Chapter 5

Validating Activity-Based Travel Demand Models Using Mobile Phone Data

Feng Liu, Ziyou Gao, Bin Jia, Xuedong Yan,
Davy Janssens and Geert Wets

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75810

Abstract

Activity-based travel demand models predict travel sequences on a day for each individual in a study region. These sequences serve as important input for travel demand estimate and forecast in the area. However, a reliable method to evaluate the generated sequences has been lacking, hampering further development and application of the models. In this chapter, we use travel behavioral information inferred from mobile phone data for such validation purposes. Our method is composed of three major steps. First, locations where a user made calls on a day are extracted from his/her mobile phone records, and these locations form a location trajectory. All the trajectories from the user across multiple days are then transformed into actual travel sequences. The sequences derived from all phone users are further classified into typical patterns which, along with their relative frequencies, define travel profiles. These profiles characterize current travel behavior in the study region and can thus be utilized for assessing sequences generated from activity-based models. By comparing the obtained profiles with statistics drawn from conventional travel surveys, the validation potential of the proposed method is demonstrated.

Keywords: mobile phone data, travel sequences, activity-based travel demand models, travel surveys, travel behavior, travel sequence classification

1. Introduction

Activity-based travel demand models view travel as demand of activity participation. In this modeling framework, travel is analyzed in relation to daily activity behavior, the context of land-use and transportation networks, as well as personal background information

(e.g., socioeconomic conditions) [1]. *Travel surveys*, which collect full daily activities and travel of a small sample of individuals during one or a few days, are also required as training sets. Once the models are built, they can generate *travel sequences* (i.e., chains of activities and travel conducted by a person during a day) of each person in the study area using the Monte Carlo simulation approach. The individual travel sequences are then accumulated across the entire population, resulting in an origin-destination (OD) matrix. In this matrix, each element describes the number of trips between each pair of the corresponding locations of the area. This matrix is further assigned to the road network based on a traffic assignment algorithm, and the number of assigned trips on each road can subsequently be used as important input for mobility-related studies in the region (e.g., travel demand prediction, emission estimate, and transport policy evaluation). **Figure 1** demonstrates the entire process of an activity-based model.

Despite the comprehensive process of activity-based models, a reliable method has been absent to validate the simulated travel sequences. Traditionally, the model results are examined at both internal and external stages of the development process (see **Figure 1**). In the *internal validation*, the statistics aggregated from the simulated sequences (e.g., the average number of trips per day) are compared with those drawn from the expanded survey data that is not used as the training set of the model development but usually collected in the same survey period. Thus, the internal validation suffers from a number of limitations that are intrinsic to the shortcomings of the survey data [2]. In contrast, the *external validation* indirectly evaluates the model results at the traffic assignment stage. The assigned traffic volumes are compared against data from external sources (e.g., traffic counts) on a number of specified roads. However, good outcomes of the compared results might have resulted from the extra processes of OD matrix aggregation and traffic assignment, thus providing no convincing evidence of the accuracy of the model itself.

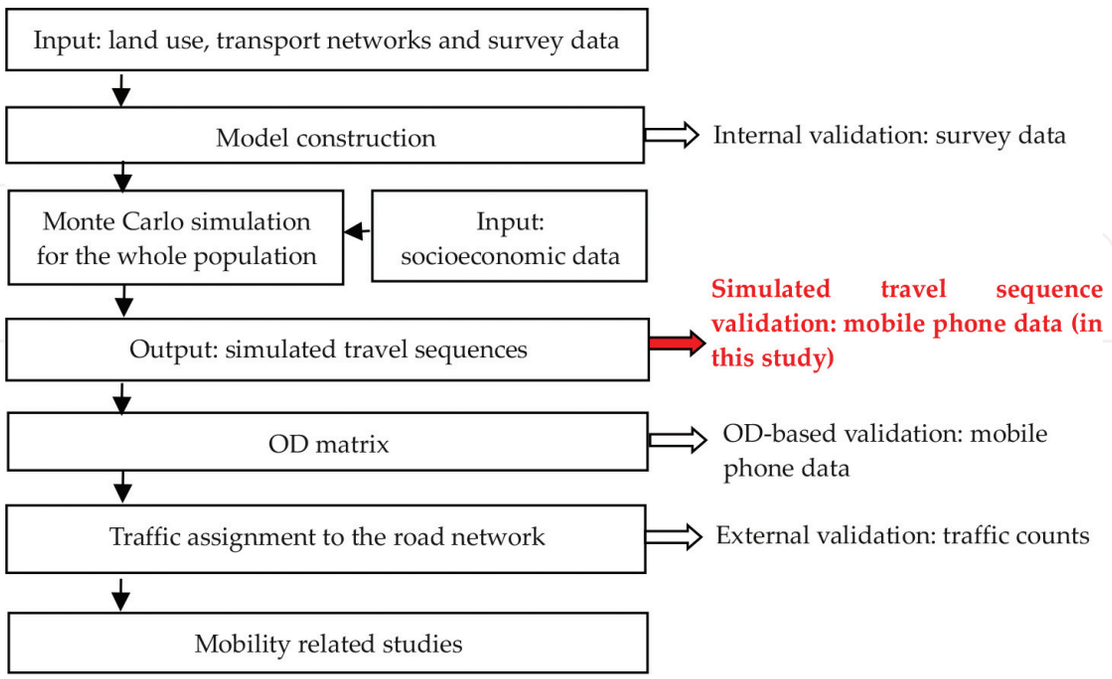

Figure 1. The entire process of an activity-based travel demand model.

Furthermore, if problems are found, it is also challengeable to trace these errors back to the model-building process. Nevertheless, despite the limitations, both internal and external validations are commonly adopted, as no methods have been developed for more accurately validating the model results [3].

The wide spread of mobile phones has offered an opportunity to use the devices as a new data collection method in transportation research. Call location data collected from the devices reflects the latest travel behavior of users, enabling up-to-date travel behavior studies on a large scale. Particularly, mobile phone data has been investigated for validating travel demand models; the study [4] can typify the state-of-art of such research. In this paper, based on the mobile phone data of 1 month recorded from 0.3 million users in Lisbon of Portugal, the two most frequent call locations (cell towers) for each of the users are first identified as home and work locations. An OD matrix is then built, aggregating commuting trips in the morning from home to work over all the users. This matrix is subsequently extrapolated according to a census survey to account for the total number of workers (1.3 million) in the city. Finally, the scaled matrix is compared with the morning travel demand that is predicted by a travel demand model developed in the study area. The results demonstrate the potential and feasibility of mobile phone data in benchmarking travel demand models (see **Figure 1**).

However, despite its advancement by adopting mobile phone data, the OD-based approach does not take into account the sequential information encoded in the call location patterns. It has been well documented that the choices of activities on a day are dependent on each other [5], as shown by the observation that the activity chain of having breakfast, travel, and working is often performed together on a working day. Furthermore, while the activity of bringing/getting people (e.g., bringing/getting children to/from schools) is usually conducted on the commuting ways, leisure is more executed in the evening. The interdependencies and temporal sequencing of daily activities have been regarded as a key factor in travel decision-making processes. The examination of how the simulated travel sequences are compatible with the sequential characteristics observed from the call location patterns is therefore important [6].

Addressing the above described limitations, our study proposes a new approach that utilizes the sequential characteristics of activity and travel behavior. Specifically, this approach first derives actual travel sequences from mobile phone data of all users. A set of *typical patterns* are then defined, each of which represents a certain class of the actual travel sequences. Profiles consisting of relative frequencies of the typical patterns in the travel sequences are subsequently computed. These profiles represent current workers' travel behavior and can thus be used to directly evaluate the simulated travel sequences yielded from activity-based models, by comparing them against the profiles obtained from the simulated sequences.

In relation to the existing OD-based method, the new approach offers the following major advantages. (1) It directly evaluates the predicted travel sequences, leading to more objective assessment and easier identification of the problems of the model system. (2) It examines the distribution of the sequences over the typical patterns, while the OD-based method looks into the number of trips across different OD pairs. In the new approach, the locations that are visited by an individual on a day are analyzed as a whole, while in the OD-based method, these locations are treated as unrelated individual activity participation. These two methods

have different perspectives and provide a complementary means of validating travel demand models.

The remainder of this chapter is organized as follows. Section 2 introduces the mobile phone data and Sections 3–5 detail the validation method. A case study is conducted in Section 6, and the obtained results are compared against real travel surveys in Section 7. Additional analysis on parameter sensitivity is performed in Section 8. Finally, Section 9 has discussions for future research and Section 10 draws major conclusions.

2. Mobile phone data

The mobile phone data consists of complete mobile communication patterns of 5 million anonymized users in Ivory Coast (i.e., 25% of this country's population) over 5 months between December 1, 2011 and April 28, 2012 [7]. The data contains the location (represented by cell ID) and time when each user conducts a call activity, including initiating or receiving a voice call or message. To address privacy concerns, the original data was divided into consecutive two-week periods; in each period, 50,000 users were randomly selected and their call records were extracted. This leads to a total of 10 datasets being generated. One of the datasets is used for this research. **Table 1** illustrates the typical call records of a user on Tuesday, December 20, 2011.

Call Time	**12:50:00**	**14:30:00**	**17:30:00**	**18:20:00**	**22:10:00**
Call location (Cell ID)	998	1520	982	956	956

Table 1. Call records of a user on a day.

3. Call location trajectory construction

A *call location trajectory* (i.e., *call-seq*) from a mobile phone user on a day can be described as a sequence of $l_1 \rightarrow l_2 \rightarrow \ldots \rightarrow l_n$, where l_i ($i = 1,\ldots,n$) is the cell ID and n is the *length* of the sequence, that is, the total number of locations that the user has reached and called that day. The *call frequency* (i.e., *call-fre*) of a set of consecutive calls made at each location l_i is denoted as k_i ($k_i > 0$), and the time for each of these calls is $T(1),\ldots, T(k_i)$. The *call interval* (i.e., *call-int*) between the first and last call time of these calls is $T(k_i) - T(1)$. Integraing the time signatures of the multiple calls, a *call-seq* can be formulated as $l_1(T(1),\ldots,T(k_1)) \rightarrow \ldots \rightarrow l_n(T(1),\ldots,T(k_n))$. From all the *call-seqs* of a user, home and work locations are first identified, and stop locations where the user has stayed for doing activities are then predicted.

3.1. Home and work locations

Two temporal points including *work start time* (i.e., *work-st*) and *work end time* (i.e., *work-et*) are estimated from the mobile phone data. The time when calls begin to substantially increase in

the morning on weekdays is considered as the *work-st*. Likewise, the moment when call activities reach the second climax in the late afternoon is selected as the *work-et*. Around this time, we assume that people start to communicate for after-work activities. Based on these two time points, a location is regarded as a home if it is the most frequent call location during the night period on weekdays between the *work-et* and *work-st* as well as throughout the entire weekend period. In contrast, a location is chosen as a work place if it is the most frequent call place in the work period between the *work-st* and *work-et* on weekdays, and if it is different from the previously identified home location for the user. In addition, the calls at a work location should occur at least 2 days a week.

3.2. Stop locations

After the home and work locations are identified for each worker, the remaining places in the *call-seqs* are either *stop locations* (i.e., *stop-locs*) where people pursue activities other than home and work activities or *non-stop locations* (i.e., *nonstop-locs*). The *nonstop-locs* can be further divided into *trip locations* (i.e., *trip-locs*) where the user travels or *false locations* (i.e., *falso-locs*) that are wrongly recorded because of location update errors. When call traffic is high in a user's location area, this location is shifted to less crowded cells for a short period of time, leading to location updates, but the user does not actually move. Furthermore, even for the previously identified home or work locations, some instances of these two locations could also be due to non-stop reasons, for example, people traveling in their work area while calling. Therefore, each location instance in the *call-seqs* should be differentiated between a *stop-loc* and *nonstop-loc*, irrespective of its activity types.

The two scenarios where *nonstop-locs* could occur can be demonstrated with the call data of two users. The first user (User298) has a trajectory of l_1(17:06,17:43) → l_2(17:51) → l_3 (17:56,19:41) → l_4 (21:55), where four locations are observed and the *call-int* is 37, 0, 105, and 0 (min) respectively. Each of these locations needs to distinguished between a *stop-loc* and a *trip-loc.* The trajectory of the second user (User64) 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 five location updates, with the *call-int* as 410, 0, 0, 0 and 65 (min), respectively. It is noted that the time difference between the first and second visits to l_2 is only 7 min. Although a small chance exists that this user may have moved to l_3 and l_4, the occurrences of these two places in such short time is most likely caused by location update errors.

In order to recognize *stop-locs* from all the possible *nonstop-locs* in a *call-seq*, a method is proposed as follows. For each l_i in the *call-seq*, the *call-int* is examined. If it is longer than a threshold $T_{call\text{-}int}$, l_i is regarded as a *stop-loc*. Otherwise, if the *call-int* is shorter than (or equal to) $T_{call\text{-}int}$ (e.g., in the case of a single call made at l_i), and if this location appears in the middle of the call trajectory, the time interval between the last call time at the previous location of l_i (i.e., $T(k_{i-1})$ at l_{i-1}) and the first call time at its next location (i.e., $T(1)$ at l_{i+1}), defined as the *maximum time boundary* (i.e., *max-boundary*), is examined. If this interval is longer than a threshold $T_{max\text{-}boundary}$, l_i is considered as a stop. However, if l_i is the first or last location of the trajectory, all the distinct stop locations already identified according to the previous steps from the user are aggregated. If l_i is among these locations, it is predicted as a stop. Otherwise, l_i is treated as either a *trip-loc* or *false-loc* and thus deleted. After the elimination of all the *nonstop-locs*, the

remaining places from the *call-seq* are stored into a *stop trajectory* (i.e., *stop-seq*). Each l_i in these trajectories is annotated with its *activity type* (i.e., *act*(l_i)), categorized into home (*H*), work (*W*), and other (*O*) activities. Travel is implicit in between each two consecutive locations. For instance, based on the above process, if 30 and 60 min are used for $T_{call\text{-}int}$ and $T_{max\text{-}boundar}$ as adopted in our case study, the obtained *stop-seqs* for User298 and User64 are $l_1 \rightarrow l_3 \rightarrow l_4$ and $l_1 \rightarrow l_2$, with the *activity types* of these locations as l_1 (*W*) $\rightarrow$ l_3 (*O*) $\rightarrow$ l_4 (*H*) and l_1 (*W*) $\rightarrow$ l_2 (*H*), respectively.

4. Trajectory transformation

Mobile phone data is event driven, in which locations are recorded only when the devices connect to the GSM network. Users' call behavior affects the number of trips and activity locations that are captured by the call data. The more active a user is in using the phone, the better his/her travel behavior is revealed by the device. The call locations can be regarded as the observed behavior at certain temporal points on a day, based on which real travel behavior of the users can be deducted. A method is therefore developed to transform the *stop-seqs* into *actual travel sequences* (i.e., *actual-seqs*) that represent real travel paths of the users on those days. This method is composed of the following steps. (1) For each user, two variables including the *actual activity duration* at a location l_i (i.e., *actual-dur*(*user,* l_i)), and the *call rate per minute* at all call locations of the user (i.e. *CallRate*(*user*)), are derived. (2) These two obtained variables are converted into a *call probability* that the user makes at least one call at l_i (i.e., *CallP*(*user,* l_i)). (3) Given a real travel sequence on a day, various *stop-seqs* could be possibly observed depending on the user's call behavior. The *conversion probability*, at which a certain *stop-seq* is generated from the *actual-seq* (i.e., *ConvertP*(*user, actual-seq, stop-seq*)), is calculated. (4) Based on the observed frequencies of all the *stop-seqs* from the user, a linear equation is built and the frequencies of the actual travel sequences are inferred.

4.1. *CallRate*(*user*) and *actual-dur*(*user,* l_i)

The *CallRate*(*user*) describes the probability that a user makes calls each minute, and it is calculated as follows:

$$CallRate(user) = \frac{\sum_{day} total - number - calls(user, day)}{\sum_{day} time - span(day)} \quad (1)$$

where, *total-number-calls* (*user, day*) and *time-span*(*day*) denote the total number of calls each day for the user and the time interval (min) of these days. *Actual-dur*(*user,* l_i) is the actual duration (min) that a user spends at l_i. Since no information on the actual stop duration is provided from the phone data, we approximate this variable using the average of the duration of all stop locations with the same activity types over all respondents that are obtained from a travel survey (see Section 7.2). The derived *average duration of locations per activity type* is referred as *actual-dur*(*act*(l_i)).

http://dx.doi.org/10.5772/intechopen.75810

4.2. *CallP*(*user, l_i*)

Given a user's call rate and the *actual-dur*(*user, l_i*) that the individual has spent at l_i, the probability that the user makes at least one call during the visit to l_i (i.e., *CallP*(*user, l_i*)) is cumputed based on the following steps. (1) The *actual-dur*(*user, l_i*) is first divided into a number of equal-length intervals. Each of these intervals is regarded as an experiment, and its length (i.e., *EpisodeL*) can be decided by the average duration that people spend on the phone each time they connect the GSM network (e.g., 2 min in our case study). (2) Under the assumption that users make calls independently in each interval and that the likelihoods of calling across different intervals are identical, *CallP*(*user, l_i*) then follows the binomial distribution. The *actual-dur*(*user, l_i*) decides the total number of intervals (i.e., independent experiments), and the call rate provides the probability of calling in each interval (i.e., the success for each experiment result). (3) *CallP*(*user, l_i*) can be computed according to Formula (2), as the probability of making at least one call (i.e., having at least one success) over the entire *actual-dur*(*user, l_i*) (i.e., the total number of experiments). *CallRate*(*user*) and *actual-dur*(*act*(l_i)) are used as the approximation of the call rate and the *actual-dur*(*user, l_i*) for a particular activity type of l_i.

$$CallP(user, l_i) = 1 - \{1 - EpisodeL \times CallRate(user)\}^{actual-dur(act(l_i))/EpisodeL} \tag{2}$$

4.3. *ConvertP*(*user, actual-seq, stop-seq*)

Let $l_1 \rightarrow l_2 \rightarrow \ldots \rightarrow l_n$ represent the *actual-seq* on a day for a user. Based on *CallP*(*user, l_i*), the probability that a certain *stop-seq* could be observed from the original sequence, that is *Convert* (*user, actual-seq, stop-seq*), is calculated. For instance, the probability of generating the *stop-seq* $l_1 \rightarrow \ldots l_{i-1} \rightarrow l_{j+1} \ldots \rightarrow l_n$ (i ≤ j) is

$$\begin{aligned} &ConvertP\ \left(user, l_1 -> l_2 \ldots -> l_n, l_1 -> \ldots l_{i-1} -> l_{j+1} -> \ldots l_n\right) \\ &\quad = \prod_{m=1}^{i-1} CallP(user, l_m) \times \prod_{m=i}^{j} \overline{CallP(user, l_m)} \times \prod_{m=j+1}^{n} CallP(user, l_m), \\ &\quad \overline{CallP(user, l_m)} = 1 - CallP(user, l_m) \end{aligned} \tag{3}$$

Where, we assume that no calls were made during the visits to the locations from l_i to l_j. We also hypothesize that users make calls independently across different location visits.

The conversion process can be demonstrated by the call records of User302. The probabilities that this user makes at least one call at home, work, and other locations are 0.81, 0.90, and 0.42, respectively. Assuming that this individual has an *actual-seq* of *HWOH* on a certain day, a total of 15 different *stop-seqs* could be possibly generated from this original sequence. The sum of the conversion possibilities of these *stop-seqs* is 1. For instance, the possibility of generating *HWH* is *ConvertP*(*user, HWOH, HWH*) = *0.34.*

4.4. *Actual-seq* derivation

Let $y_1, y_2, \ldots, y_k$ represent the frequencies of all k different *stop-seqs* of $s_1, s_2, \ldots, s_k$ constructed from a user's call records. These *stop-seqs* are sorted by their *length* in a descending order, with

s_1 having the largest number of locations. Assume that the corresponding *actual-seqs* of the user also occur among $s_1, s_2, \ldots, s_k$; the frequencies of the *actual-seqs* (i.e., $x_1, x_2, \ldots, x_k$) can be estimated based on Formula (4). Note that the parameter user in *ConvertP* is left out.

$$\begin{aligned} & x_1 \times ConvertP(s_1, s_1) = y_1 \\ & x_1 \times ConvertP(s_1, s_2) + x_2 \times ConvertP(s_2, s_2) = y_2 \\ & \ldots \\ & x_1 \times ConvertP(s_1, s_k) + x_2 \times ConvertP(s_2, s_k) + x_2 \times ConvertP(s_k, s_k) = y_k \end{aligned} \tag{4}$$

An additional constraint $\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} y_i$ is added to the above formula, in order to ensure that the total number of the derived sequences and that of the observed trajectories are equal. This leads to a model with $k + 1$ equations and k unknown variables $x_1, x_2, \ldots, x_k$. To find the optimal solution to the unknown variables, the Linear Least Square Method is employed. This method searches for the answer by minimizing the sum of the squares of *residuals* that are the differences between the observed frequencies and the corresponding estimated frequencies by the model. Specifically, let the estimators of $x_1, x_2, \ldots, x_k$ as $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_k$; the *residual* for the ith equation (i.e., $residual_i$, $i = 1,\ldots,k$) is calculated as follows.

$$\begin{aligned} & residual_1 = \widehat{x}_1 \times \mathrm{ConvertP}(s_1, s_1) - y_1 \\ & \ldots \\ & residual_k = \widehat{x}_1 \times \mathrm{ConvertP}(s_1, s_k) + \widehat{x}_2 \times \mathrm{ConvertP}(s_2, s_k) + \ldots + \widehat{x}_k \times \mathrm{ConvertP}(s_k, s_k) - y_k \end{aligned} \tag{5}$$

With $\widehat{x}_k$being replaced with $\widehat{x}_k = \sum_{i=1}^{k} y_i - \sum_{i=1}^{k-1} \widehat{x}_i$, the last equation is converted as

$$residual_k = \widehat{x}_1 \times \mathrm{ConvertP}(s_1, s_k) + \ldots + \left(\sum_{i=1}^{k} y_i - \sum_{i=1}^{k-1} \widehat{x}_i \right) \times \mathrm{ConvertP}(s_k, s_k) - y_k \tag{6}$$

The total sum of the squared residuals (i.e., *residual-sum*) is computed, and the minimum of the *residual-sum* is found by setting its partial derivatives to zero as follows.

$$residual - sum = \sum_{i=1}^{k} (residual_i)^2, \quad \frac{\partial (residual - sum)}{\partial \widehat{x}_i} = 0, i = 1, \ldots k-1 \tag{7}$$

This results in the computation of $\widehat{x}_1, \widehat{x}_2, \ldots \widehat{x}_{k-1}$; $\widehat{x}_k$ is obtained as $\widehat{x}_k = \sum_{i=1}^{k} y_i - \sum_{i=1}^{k-1} \widehat{x}_i$.

In the case of User302, the *stop-seqs* derived from his/her call records of 10 days are *HWOH, WH, OH, W,* and *H*, with the frequencies as 1, 3, 2, 1, and 3, respectively. The original frequencies of these sequences are estimated as $\widehat{x}_1 = 1.17, \widehat{x}_2 = 3.74, \widehat{x}_3 = 4.89, \widehat{x}_4 = 0.07, \widehat{x}_5 = 0.14$, and the *residual-sum* is 0.74.

During the above described process, we assume that the actual travel sequences *actual-seqs* could only occur within the set of the observed location trajectories, that is, *Set* = $\{s_1, s_2, \ldots, s_k\}$. This is

based on the well-established findings that human activity and travel behavior exhibit a high degree of spatial and temporal regularities as well as sequential ordering. A limited variety of travel sequences for a user can be observed during a certain time period. In addition, the optimal solution of the frequencies of the *actual-seqs* would be most likely found within the *Set*. This is due to the fact that if an *actual-seq* s_p is not in the *Set*, the optimal estimator $\widehat{x}_p$ for the frequency of s_p would be a value less than or equal to zero. For instance, for User302, if s_p is longer than any trajectory in the *Set*, e.g. s_p = *HWOWH*, the equation $x_p \times ConvertP(HWOWH, HWOWH) = 0$ is obtained. From this equation, we obtain $\widehat{x}_p \approx 0$. Similarly, if s_p is shorter than certain trajectories in the *Set*, for example, $s_p = HWO$, the equation $x_1 \times ConvertP(HWOH, HWO) + x_p \times ConvertP(HWO, HWO) = 0$ is constructed, leading to $\widehat{x}_p < 0$.

5. Workers' travel sequence classification

Figure 2 describes travel sequences for workers, in which a sequence is divided into four parts, including before-work, commute, work-based, and after-work parts. They respectively represent the activities and travel undertaken before leaving home to work (indicated by the arrow a, e.g., *HOH*), between home and work commutes (by b and d, e.g., *HOW* or *WOH*), work-based (by c, e.g., *WOW*), and after arriving home from work (by e, e.g., *HOH*).

A *home based tour* (i.e., *tour*) is defined as a chain of locations that starts and ends at home and accommodates at most two work location visits. For a working day, a *tour* can be classified into the patterns of *HWH, HOWH, HWOH, HWOWH, HOWOH, HOWOWH, HWOWOH,* and *HOWOWOH*, where *O* represents one or a chain of visits to several other different locations. On a non-working day, a *tour* is described with *H* or *HOH*. All the above 10 patterns characterize the *tours* for worker's travel behavior, and they are defined as *home-based tour classification* (i.e., *tour-class*). Each pair of these patterns is then merged, leading to 81 combinations that can be used to classify an entire day's sequences containing two *tours*. For instance, the combination of *HWH* and *HOWH* results in the class *HWHOWH*. The daily sequences that have more than two work location visits in a *tour* (e.g., *HWOWOWH*) or that contain more than two *tours* (e.g., *HWHWHWH*) are each formed into one additional category. Thus, all the combinations along with the original 10 *tour* patterns that describe daily sequences with only one *tour* lead to a total of 93 patterns. These patterns underlie workers' daily travel behavior,

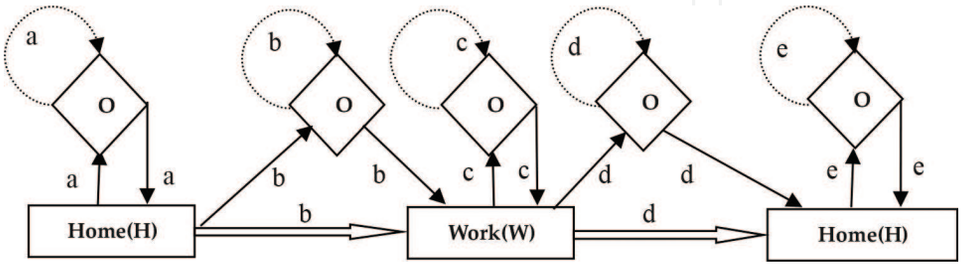

Figure 2. Workers' travel sequence representation. Note: H and W denote home and work locations respectively, while O refers to all other places. Each arrow from the start and end locations represents the related travel between these two places, and the arrow from the location O to itself indicates a chain of consecutive visits to different O locations.

and they are denominated as *day sequence classification* (i.e., *day-class*). Given a group of users, their travel sequences can be categorized according to the *tour-class* and *day-class*, respectively. The relative frequency of each of the patterns over the total occurrences of all the corresponding patterns among the travel sequences forms a *home-based-tour-profile* (i.e., *tour-profile*) and *day-sequence-profile* (i.e., *day-profile*) among these individuals.

Based on these two types of classification, all the *stop-seqs* and *actual-seqs* previously constructed from the mobile phone data are grouped. During this process, a home location H is added at the beginning and end of a sequence if it is absent from the sequence, under the assumption that each user starts and ends a day at home. After classification, two types of profiles, that is, the *tour-profiles* and *day-profiles*, are derived from both the *stop-seqs* and *actual-seqs*, respectively.

The Pearson correlation coefficient r (see Formula (8)) is used to measure the relation between the corresponding profiles derived from the different sets of sequences. It reveals the strength of relationship between the compared sets; the closer r is to 1, the stronger the relationship is.

$$r = \frac{\sum_{i=1}^{d}\left(\frac{A_i - \overline{A}}{S_A}\right)\left(\frac{B_i - \overline{B}}{S_B}\right)}{d-1}, \overline{A} = \frac{\sum_{i=1}^{d} A_i}{d}, \overline{B} = \frac{\sum_{i=1}^{d} B_i}{d}$$
$$S_A = \sqrt{\frac{\sum_{i=1}^{d}\left(A_i - \overline{A}\right)^2}{d}}, S_B = \sqrt{\frac{\sum_{i=1}^{d}\left(B_i - \overline{B}\right)^2}{d}} \tag{8}$$

where A and B denote the two compared profiles, A_i and B_i are the frequencies of the pattern i in these two profiles and d denotes the total number of the patterns in each profile.

6. Case study

In this section, adopting the proposed approach and using the mobile phone data described in Section 2, we carry out a case study. In this process, *stop-seqs* are first constructed and *actual-seqs* are then derived.

6.1. *Stop-seq* construction

Figure 3 describes the distribution of the number of calls in each hour of the weekdays, showing that the peaks of calls in the morning and in the afternoon occur at 9 am and 18 pm, respectively. These two temporal points are chosen as the *work-st* and *work-et*. Based on the criteria for home and work locations, 49,421 (i.e., 98.8% of the total) users have their home identified, and 8016 users (i.e., 16.2% of the total) are screened out as employed people who work between 9 am and 18 pm at least two weekdays per week. All the call records of these workers on weekdays are extracted, resulting in a total of

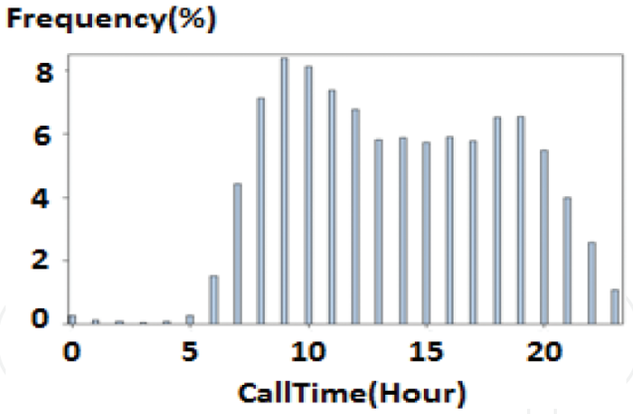

Figure 3. The distribution of call time.

69,536 *call-seqs*. From these sequences, 40.3% of the locations are removed as non-stop ones, using the thresholds $T_{call\text{-}int}$ and $T_{max\text{-}boundary}$ set as 30 and 60 min respectively. The remaining locations form the *stop-seqs*, with the average length of these sequences as 3.3 (locations).

6.2. *Stop-seq* transformation

The *time-span* (*day*) is specified as the period from 6 am to 12 pm. From each user, all the calls made during this period each day across the two survey weeks are counted, and the *CallRate*(*user*) is computed according to Formula (1). The average *CallRate*(*user*) over all the workers is 0.0073. The *actual-dur*(*act*(l_i)) is estimated from the travel survey conducted in Belgium which will be described in Section 7.2. From this survey, the duration is 222, 317 and 75 (min) for home, work, and other activities, respectively. The variable *EpisodeL* specifies the time window by which the *actual-dur*(*act*(l_i)) is split into a number of intervals, that is, experiments. To obtain this interval length, the total number of voice calls and the total duration of these calls over the 5-month mobile phone data in Ivory Coast are extracted. The ratio between these two variables leads to the average call duration as 1.92 min, and 2 min is thus taken as the estimation of *EpisodeL*. Based on all the above parameter settings, the call probabilities *CallP*(*user*, l_i) at home, work, and other locations for each user are respectively derived based on Formula (2). The average of *CallP*(*user*, l_i) over all the users for these three types of locations is 0.81, 0.88, and 0.41, respectively. The obtained call probabilities of each user, combined with the observed frequencies of the *stop-seqs* for the person lead to the calculation of the number of the *actual-seqs*, using the method described in Sections 4.3 and 4.4.

7. Result comparison with travel surveys

To examine the validation ability of the proposed method, the results inferred from the phone data are compared with the statistics drawn from real travel surveys. However, no

official surveys have been conducted in Ivory Coast, necessitating the use of data collected from other countries including South Africa and Belgium for this purpose. We acknowledge that the travel behavior in Ivory Coast and that in these two travel-surveyed countries are most likely different. Consequently, the comparison is to examine the validation potential of this approach but not to infer travel behavior relationship among these countries.

7.1. The two travel surveys

The South Africa National Household Travel Survey (*NHTS*) was based on a sample of 50,000 households over a period of 2 months between May and June in 2003 [8]. The collected information includes the number of trips on a typical weekday and the travel time and purposes of these trips. According to the survey results, the majority of the respondents can access most of the activity services in this country (e.g., train and bus stops as well as shops and post offices) within half an hour, and the average number of activity locations visited by a worker on a day is between 3.46 and 4.06.

The Belgian survey (*SBO*) was carried out on 2500 households between 2006 and 2010. This survey collects trip information of the respondents during the course of 1 week, such as trip origin and destination (i.e., activity locations), trip start and end time, and purposes of the trips (activity types). Activity locations are described with statistical sectors; the size of these sectors varies from a few hundred meters to a few thousands in radius, comparable to the spatial granularity of cells in a GSM network. Based on this survey, the average travel time is 24 min, 6 min shorter than a typical travel in South Africa. **Table 2** illustrates a representative diary of the respondent 'HH4150GL10190' on Tuesday, May 9th, 2006.

From all the respondents in the *SBO* survey, 342 individuals who work at least 2 days a week are selected, and the corresponding travel sequences are constructed. The duration of each location in the travel sequences is estimated as follows. If the location is not the first and the last one of the day, the duration is between the arrival time of the current trip at the location and the leaving time of the next trip from the location. Otherwise, the time of 6 am is used as the start time of the location if it is the first one of the day, and the time of 12 pm is adopted as the end time of the location if it is the last one on the day. For instance, the respondent demonstrated in **Table 2** has a sequence of *HWOH*, with the location duration as 165, 540, 25, and 255 (min), respectively. All the obtained location duration is averaged per activity type over all these individuals, leading to the estimate of the *actual-dur*($act(l_i)$), which has been used to derive the *actual-seqs* in the case study in Section 6.

Trip ID	Start Time	End Time	Origin	Destination	Purpose
1	08:45:00	09:00:00	34,137	34,145	Work
2	18:00:00	18:15:00	34,145	34,849	Shopping
3	18:40:00	19:05:00	34,849	34,637	Home

Table 2. Diary data.

7.2. Average length of sequences

Table 3 summarizes the average length of the sequences including *call-seqs*, *stop-seqs*, and *actual-seqs* derived from the mobile phone data as well as of the sequences constructed from the *NHTS* and *SBO* diaries. It shows that the average length of the sequences first decreases from 5.69 for the *call-seqs* to 3.3 for the *stop-seqs* and then rises back to 4.02 for the *actual-seqs* that is the closest to the length of both the *NHTS* and *SBO* diaries. The length differences suggest the importance of the process from the identification of stop locations to the inference of complete travel sequences, when travel behavior is analyzed based on mobile phone data.

7.3. *Tour-profiles*

Based on the classification method described in Section 5, two types of profiles, including the *tour-profiles* and *day-profiles*, are derived from the *stop-seqs*, *actual-seqs* and *SBO* diaries, respectively. **Table 4** shows the frequency of each pattern in the *tour-profiles*; the differences in the frequencies of the corresponding patterns between the *stop-seqs* and *actual-seqs* as well as between the *actual-seqs* and *SBO* diaries are also presented.

Due to the data collection nature of mobile phone data, when the *stop-seqs* are converted into the *actual-seqs*, two important characteristics are expected. (1) A *stop-seq* is generated not only from an *actual-seq* that is identical to this trajectory (e.g., when calls were made at each of the locations actually visited), but more likely from a sequence that is longer than this observed one (i.e., some of the real locations being missed if no calls were made there). Thus, when the *stop-seqs* are transformed into the *actual-seqs*, the number of long patterns increase, while that of short patterns decrease. (2) If the probability of making calls at a location is lower, the frequency for the derived *actual-seqs* that contain this location tends to be higher. These two features are well observed in **Table 4**. For instance, when the *actual-seqs* are compared with the *stop-seqs* (see the 4th row), the frequencies for the short patterns *H* and *HWH* decrease by 4.6 and 11.2%, while the frequency for the long one *HWOWH* increases by 0.7%. Furthermore, as the average call probability at the location *O* is the lowest among all the three activity types, an 8.3% rise is obtained for the pattern *HOH* among the *actual-seqs*. This forms a contrast with the pattern *HWH* that has an 11.2% decrease.

When the profiles drawn from the *actual-seqs* and *SBO* diaries are compared, a correlation coefficient of 0.99 is obtained (i.e., higher than the coefficient of 0.93 between the *stop-seqs* and *SBO* diaries). The high coefficient shows an overall high level of similarities across the patterns in the *tour-class* between these two types of sequences. Nevertheless, as previously indicated, due to the contextual deviations between Belgium and Ivory Coast, the real travel behavior between two countries can be very different. According to **Table 4** (see the 5th row), the differences in the frequencies over all the patterns between the *actual-seqs* and *SBO* diaries range from −6.2 to

Call-seqs	Stop-seqs	Actual-seqs	NHTS	SBO
5.69	3.30	4.02	3.46–4.06	3.96

Table 3. Average length of sequences.

	H	HWH	HOH	HOWH	HWOH	HWOWH	HOWOH	HOWOWH	HWOWOH	HOWOWOH	More than 2 W
SS	9.0	50.3	18.0	5.1	8.2	3.4	2.5	0.7	1.4	0.5	1.0
AS	4.4	39.1	26.3	6.7	10.3	3.8	4.1	1.0	2.1	0.8	1.3
SBO	6.4	42.9	32.5	3.1	10.8	1.6	1.9	0.2	0.5	0.1	0.2
AS - SS	−4.6	−11.2	8.3	1.6	2.1	0.4	1.6	0.3	0.7	0.3	0.3
AS – SBO	−2.0	−3.8	−6.2	3.6	−0.5	2.2	2.2	0.8	1.6	0.7	1.1

Note: The column represents the patterns, while the row denotes each single set of sequences (in the first three rows) and the differences in frequencies between pair sets of sequences (in the last two rows). SS, AS, and SBO refer to the stop-seqs, actual-seqs, and SBO diaries respectively.

Table 4. Tour-profiles (%).

3.6%. In addition, the increases in frequencies for short patterns *H*, *HWH* and *HOH* from the *SBO* survey by 2, 3.8 and 6.2%, respectively, could also be caused by the under-reporting of short trips or short-duration activities that typically occur in travel surveys. This leads to travel sequences obtained from the surveys being shorter than they actually are [1].

7.4. *Day-profiles*

Figure 4 describes the correlation between the frequencies of corresponding patterns in the *day-profiles*, where the x-axis represents the frequencies for the *stop-seqs* (**Figure 4a**) and *SBO* diaries (**Figure 4b**), respectively, while the y-axis denotes the frequencies for the *actual-seqs*. The line of $y = x$ is presented as a reference. From **Figure 4(a)**, it is noted that the majority patterns follow similar frequency distributions, with a coefficient as 0.91. However, there exist a few outliers that can be further divided into two groups. (1) The group of *HWH, H* and *HWHWH* with 14.3, 5.7, and 1.5% increases in frequencies for the *stop-seqs*, respectively. (2) The other group of *HOH, HOWOH*, and the patterns composed of more than two *tours*, showing 3.5, 2.4, and 2.2% higher for the *actual-seqs*. This further demonstrates that, in relation to the *stop-seqs*, the *actual-seqs* are likely to have a high percentage for long patterns and for patterns that contain locations featured with low call probabilities (e.g., the location *O*). In contrast, a lower proportion is expected among the *actual-seqs* for short patterns and for patterns consisting of locations featured with a high call rate (e.g., the location *W*).

In **Figure 4(b)**, the *day-profiles* obtained from the *actual-seqs* and the *SBO* survey are compared. It shows that the majority of the patterns have higher frequencies for the *actual-seqs* than for the *SBO* data (i.e., the points above the line). Nevertheless, a few patterns show higher occurrences for the *SBO* diaries (i.e., the points below the line) and they mainly consist of short patterns, for example, *HWH, HOH* and *HWOH* with 7.3, 7.1, and 3.2% increases, respectively. Apart from the deviations in travel behavior between these two countries, this figure demonstrates again the possibly missing records for short-duration trips or activities in travel surveys, resulting in a high frequency for short patterns. On top of that, further investigation reveals that out of all 93 patterns in the *day-profiles*, 57 (i.e., 61.3%) have zero frequencies for the *SBO* data; while for the *stop-seqs* and *actual-seqs*, only 18 patterns (i.e., 19.4%) are not present. It indicates that the sequences built from the mobile phone data are more diverse and representative in travel

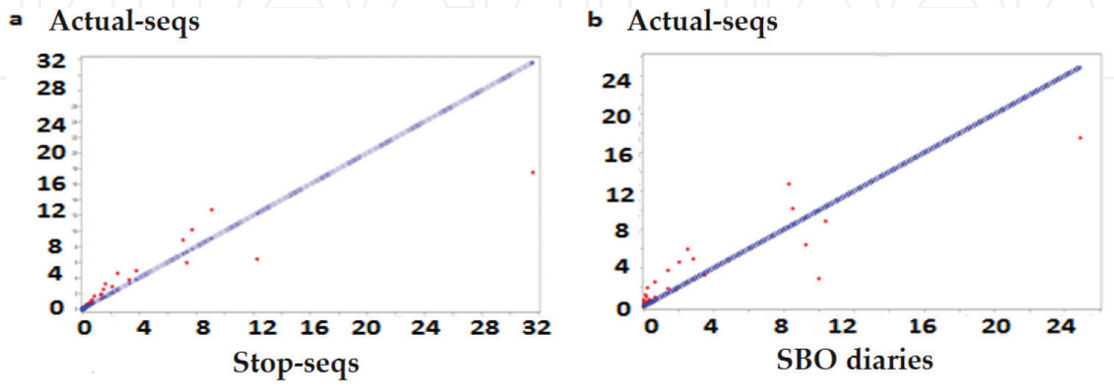

Figure 4. Correlation between the frequencies of corresponding sequences.Correlation between the frequencies of the actual-seqs and those of the stop-seqs (a) and SBO diaries (b).

behavior than the ones from the survey, further underlying the potential value of using mobile phone data for travel behavior analysis.

Despite the differences in each particular pattern, the coefficient is 0.89 between the whole *day-profiles* obtained from the *actual-seqs* and *SBO* survey. It shows that the profile inferred from the mobile phone data is comparable to the one extracted from a real travel survey. This further suggests that the derived profile can sufficiently represent workers' travel behavior in a study area, and therefore capable of validating the sequences generated from activity-based models.

8. Sensitivity analysis

In the proposed approach, several parameters including $T_{call\text{-}int}$, $T_{max\text{-}boundary}$ and *actual-dur*(*act* (l_i)) have been defined. Final investigation into how these parameters affect the derived results is conducted in two aspects. (1) The average length of the *stop-seqs* and *actual-seqs* (i.e., *SS-length* and *AS-length*). (2) The coefficients between the *stop-seqs* and *actual-seqs* as well as between the *actual-seqs* and *SBO* diaries (i.e., r_1 and r_2 respectively).

8.1. $T_{call\text{-}int}$ and $T_{max\text{-}boundary}$

$T_{call\text{-}int}$ defines the minimum time duration above which a call location is considered as a stop. The larger this value, the longer the duration of a stop is required, and the shorter the average length of the obtained sequence tends to be. This is well reflected in **Figure 5(a)**. However, the length of the sequences decreases slowly and enters into a constant level when this parameter passes the 30-min threshold (i.e., the value adopted in our case study). Similarly, r_1 and r_2 reach a stable level at the same 30-min threshold.

The parameter $T_{max\text{-}boundary}$ specifies a minimum value, such that given the current location under investigation and the trip to this location, if the time interval between the start of its next trip and the end of its previous trip is longer than this threshold, the current location is predicted as a stop. From **Figure 6a**, it is observed that as $T_{max\text{-}boundary}$ increases, both *SS-length*

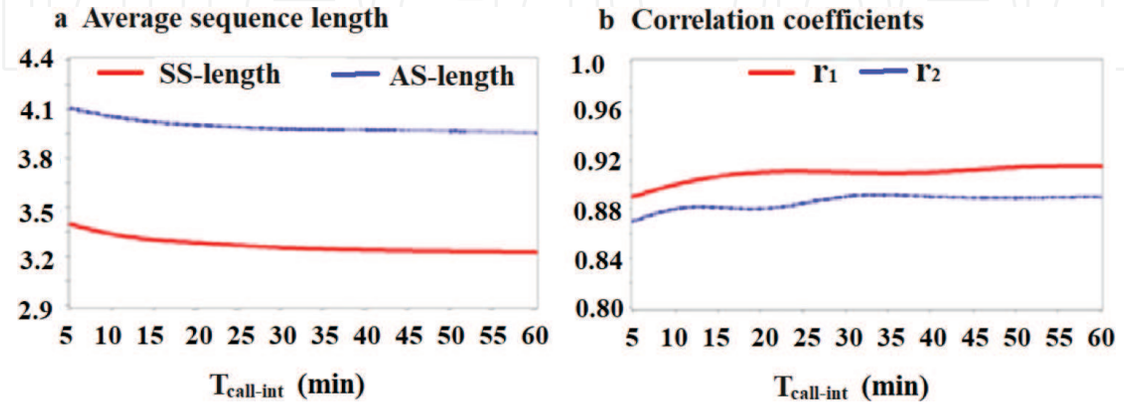

Figure 5. The relation between $T_{call\text{-}int}$ and the average sequence length (a) and the coefficients (b).

and *AS-length* continuously decrease without stopping at a certain value. In contrast, r_1 and r_2 fall into a steady state when $T_{max\text{-}boundary}$ passes to a certain value (e.g., 60 min) (see **Figure 6b**). This can be due to the possibility that the dismissed stop locations caused by the increase in $T_{max\text{-}boundary}$ are likely distributed randomly across various types of patterns, leading to the relative frequencies of these patterns remaining unchanged.

8.2. *Actual-dur*(*act*(l_i))

Figure 7 depicts the relation between the threshold *actual-dur*(*act*(l_i)) for work activities and the derived results. It shows that, when this parameter passes a certain point, e.g., 317 min specified in this study, the changes in *AS-length* as well as r_1 and r_2 disappear. This phenomenon can be explained by the binomial model used to estimate the call probability *CallP*(*user*, l_i). According to this model, when the *actual-dur*(*act*(l_i)) is longer, *CallP*(*user*, l_i) becomes larger. But the call probability eventually enters into a constant value of 1 at a certain point of the *actual-dur*(*act*(l_i)) (see **Figure 8**).

The above sensitivity analysis shows that, except $T_{max\text{-}boundary}$ that exhibits a certain level of influences on the average length of the sequences, a certain amount of changes in these

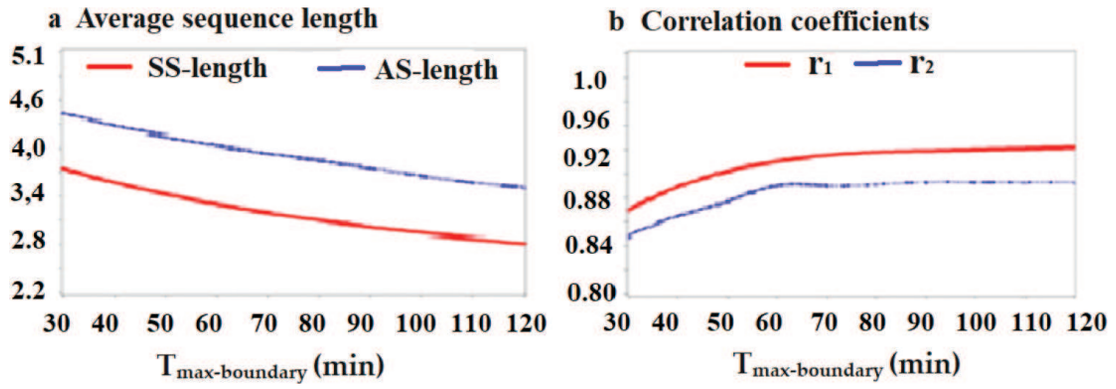

Figure 6. The relation between $T_{max\text{-}boundary}$ and the average sequence length (a) and coefficients (b).

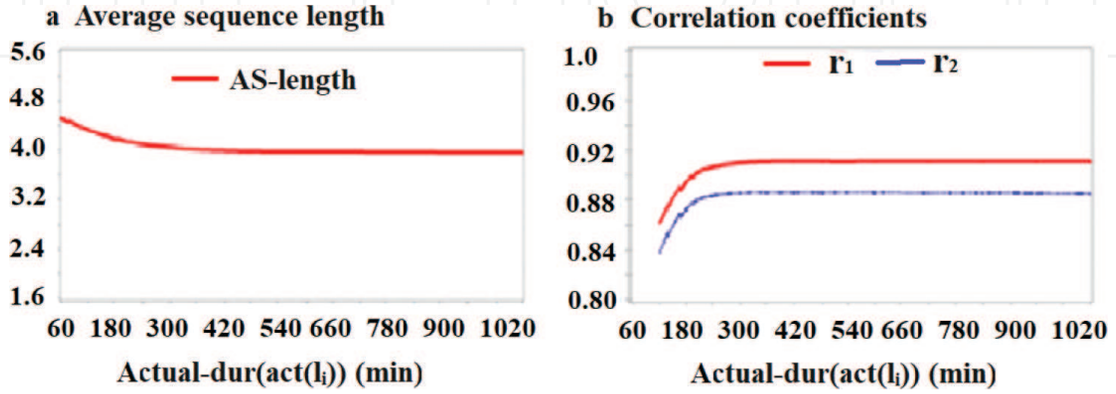

Figure 7. The relation between the *actual-dur*(*act*(l_i)) for work activities and the AS-length (a) and the coefficients (b).

parameters does not cause a substantial discrepancy in the sequence length and the profiles. This indicates that the profiles constructed from the mobile phone data are stable and consistent in characterizing workers' travel behavior; a minor change in these parameters will not result in significantly different outcomes.

9. Discussions

A number of areas can be enhanced in the future research. (1) In the prediction of home and work locations, a fixed work period (i.e., the time interval between 9 am and 18 pm on weekdays) is assumed. Under this assumption, people who work in night shifts or at weekends are ignored. The prediction of these two places could be improved by first deriving possible work regimes of the users. Flexible work periods can then be adopted corresponding to the different regimes. (2) During the process of stop location identification, rather than using general thresholds of 30 and 60 min for $T_{call\text{-}int}$ and $T_{max\text{-}boundary}$, these two parameters should be tailored to particular cells and individuals' travel speeds. For instance, smaller values than the current settings should be used for cells in a smaller size and for individuals with a higher travel speed (e.g., by car or by train). (3) With respect to the process of converting *stop-seqs* into *actual-seqs*, the estimation of the *actual-dur*(*act*(l_i)) should be separated among different social-economic groups, as the work duration for full-time workers is longer than that for part-time ones. Moreover, rather than using an identical *CallRate*(*user*) for all activities conducted by a user, the call rate could be differentiated across different activity types, as the likelihood of making calls may vary depending on the activity context. (4) When examining the validation potential of this method, travel surveys stemmed from different geographic areas than that of mobile phone data are used. However, as discussed in Section 7, deviations exist in terms of land use, transport networks and social-economic conditions of individuals across different regions and countries, and travel behavior is shaped by all of these factors. Thus, in the future, the proposed method must be validated using a real travel survey performed in the same or similar context to where the mobile phone data is recorded. Such surveys will provide more relevance to the current method by identifying the optimal parameters as well as assessing the results. (5) With the rapid

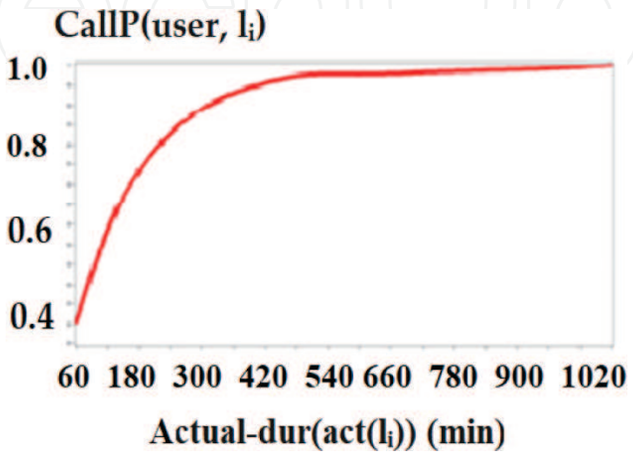

Figure 8. The relation between the *actual-dur*(*act*(l_i)) for work and *CallP*(*user*, l_i).

development of mobile application, mobile phone data will be collected not only when people make calls but when they use the application. The location data will thus reveal more activities and travel episodes, enabling more accurate travel sequence derivation based on the proposed approach. This will lead to an even more reliable activity-based model validation method as well as to improved travel behavior analysis in general.

10. Conclusions

The proposed method can be integrated into activity-based models at the stage of 'simulated travel sequences' (see **Figure 1**). Specifically, during this validation, the set of simulated travel sequences yielded from the activity-based models is compared with the set of the *actual-seqs* derived from the mobile phone data in the following two aspects: (1) The average number of location visits per day, that is, the average length of the sequences. (2) The sequential order of the activities, reflected by the correlation between the corresponding profiles (i.e., the *tour-profiles* or *day-profiles*) constructed from each set of the sequences. If a large difference in the average length of the sequences or a low coefficient between these profiles is found, it would suggest mismatches between the simulated results and the travel patterns represented by the call data. This thus signals possible problems and calls immediate action into the examination of the activity-based models, prior to the utilization of the model results for further traffic assignment and mobility-related analysis.

Apart from the initial goal of developing a new validation method, this study also designs a novel process for stop location identification and *actual-seqs* derivation. This process integrates daily activities and travel with call activities; both types of activities occur concurrently and are performed by the same individuals. This method presents a solution to the challenge that is pertinent to mobile phone data research and application in a variety of fields, for example, in urban planning and location-based services. Due to the event-driven nature of mobile phone data collection, the locations are recorded only when a user connects the GSM network. What the user is doing (e.g., traveling or doing activities) is not known. Moreover, the places, where the person has stayed but without calling, are also dismissed. The location update errors which result in wrong documentation of user's actual locations raise another issue on the data collection. The results from our case study suggest a decrease by 42% in the number of location visits per day, when *stop-seqs* are constructed from the raw phone data. This number increases by 22% when the dismissed places are interpolated and the complete *actual-seqs* are formed. Such scales of changes signify the importance of the integration between the existing research using mobile phone data and this process in this study.

Acknowledgements

The authors would like to acknowledge the support of the European Union through the project of DataSim. We also thank the Orange Data for Development (D4D) challenge committee for the provision of mobile phone data.

Author details

Feng Liu[1,2]*, Ziyou Gao[1], Bin Jia[1], Xuedong Yan[1], Davy Janssens[2] and Geert Wets[2]

*Address all correspondence to: feng.liu@uhasselt.be

1 MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing, China

2 Transportation Research Institute (IMOB), Hasselt University, Diepenbeek, Belgium

References

[1] Cui JX, Liu F, Hu J, Janssens D, Wets G, Cools M. Identifying mismatch between urban travel demand and transport network services using GPS data: A case study in the fast growing Chinese city of Harbin. Neurocomputing. 2016;**181**:4-18

[2] Cui JX, Liu F, Janssens D, An S, Wets G, Cools M. Detecting urban road network accessibility problems using taxi GPS data. Journal of Transport Geography. 2016;**51**:147-157

[3] Hartgen DT. Hubris or humility? Accuracy issues for the next 50 years of travel demand modeling. Transportation. 2013;**40**(6):1133-1157

[4] Shan J, Viña-Arias L, Ferreira J, Zegras C, González MC. Calling for validation, demonstrating the use of mobile phone data to validate integrated land use transportation models. In: Proceedings 7VCT. 2011. 2011. http://hdl.handle.net/2099/15544, ISBN: 978-972-96524-6-2, 978-989-97510-0-2

[5] Liu F, Janssens D, Cui JX, Wets G, Cools M. Characterizing activity sequences using profile hidden Markov models. Expert Systems with Applications. 2015;**42**(13):5705-5722

[6] Liu F, Cui JX, Janssens D, Wets G, Cools M. Semantic annotation of mobile phone data using machine learning algorithms. In: Smartphones from an Applied Research Perspective. Intech; 2017. DOI: 10.5772/intechopen.70255. https://www.intechopen.com/books/smartphones-from-an-applied-research-perspective/semantic-annotation-of-mobile-phone-data-using-machine-learning-algorithms

[7] Blondel VD, Esch M, Chan C, Clerot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C. Data for development: The D4D challenge on mobile phone data. Computer Science. 2012. https://www.researchgate.net/publication/231513146_Data_for_Development_the_D4D_Challenge_on_Mobile_Phone_Data

[8] Department of Transport. Key Results of the National Household Travel Survey – Final Report. Department of Transport, Pretoria, South Africa. 2005. Available from: http://www.arrivealive.co.za/pages.aspx?nc=household