Profiling workers' activity-travel behavior based on mobile phone data

Feng Liu[a], Davy Janssens[b], Geert Wets[b], Mario Cools[c]

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

[c] TLU+C (Transport, Logistique, Urbanisme, Conception) 1, Chemin des Chevreuils Bât B.52/3, 4000 Liège, Belgium

E-mail address: feng.liu@uhasselt.be (F. Liu), davy.janssens@uhasselt.be (D. Janssens), 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 models typically predict 24-hour activity-travel patterns for each individual in a study area. These patterns reflect the characteristics of the available transportation infrastructure and land-use system as well as individuals' lifestyles and needs. However, the lack of a reliable benchmark to evaluate the generated patterns has been a major concern. To address this issue, we explore the possibility of using 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 movements revealed by the call data and the real traces that the user has made, the daily trajectories are then transformed into travel sequences. Finally, all the inferred travel sequences are classified into typical activity-travel patterns which, in combination with their relative frequencies, define a profile. The established profile represents the activity-travel behavior in the study area, and thus can be used as a benchmark for the validation of the activity-based models.

By comparing the benchmark profiles derived from the call data with statistics that stem from 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.

1. Introduction

1.1 Micro-simulation model of travel behavior

The main premise of *activity-based micro-simulation models* is the treatment of travel behavior as a derived demand of activity participation. In this modeling paradigm, travel is generally analyzed through daily patterns of behavior related to and derived from the context of the land-use and transportation network in a study region and of the personal characteristics such as social-economic background, lifestyles, and needs of the individuals in the area (e.g. Axhausen & Gärling, 1992; Bhat & Koppelman, 1999; Davidson et al., 2007; Lemp et al., 2007). As such, the modeling system is calibrated using land-use and transportation network information as well as a dataset stemming from household travel surveys which document the full daily activity-travel sequences of individuals during one or a few days. All the input data are analyzed and translated into heuristic decision making strategies which represent the scheduling of activities and travel by the individuals (e.g. Arentze & Timmermans, 2004). Once established, these strategies are used as the probabilistic basis for a micro-simulation process, in which complete daily activity-travel sequences for each individual in the whole population of the region are synthesized, using Monte Carlo simulation.

The synthesized individual activity-travel sequences are afterwards aggregated into origin-destination (OD) matrix, i.e. *a matrix that represents the number of trips between all the different locations of the region*. This matrix, after being assigned to road network, can subsequently serve as input for travel analysis in the region, such as travel demand forecasting, emission estimates and evaluation of emerging effects caused by different transportation policy scenarios. Figure 1 illustrates the entire process of a micro-simulation model.

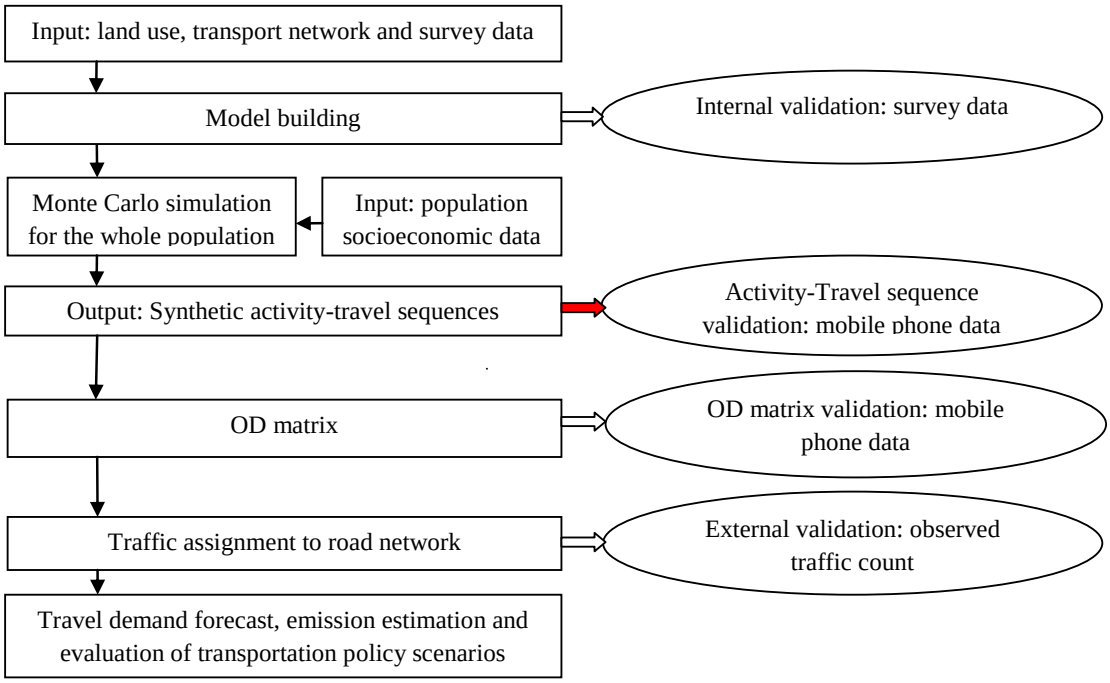

Figure 1. The entire process of a micro-simulation model

1.2. Problem statement

Despite comprehension and advancement of the activity-based modeling system, the lack of reliable data in sufficient size does not enable one to have a decent benchmark and evaluation criterion of 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 Figure 1 (e.g. Bellemans et al., 2010; Yagi and Mohammadian 2007; Yagi & Mohammadian 2010). The internal validation involves the comparison of the estimation results with expanded survey data which are not used in the training phase of the model but usually collected in the same survey period. 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 are required for an objective and up-to-date validation measure. In addition to this time limitation, the issue of budgetary constraints related to the financial cost associated with travel surveys, make it a challenge to collect samples that are sufficiently large to provide a good representation of the activity-travel behavior of a population. Moreover, travel surveys usually query information of only one or two days, to limit the negative effects associated to the respondent burden that is imposed by this type of surveys. This tends to obfuscate the less common activities which occur with a lower frequency (e.g. once a week or once a month), such as sports or telecommuting activities. These shortcomings have been well reported in the literature (e.g. Asakura & Hato, 2006; Cools et al., 2009; Wolf et al., 2001).

In contrast to the internal validation, the external validation consists of indirect evaluations of the model output at a later phase, i.e. traffic assignment stage (see Figure 1). The traffic volumes estimated by the model and assigned to transport network are compared against

commonly available external information sources, such as traffic counts collected by inductive loops detectors deployed on the road network.

However, this external validation process encompasses an aggregation step to compose the OD matrix which is assigned to the road network. Valuable information may be lost in this process. A major limitation that results from this loss of information is that positive outcomes of the comparison 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 identify the causes, thus limiting remedies to improve model construction. Despite such limitations, at the present, indirect evaluation is essentially the only option for model quality assessment, as no well-established methods are known for operating closer to the model itself. This is a problem that seriously hampers further model development and model application. Having useful and reliable benchmark and evaluation criteria for activity-based micro-simulation models thus is a major concern. Nonetheless, to a large extent, this aspect is neglected in currently available benchmarking standards.

The wide deployment of mobile phone devices provides a very promising source of information on measuring people's transfer phenomenon. Mobile phone data reflect up-to-date travel patterns on significantly large samples of population – in terms of both spatial coverage and temporal extension, making them a natural candidate for the analysis of activity-travel behavior. The importance of mobile phone data in traffic related researches has been manifested by extensive studies on the development and application of the data (e.g. Hansapalangkul et al., 2007; Liu et al., 2013; Ratti et al., 2006; Rose 2006; Steenbruggen et al., 2011). Especially, OD matrices have been constructed based on mobile phone data in a number of regions and countries (e.g. Calabrese et al., 2011; Sohn & Kim, 2008; White & Wells, 2002), and they can be used for travel demand analysis after being allocated to a specific road network. Besides, these matrices can also be utilized for the examination of ODs generated by the simulated travel sequences, as indicated in Figure 1. The feasibility of the benchmarking approach at the ODs level based on mobile phone data has been explored in a recent European project 'DATA science for SIMulating the era of electric vehicles', namely DataSim (http://www.datasim-fp7.eu/). However, while moving the validation process one step closer to the simulated travel sequences, the comparison with ODs still involves an extra procedure of the calculation of the OD matrices. Consequently, the validation is unable to provide a direct assessment on the simulated sequences themselves, and therefore still does not fully address the problems which are related to the external validation measures.

1.3. Research contributions

Extending the current research on the application of the massive mobile phone data in traffic demand analysis, and particularly addressing the above mentioned limitations in having reliable evaluation measures for travel behaviour simulation models, our study proposes a new approach which 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 be used to directly evaluate the sequences yielded from the simulation models by comparing it against the frequencies of the corresponding pattern classes obtained from the simulated sequences (see Figure 1). This comparison is done at the level of the generated activity-travel sequences, thus capable of

detecting problems that are directly caused by the model itself and providing immediate feedback for the enhancement of the model.

Compared with existing validation measures, this approach offers the following advantages. (i) It monitors current activity-travel behavior in a large proportion of population and provides a more representative and up-to-date validation measure. (ii) Through a long recording period of the mobile phone data, inter- and intra- personal variations of travel behavior as well as weekday/weekend and seasonal deviations can be more efficiently captured. (iii) It can offer immediate response to problems directly linked to the model system, allowing problems to be addressed at an earlier stage of the modeling process before they are propagated into further analyses. (iv) It aims at generating a novel measure for evaluating and benchmarking activity-based micro-simulation models, filling in the gap between the development of the comprehensive model system and the lack of a good and widely accepted evaluation procedure. (V) Apart from the above described technical aspects, the mobile phone data is a by-product of phone companies, requiring no extra cost for data collection, thus providing another appeal in terms of financial consideration.

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 outcome 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 as well as discretionary activities e.g. social visit, sports or going to restaurant (e.g. Arentze & Timmermans, 2004; Bradley and 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 the fixity of the work activity. In contrast, no such obvious periodicity is present in the case of non-workers (Spissu et al., 2009). 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. The representation of their daily sequences is described in Figure 2. In this representation, 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.

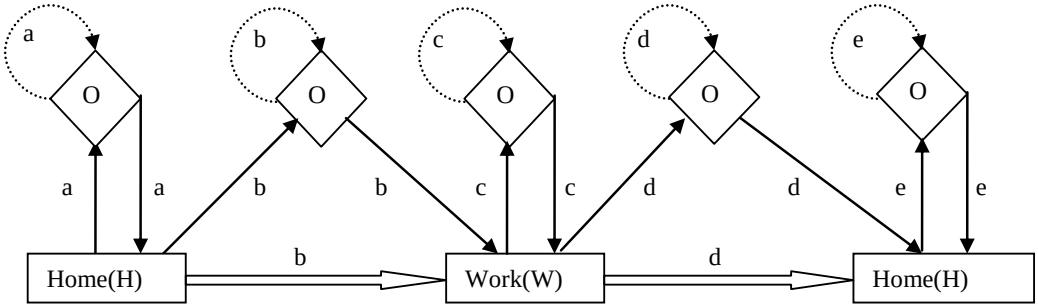

Figure 2. The representation of workers' activity-travel sequences
Note: Each 'rectangular' indicates the home or work location, while the 'diamond' represents a non-mandatory activity location. Each 'arrow' from a home, work or non-mandatory activity location to the other represents the related travel, 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, comprised of a chain of trips (locations) 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.*

All the above pattern classes (excluding H) are then merged in pair, leading to 81 combinations which represent daily sequences accommodating maximum 2 home-based tours. For instance, the combination of HWH and HOWH results in the sequence HWHOWH. In addition these pairwise combinations, sequences that contain more than 2 home-based tours, e.g. HWHWHWH, or those that have more than 2 work activity locations in a home-based tour, e.g. HWOWOWH, are each assigned into one additional category. By contrast, a daily sequence can also accommodate only a single home-based tour, e.g. HWH. All these scenarios lead to a total of 93 patterns which underlie workers' activity-travel behavior, and which are denoted 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 data was collected by a mobile phone company for billing and operational purposes. The 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 (Vincent et al., 2012). The data contain 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 of human travel behavior and for the drawing of 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 when this individual was connecting to the GSM network and the 'Antenna_id' as the cell area where he/she is located.

4. Construction of call location trajectories

A *raw_call_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 reached when using his/her phone on that day, and $l_i (1 \le i \le n)$ 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 $k_i (k_i \ge 1)$; 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_call_location_trajectory* can be represented as $l_1(T(l_1,1),T(l_1,2),...,T(l_1,k_1)) -> ... -> l_n(T(l_n,1),T(l_n,2),...,T(l_n,k_n))$, simplified as $l_1(T(1),T(2),...,T(k_1)) -> ... -> l_n(T(1),T(2),...,T(k_n))$.

Given the *raw_call_location_trajectories* constructed from the mobile phone data, the home and work locations are first predicted. This is followed by the identification of stop locations where activities are 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 (e.g. Becker et al., 2011; Calabrese et al., 2011), mainly based on the visited frequency of a location during a particular time period. However, different time windows have been specified in these studies, depending on the context of the study area. In this paper, 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 w*ork_start_time.* Secondly, the moment when the second peak of call activities start to appear in late afternoon is considered as the work end time, referred as w*ork_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 w*ork_end_time* and w*ork_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 w*ork_start_time* and w*ork_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 occasionally accessed places for home or work activities are regarded as a stop for non-mandatory activities. In addition, only individuals who work at different locations than their home location areas 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_call_location_trajectories* are either *stop locations* where people pursue activities, i.e. non-mandatory activities, or non-stop ones. Each of these non-stop locations could be either a *trip location* where the user is traveling, or a 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 to the locations which are neither home nor work locations in the *raw_call_location_trajectories* and which need to be differentiated between stop and non-stop visits, the identified home or work locations are also not constantly reached for activity purposes, some occurrences of these locations could be caused by the non-stop reasons. The necessity to identify stop location from non-stop ones can be illustrated with the call records of two typical users.

The trajectory from the first user identified as User265 on a Friday is $l_1(17:06pm,17:43pm)->l_2(17:51pm)->l_3(17:56pm,19:41pm)->l_4(21:55pm)$, where 4 locations are observed, with each lasting 37, 0, 105 and 0 minutes respectively. From this trajectory, a distinction needs to be made to identify stop visits from possible trip visits.

The location update errors can be demonstrated using the call location trajectory of a second user of User72, which is $l_1(13:21pm,20:11pm)->l_2(22:00pm)->l_3(22:02pm)->l_4(22:06pm)->l_2(22:21pm,23:12pm)$. This user has 5 location updates, with the *call_location_duration* as 410, 0, 0, 0 and 51 minutes respectively. However, the time interval between the first visit and the second one to location l_2 is only 21 minutes. The temporary interruption of l_2 by the extra locations l_3 and l_4 in such a short interval 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 22:02pm and 22:06pm 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 visit from a momentary access due to traveling or due to location update errors. In their approach, the interval between the first login of the location l_i under investigation and that of the next one 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 minutes 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 minutes, shorter than this 60-minute limit, suggesting that the location l_1 would be for trip purposes. This may be true if this individual has made a long trip of at least 37 minutes 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, the location l_1 is misclassified by the existing method.

In order to accommodate all the possible stop locations, we propose a new approach consisting of the following steps. (1) For each location visit l_i, the *call_location_duration* is first examined. If it is longer than a certain time limit, denoted as $T_{call\_location\_duration}$, this location is considered as a stop location. (ii) Otherwise, if the condition does not hold e.g. when only a single call being made at the location, 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 the 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 $T_{\max imum\_time\_boundary}$, the location l_i is perceived as a stop visit. (iii) When the location is in the first or last position of a trajectory and the *call_location_duration* is shorter than $T_{call\_location\_duration}$, there is no sufficient information to estimate the maximum possible time for this visit. Thus, all the distinct locations where the user has stayed at least once for carrying an activity, are collected. These locations are considered as potential stop locations that are on the individual's daily activity agenda and that are visited routinely or once in a while. If the first or last visit of a day is to

these locations, it is assumed to be a stop for activity purposes. On the contrary, if this visit is to the place where the individual has not been observed doing activities, it is considered as a passing-by place or being recorded as a localization error and therefore removed.

To exemplify the procedure, we return to the examples of User265 and User72. For User265, based on the parameters of $T_{call\_location\_duration}$ and $T_{\max imum\_time\_boundary}$ which are set up as 30 and 60 minutes respectively in our experiment, l_1 and l_3 are predicted as stop locations, while l_2 is as a trip location due to the short *call_location_duration* (0 min) and *maximum_time_boundary* (13 min). Although only a single call is made at the last location l_4, knowledge gathered from other days has shown that this location has been a regular activity place for this individual. Consequently, this location is labeled as a stop visit. The finally obtained trajectory of stop locations for this user is $l_1 -> l_3 -> l_4$. For User72 this would imply that the locations l_3 and l_4 are deleted as a result of the identification process, and that the divided parts of location l_2 are merged together into a stop location. 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 reached by an individual on a day are formed into a *call_stop_location_trajectory*. Each location l_i in these 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 location trajectories

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 *call_stop_location_trajectories* into the sequences that mirror the real picture of people's activity-travel behavior.

During this transformation, we first derive for all the users 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 individuals make at least one call when they visit the location and which thus indicates to what extent their call records reveal their actual movement. Given a real daily activity-travel sequence, various *call_stop_location_trajectories* could be possibly observed from call data. Next, the probability under which a certain *call_stop_location_trajectory* is generated from the original travel sequence is calculated based on the call probabilities at these locations in the travel

sequence. Finally, given the observed frequencies of the *call_stop_location_trajectories*, 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 an individual 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(individual, day),* and the time span of the day (measured in minutes), denoted as *time_span(day),* as follows.

$$call\_\text{int}ervene(individual) = \frac{\sum_{day} time\_span(day)}{\sum_{day} total\_number\_calls(individual, day)}$$

The average call intervene across all the users is obtained as

$$average\_call\_\text{int}ervene = \frac{\sum_{individual} call\_\text{int}ervene(individual)}{total\_number\_of\_individuals}.$$

Based on the *average_call_intervene,* the variable of *call_rate* which describes the probability that the individuals makes calls each minute, can be calculated as

$$average\_call\_rate = \frac{1}{average\_call\_\text{int}ervene}$$

The other important variable, defined as $actual\_location\_duration(individual, l_i)$, specifies the actual activity duration *(in minutes)* at a location l_i for an individual. This variable is simplified by the average duration over all individuals across all locations with the same activity purposes as follows.

$$average\_actual\_location\_duration(activity) = \frac{\sum_{individual} \sum_{l_i=activity} actual\_locationy\_duration(individual, l_i)}{\sum_{individual} total\_number\_visit(individual, activity)}$$

Where the *total_number_visit(individual,activity*) represents the total number of actual visits by the individual to the locations with the particular *activity* purposes, such as home, work or non-mandatory activities.

5.2 Call probability at a location

Given a user's call rate and the duration of a location l_i where the individual has actual spent, the probability of making at least a call during the entire period of the visit to the location, defined as $CallP(l_i)$, can be estimated in the following manner. The location duration is first divided into episodes with an equal interval referred as *episode_length*, e.g. 5 min, each of which can be regarded as an experiment. 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, $CallP(l_i)$ can then be modeled as 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 $CallP(l_i)$ as the probability of having at least one success (making calls) over the total number of experiments, in this case, over the total location duration.

In this study, the previously derived two variables including the *average_call_rate* and the *average_actual_location_duration(activity)* are used as the approximation of the call rate for each individual and the duration for a location with a particular activity purpose, respectively. The probability $CallP(l_i)$ is then obtained as follows.

$$CallP(l_i) = CallP(activity) =$$

$$1-\{1-episode\_length \times average\_call\_rate\}^{average\_actual\_location\_duration(activity)/episode\_length}$$

5.3 Sequence conversion probability

After the probability of making calls at a location of home, work or non-mandatory activities is 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. The sequence $l_1 -> l_2 -> \ldots -> l_n$ is defined as the *actual_travel_sequence*, and the call probability at each location l_i as $CallP(l_i)$. In constrast, $\overline{CallP(l_i)}$ is used to denote the probability that no calls are made at location l_i, $\overline{CallP(l_i)} = 1 - CallP(l_i)$. Based on these probabilities, the likelihood of various *call_stop_location_trajectories,* that could be observed from the *actual_travel_sequence*, defined as *ConversionP*, can be calculated as follows. The probability that the original full travel sequence can be revealed by the call records is

$$\text{ConversionP}(actual\_travel\_sequence, call\_stop\_location\_trajectory)$$

$$= \text{ConversionP}(l_1 -> l_2 -> \ldots -> l_n, l -> l_2 -> \ldots -> l_n) = \prod_{i=1}^{n} callP(l_i). \quad (1)$$

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

$$\text{ConversionP}(l_1 -> l_2 -> \ldots -> l_n, l -> \ldots l_{i-1} -> l_{j+1} \ldots -> l_n)$$

$$= \prod_{m=1}^{m-1} callP(l_m) \times \prod_{m=i}^{j} \overline{callP(l_m)} \times \prod_{m=j+1}^{n} callP(l_m). \quad (1)$$

Where we assume that locations from x_i to x_j ($i \le j$) are missing since no phone communications have been made during the visits to these locations.

Suppose that the probabilities to make at least one call at the locations of home, work and non-mandatory activities are 0.805, 0.903 and 0.424, respectively. For the sequence of HWOH which represents the actual activity-travel behavior of a user identified as User121, there could be various location traces generated by this original travel sequence under certain

probabilities. For instance, the possibilities to emanate trajectories HWOH, HWH and H are as follows:

$$ConversionP(HWOH,HWOH) = CallP(H) \times CallP(W) \times CallP(O) \times CallP(H) = 0.248$$

$$ConversionP(HWOH,WH) = \overline{CallP(H)} \times CallP(W) \times \overline{CallP(O)} \times CallP(H) = 0.082$$

$$ConversionP(HWOH,H) = CallP(H) \times \overline{CallP(W)} \times \overline{CallP(O)} \times CallP(H) +$$
$$2 \times CallP(H) \times \overline{CallP(W)} \times \overline{CallP(O)} \times \overline{CallP(H)} = 0.054$$

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 *call_stop_location_trajectories* $s_1, s_2, \cdots s_m$ are constructed from a user's call records, sorted by the length of these sequences, i.e. $length(s_1) \geq length(s_2) \geq \ldots \geq length(s_m)$. Let the frequencies of these observed trajectories as $y_1, y_2, \cdots y_k$ respectively; the original occurrences of the corresponding travel sequences, denoted as $x_1, x_2, \cdots x_k$, can be estimated by the following linear equation.

$$x_1 \times \mathrm{ConversionP}(s_1, s_1) = y_1$$

$$x_1 \times \mathrm{ConversionP}(s_1, s_2) + x_2 \times \mathrm{ConversionP}(s_2, s_2) = y_2$$

...

$$x_1 \times \mathrm{ConversionP}(s_1, s_k) + x_2 \times \mathrm{ConversionP}(s_2, s_k) \ldots + x_k \times \mathrm{ConversionP}(s_k, s_k) = y_k$$

From the above equation, the variables $x_1, x_2, \cdots x_k$ can be solved as follows.

$$x_1 = \frac{y_1}{ConversionP(s_1, s_1)}$$

$$x_2 = \frac{y_2 - x_1 \times ConversionP(s_1, s_2)}{ConversionP(s_2, s_2)}$$

...

$$x_k = \frac{y_k - x_1 \times ConversionP(s_1, s_k) - \ldots - x_{k-1} \times ConversionP(s_{k-1}, s_k)}{ConversionP(s_k, s_k)}$$

In the case of the User121, apart from a daily sequence of HWOH, four other *call_stop_location_trajectories* are revealed by this user's call records, including WH, OH, W and H, with the occurrences as 3, 2, 1 and 3 respectively. The original frequencies of these sequences, i.e. $x_1 - x_5$, can be solved in the following equation:

$$x_1 \times ConversionP(HWOH, HWOH) = 1$$

$$x_1 \times ConversionP(HWOH, WH) + x_2 \times ConversionP(WH, WH) = 3$$

$$x_1 \times ConversionP(HWOH, OH) + x_3 \times ConversionP(OH, OH) = 2$$

$$x_1 \times ConversionP(HWOH, W) + x_2 \times ConversionP(WH, W) + x_4 \times ConversionP(W, W) = 1$$

$$x_1 \times ConversionP(HWOH, H) + x_2 \times ConversionP(WH, H) +$$
$$x_3 \times ConversionP(OH, H) + x_5 \times ConversionP(H, H) = 3$$

From this equation, we obtain $x_1 = 4.03, x_2 = 3.67, x_3 = 5.78, x_4 = 0.30, x_5 = -0.23 \cdot$

The obtained results then undergo two further processes. First, a zero is assigned to the variables which have negative values, e.g. x_5 for the sequence H in the above case. The

negative frequency for a travel sequence suggests that the actual occurrence probability of the potential travel sequence could be very low, and that the corresponding observed identical call location trajectory, e.g. H in this example, is likely to be generated by other longer travel sequences, such as HWOH, WH and OH. These negative frequencies are thus dismissed by setting the corresponding variables to zero.

The second process is to normalize the obtained results for each individual, such that the total frequency of the derived travel sequences amounts to the observed sum of the call location trajectories. For the User121, the sum of the observed trajectories is 10, but that of the derived ones reaches 13.78, a ratio of these two numbers is used as the scaling factor, leading to the final solution as $x_1 = 2.92, x_2 = 2.67, x_3 = 4.19, x_4 = 0.22, x_5 = 0$.

From the call location trajectories for this user, a total of 10, 5 and 3 location visits for home, work and non-mandatory activity purposes respectively, have been observed; while for the derived travel sequences, the corresponding number changes to 12.7, 5.8 and 7.1, respectively. The ratio of the total locations between these two types of sequences is 0.79, 0.86 and 0.42 for these three activity classes respectively, close to the call location probabilities which are initially used for this derivation process. This further demonstrates that the derived travel sequences not only maintain the sequential order of the activity locations which are imbedded in the call location trajectories, but that they also preserve the call probabilities at individual locations as a whole.

It can be noted that during the entire procedure of seeking the solutions, we assume that the original travel sequences could only occur within the space of observed call location trajectories $S = \{s_1, s_2, \ldots s_m\}$. In theory, however, there could be a chance that an observed call location trajectory is produced by many other potential travel sequences, rendering the solution space to become infinite. However, for a possible travel sequence s_p which is not in the observed sequence space S, i.e. the frequency of the corresponding call location trajectory y_p being zero, a value less than or equal to zero would be obtained as the actual frequency x_p of this travel sequence. This implies that the positive frequencies of a travel sequence can only be found if this sequence is within the limited space S. For instance, for the User121, if the potential travel sequence is longer than any trajectory in S, i.e. $length(s_p) \geq length(s_1)$, assume $s_p = \text{HWOWH}$, we obtain the following equation: $x_p \times \text{ConversionP}(\text{HWOWH}, \text{HWOWH}) = 0$. From this equation, we have $x_p = 0$. Otherwise, if the length of this travel sequence is shorter than certain observed trajectories in S, e.g. $s_p = \text{HWO}$, we have $x_1 \times \text{ConversionP}(\text{HWOH}, \text{HWO}) + x_p \times \text{ConversionP}(\text{HWO}, \text{HWO}) = 0$, from which a value of $x_p < 0$ would be derived.

6 Classification

All the obtained *call_stop_location_trajectories* and *actual_travel_sequences* are subsequently classified according to the *home-based-tour-classification* and *daily-sequence-classification*, which have been previously established for workers' activity-travel behavior. During this classification, a home location H is added at 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 is used to measure the relation of the corresponding profiles between these two types of sequences. The correlation coefficient, denoted by *r*, is a measure of the strength of linear relationship between two variables. It takes on values ranging between 1 and -1, with 1 indicating a perfect positive linear relationship: as one variable increases in its values, the other variable increases as well. The closer the value is to 1, the stronger the relationship is.

For two matrices, denoted as *A* and *B*, and let *d* as the total number of the matrix elements, the *r* is computed as follows:

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

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

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 *call_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 *call_stop_location_trajectories*

7.1.1 *Work_start_time* and *work_end_time*

Figure 3 describes the distribution of the frequency of calls made in each hour of the day during 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 start to occur. These two temporal points are chosen as the w*ork_start_time* and w*ork_end_time*, respectively.

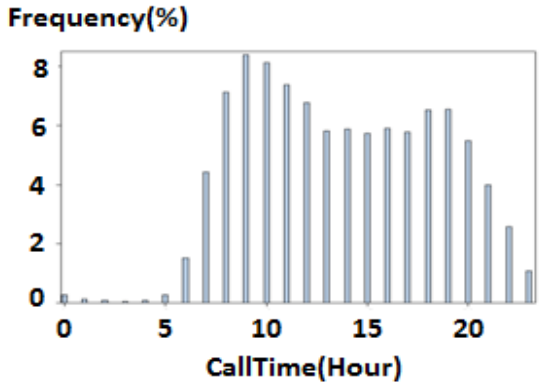

Figure 3. The distribution of the time of calls

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, and 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 17:59pm at least two weekdays per week. By contrast, those who work at night shifts or at weekends, or who work less than two days a week, or who make few calls at work, are not identified as workers.

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_call_location_trajectories* constructed for further analysis.

7.1.2 *Call_location_duration* and *maximum_time_boundary*

Two parameters characterize the stop location identification process. The first one, *call_location_duration,* determines the time limit above which the location is defined as a stop. This parameter depends on the minimum time required to possibly pursue an activity as well as the time period needed for traversing across the area. The other parameter, *maximum_time_boundary,* measures the time interval between the last call time at the previous location and the first call time at the next location, relative to the current place under investigation. Similar to *call_location_duration,* this parameter must be longer than a combination of the possible activity duration and the travel time needed going from the previous cell, passing the current one, and to the next area. In addition, it should also be able to detect location update errors which usually occur in a short time interval.

In this experiment, $T_{call\_location\_duration}$ and $T_{\max imum\_time\_boundary}$ are set as 30 minutes and 60 minutes respectively. Under these thresholds, 40.3% of all locations from the *raw_call_location_trajectories* are removed; the remaining locations in these sequences form the set of *call_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 minutes, 67.6% of all the raw call locations are dismissed, with the average length of the retained sequences as 2.33.

7.2 Conversion from *call_stop_location_trajectories* into *actual_travel_sequences*

7.2.1 *Average_call_intervene* and a*verage_call_rate*

When estimating these two variable values, all the calls made by the identified workers, including the ones that may be made on a road or have false location IDs due to localization errors, are all considered. This results in the *average_call_intervene* as 192 min over a full day of 24 h. However, as demonstrated in Figure 3, the occurrence of calls is 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. This reduces the call intervene to 137 min; accordingly, the *average_call_rate* is 0.0073.
In an existing study (Calabrese et al., 2011), however, a 260 min of call intervene is derived; this difference 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 *Actual_location_duration*
This variable value is approximated by a real activity-travel behavior survey that was conducted in Belgium which will be described later. From this survey, the average location duration $average\_actual\_location\_duration(activity)$ are 222, 317 and 75 min for home, work and non-mandatory activity locations, respectively.

7.2.3 E*pisode_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 an episode should be independent of that in its 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 duration for all voice calls between each two cells in Ivory Coast. 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 probability at a location is derived, and it is 0.805, 0.903 and 0.424 for home, work and non-mandatory activity locations, respectively. These obtained probabilities, combined with the observed frequencies of the *call_stop_location_trajectories* for each user, lead to the prediction of the number of the *actual_travel_sequences* for the individual, using the method described in Section 5.3 and 5.4.

8 Comparison of results from mobile phone data with real activity-travel diary data
To illustrate the practical ability of our approach to really serve as a benchmark method, 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, have been adopted for this purpose. The authors acknowledge that the real travel behavior in Ivory Coast most likely is considerably different to 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 Travel survey in South African

The South Africa National Household Travel Survey (NHTS) was the first national survey of travel habits of individual 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 (http://www.arrivealive.co.za/pages.aspx?nc=household).

The information recorded by the survey includes the travel time to various public transport services, e.g. trains and buses, as well as to activity services, 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 activity location visited by a worker on a weekday is estimated at between 3.46 and 4.06 (http://www.arrivealive.co.za/document/household.pdf).

8.2 Travel survey in Belgium

Despite the relative geographic proximity between South Africa and Ivory Coast, the information on the NHTS is nevertheless limited. Moreover, 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 **S**trategic **B**asic **R**esearch project on transportation modeling and simulation, and it was conducted on 2500 households between 2006 and 2007 in Belgium. In the 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' on May, 9th, 2006. Only the variables that are relevant for the current study are presented in this table; a more detailed variable list and elaboration on this survey can be found in (Cools et al., 2009).

TABLE 2. 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	17:30:00	34349	34337	Home

From the dataset, the diaries on weekdays from 372 individuals who work at least two days a week 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 travel is not the last movement of a day. Otherwise, for the last trip, the activity end time at the travel destination is unknown. Another unknown factor is the activity start time at the origin of the first travel of a day. These two times are thus approximated by 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). The average of all the obtained duration at locations with an identical activity motivation over all the individuals is stored in the variable *average_actual_location_duration* 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_call_location_trajectories*, *call_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	RCLT	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_call_location_trajectories (RCLT), call_stop_location_trajectories (CSLT), actual_travel_sequences (ATS), NHTS and SBO surveys, respectively.

It was noted from Table 3 that the average length of sequences first drops from initial 5.69 for the *raw_call_location_trajectories* to 3.3 for the *call_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 *call_stop_location_trajectories*, the *actual_travel_sequences* and the SBO diaries, respectively. The differences in the percentages of corresponding pattern classes between these each two types of sequences are also listed.

Table 4. Home_based_tour_profile (%)[a]

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

[a] The columns from left to right represent the typical patterns, the call_stop_location_trajectories (CSLT), the actual_travel_sequences (ATS), the differences between ATS and CSLT, the SBO diaries (SBO), the differences between ATS and SBO, and the differences between CSLT and SBO, respectively.

Table 4 indicates that, when the *call_stop_location_trajectories* are converted into the *actual_travel_sequences*, the percentage of shorter patterns, e.g. H and HWH, increases; while that of longer patterns, e.g. HWOWOH, decreases. During this sequence conversion process, an observed call location trajectory is expected to be generated not only from an travel sequence that is identical to this observed trajectory, but more likely from a sequence that is longer than this observed one due to the fact that not every visited location is exposed by the mobile phone data. For instance, although 9.0% of the total call locations trajectories belong to the pattern of H, only 4.4% is estimated to be the days when the individuals do not make any trips but staying at home. The remaining 4.6% is probably generated from other longer travel sequences where the missing locations are as a result of the nature of the mobile phone data.

Another feature in the conversion process is that, 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 call location trajectories that amount to the observed frequency of the trajectories. This can be further illustrated by the pattern HOH. Although this pattern is as short as HWH, the probability at a non-mandatory activity location O, e.g. 0.424 in this experiment, is the lowest among all the three activity types. This leads to a prediction of high frequency of this pattern for the derived travel sequences.

When the patterns obtained from the derived travel sequences are compared with the ones drawn from the SBO diaries, it was observed that the major contrast resides in the difference between the group of short sequences and the other group accommodating long patterns. The SBO data has higher frequencies in short sequences while lower occurrences for long patterns, than the derived travel sequences. This tendency remains when the SBO data is compared with the *call_stop_location_trajectories*. While apart from the likelihood that people in Belgium may conduct less activities on average than in Ivory Coast, this also demonstrates the possibility that the diaries under-represent people's activity-travel behavior, especially for short period of activities. The shortcoming has been well documented in literatures (e.g. Cools et al., 2009).

8.5 Daily_sequence_profile

Figure 4(a) depicts the correlation between the relative frequency of each pattern class in the daily_sequence_profile obtained from the *call_stop_location_trajectories* and the *actual_travel_sequences*. It was noted that, the majority pattern classes follow a similar distribution in relative frequencies between these types of sequences. The few outliers can be divided into two groups: the group of HWH, H and HWHWH which are 14.1%, 5.9% and 1.4% higher for the *call_stop_location_trajectories*, and the other group consisting of HOH, the patterns with more than 2 home-based tours, and HOWOH, which show a 3.7%, 2.5% and 2.1% higher frequency for the *actual_travel_sequences*, respectively. This further demonstrates that, compared to the *call_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. non-mandatory activity locations O. In contrast, a lower percentage is anticipated for short patterns and for patterns containing locations with high call probabilities, e.g. work places W, after the sequence conversion process.

In Figure 4(b) which describes the correlation between the daily_sequence_profiles obtained from the *actual_travel_sequences* and the SBO diaries, we found that most patterns have a moderately higher frequency for the *actual_travel_sequences* than the SBO data. However, a few outliers show remarkably higher occurrences for the SBO diaries, e.g. HWOH and HWHOH accounting for a 7.3% and 7.1% higher percentage, respectively. It suggests that compared to Ivory Coast, people in Belgium may carry out more non-mandatory activities on the way from work back to home as well as in the evening period after arriving at home. In addition, 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 *call_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 characters of travel behavior.

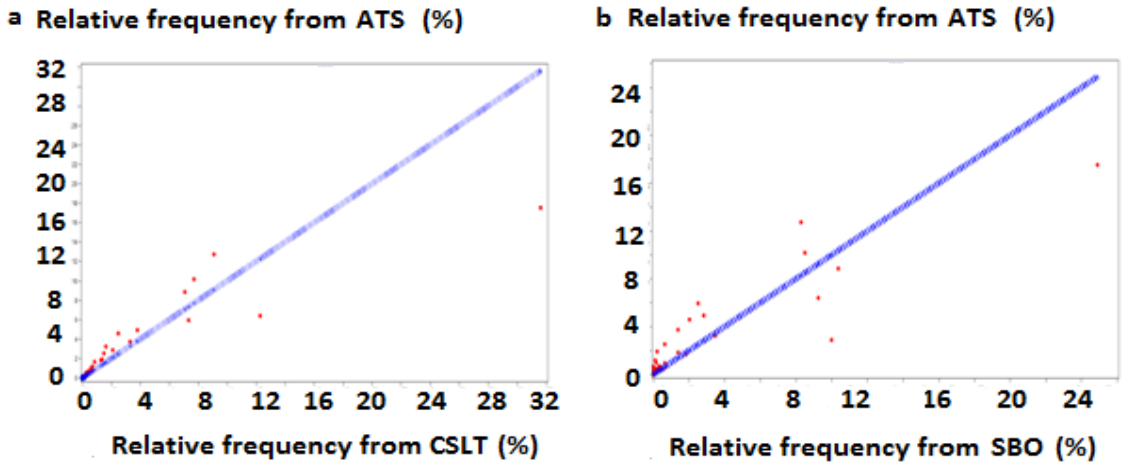

Figure 4. Correlation between the relative frequency of each corresponding pattern class
Note: x- and y-axis represent the relative frequency of each corresponding pattern class obtained from the call_stop_location_trajectories (CSLT) and the actual_travel_sequences (ATS) (a), and the SBO diaries and the actual_travel_sequences (b). The line of y=x is also presented as a reference line.

The correlation r between the *call_stop_location_trajectories* and the *actual_travel_sequences* as well as between the *actual_travel_sequences* and the SBO data is 0.91 and 0.89, respectively. The high correlation shows that the profile derived from the estimated travel sequences has an overall close relationship to that obtained from the call stop location trajectories, and in the meantime the profile of the travel sequences also accounts for the deviation in frequencies for each particular pattern which are caused by the discrepancy between the call behavior and the actual activity-travel behavior. In addition, the derived profile also resembles the frequency distribution of travel sequences from a real travel behavior survey. These results suggest the derived profile of travel sequences can properly represent workers' travel behavior in a studied area, and therefore capable of being used to validate the simulated sequences generated from travel behavior 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 described in the introduction, 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 the activity-travel behavior profiling approach based on the mobile phone data would bring even better results to current experimental outcome.

9. Sensitivity analysis

Throughout the profiling process, several parameters including *call_location_duration*, *maximum_time_boundary* and *actual_location_duration*, have been defined. This prompts to have a final investigation into how the thresholds of these parameters affect the predicted results, including the average length of *call_stop_location_trajectories* and the *actual_travel_sequences*, referred as CSLT_*length* and *ATS_length* respectively, as well as the coefficients between the *call_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 Call_location_duration and maximum_time_boundary

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

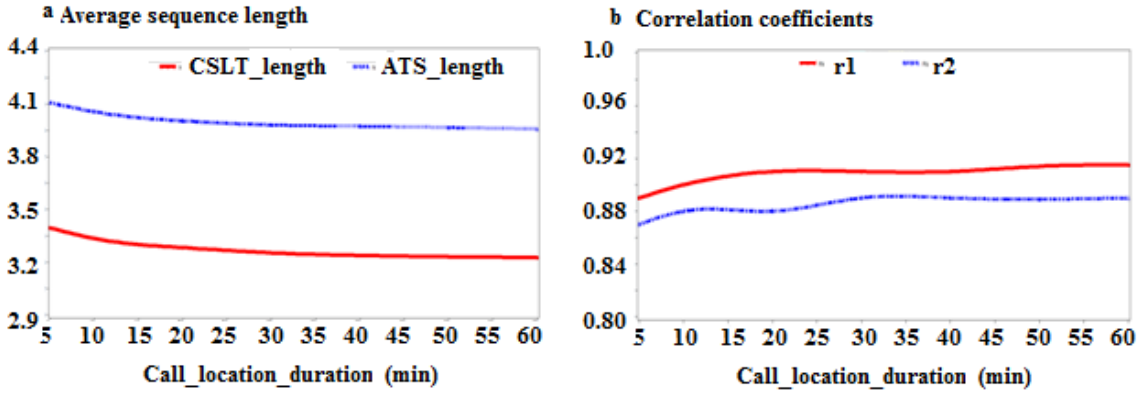

Figure 5. Correlation between the threshold of call_location_duration and the results
Note: x-axis stands for the threshold of call_location_duration, and y-axis for the sequence length of CSLT_*length* and *ATS_length* respectively (a) and the coefficients *r1* and *r2* respectively (b).

Figure 6(a) and 6(b) show how the results evolve with the threshold $T_{\max imum\_time\_boundary}$ for the parameter of *maximum_time_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 Figure 6(a). However, this does not bring about the same amount of change to the coefficients; especially when $T_{\max imum\_time\_boundary}$ increases to a certain value, e.g. 60 minutes adopted in our experiment, both *r1* and *r2* develop into a stable level. This suggests that, although the number of disclosed stop locations diminishes as this duration limit becomes stricter, the disregarded potential stop locations are likely distributed randomly across various types of pattern classes. As a result, the coefficients which reflect the relative frequency of these patterns remain almost the same, regardless of the minor changes that could arise from these parameter settings.

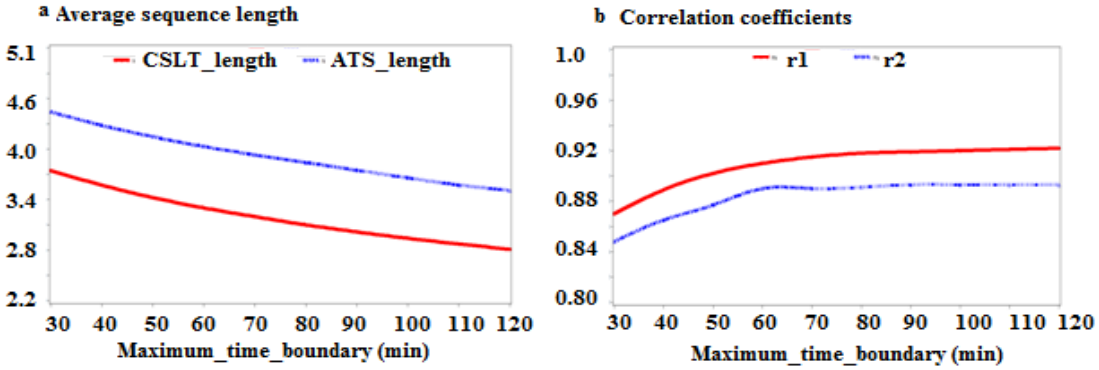

Figure 6. Correlation between the threshold of maximum_time_boundary and the results
Note: x-axis stands for the threshold of maximum_time_boundary, and y-axis for the sequence length of CSLT_length and ATS_length respectively (a) and the coefficients r1 and r2 respectively (b).

9.2 Actual_location_duration

Figures 7 describes the relation between the parameter *actual_location_duration* for work activities and the estimated results. It indicates that, as this duration becomes longer, the ATS_length2 for the derived travel sequences decreases while r1 and r2 increases, but these changes disappear when this duration pass a certain point, e.g. 240 minutes.

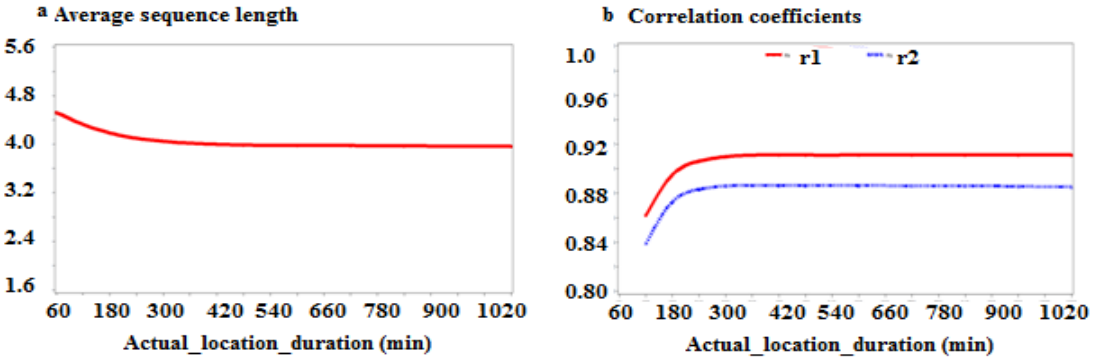

Figure 7. Correlation between the actual_location_duration for work activities and the derived results
Note: x-axis stands for the actual_location_duration for work activities, and y-axis for the sequence length of ATS_length (a) and the coefficients r1 and r2 (b).

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 $CallP(l_i)$ becomes higher, as demonstrated by Figure 8(a). However, the amount of increases in the probabilities as the activity duration extends e.g. one hour longer, is not evenly distributed, which can be manifested in Figure 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 Figure 7.

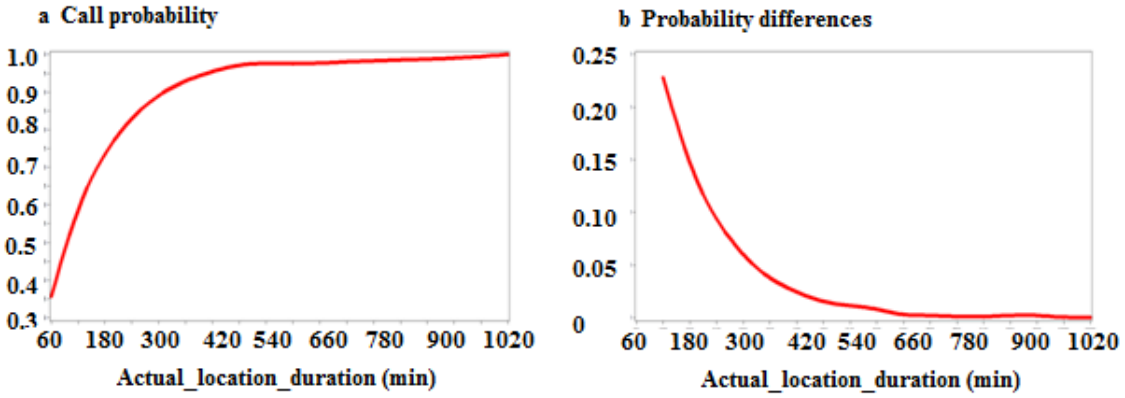

Figure 8. Correlation between the actual_location_duration and the call probability at a location
Note: x-axis stands for the actual_location_duration for work activities, and y-axis for the call probability at a location (a) and for the difference between the probability obtained from the corresponding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

All these above analysis shows that, except that the increase in $T_{\max imum\_time\_boundary}$ reduces the number of identified stop locations, a certain amount of changes in these parameters do not incur a significant deviation in the results of both the average length of sequences as well as the profiles. This suggests that the profiles built upon the mobile phone data are stable and consistent in revealing people's activity-travel behavior; a minor change in these parameters that are required in the profiling process will not lead to a substantially different outcome.

10. Conclusions and discussion

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 simulated activity-travel sequences yielded from micro-simulation models of travel behaviour. The advantage of using this method is that it does not depend on conventional diaries, the data requirement is fairly simple and its collection cost is low. More importantly, the massive mobile phone data monitors current activity-travel behavior in a large proportion of population expanding over a long time period, the profile derived from the data is thus capable of providing a more representative and objective validation measure.

Experiments on this approach by using data collected from people's natural mobile phone usage have demonstrated an overall high correlation coefficient between the profiles derived from the observed call location trajectories as well as from the derived travel sequences. The

relative frequency of each corresponding pattern class between these two profiles, however, shows a certain level of differences, a reflection of the deviation between the movement revealed by the call behavior and the real path that the individuals have experienced. In addition, the derived travel sequences also show reasonable outcome when they are compared to the statistics drawn from real travel surveys conducted in South Africa and Belgium, respectively. Furthermore, the examination into this method's sensitivity demonstrates its consistence and stability in drawing the real picture of activity-travel behavior over various parameter settings.

Beyond the initial goal of building a new measure for travel behavior simulation models, the proposed method for stop location identification and subsequent actual travel sequences derivation provides a broad use for the application of the massive mobile phone data. For instance, in the process of building OD matrices (Calabrese et al., 2011), only the stop locations revealed by the phone data are used; places where no calls are made are thus ignored. The results in our experiments suggest of an average of 21.8% increase from the initially obtained call stop locations to the derived complete location visits. Consequently, the OD matrices based on the mobile phone data reflect only a part of the whole picture of people's transfer phenomena, as acknowledged by the authors of the study. Based on our method, the real travel sequences could be derived first and a more accurate OD matrix could be anticipated.

The proposed approach can also be adopted for the characterization of non-workers' travel behavior. No work activities dominate the individuals' activity-travel sequences, but more home-based tours for non-mandatory activity purposes could be considered.

Nevertheless, despite the promising experiment results, there are still certain areas which need to be improved in the future research. First, when calculating the call probability at a location, we simply use a universal call rate which is derived from the mobile phone data across all types of activity locations and all individuals. But people communicate with others not at a same pace, and they may also call at different frequencies depending on what they are doing. Like the call rate, the use of an average actual location duration for each activity purpose across all users leaves a second possibility for improvement, as the activity duration across different individuals is likely to differ. The proposed method will be undoubtedly strengthened if both the call rate and the activity duration is considered at individual level and across each category of activity locations. Third, the method to identify home and work places could also be enhanced through machine learning techniques, as explored by a recent study (Liu et al., 2013).

While being faced with the challenge of acquiring both the mobile phone data and the real travel survey from a same or similar study region, in this study we use the travel surveys which are conducted in different environments than 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.

11. References

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.

Axhausen, K., & Gärling, T. (1992). Activity-based approaches to travel analysis: conceptual frameworks, models and research problems. Transport Reviews, 12, 324-341.

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, 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, 2175, 111-119.

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

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.

Cools, M., Moons, E., Bellemans, T., Janssens, D. & Wets G. (2009). Surveying activity-travel behavior in Flanders: Assessing the impact of the survey design. In C. Macharis, and L. Turcksin (eds.), 369 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.

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.

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, F., Janssens, D., Wets, G. & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Applications. http://dx.doi.org/10.1016/j.eswa.2012.12.100.

Ratti, C., Pulselli, R. M., Williams, S., & Frenchman, D. (2006). Mobile landscapes: Using location data from cellphones for urban analysis. Environment and Planning B—Planning and Design, 33(5), 727–748.

Rose, G. (2006). Mobile phones as traffic probes: Practices, prospects and issues. Transport Reviews, 26(3), 275–291.

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.

Sohn, K., & Kim, D. (2008). Dynamic origin-destination flow estimation using cellular communication system. IEEE Transactions on Vehicular Technology, 57(5), 2703–2713.

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. Journal Transportation, 36(5), 483-510.

Steenbruggen, J., Borzacchiello, M. T., Nijkamp P., & Scholten, H. (2011). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities, GeoJournal.

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.

White, J., & Wells, I. (2002). Extracting origin destination information from mobile phone data. In Proceedings from 11th international conference on road transport information and control, 486, 30–34.

Wolf, J. L., Guensler, R., & Bachman, W. H. (2001). Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In Journal of the Transportation Research Record, 1768, 125-134.

Yagi, S., & Mohammadian, A. (2010). An Activity-Based Microsimulation Model of Travel Demand in the Jakarta Metropolitan Area. Journal of Choice Modeling, 3(1).

Yagi, S. & Mohammadian, A. (2007). Validation of an Activity-Based Microsimulation Model of Travel Demand. Proc. of 11th World Conference on Transport Research Society, (DVD), Berkeley, CA.