the remaining ones conducted outside home are categorized into *mandatory* activities (e.g. working or studying) and *non-mandatory* activities that include maintenance activities (e.g. shopping, banking or visiting doctors) and discretionary activities (e.g. social visit, sports or going to restaurant) (e.g. Arentze & Timmermans, 2004; Bradley & Vovsha, 2005). The home, mandatory and non-mandatory activities are represented as 'H', 'W' and 'O', respectively.

The sequence of activities and travel that a person undertakes during a day is referred as the individual's *activity-travel sequence* for that day. A critical difference is imbedded in activity-travel sequences between workers and non-workers; the sequences of workers mostly rely on the regularity and fixity of the work activity. In contrast, no such obvious periodicity is present in the case of non-workers. This motivates the development of separate representations for these two types of individuals' behavior. In this study, only the activity-travel behavior of workers is analyzed, and the representation of their daily sequences described in the research by Spissu et al. (2009) is adopted. In this representation (see Fig. 2), an activity-travel sequence is divided into four different parts, including: (*i*) before-work sub-sequences which represent the activities and travel undertaken before leaving home to work (as indicated in arrows 'a'), 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 (in arrows 'b' and 'd'), e.g. HOW or WOH; (*iii*) work-based sub-sequences which accommodate

all activities and travel undertaken from work (in arrows 'c'), e.g. WOW; (*iv*) after-work sub-sequence which comprises the activities and travel engaged after arriving home from work (in arrows 'e'), e.g. HOH.

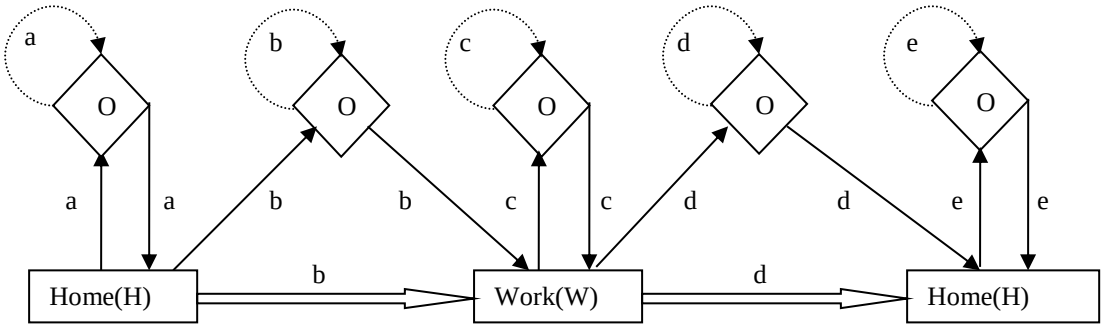

Fig. 2. The representation of workers' activity-travel sequences
Note: Each 'rectangular' indicates the home or work location, while the 'diamond' represents a non-mandatory activity location. Each 'arrow' from a home, work or non-mandatory activity location to the other location represents the related travel, and the 'arrow' from a non-mandatory activity location to itself indicates the chain of consecutive visits to different non-mandatory activity locations.

According to the above characterization, a *home-based-tour*, which is defined as a chain of locations (trips) that start and end at home and accommodates at most two work location visits, can be classified into the following patterns: HWH, HOWH, HWOH, HWOWH, HOWOH, HOWOWH, HWOWOH, HOWOWOH, where each H or W stands for a home or work location while each O represents one or a chain of visits to several non-mandatory activity locations. The days when an individual does not go to work, can be characterized with 2 additional patterns, namely H and HOH. In total, 10 classes are formed to identify each home-based-tour in a worker's daily activity-travel sequence, and they are defined as *home-based-tour-classification.*
Every pair of the above pattern classes (excluding H) is then merged, leading to 81 combinations of daily sequences which contain 2 home-based-tours. For instance, the combination of HWH and HOWH results in the sequence HWHOWH. The daily sequences that represent only a single home-based-tour with maximum two work location visits, e.g. HWOWH, are also clustered, according to the home-based-tour-classification. Finally, the remaining sequences which contain more than 2 home-based-tours (e.g. HWHWHWH) or which have more than 2 work activity locations in a home-based-tour (e.g. HWOWOWH), are each assigned into one additional category. Thus, all the above classification leads to a total of 93 patterns which underlie workers' activity-travel behavior, and are denominated as the workers' *daily-sequence-classification.* Given a group of individuals, their activity-travel sequences can be attributed to the corresponding pattern classes. The relative frequency of each of the pattern classes over the total number of activity-travel sequences forms the *profile* of activity-travel behavior among these people.

3. Mobile phone data description

The mobile phone dataset consists of full mobile communication patterns of around 5 million users in Ivory Coast over a period of 5 months between December 1, 2011 and April 28, 2012 (Blondel et al., 2012). The dataset contains the location and time when each user conducts a call activity, including initiating or receiving a voice call or 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. 25%) of this country's 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, 50,000 of all the 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 10 randomly sampled datasets, each of which contains communication records of 50,000 users over two weeks. One of the datasets is selected for this study. Table 1 illustrates typical call records of an individual identified as User2 on Monday, December 12th, 2011.

Table 1. The typical call data of an individual[a]

Time	11:57:00	13:40:00	16:59:00	17:43:00	21:28:00
Antenna_id	898	1020	972	926	926

[a] The 'time' represents the moment (i.e. the hour, minute and second) when this user was connecting to the GSM network and the 'Antenna_id' as the cell area where he/she is located.

4. Construction of call stop location trajectories

A *raw-location-trajectory* from a mobile phone user during a day 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, referred as *call-frequency*, denoted as k_i $(k_i \geq 1)$; the time for each of the calls is 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 raw-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 raw-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 being carried out.

4.1. 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 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 at locations different from their home locations and who work at least two days per week, are included for the analysis of workers' travel behavior.

4.2. Identification of stop locations

After the identification of the distinct home and work locations for each worker, the remaining locations in the raw-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 classified 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; Schlaich et al., 2010).

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 raw-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 discussed above 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. From this trajectory, a distinction needs to be made to identify stop visits from possible trip visits at each of these locations. The trajectory of the second user known as 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 of 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.

4.2.1. Identification process

Schlaich et al. (2010) have proposed a method to distinguish a stop-location from a non-stop one. In their approach, the interval between the first logins of two adjacent locations l_i and l_{i+1}, i.e. $T(l_{i+1},1) - T(l_i,1)$, is examined. If this interval is longer than a time limit, e.g. 60 min in their experiment, l_i is considered as a stop-location. However, this method is likely to overlook stop-locations where calls are made just before the departure of the locations. In this situation, the time interval can be very short, despite the possibility that users may spend a considerable time period at the locations. This can be further illustrated with the case of User265. The interval between the two first time signatures of locations l_1 and l_2 is 45 min, shorter than this 60-min limit, suggesting that location l_1 would be for trip purposes. This may be true if this individual has made a long trip of at least 37 min within l_1 and made calls at the start and end of this travel. However, if this individual has stayed there doing activities for a long time, e.g. a few hours, and he/she made calls later in this sojourn period, location l_1 is then misclassified by the existing method.

In order to identify all the possible stop-locations, we propose a new approach consisting 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 occurs 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.

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, the obtained trajectory of stop locations for User265 and User72 are l_1 -> l_3 -> l_4 and l_1 -> l_2 respectively. In comparison, using the existing method which only considers the first temporal logins of two consecutive locations (Schlaich et al., 2010), only one single location would be derived for each of these users, which is l_3 for User265 and l_1 for User72.

After the removal of locations that are either trips or stemming from localization errors, all the remaining locations from a raw-location-trajectory are regarded as stops and stored into a *stop-location-trajectory*. Each location l_i in these stop-location-trajectories is complemented with its function, categorized into home, work and non-mandatory activities, denoted as *activity*(l_i). Travel is implicit in between each two consecutive locations of these sequences.

5. Transformation of call stop location trajectories into actual travel sequences

The considered mobile phone dataset is event driven, in which location measurements are only available when the devices make GSM network connections. Consequently, users' call behavior can affect the possibility of capturing a larger or smaller number of trips and/or activity locations. In general, the more active a user is in communicating electronically with others, the better his/her activity-travel behavior is revealed by his/her call records. The call locations can be seen as the observed behavior at certain temporal sampling moments during a day, and the characteristics of the real travel behavior must be deduced. A transformation therefore should be made from the previously derived stop-location-trajectories into the sequences that mirror the real picture of people's activity-travel behavior.
During this transformation, we first derive for each user the actual activity duration as well as the call rate at each minute. These two variables are then translated into the call probability at each location, which describes how likely the individual makes at least one call when he/she visits the location and which thus indicates to what extent his/her call records reveal his/her actual movement. Next, given a real daily activity-travel sequence, various stop-location-trajectories could be possibly observed from the call data. The probability, under which a certain stop-location-trajectory is generated from the original travel sequence, is calculated based on the call probabilities at the actually visited locations. Finally, given the observed frequencies of the stop-location-trajectories derived from the call data, a linear equation is built and the frequencies of the original travel sequences are inferred.

5.1. Call rate and actual location duration

Call-intervene for a user measures the time interval between each two calls, and it is calculated as the ratio between the total number of calls each day, denoted as *total-number-calls*, and the time span of the day (measured in min), denoted as *time-span*, as follows:

$$call-\text{int}ervene(user)=\frac{\sum_{day} time-span(day)}{\sum_{day} total-number-calls(user,day)}$$

Based on the *call-intervene*, the call rate defined as *CallRate*, which describes the probability that a user makes calls each minute, can be calculated as follows:

$$CallRate(user)=\frac{1}{call-\text{int}ervene(user)}$$

Let the variable *actual-location-duration*(*user*,l_i) represent the actual activity duration (in min) which a user spends at location l_i. Given that this information is unknown from the phone data, we thus turn to activity-travel surveys to obtain the real behavioral data. This duration variable is approximated by the average duration over all respondents in a survey across all locations with the same activity purposes, defined as *average-location-duration*(*activity*(l_i)).

5.2. Call probability at a location

Given a user's call rate and the duration that the individual has actual spent at l_i, the probability of making at least a call during the entire period of the visit to the location, defined as *CallP*(*user*,l_i), can be estimated in the following manner. The location duration is first divided into a number of equal-interval episodes, and each of the episodes can be regarded as an experiment. The length of the episodes, referred as *EpisodeL*, can be estimated,

e.g. by the average time that people spend on the phone each time when they are in the connection of the GSM network, e.g. 2 min for voice calls and a few seconds for the messages. Under the assumption that the user makes calls (including both initiating and receiving voice calls and messages) independently in each episode, and that the probability of making calls across different episodes at the location is identical, *CallP(user,l_i)* can then be modeled as the binomial distribution. The actual location duration delimits the total number of episodes, i.e. the number of independent experiments. While the call rate provides the probability of success for each experiment result, that is the probability of making a call in each episode. This leads to the final estimation of the probability *CallP(user,l_i)* as the probability of having at least one success (making at least one call) over the total number of experiments, in this case, over the total location duration.
In this study, the previously derived two variables CallRate and average-location-duration are used as the approximation of the call rate for a user and the duration for a location with a particular activity purpose, respectively. The probability *CallP(user,l_i)* is then obtained as follows:

$$CallP(\ user, l_i\ ) = CallP(user, activity(l_i)\ ) =$$
$$1 - \{1 - EpisodeL \times CallRate(user)\}^{average-location-duration(activity(l_i))/EpisodeL}$$

5.3. Sequence conversion probability
After the probabilities of making calls at a location of home, work or non-mandatory activities for a user are known, the likelihood that a call location trajectory is generated from an actual activity-travel sequence can be derived. In addition to the assumption that users make calls independently in each episode during a location visit, we also hypothesize that they make calls independently across each location visit. Let the sequence l_1 -> l_2 -> ... -> l_n represent the *actual-travel-sequence* for a user. Based on the previously derived probabilities *CallP(user,l_i),* the likelihood of various stop-location-trajectories that could be observed from the actual-travel-sequence, defined as the conversion probability *ConversionP*, can be calculated as shown below. The probability that the original complete travel sequence can be revealed by the call records is:

$$\text{ConversionP}(user, actual-travel-sequence, stop-location-trajectory)$$
$$=\text{ConversionP}(user, l_1 -> l_2 -> ... -> l_n, l -> l_2 -> ... -> l_n) = \prod_{i=1}^{n} callP(user, l_i).$$

While the probability that only a part of the travel sequence is observed, is

$$ConversionP\left(user, l_1 \to l_2 \to \cdots \to l_n, l_1 \to \cdots l_{i-1} \to l_{j+1} \cdots \to l_n\right)$$
$$= \prod_{m=1}^{i-1} callP\left(user, l_m\right) \times \prod_{m=i}^{j} \overline{callP\left(user, l_m\right)} \times \prod_{m=j+1}^{n} callP\left(user, l_m\right),$$
$$\text{with } \overline{callP\left(user, l_i\right)} = 1 - callP\left(user, l_i\right),$$

and where, we assume that no phone communications have been made during the visits to locations from x_i to x_j ($i \leq j$).

The above described sequence conversion process can be illustrated by the call data of User121 in our dataset. The probabilities to make at least one call at the locations of home, work and non-mandatory activities for this user are 0.805, 0.903 and 0.424, respectively. Suppose the sequence of HWOH represents the actual activity-travel behavior of this

individual on a certain day, there could be a total of 15 various call location traces generated by this travel sequence, and the sum of the corresponding conversion probabilities is 1. For instance, the possibility to emanate the call trajectory of HWH is *ConversionP(user121, HWOH, HWH) = 0.34*.

5.4. Derivation of activity-travel sequences

Based on the previously obtained conversion probabilities and the frequencies of the observed call location trajectories, the occurrences of original activity-travel sequences can be ultimately derived. Suppose that *m* different stop-location-trajectories $s_1,s_2,\ldots,s_m$ are constructed from a user's call records with the observed frequencies as $y_1,y_2,\ldots,y_k$ respectively, and that they are sorted by the length of these sequences, i.e. $length(s_1) \geq length(s_2) \geq \ldots \geq length(s_m)$. The original occurrences of the corresponding travel sequences for the user, denoted as $x_1,x_2,\ldots,x_k$, can be estimated by the following linear equations; to simplify, the parameter of *user* in the function *ConversionP()* is omitted here:

$$\begin{cases} x_1 \times ConversionP(s_1,s_1) = y_1 \\ x_1 \times ConversionP(s_1,s_2) + x_2 \times ConversionP(s_2,s_2) = y_2 \\ \cdots \\ x_1 \times ConversionP(s_1,s_k) + x_2 \times ConversionP(s_2,s_k) + \ldots + x_k \times ConversionP(s_k,s_k) = y_k \end{cases} \quad (1)$$

An additional constraint is added to the unknown variables $x_1,x_2,\ldots,x_k$, in order to ensure that the total number of the derived travel sequences is equal to that of the observed call trajectories, i.e. $\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} y_i$. From this equation, variable x_k in the last equation in formula (1) is then substituted with the new value of this variable ($x_k = \sum_{i=1}^{k} y_i - \sum_{i=1}^{k-1} x_i$), resulting in the formation of formula (2) as follows:

$$\begin{cases} x_1 \times ConversionP(s_1,s_1) = y_1 \\ \cdots \\ x_1 \times ConversionP(s_1,s_{k-1}) + x_2 \times ConversionP(s_2,s_{k-1}) + \ldots + x_{k-1} \times ConversionP(s_{k-1},s_{k-1}) = y_{k-1} \\ x_1 \times \left[ConversionP(s_1,s_k) - ConversionP(s_k,s_k) \right] + \ldots + x_{k-1} \times \left[ConversionP(s_{k-1},s_k) - ConversionP(s_k,s_k) \right] \\ = y_k - ConversionP(s_k,s_k) \times \sum_{i=1}^{k} y_i \end{cases} \quad (2)$$

Formula (2) is a model with *k* equations and *k–1* unknown variables. To find the optimal solution, we use Linear Least Square Methods which are a standard approach to find the solution to a set of unknown factors from a model that has more equations than the unknowns (Chen & Plemmons, 2009; Van de Geer, 2000). This approach searches for the answer by minimizing the sum of the squares of errors (or residuals) made in the results of every single equation. A *residual* is the difference between an observed value and the fitted value provided by the estimated model. As well as minimizing the total sum of squared residuals, the obtained results also maximize the likelihood of the observed values, i.e. the frequencies of the observed call location trajectories in this study.

Specifically, to solve the equations in formula (2), we assume the least square estimators are: $\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{k-1}$; the residuals for each of the equations are then calculated as follows:

$$residual_1 = \hat{x}_1 \times \text{ConversionP}(s_1, s_1) - y_1$$

...

$$residual_k = \hat{x}_1 \times [\text{ConversionP}(s_1, s_k) - \text{ConversionP}(s_k, s_k)] ...$$

$$+ \hat{x}_{k-1} \times [\text{ConversionP}(s_{k-1}, s_k) - \text{ConversionP}(s_k, s_k)] - [y_k - \text{ConversionP}(s_k, s_k) \sum_{i=1}^{k} y_i]$$

The total sum of the squared residuals, denoted as *Sum*, can be obtained as $Sum = \sum_{i=1}^{k} (residual_i)^2$.

The minimum of *Sum* is then found by setting its partial derivatives to zero, which is $\frac{\partial Sum}{\partial \hat{x}_i} = 0, i = 1, ...k-1.$

The above models have the equal number of (i.e., *k-1*) equations and unknown variables, thus leading to the final resolution of the estimators $\hat{x}_1, \hat{x}_2, \cdots \hat{x}_{k-1}$, as well as of the estimator $\hat{x}_k$ as $\hat{x}_k = \sum_{i=1}^{k} y_i - \sum_{i=1}^{k-1} \hat{x}_i$.

In the case of User121, five different stop-location-trajectories are revealed by his/her call records over a total of 10 days, including HWOH, WH, OH, W and H, with the occurrences as 1, 3, 2, 1 and 3 respectively. The original frequencies of these sequences, i.e. x_1-x_5, are estimated based on the above described methods as $\hat{x}_1 = 1.17, \hat{x}_2 = 3.74, \hat{x}_3 = 4.89, \hat{x}_4 = 0.07, \hat{x}_5 = 0.14$, with the total sum of the squared residuals between the modeled frequencies and the real observed values is *Sum = 0.74*.

It can be noted that during the entire procedure of seeking the optimal solution, we assume that the original travel sequences could only occur within the space of observed call location trajectories, i.e. *Space = {*s_1,s_2,...,s_m*}*. In theory, however, there could be a chance that an observed call location trajectory is generated by many other potential travel sequences, rendering the solution space infinite. The definition of the search space of Space can be explained as follows. *(i)* Based on the well-established findings that human activity-travel behavior exhibits a high degree of spatial and temporal regularities as well as sequential ordering (e.g. Joh et al., 2008; Shoval & Isaacson, 2007; Wilson, 2008), a limited variety of real travel sequences for an individual can be assumed during a certain time scale. *(ii)* For a possible actual travel sequence, e.g. s_p, which is not in the observed Space, the optimal estimator of this sequence's actual occurrence x_p would be a value less than or equal to zero, due to the fact that the observed frequency of this sequence in its intact form is zero. This can be further demonstrated as follows. For User121, if the considered travel sequence s_p is longer than any trajectory in the Space of this user, i.e. *length(*s_p*)* $\geq$ *length(*s_1*)*, assume s_p = *HWOWH*, we obtain the equation as x_p × *ConversionP(HWOWH,HWOWH) = 0*. From this equation, we have the optimal solution as $\hat{x}_p \approx 0$. Similarly, if the length of s_p is shorter than certain observed trajectories in the Space, e.g. s_p = *HWO*, we have the equation of x_1 × *ConversionP(HWOH,HWO)* + x_p × *ConversionP(HWO,HWO) = 0*, from which a value of $\hat{x}_p < 0$ would be preferable. Based on the above two considerations, the optimal solution of the frequencies of original travel sequences would most likely found within the space of Space.

6. Classification

All the obtained stop-location-trajectories constructed directly from the call records as well as the actual-travel-sequences that undergo a transformation process, are subsequently classified according to the previously established home-based-tour-classification and daily-sequence-classification, respectively. During this classification, a home location H is added at the beginning and the end of a sequence if it is absent from this sequence, based on the assumption that each individual starts and ends a day at home. For each of these two types of sequences, two corresponding profiles are obtained and they are stored into matrices, namely *home-based-tour-profile* and *daily-sequence-profile*.

The Pearson correlation coefficient r is used to measure the relation between the corresponding profile matrices built from different sets of sequences. It reveals the strength of linear relationship between two matrices; the closer the value is to 1, the stronger the relationship is. The coefficient is computed as follows.

$$\overline{A} = \frac{\sum_{i=1}^{d} A_i}{d}, \overline{B} = \frac{\sum_{i=1}^{d} B_i}{d},$$

$$S_A = \sqrt{\frac{\sum_{i=1}^{d} (A_i - \overline{A})^2}{d}}, S_B = \sqrt{\frac{\sum_{i=1}^{d} (B_i - \overline{B})^2}{d}}, r = \frac{\sum_{i=1}^{d} \left(\frac{A_i - \overline{A}}{S_A} \right) \left(\frac{B_i - \overline{B}}{S_B} \right)}{d - 1}.$$

$$\overline{A} = \frac{\sum_{i=1}^{d} A_i}{d}, \overline{B} = \frac{\sum_{i=1}^{d} B_i}{d},$$

$$S_A = \sqrt{\frac{\sum_{i=1}^{d} (A_i - \overline{A})^2}{d}}, S_B = \sqrt{\frac{\sum_{i=1}^{d} (B_i - \overline{B})^2}{d}}, r = \frac{\sum_{i=1}^{d} \left(\frac{A_i - \overline{A}}{S_A} \right) \left(\frac{B_i - \overline{B}}{S_B} \right)}{d - 1}.$$

Where, A_i and B_j represent the matrix elements of the two concerned matrices A and B, respectively, with d as the total number of the matrix elements.

7. Case study

In this section, adopting the proposed profiling approach and using the mobile phone data described in Section 3, we carry out an experiment. In this process, a set of stop-location-trajectories are first constructed, followed by the translation of the trajectories into actual-travel-sequences. Each step of this process is highlighted with the examination of some particular parameters.

7.1. Construction of stop-location-trajectories

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

Fig. 3 describes the distribution of the frequencies of calls made in each hour of the weekdays, showing that from 9am in the morning, calls reach to their peak level; while from 18pm in the late afternoon, a second climax of call activities starts to occur. These two temporal points are chosen as the work-start-time and work-end-time, respectively.

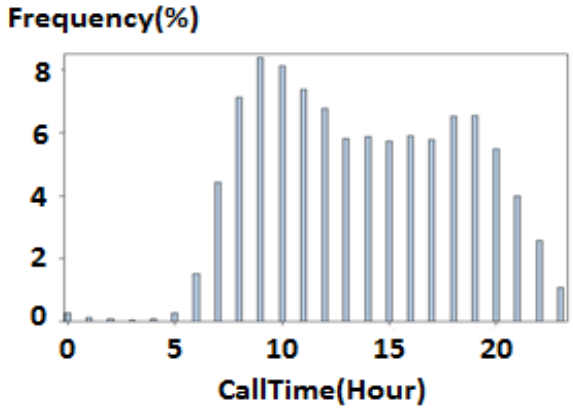

Fig. 3. The distribution of the time of calls

Based on the pre-defined criteria for home and work location identification, 49436 (98.9% of the total) users have their home locations discovered. The remaining 1.1% are those who made no calls at weekend or in the night period from 18pm to 9am across the two surveyed weeks. As a result, their homes cannot be spotted by these rules. Meanwhile, 9,458 users (18.9% of the total) are screened out as employed people, if they work between 9am and 18pm at least two weekdays per week. By contrast, those 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.
For those who have both predicted home and work locations, we further remove nearly 15% of the individuals who have unknown cell IDs for the identified home or work locations due to technical reasons that occur in the mobile phone data collection process. This results in a final dataset of 8,027 workers who represent 16% of the total users in the selected dataset. All the call records of these individuals during weekdays are extracted, and the consecutive calls made at a same location are aggregated. This leads to 69,578 raw-location-trajectories constructed for further analysis.

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

For each location in the above obtained raw-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, 40.3% of all locations from the raw-location-trajectories are removed; the remaining locations in these sequences form the set of stop-location-trajectories. The average length of these trajectories is 3.3. In comparison, using the existing method which defines as a stop location if the interval between the first login of the location and that of its next location is longer than 60 min (Schlaich et al., 2010), 67.6% of all the call locations are dismissed, with the average length of the retained sequences as 2.33.

7.2. Transformation from stop-location-trajectories into actual-travel-sequences

7.2.1. Call-intervene and CallRate

For each user, all the calls made over the two survey weeks are counted, and the call-intervene and CallRate are computed based on the formulas in Section 5.1. The average call-intervene over all the identified workers is 192 min for a full day of 24 h. However, as demonstrated in Fig. 3, the occurrences of calls are not equally distributed, more calls are observed during the day than at night, the inclusion of the night period would bias the real call intervene time during the daytime period. In this study, only the period of 6am-12pm is thus taken into account as the time-span of a day. This reduces the average call intervene to 137 min; accordingly, the average call rate is 0.0073.

In the study reported by Calabrese et al. (2011), a 260 min of call intervene for an entire day is derived. The difference between the call intervene reported in this study and the one estimated by Calabrese et al. could be caused by the following factors. (*i*) Only workers are considered in our study. (*ii*) The mobile phone data in this experiment is more recent than the data used in the existing study. (*iii*) People could make more calls in Ivory Coast than in Massachusetts in the United States where the existing study is performed.

7.2.2. Average-location-duration

This variable value is approximated by the activity-travel survey conducted in Belgium which will be described in Section 8.2. From this survey, the average-location-durations are 222, 317 and 75 min for home, work and non-mandatory activities, respectively.

7.2.3. Episode length

This variable specifies the time window by which the location duration is split into a number of episodes, i.e. experiments. The length of this window is decided such that the call behavior of users in one episode should be independent of that in the next episode. To obtain such an episode length, the average voice call duration of users is considered, which is derived from an additional dataset that records the total number of voice calls as well as the total duration for these calls each hour between each two cells in the GSM network in Ivory Coast, over the 5-month data collection period. The resultant average call duration is 1.92 min, a 2-min interval is thus taken as the estimation of this episode length.

Based on all the above parameter settings, the call probabilities at a location of home, work and non-mandatory activities for a user are respectively derived; the average call probabilities over all the individuals for the three types of activity locations are 0.81, 0.88 and 0.41, respectively. These obtained probabilities of each user, combined with the observed frequencies of the stop-location-trajectories for the individual, lead to the prediction of the number of the actual-travel-sequences, using the method described in Sections 5.3 and 5.4.

8. Comparison of the results derived from mobile phone data with real activity-travel surveys

To illustrate the practical ability of our approach to really serve as a benchmark, we compare the results derived from the mobile phone data with the statistics drawn from real activity-travel surveys. Unfortunately, no official activity-travel surveys have been documented in Ivory Coast. Therefore, data stemming from other countries, including South Africa and Belgium, has been adopted for this purpose. The authors acknowledge that the real travel behavior in Ivory Coast most likely is considerably different from the one reported in South Africa and Belgium. Consequently, the illustration serves to underline the applicability of the approach, not to infer the travel behavioral relationships in this particular case. 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 activity locations, e.g. the home-based-tour-profile and the daily-sequence-profile.

8.1. The travel survey in South Africa

The South Africa National Household Travel Survey (NHTS) was the first national survey of travel habits of individuals and households, aimed at making significant improvements in public transport services. The survey was based on a representative sample of 50,000 households throughout South Africa and undertaken between May and June in 2003 (Department of transport, 2003)

The information recorded by the survey includes the travel time to various public transport modes, e.g. trains and buses, as well as to activity locations, e.g. shops and post offices. The number of trips and the purposes for these trips are also documented for each individual on a typical weekday. The survey results reveal that the majority of the respondents can access to most of the activity services within half an hour (i.e. the travel time), and the average number of activity locations visited by a worker on a weekday is estimated between 3.46 and 4.06.

8.2. The travel survey in Belgium

Despite the relative geographic proximity between South Africa and Ivory Coast, the information on the NHTS survey is nevertheless limited. Particularly, the detailed travel patterns for each individual are not accessible for us. This necessitates the use of a second survey that provides activity-travel sequences on entire days and will be used as a reference for the illustration of the derived profiles.

The survey, namely SBO, stems from a large scale Strategic Basic Research Project on transportation modeling and simulation, and it was conducted on 2500 households between 2006 and 2007 in Belgium. In this survey, the respondents recorded trip information during the course of one week, such as trip start time and end time, purpose of the trip (e.g. activity type), and trip origin and destination (e.g. activity location). The average travel time is 24 min, comparable to the 30 min for a typical travel in South Africa.

In the SBO survey, activity locations are represented with statistical sectors, each of which ranges from a few hundred meters to a few thousands in radius, similar to the spatial granularity level of cell locations in GSM network. Table 2 illustrates a typical diary of respondent identified as 'HH4123GL10089'. Only the variables that are relevant for the current study are presented in this table; a more detailed variable list and elaboration on the survey can be found in (Cools et al., 2009).

Table 2. Activity-travel diary data

Respondent ID	Date	Trip Start Time	Trip End Time	Trip Origin	Trip Destination	Trip Purpose
HH4123GL10089	09/05/2006	07:45:00	08:00:00	34337	34345	Work
HH4123GL10089	09/05/2006	17:00:00	17:15:00	34345	34349	Shopping (non-mandatory)
HH4123GL10089	09/05/2006	17:40:00	18:05:00	34349	34337	Home

From the dataset, the diaries from 372 individuals who work at least two days a week are extracted. Note that only the activity-travel sequences recorded on weekdays are extracted. Activity duration at the destination of a trip is estimated as the time interval between the end time of the trip and the start time of its next trip, if the activity is not the first and the last one of a day. Otherwise, the duration is approximated in combination with the typical time for getting up in the morning and going to sleep in the evening in Belgium, which are estimated as 6am and 12pm, respectively (Hannes et al., 2012). An assumption that respondents start

and end a day at home, is also made, when the activity duration is calculated. Based on the above process, the duration of each activity for each individual on each weekday is computed. For instance, the previously demonstrated respondent has a daily activity sequence of HWOH, with the activity duration as 105, 540, 25 and 355 in min, respectively. All the obtained activity durations are averaged per activity type over all the individuals, and stored in the variable average-location-duration(activity), which has been previously used in the experiment to derive the actual-travel-sequences.

8.3. Statistics on the average length of sequences

Table 3 summarizes the statistics on the average number of locations visited each day, i.e. the average length of sequences, derived from the sequences of raw-location-trajectories, stop-location-trajectories and actual-travel-sequences which have been previously built based on the mobile phone data. The results drawn from both the NHTS and SBO surveys are also presented alongside as a comparison.

Table 3. Statistics on the average length of sequences[a]

Sequences	RLT	CSLT	ATS	NHTS	SBO
Average length of sequences	5.69	3.30	4.02	3.46-4.06	3.96

[a]The columns from left to right represent the raw-location-trajectories (RLT), call stop-location-trajectories (CSLT), actual-travel-sequences (ATS), NHTS and SBO surveys, respectively. The same abbreviation for each type of sequences will be used throughout the remaining tables and figures in this paper.

From Table 3, it was observed that the average length of sequences first drops from initial 5.69 for the raw-location-trajectories to 3.3 for the stop-location-trajectories, and then rises again to 4.02 for the estimated travel sequences which is the closest to the number observed in both NHTS and SBO surveys. In addition, the differences in this variable value imply the importance of the process from the identification of stop locations to the inference of complete travel sequences proposed by our approach, when analyzing activity-travel behavior based on the mobile phone data.

8.4. Home-based-tour-profile

Table 4 shows the relative frequency of each pattern class in the home-based-tour-classification, obtained from the stop-location-trajectories, the actual-travel-sequences and the SBO diaries, respectively. The differences in the frequencies of each pair of corresponding pattern classes are also listed.

Table 4. Home-based-tour-profile (%)[a]

Typical patterns	CSLT	ATS	SBO	ATS - CSLT	CSLT - SBO	ATS - SBO
H	9.0	4.4	6.4	-4.6	2.6	-2.0
HWH	50.3	39.1	42.9	-11.2	7.4	-3.8
HOH	18.0	26.3	32.5	8.3	-14.5	-6.2
HOWH	5.1	6.7	3.1	1.6	2.0	3.6
HWOH	8.2	10.3	10.8	2.1	-2.6	-0.5
HWOWH	3.4	3.8	1.6	0.4	1.8	2.2
HOWOH	2.5	4.1	1.9	1.6	0.6	2.2
HOWOWH	0.7	1.0	0.2	0.3	0.5	0.8
HWOWOH	1.4	2.1	0.5	0.7	0.9	1.6
HOWOWOH	0.5	0.8	0.1	0.3	0.4	0.7
More than 2 work activities	1.0	1.3	0.2	0.3	0.8	1.1

[a] The columns from left to right represent the typical patterns, the frequency of each pattern class relative to the total number of sequences within each type of CSLT, ATS and SBO, and the pairwise differences in the frequencies for each pattern among these three types of sequences, respectively.

Because of the event-driven nature of the mobile phone data, two major characteristics are expected when the stop-location-trajectories are converted into the actual-travel-sequences. First, an observed call location trajectory is generated not only from a travel sequence that is identical to this observed trajectory, but more likely from a sequence that is longer than this observed one. As a result, the number of short patterns should decrease while that of long patterns should increase, when the stop-location-trajectories are transformed into the actual-travel-sequences. Secondly, the lower the probability that people make calls at a location, the higher the frequency of the derived travel sequence that contains this location tends to be. These two important features are well reflected in the results shown in Table 4. For instance, on the one hand, a frequency reduction by 4.6% and 11.2% for the short patterns of H and HWH is observed, while for the long pattern of HWOWOH, a 0.7% increase is obtained, after the conversion process. On the other hand, despite that the pattern HOH is as short as HWH, the average call probability at the non-mandatory activity location O is the lowest among all the three activity types, according to our experiment results. This leads to a prediction of high percentage of the same pattern for the derived travel sequences, e.g. a 8.3% rise in this case.
When the patterns drawn from both the stop-location-trajectories and the derived actual-travel-sequences are compared with the ones characterizing the SBO diaries, correlation coefficients of 0.93 and 0.99 are obtained, respectively. The high correlations show an overall high level of similarities in terms of the frequency distribution of the home-based tour patterns between these types of sequences. Nevertheless, as stated before, the real travel behavior between Ivory Coast and Belgium can be very different due to the contextual variation between these two countries. The differences are particularly revealed in Table 4 by the frequency deviation for each specific pattern class between the derived travel sequences and SBO survey, which ranges between 0.5%-6.2% over all patterns. In addition, the increase in frequency for short patterns from the SBO survey, e.g. a 2%, 3.8% and 6.2% rise for H, HWH and HOH respectively, could also result from the problems of under-reporting of short trips or short-duration activities which typically occur in travel surveys (e.g. Cools et al., 2009).

8.5. Daily-sequence-profile

Fig. 4(a) depicts the correlation between the relative frequency of each pattern class in the daily-sequence-profile obtained from the stop-location-trajectories and the actual-travel-sequences. It shows that the majority pattern classes drawn from each of these types of sequences follow a similar distribution in relative frequencies. The few outliers can be divided into two groups: *(i)* the group of HWH, H and HWHWH with a 14.1%, 5.9% and 1.4% increase for the stop-location-trajectories, respectively; *(ii)* the other group consisting of HOH, HOWOH, and the patterns with more than 2 home-based-tours, being 3.7%, 2.5% and 2.1% higher for the actual-travel-sequences, respectively. This further demonstrates that, compared to the stop-location-trajectories, the derived travel sequences tend to have a high proportion for long patterns and for patterns which accommodate locations with low call probabilities, e.g. the non-mandatory activity locations. In contrast, a lower percentage is anticipated for short patterns and for patterns containing locations with high call probabilities, e.g. the work places, after the sequence conversion process.
Fig. 4(b) compares the daily-sequence-profiles obtained from the actual-travel-sequences with the one from the SBO diaries. In this figure, we found that most patterns have a moderately higher frequency for the actual-travel-sequences than the SBO data, except for a few pattern classes which show remarkably higher occurrences for the SBO diaries and which are mainly for short patterns, e.g. HWH, HOH and HWOH with a 7.3%, 7.1% and 3.2% rise, respectively. Apart from the inherent differences in travel behavior between these two

countries, this figure demonstrates again the possible missing records for short-duration trips or activities in travel surveys, which could cause the derived travel sequences to be shorter than they actually are, resulting in a relatively high frequency for short pattern classes. In addition, the high share for HOH could also be explained by the fact that the number of days when people work at home (telecommuting) is higher in Belgium than in Ivory Coast, as reported (Ruth & Chaudhry, 2008). In the case of the pattern HWOH, its high frequency might suggest that people in Belgium carry out more non-mandatory activities on the way from work back to home, which, nevertheless, needs to be further investigated. Finally, a further examination reveals that out of all 93 pattern classes in the daily-sequence-profile, 59 (63.4%) are zero frequencies for the SBO data; while for the stop-location-trajectories and actual-travel-sequences, only 16 patterns (17.2%) are not represented. It reflects that the sequences derived from the mobile phone data are more representative in travel behavior than the survey data, further underlying the significance of using mobile phone data to explore the characteristics of travel behavior.

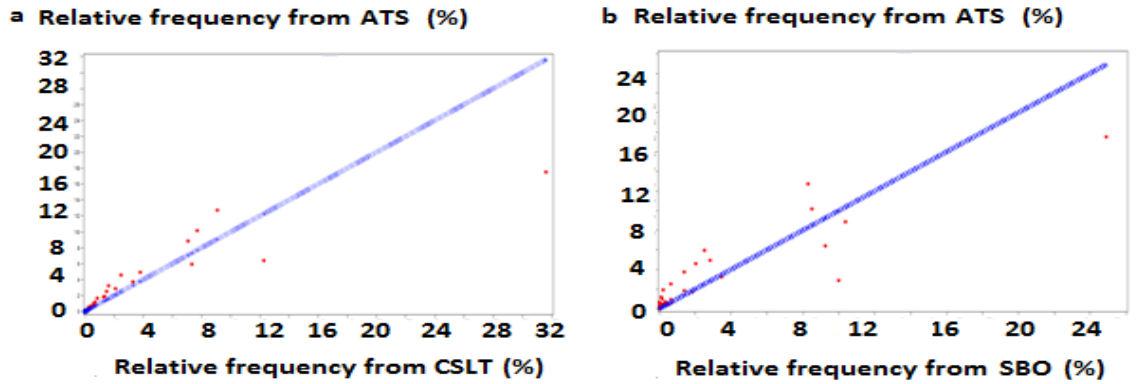

Fig. 4. Correlation between the relative frequency of each corresponding pattern class
Note: y-axis represent the relative frequency of each pattern class obtained from ATS; while x-axis denotes the relative frequency of the corresponding patterns obtained from CSLT (a) and SBO (b), respectively. The line of y=x is presented as a reference.

A correlation coefficient of 0.91 is obtained between the profile derived from stop-location-trajectories and that from the actual-travel-sequences. It shows that, while the profile of the travel sequences has accounted for the deviation in frequencies for each particular pattern class which are caused by the discrepancy between the call behavior and the actual activity-travel behavior, the two profiles have an overall close relationship. In addition, the correlation between the actual-travel-sequences and the SBO data is 0.89, suggesting that the derived profile is also comparable to the one drawn from a real travel behavior survey. These results suggest the derived profile of travel sequences can properly represent workers' travel behavior in a study area, and therefore capable of being used to validate the sequences generated from activity-based transportation models.
Nevertheless, in this case study, we used the surveys conducted in South Africa and Belgium as an illustration for the results derived by our approach. However, variation exists across different regions and countries. As previously described, travel behavior is shaped by the conditions of land use and transportation network as well as the social-economic background of individuals. Besides, several years of time differences when these datasets were collected, as well as the fact that the surveys, especially the SBO survey, were based on a small set of samples, all contribute to the deviation exposed in this experiment results between the derived travel sequences and the survey data. With a real travel survey conducted in the same or similar context to where the mobile phone data is obtained, it is believed that even better results than the current experimental outcomes can be reached.

9. Sensitivity analysis

Throughout the profiling process, several parameters including $T_{call\text{-}location\text{-}duration}$, $T_{maximum\text{-}time\text{-}boundary}$ and actual-location-duration, have been defined. This prompts to have a final investigation into how the parameter settings affect the predicted results. The results will be examined in the following aspects. (*i*) The average length of the stop-location-trajectories and actual-travel-sequences, referred as *CSLT-length* and *ATS-length* respectively. (*ii)* The coefficients between the stop-location-trajectories and the actual-travel-sequences as well as between the actual-travel-sequences and the SBO diaries, simplified as *r1* and *r2*, respectively.

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

In the process of stop location identification, when $T_{call\text{-}location\text{-}duration}$ increases, the minimum time duration required to consider a location as a stop becomes longer, leading to a decrease in the number of discovered daily locations. This is well reflected in Fig. 5(a). However, the rate of reduction is very slow; particularly, when this parameter reaches a certain threshold, e.g. 30 min set up in this experiment, the lengths of both types of sequences enter into a nearly constant level. A similar stabilization is observed in Fig. 5(b) when this parameter passes the 30 min threshold.

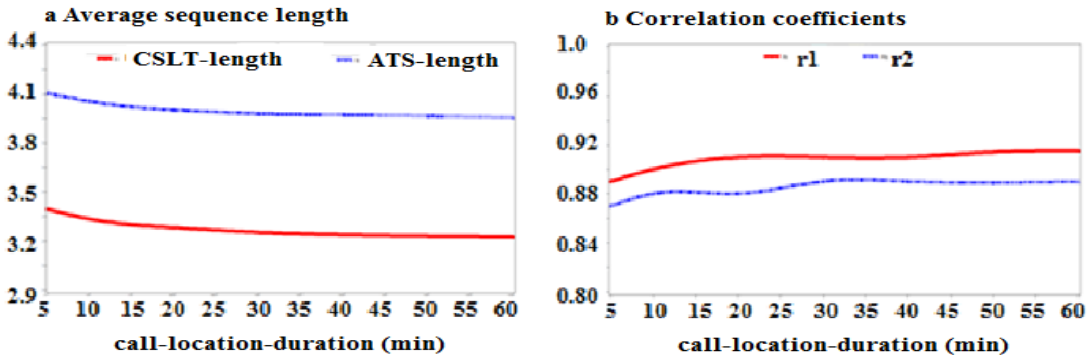

Fig. 5. Correlation between the threshold of call-location-duration and the results

Note: x-axis stands for the threshold of call-location-duration, and y-axis for *CSLT-length* and *ATS-length* respectively (a) and for the coefficients *r1* and *r2* respectively (b).

Fig. 6(a) and 6(b) show how the results evolve with $T_{maximum\text{-}time\text{-}boundary}$. As expected, when the maximum available time needed for a possible stop location sets longer, the number of identified stop locations drops, as shown in Fig. 6(a). However, this does not bring about the same amount of changes to the coefficients; especially when this parameter increases to a certain value, e.g. 60 min adopted in our experiment, both *r1* and *r2* develop into a stable level (see Fig. 6(b)). This can be explained by the fact that the dismissed stop locations are likely distributed randomly across various types of pattern classes, thus leading to the relative frequency of these patterns nearly the same.

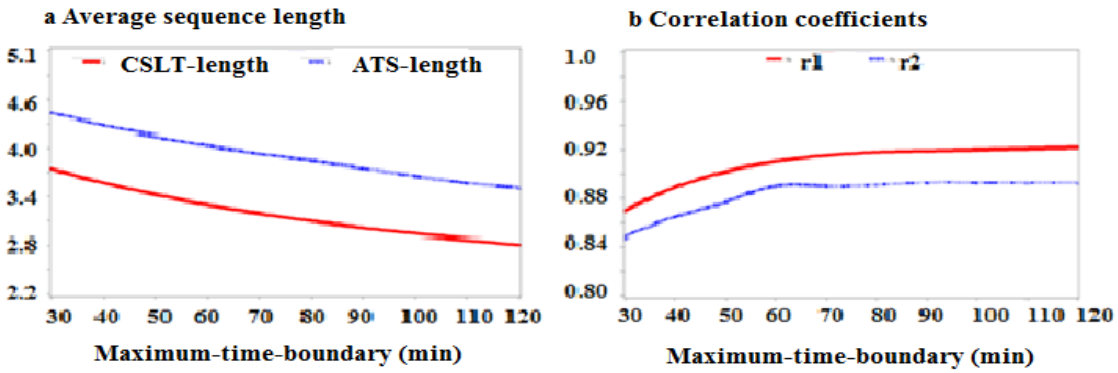

Fig. 6. Correlation between the threshold of maximum-time-boundary and the results
Note: x-axis stands for the threshold of maximum-time-boundary, and y-axis for *CSLT-length* and *ATS-length* respectively (a) and for *r1* and *r2* respectively (b).

9.2. Actual-location-duration

Fig. 7 describes the effects of the parameter actual-location-duration for work activities on the estimated results. It indicates that, as this duration becomes longer, *ATS-length* decreases while *r1* and *r2* increases, but these changes disappear when this duration pass a certain point, e.g. 240 min.

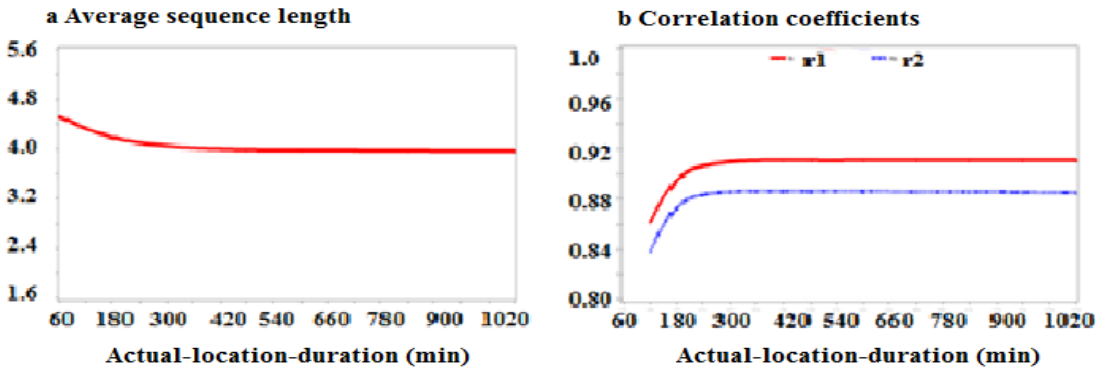

Fig. 7. Correlation between the actual-location-duration for work activities and the derived results
Note: x-axis stands for the actual-location-duration for work activities, and y-axis for ATS-length (a) and for r1 and r2 respectively (b).

This phenomenon can be explained by the binomial model employed to estimate the call probability at a location. According to this model, when the actual-location-duration is longer, the probability at a location $CallP(user,l_i)$ becomes higher, as demonstrated by Fig. 8(a). The amount of increases in the call probabilities as the activity duration extends, is not evenly distributed, which can be further evidenced by Fig. 8(b). It shows that as the activity duration becomes longer, the amount of growth in the location probabilities diminishes until to a nearly zero level. This explains the occurrence of the flat curves observed in Fig. 7.

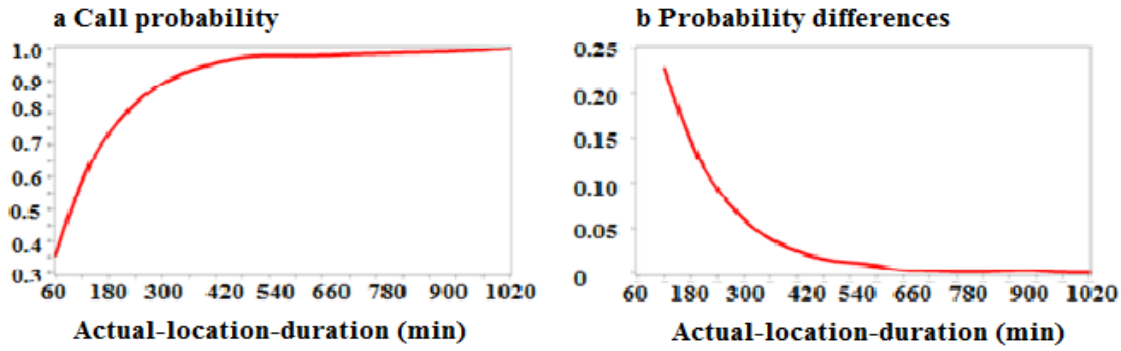

Fig. 8. Correlation between the actual-location-duration and the call probability at a location
Note: x-axis stands for the actual-location-duration for work activities, and y-axis for the call probability at a location (a) and for the difference between the call probability at the duration corresponding to the x-axis and the other probability obtained from a duration which is 60 min longer than this current actual-location-duration (b).

All these above analysis shows that, except that the increase in $T_{maximum\text{-}time\text{-}boundary}$ reduces the number of identified stop locations, a certain amount of changes in these parameters does not incur a significant deviation in the results of both the average length of the sequences and the profiles. This suggests that the profiles built upon the mobile phone data are stable and consistent in representing people's activity-travel behavior; a minor change in these parameter settings will not lead to a substantially different outcome.

10. Discussion and conclusion

The approach of profiling workers' travel behavior based on mobile phone data is both unique and important in that it builds a new measure which can be used to directly evaluate the activity-travel sequences simulated by activity-based transportation models. The advantage of using this approach is that it does not depend on conventional travel data survey methods. Thus, 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 relative long time period. The profile derived from the data is thus capable of providing a more representative and objective validation measure. Apart from the benefits that are realized by the use of mobile phone data, this approach also provides added value in taking into account the sequential constraints of activity-travel patterns into the evaluation.
The developed measure can be integrated into existing activity-based simulation models to assess the predicted sequences in two aspects. First, the evaluation can be done in terms of the average daily number of location visits that is encoded in the predicted sequences as well as in the actual-travel-sequences derived from the mobile phone data. Secondly, the assessment can be conducted on the temporal ordering of the activity locations, by comparing the corresponding profiles (i.e. the home-based-tour-profile or daily-sequence-profile) of workers' travel behavior built from both sets of sequences. High correlation coefficients between the profiles would suggest a high level of similarities between the sequences in terms of the sequential aspect. In contrast to SAM validation measures, which compare each single observed sequence with its predicted equivalent, the validation via the profiles compares the distribution of different pattern classes of the predicted sequences with that of the actual-travel-sequences, thus accounting for the variation of individuals' activity behavior among different days. If a mismatch in the daily number of location visits or a low correlation coefficient is found, this would suggest inconsistence between the predicted results and the travel behavior reflected by the mobile phone data, thus signaling possible problems and prompting immediate examination into the simulation model before the potential problems are propagated to the subsequence traffic assignment and travel demand analysis.

Besides the initial goal of building a new benchmark measure for activity-based models, the proposed method for stop location identification and subsequent actual-travel-sequence derivation can be employed in a wide range of applications where location path information of cell phone users plays a central role. The applications do not only include travel behavior and transportation modeling related research, such as mobility pattern discovery (Do & Gatica-Pereza, 2013; Schneider et al., 2013), transportation modelling and traffic analysis (Angelakis et al., 2013; Berlingerio et al., 2013; Calabrese et al., 2011), and urban planning (Becker et al., 2011; Jiang et al., 2012), but they also cover context-awareness services where user-centric assistance is provided based on users' specific location and activity context (García-Sánchez el al., 2013; Lee & Cho, 2013), and location tracking systems where knowledge of individuals' real-time locations and related routine activities is used in tools that provide support for industry, childcare, elderly health care and emergency rescue (Horng et al., 2011; Zhang et al., 2013; Zhou et al., 2014).

Despite the multitude of possible applications, there are also challenges that are pertinent to the data, as acknowledged by some of the existing studies (e.g. Calabrese et al., 2011; Jiang et al., 2013; Rose, 2006). 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. Whether the person is travelling or conducting an activity is not disclosed. Moreover, the places, where the individual has stayed but where no calls were made, are also unknown. The location update errors, under which a user's real location area is wrongly documented, further complicate the data collection issue. The results in our experiment suggest an average of 42.0% decrease in the number of visited locations per day, when the actual stop locations are singled out from the original call records. This number increases by 21.8% back, when the missing locations are interpolated into the identified stop-location-trajectories to form the complete activity-travel patterns. Such scales of changes in the daily number of visited locations signify the importance of the methods for accurately identifying stop locations and inferring missing places. Thus, if the proposed approach in this research is incorporated into the existing studies, in which the complete travel patterns of mobile phone users are first constructed, more accurate results that are derived from the whole picture of people's transfer phenomena could be subsequently reached.

Besides the added value in terms of potential applications, this study also makes important contributions from theoretical point of view. Particularly, in this paper, a novel process to derive actual travel location sequences from the stop-location-trajectories observed by means of mobile phone data has been constructed. This process integrates basic characteristics of human activity-travel behavior with statistical modelling, and links the daily activities and travels with call activities as these two behaviors occur at the same time and are carried out by the same individuals. In terms of methodological soundness it should be stressed that the process is based on the well-established findings that human activity-travel behavior exhibits a high degree of spatial and temporal regularities as well as sequential ordering (e.g. Joh et al., 2008; Shoval & Isaacson, 2007; Wilson, 2008).

With regard to the performance of the proposed approach, data collected from people's natural mobile phone usage has been used, and the real travel surveys conducted in South Africa and Belgium have been adopted as a reference. In this experiment, several comparisons have been carried out, and the results show the strengths of the proposed method. (*i*) When the profiles built from the stop-location-trajectories are compared with those constructed from the actual-travel-sequences, it was observed that the frequencies of short patterns decrease while those of long patterns increase. In addition, the frequencies of patterns, which accommodate locations (e.g. the non-work locations in this experiment) characterized with a lower call probability than at other types of places, also rise. These observations reflect well

the two statistical characteristics of the sequence conversion process. First, a stop-location-trajectory is generated more likely from a sequence that is longer than the observed one. Secondly, the lower the probability that people make calls at a location, the higher the frequency of the derived travel sequence that contains this location tends to be, in order to give rise to the same amount of sequences that can be observed through the call records. (*ii*) When the profiles built from the actual-travel-sequences are compared with the ones drawn from the travel survey conducted in Belgium, Pearson correlation coefficients of 0.99 and 0.89 are obtained for the home-based-tour-profile and daily-sequence-profile, respectively. The high correlation coefficients suggest that the derived profiles are comparable to the ones drawn from a real travel survey. (*iii*) A further examination into the daily-sequence-profiles reveals that, out of all 93 pattern classes in the profiles, 59 (63.4%) are zero frequencies for the one from the survey data, while 82.8% are represented in the profiles built from both the stop-location-trajectories and actual-travel-sequences, respectively. The larger variability of pattern classes exhibited by the latter trajectories and sequences reflects that the sequences derived from the mobile phone data are more representative in travel behavior than that obtained from the survey. (*iv*) Finally, the sensitivity analysis of various parameter settings demonstrates the consistency and stability of the derived results in representing travel behavior of the workers.

Despite the promising experiment results of this method, there are still certain areas which need to be enhanced in the future research. First, by considering a fixed work period (e.g. time interval between 9am and 18pm on weekdays in this experiment), individuals who work during night shifts are ignored. The prediction accuracy of residence and work locations could be improved by taking into account the information on individuals' work regime. In addition, the integration with a more comprehensive approach proposed by Liu et al. (2013), which annotates mobile phone locations with activity purposes based on the combination of machine learning techniques with the characteristics of underlying activity-travel behavior, would also increase the prediction precision. Secondly, in the process of stop location identification, the settings of two important parameters, namely $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ can be refined. $T_{call\text{-}location\text{-}duration}$ defines the maximum time duration needed to traverse a cell area. If the time interval between the first and last call time in a set of consecutive calls at a location area (i.e. the *call-location-duration*) is longer than this parameter value, the location is considered as a possible stop where the corresponding individual has performed activities at least during the extra time. Instead of using an overall threshold value of 30 minutes, as is done in this experiment, this parameter should be tailored to each particular cell and individual. $T_{maximum\text{-}time\text{-}boundary}$ considers the call times at three consecutive locations, and it estimates the total time required for the travel from the previous cell to the current one and from the current one to the next cell. If the time interval between the last call time at the previous location and the first call time at the next location (i.e. the *maximum-time-boundary*), is longer than this parameter value, the current location is regarded as a stop. Similar to $T_{call\text{-}location\text{-}duration}$, $T_{maximum\text{-}time\text{-}boundary}$ should also be adjusted to each particular individual and cell pairs, through the use of the proportion between the total distance of the cell pairs and the individual's travel speed. Thirdly, concerning the process of converting stop-location-trajectories into actual-travel-sequence, improvement can also be made in terms of the use of the actual location duration and the call rate. The estimation of the actual location duration could be separated by different social-economic groups. Regarding the call rate, instead of using a single call rate for each individual the call rate could be distinguished among different types of activities, as the probabilities that people make calls may differ depending on what they are doing.

While 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 travel survey data was used,

which stemmed from different environments than the settings of the phone data, as the reference to compare and illustrate the results. Nevertheless, in the future research, the proposed method must be applied to a real travel survey which is sampled in a similar context to where the phone data is obtained. Such surveys thus provide another possibility of enhancement by bringing more relevance to this method in terms of tuning up the parameters as well as validating the results.

With the rapid development of mobile phone based services in the future (e.g. Liu & Chen, 2013; Monares et al., 2013; Rodriguez-Sanchez et al., 2013; Zhang et al., 2013; Zhou 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. This study can thus be seen as a baseline, above which an even more accurate stop location identification and real travel sequence derivation based on the phone data will be reached, leading to an even better activity-based model validation standard as well as to an improved human location path analysis in general.

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 Orange Data for Development (D4D) challenge committee for the provision of mobile phone data, as well as the anonymous reviewers for their valuable comments on this paper.

References

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16(3), 471-494.

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.
Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., & Ziemlicki, C. (2012). Data for Development: the D4D Challenge on Mobile Phone Data. *Computer Science*.
Bradley, M., & Vovsha, P. (2005). A model for joint choice of daily activity pattern types of household members. *Transportation,* 32, 545-571.
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.
Chen, D., & Plemmons, R. (2009). Nonnegativity Constraints in Numerical Analysis. *In The Birth of Numerical Analysis.* World Scientific Press. Bultheel, A. & Cools, R. Eds., 109-140.
Cherchi, E., & Cirillo, C. (2010). Validation and Forecasts in Models Estimated from Multiday Travel Survey. *Transportation Research Record: Journal of the Transportation Research Board,* 2175, 57-64.
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.
Cools, M., Moons, E., & Wets, G. (2010a). Calibrating Activity-Based Models with External Origin-Destination Information: Overview of Possibilities. *Transportation Research Record: Journal of the Transportation Research Board,* 2175, 98–110.
Cools, M., Moons, E., & Wets, G. (2010b). Assessing the Quality of Origin-Destination Matrices Derived from Activity Travel Surveys: Results from a Monte Carlo Experiment. *Transportation Research Record: Journal of the Transportation Research Board,* 2183, 49–59.
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.
Department of Transport. (2005). Key Results of the National Household Travel Survey – Final Report. *Department of Transport*, Pretoria, South Africa. http://www.arrivealive.co.za/pages.aspx?nc=household.
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
Fan, Y., & Khattak, A. (2012). Time Use Patterns, Lifestyles, and Sustainability of Nonwork Travel Behavior. *International Journal of Sustainable Transportation*, 6(1), 26–47.
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.
GarcíA-SáNchez, P., GonzáLez, J., Mora, A.M., & Prieto, A. (2013). Deploying intelligent e-health services in a mobile gateway. *Expert Systems with Applications*, 40(4), 1231-1239.
González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*, 453, 779-782.
Hannes, E., Liu, F., Vanhulsel, M., Janssens, D., Bellemans, T., Vanhoof, K., & Wets, G. (2012). Tracking Household routines using scheduling hypothesis embedded in skeletons (THRUSHES). *Transportmetrica, Special Issue "Universal Design",* 8(3), 225-241.
Hansapalangkul, T., Keeratiwintakorn, P., & Pattara-Atikom, W. (2007). Detection and estimation of road congestion using cellular phones. *In Proceedings from 7th International conference on intelligent transport systems telecommunications*, 143–146.

Hartgen, D.T. (2013). Hubris or humility? Accuracy issues for the next 50 years of travel demand modeling. *Transportation*, 40(6), 1133-1157.
Horng, S. J., Chen, C., Ferng, H. W., Kao, T. W., & Li, M. H. (2011). Enhancing WLAN location privacy using mobile behavior. *Expert Systems with Applications*, 38(1), 175–183.
Huang, C. M., Hong, T. P., & Horng, S. J. (2009). Discovering mobile users' moving behaviors in wireless networks. *Expert Systems with Applications*, 36(8), 10809–10814.
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.
Jiang, S., Joseph Ferreira, J., & Gonzalez, M. C. (2012). Discovering urban spatial-temporal structure from human activity patterns. *In Proceedings of the Proceedings of the ACM SIGKDD International Workshop on Urban Computing (Beijing, China2012)*, ACM, 2346512, 95-102. http://dx.doi.org/10.1145/2346496.2346512.
Jiang, S., Yang, Y., Fiore, G., Ferreira J., Frazzoli, E., & González, M. C. (2013). A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, "Best Paper Award".
Joh, C. H., Arentze, T. A., & Timmermans, H. J. P. (2007). Identifying Skeletal Information of Activity Patterns by Multidimensional Sequence Alignment. *Transportation Research Record: Journal of the Transportation Research Board*, 2021, 81-88.
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.
Kopp, C., Kochan, B., May, M., Pappalardo, L., Rinzivillo, S., Schulz, D., & Simini, F. (2013). Evaluation of Spatio-temporal Microsimulation Systems. In Janssens, D., Yasar, A., Knapen, L. editor(s), *Data on Science and Simulation in Transportation Research*, *IGI Global*.
Lee, Y. S., & Cho, S. B. (2013). Mobile context inference using two-layered Bayesian networks for smartphones. *Expert Systems with Applications*, 40(11), 4333–4345.
Lemp, J., McWethy, L., & Kockelman, K. (2007). From Aggregate Methods to Microsimulation: Assessing Benefits of Microscopic Activity-Based Models of Travel Demand. *Transportation Research Record: Journal of the Transportation Research Board*, 1994, 80–88.
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., 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.
Rasouli, S., & Timmermans, H. (2013). Uncertainty in Predicted Sequences of Activity Travel Episodes: Measurement and Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2382, 46–53.
Rodriguez-Sanchez, M. C., Martinez-Romo, J., Borromeo, S., & Hernandez-Tamames, J. A. (2013). GAT: Platform for automatic context-aware mobile services for m-tourism. *Expert Systems with Applications*, 40(10), 4154–4163.
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.

Rose, G. (2006). Mobile phones as traffic probes: Practices, prospects and issues. *Transport Reviews*, 26(3), 275–291.
Ruth, S., & Chaudhry, I. (2008). Telework: A Productivity Paradox? *IEEE*, 12(6), 87-90.
Sammour, G., Bellemans, T., Vanhoof, K., Janssens, D., Kochan, B., & Wets, G. (2012). The usefulness of the Sequence Alignment Methods in validating rule-based activity-based forecasting models. *Transportation*, 39(4), 773-789.
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.
Schlaich, J., Otterstätter, T., & Friedrich, M. (2010). Generating Trajectories from Mobile Phone Data. *TRB 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies*, Washington, D.C., USA.
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), 20130246.
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*.
Shoval, N., & Isaacson, M. (2007). Sequence Alignment as a Method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers*, 97(2), 282–297.
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.
Van de Geer, S. A. (2000). Empirical Processes in M-estimation: Applications of empirical process theory. *Volume 6 of Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
Wegener, M. (2013). The Future of Mobility in Cities: Challenges for Urban Modelling. *Transport Policy*, 29, 275–282.
Wilson, C. (1998). Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, 30(6), 1017–1038.
Wilson, C. (2008). Activity patterns in space and time: calculating representative Hagerstrand trajectories. *Transportation*, 35(4), 485-499.
Yagi, S., & Mohammadian, A. (2010). An Activity-Based Microsimulation Model of Travel Demand in the Jakarta Metropolitan Area. *Journal of Choice Modeling*, 3(1), 32-57.
Zhang, L., Liu, J. C., Jian, H. B., & Guan, Y. (2013). SensTrack: Energy-Efficient Location Tracking With Smartphone Sensors. *Sensors Journal, IEEE*, 13(10), 3775-3784.
Zhou, M., Tian, Z. S., Xu, K. J., Yu, X., Hong, X., & Wu, H. B. (2014). SCaNME: Location tracking system in large-scale campus Wi-Fi environment using unlabeled mobility map. *Expert Systems with Applications*, 41(7), 3429–3443.
Zhou, M., Tia, Z. S., Xu, K. J., Yu, X., & Wu, H. B. (2013). Theoretical entropy assessment of fingerprint-based Wi-Fi localization accuracy. *Expert Systems with Applications*, 40(15), 6136–6149.