Estimating activity patterns using spatio-temporal data of cell phone networks

Seyedmostafa Zahedi and Yousef Shafahi

Civil and Environmental Engineering Department, Sharif University of Technology, Tehran, Iran

ABSTRACT

The tendency towards using activity-based models to predict trip demand has increased dramatically over recent years. However, these models have suffered from insufficient data for calibration, and the intrinsic problems of traditional methods impose the need to search for better alternatives. This paper discusses ways to process cell phone spatio-temporal data in a manner that makes it comprehensible for traffic interpretations and proposes methods on how to infer urban mobility and activity patterns from the aforementioned data. The movements of each subscriber are described by a sequence of stops and trips, and each stop is labelled by an activity. The types of activities are estimated using features such as duration of stop, frequency of visit, arrival time to that activity and its departure time. Finally, the chains of the trips are identified, and different patterns that citizens follow to participate in activities are determined. These methods have been implemented on a dataset that consists of 144 million records of the cell phone locations of 300,000 citizens of Shiraz at five-minute intervals.

© 2017 The Institute of Urban Sciences

CONTACT  Seyedmostafa Zahedi  Zahedi.ast16@yahoo.com

KEYWORDS  Cell phone spatio-temporal data; activity-based models; activity patterns; big data

ARTICLE HISTORY

Received 15 August 2016  Accepted 5 May 2017

1. Introduction

Transportation planners need to understand human movements to better design networks that suit the needs of citizens and their demands for travel. There are two main ways to model trips in an urban area: trip-based models and activity-based models. The required data to calibrate such models are normally gathered by interviews or surveys. These traditional methods of obtaining data have several inconveniences. The process of sending out questionnaires and recollecting them is quite costly. Entering or importing the data into an appropriate database is very time-consuming. The infeasibility of surveying large percentages of the society has introduced errors into the analyses of engineers. Moreover, not all citizens are comfortable replying to such detailed questions about their daily trips, and they might answer the questionnaires incorrectly. Some people do not even take the time to answer the interviews or questionnaires. Usually, the educated portion of a society understands the importance of these surveys and cares to participate in the procedure, which results in biased data. All of these inconveniences and problems hinder the frequent use of such traditional methods and have prompted scholars to come up
with better ways to acquire sufficient data to calibrate their models. Between the two prevalent models, the trip-based ones have long been used and calibrated by survey data. The activity-based models, however, face problems when used with survey data. Their multi-level nested nature along with their numerous alternatives demand for huge data if they are to correctly model and predict trips. As location-aware technologies grew and developed, scientists paid more and more attention to the feasibility of using them to observe urban mobility. Among these technologies, cell phone networks stand out as a more promising way of collecting data. Cell phone networks have a built-in capability of recording the location of their subscriber’s cell phone without the need of any additional infrastructure. They try to always be aware of the whereabouts of their subscribers to be able to calculate the cost of making a call and maintain readiness for a fast connection. Cell phone networks are capable of tracking a large portion of citizens for as long a period as required and for as many times as needed. There are four events that result in registering a cell phone’s location: when a cell phone leaves a location area; during a call; when a cell phone is turned on; and a periodic location update. The location updates are passive, and subscribers do not take any actions to make them happen. When we use the data obtained by cell phone networks to represent the movements of citizens, the participants are unaware of being in the sample, which increases the reliability of the data. On the other hand, some issues arise when we try to elicit traffic-related information from the spatio-temporal data from cell phones. The accuracy of the locations might be insufficient. The accuracy by which the networks estimate the location of a subscriber may vary between 50 metres to several kilometres. In addition, we assume that the cell phone networks record the location of the nearest BTS (Base Transceiver Station) to the subscriber’s cell phone, which is not necessarily always true. A still cell phone may connect to several nearby BTSs and have the network record different locations for it when in fact it has been stationary the whole time, a phenomenon known as the Ping-Pong handover. Scholars have tried to overcome these issues and provide frameworks for practical use of these spatio-temporal data. The next section presents a review of the literature on this topic.

2. Related work

Scholars have made efforts to develop methods to use cell phone data to understand and estimate urban mobility and trip patterns. Studies were initially simple and mainly addressed the feasibility of using mobile phones to measure traffic variables. Ygnace, Drane, Yim and de Lacvivier (2000) tried to find out if it is possible to use cell phones as a probe to estimate travel times (Ygnace et al., 2000). Cayford and Johnson (2003) studied the effective parameters on the practical use of cell phone traces for generating traffic information (Cayford & Johnson, 2003). One of the obvious outcomes was spatial accuracy. Afterwards, Asukara and Hato undertook more complicated research to elicit travel behaviour from cell phone data (Asakura & Hato, 2004). They tried to actually use this data for traffic interpretations. They enumerated some advantages for using cell phone data; advantages such as a higher accuracy in comparison to traditional methods, the all-day-long coverage of cell phone networks, and the reachability of data in all sorts of weather conditions. However, their main contribution was developing a label-setting algorithm that distinguishes stop locations from on-move locations. Recent studies are more precise in their objectives. In 2010, Song et al. essayed to find the
degree to which human mobility patterns are predictable. They explored the limits of predictability in human dynamics by studying the mobility patterns of anonymized mobile phone users (Song, Qu, Blumm & Barabási, 2010). In 2013, using the idea that Anonymous location data from cellular phone networks can shed light on how people move around on a large scale, Becker et al. published a paper titled ‘human mobility characterization from cellular network data’ (Becker et al., 2013). Later, Xu et al. (2015) tried to understand aggregate human mobility patterns using passive mobile phone location datasets from Shenzhen, China (Xu et al., 2015). Their study presented a home-based approach to find human movement patterns that considered the homes of individuals as anchor points and references to analyze those individuals’ activities. Then, they categorized people based on their approximate home locations to obtain aggregate mobility patterns for each BTS. Finally, they used a multilevel hierarchical clustering algorithm to classify regions that showed similar mobility patterns. Song Gao (2015) presented an analytical framework using the detailed records of cell phone calls in a city to explore human mobility patterns and intra-urban communication dynamics (Gao, 2015). Allahviranloo and Will (2015) conducted research to mine activity pattern trajectories and allocate activities in different parts of the network (Allahviranloo & Will, 2015). Although their data was obtained from GPS, the method they proposed can be used for mobile phone data as well. They tried to infer the types of activities in which each individual has engaged at different locations by features such as the duration of stops and the distances from home. Later they used a Markov chain with conditional random fields to find the relations between an individual’s socio-economic attributes and activity sequencing and his or her spatio-temporal trajectory of activities. In the same year, Widhalm et al. made an effort to discover urban activity patterns in cell phone data (Widhalm, Yang, Ulm, Athavale, & González, 2015). In order to do so they developed a two-staged method. In the first stage, they detected stops and extracted geocoded time stamps that formed trip chains. In the second stage, they combined stops with land-use data to cluster activities. They modelled the dependencies, activity type, trip-scheduling, and land-use type via a relational Markov network. They tested their method by a CDR (see section 4) dataset on Boston and claimed that the results agreed with the city surveys. Jiang et al. used data from Singapore to infer activity-based human mobility patterns from mobile phone data (Jiang, Ferreira, & González, 2015). The patterns are supposed to be suitable for the activity-based modelling approach. By parsing trajectories to extract stops and monitoring the most frequently communicated tower during night, they were able to detect the homes of individuals. They used the concept of ‘motifs’ as a representative of the numerous and complex activity patterns and chains of trips and provided an algorithm to identify the daily motifs for each individual. Motifs are very simple sketches of human movements. Several different movement patterns can be shown by one motif. Later, they interpreted the results and demonstrated the spatial patterns of human mobility for that data.

3. Area of study

Shiraz is one of the major cities in Iran and is located in the southern regions of the country. More than one and a half million residents live in the city. Shiraz is mainly aligned directionally from the east to the west and is bounded by mountains in the
north and south. Shiraz has a landed area of 224 square kilometres. The three modes of transportation in the city are cars, busses, and subways. Two service providers sell SIM cards in Shiraz one of which is normally used by the young people but the other one is uniformly distributed among different levels and ages in the society.

4. Data

Previous scholarly works have been more or less based on Call Detailed Records (Tettemanti, Demeter, & Varga, 2012; Vajakas, Vajakas, & Lillemets, 2015). CDR data are location-plus-time records that happen during a call. When a subscriber receives or makes a call, his or her location is recorded in the networks data base along with the time of the event. The problem with CDR data is that they are very sparse and sporadic. Not everyone makes or receives enough phone calls during a day such that their movements can be observed, and even if they do there is no certainty that the data catches all the participated activities. There might be stops and activity locations that remain hidden to researchers because there is no record of them. Our data, however, are records of periodic location updates. Three hundred thousand citizens in Shiraz were tracked every five minutes, and their locations were recorded for 40 h. This constitutes tracking of nearly one fifth of the city’s population for almost two entire days and nights and comprises 144 million records. These records are anonymous so that the privacy of the citizens is not violated. The precision of the recorded location in the network differs regarding the density of the BTSs in various parts of the city, however, in crowded zones with a dense presence of BTSs, the network can locate its subscribers within 50 metres. Figure 1 shows the distribution of BTSs in the city and the traffic zones. Some scholars have tried to improve the precision of such data. For instance, Hoteit et al. used the cell phone data collected by Airsage (http://www.airsage.com/) that calculated the spatial information of cell phones with a triangulation algorithm based on the signal strength received from three nearby towers (Hoteit, Secci, Sobolevsky, Ratti, & Pujolle, 2014). The 300,000 participants’ cell phones were scanned the night before to make sure that they resided within the urban areas of the city, however, some

![Figure 1. Shiraz traffic zones and the distribution of BTS towers in the city.](image-url)
of them made trips outside the city during the 40-hour period. We used ArcGIS to filter out those who had spent a fair share of their time outside the city. Other than cell phone data, 4% of the citizens were surveyed during the 40-hour period. Twenty thousand questionnaires were given to random households, and their daily chains of trips were gathered by these questionnaires and can be used to verify the methods presented in this paper.

5. Methodology

The data in this paper is intractable and quite difficult to handle. Different phenomena such as the Rayleigh fading effect and Ping-Pong handover distort the images seen by the network of a subscriber’s movements. Therefore, some pre-processing needs to be done in order to convert the raw data into meaningful trips and stop locations that show movements and participation in activities. Widhelm et al (Widhelm et al., 2015) used a low-pass filter to eliminate the outliers and smooth the movements with a velocity higher than the acceptable range. Then they used an incremental clustering algorithm to detect stops and convert the raw cell phone track into a sequence of visited places. Jiang et al (Jiang et al., 2015) applied an outlier detection algorithm based on time intervals and distances between consecutive points to eliminate the two types of noise that existed in the raw data. Other than eliminating signal jumps and outliers, they also agglomerated points that were spatially close but not necessarily adjacent in temporal sequence to obtain one unique location that represented the location of an activity. These methods may work well for CDR, but are not practical for PLU (periodic location update) data. In CDR data, there might be several records for a single user in a matter of minutes, but on the other hand the network may be unaware of its subscriber for hours. This is when a clustering algorithm comes in handy. But in PLU data, where there are records in predefined intervals (in our case, five minutes), Ping-Pong handover and frequent signal jumps between towers dominate the data and insert fake movements in the trajectory of subscribers. Since the distance between the towers that Ping-Pong occurs varies significantly, a clustering algorithm that uses distance as a similarity feature is not effective. The following sections describe how to distinguish Ping-Pong handover from real displacements and provide instructions to discern stops that were motivated by an activity from points that were simply passed by.

5.1. Distinguishing ping-Pong handover from real movements

Cell phones connect to the BTS from which they receive the strongest signal. Since signal strength decays exponentially as the distance between a cell phone and tower increases, a logical assumption would be that cell phones connect to the nearest tower that will later inform us of the whereabouts of the subscriber. However, this is not always the case. In telecommunication science there is a phenomenon known as the Rayleigh fading effect. When a signal is generated and propagates through an environment, due to the reflections and refractions that happen in its way, the signal power gets stronger or weaker randomly. These random fluctuations in signal strength change the BTS to which the cell phone connects and, therefore, change the represented location of the cell phone in the network database. As a result, within consecutive records a cell phone may be represented at different locations while its actual location has not changed at all. Sometimes the presence of tall
buildings blocks the way of a signal and the cell phone connects to a more accessible tower. Then, after a slight displacement, the obstacle is removed and the cell phone changes its connection to a closer BTS. Sometimes a tower is fully loaded and a nearby cell phone is not welcome and has to connect to a more distant BTS, and then shortly afterwards, as the load decreases, the former tower is suddenly able to accommodate the cell phone. Sometimes the topology of the region causes cell phones to connect to farther towers that have a similar altitude as the cell phone. All of these examples cause frequent back and forth movements of cell phones between towers that resemble a game of Ping-Pong, and, therefore, this phenomena is called the ‘Ping-Pong handover.’ The Ping-Pong handover exists all over the data and mistakenly increases displacements of each and every individual. It assigns invalid movements to cell phone users, even those that are not moving. There is no discernable pattern for Ping-Pong handover. It can happen between any numbers of towers for a single cell phone. It can happen every now and then at unknown intervals. Sometimes towers handover a cell phone around every minute, and sometimes a cell phone lingers on the same tower for hours.

Despite the fact that Ping-Pong handover does not follow any regular pattern, it defies logical human movements. It is very unusual for a subscriber to oscillate between two or more points for a notable period. Humans follow a purpose when they travel and try to minimize the trips necessary to achieve these purposes. Several attempts have been made to detect such noise and handovers (Lee & Hou, 2006; Yoon, Noble & Liu, 2006). Lasonen et al (Laasonen, Raento, & Toivonen, 2004). studied these oscillations and came up with three conditions that detect if Ping-Pong handover is occurring among a set of towers. The first condition requires that all towers in a set are close to each other. The second condition checks if oscillation is happening, which is when the average time spent visiting a set is larger than the sum of the individual times on each cell. The third condition makes sure that no subset of the towers satisfy the second condition and minimizes the number of members in each set. The problem with this algorithm is that if a cell phone is in the vicinity of a set of towers that are causing Ping-Pong handovers and then moves to another set that has an intersection with the previous one, the algorithm erroneously assumes that the cell phone did not move at all. Imagin a cell phone is oscillating among three towers, say A, B and C. It then moves to another point and starts Ping-Ponging among different cells that have an intersection with the previous ones, for example C, D and E. In this case, the algorithm will combine all of the towers and outputs into a single location for the cell phone, representing that no actual movement has occurred, which is not true. Hong and Kim (Hong & Kim, 2009) noticed that the same Ping-Pong handover happens in WLAN traces, and that 90% of transitions are irrelevant to actual user movements. They tried to filter out the false transitions. As a result, they consider a transition to be Ping-Pong if it satisfies two conditions: A) The transition should be among \( l \) recently associated BTSs. B) There should be at least \( p_{nu} \) current transitions among these \( l \) BTSs. Readers can refer to (Dash et al., 2015) for a more complete explanation. We have employed the method provided by Hong and Kim to distinguish Ping-Pong handover from real movements. We chose \( l \) to be three and \( p_{nu} \) to be two. However, their method has some deficiencies. First, their conditions detect Ping-Pong handover with a delay. It takes at least three handovers for their method to detect a Ping-Pong effect. Second, there were a finite number of situations where their method couldn’t detect a Ping-Pong handover. We enumerated all the possible situations of
occurrence of a Ping-Pong handover for \( l = 3 \) and \( p_{nu} = 2 \) where the Hong and Kim algorithm failed to detect them. Moreover, we found the source of the delay and noticed that with our chosen \( l \) and \( p_{nu} \), if the algorithm detects a Ping-Pong handover after several real movements, the three previous displacements before a detection were definitely a Ping-Pong handover. We completed their method, added those undetected conditions, removed the delay and inserted some modifications to better distinguish Ping-Pong handover from real movements. Table 1 demonstrates what is meant by a delay and how the algorithm was modified to compensate for this delay. The red cells show where Ping-Pong handovers have occurred but the algorithm failed to detect them. Finally, the above-mentioned algorithm solely detects Ping-Pong handovers and fake transitions. However, it does not comment on the actual whereabouts of the cell phone. If a transition made by a cell phone is detected to be fake, this means that the cell phone is staying stationary somewhere between the Ping-Ponging towers. Since the distance between the towers contributing to a Ping-Pong handover may be several kilometres, it is crucially important to have an estimate of the cell phone’s location. Hence, considering the times spent on each BTS as weights, a weighted average among all the Ping-Ponging towers’ locations for a cell phone was calculated that resulted in one representative location that is the best estimator of the cell phone’s actual whereabouts.

5.2. Distinguishing stops and activity locations from on-move points

After eliminating false transitions, it is now time to determine at which points a user has stopped to make an action and at which points he or she was moving to get to the places he or she wanted to visit. Correctly detecting stops is of paramount importance when it comes to extracting activity patterns; it is also crucial to other transportation uses of cell phone data, such as mode choices (Kalatian & Shafahi, 2016) or route assignments (Taghipour & Shafahi, 2016). Failing to detect a stop can change the patterns and trip chains dramatically. For example, it can drop a three-node activity pattern into a two-node one and mislead us to erroneous results. Also, by missing a stop point, the path that a traveller moves along can be unusually long. A logical explanation as to why travellers use such paths rarely exists. As discussed earlier, Widhelm et al. used an incremental clustering algorithm to detect such stops. In a time-stamped location sequence, Jiang et al. clustered points that were spatially close (within the threshold of \( \Delta d \)) and temporally adjacent. But, as mentioned earlier, these methods are not efficacious for PLU data. Dash et al (Dash et al., 2015) used two thresholds to detect stops. They postulated that a person is

| BTS | \( B_{nu} \) | \( P_{nu} \) | Ping-Pong detection |
|-----|-------------|-------------|---------------------|
| A   | (A)         | 0           | FALSE               |
| B   | (A,B)       | 0           | FALSE               |
| C   | (A,B,C)     | 0           | FALSE               |
| D   | (B,C,D)     | 0           | FALSE               |
| E   | (C,D,E)     | 0           | FALSE               |
| F   | (D,E,F)     | 0           | FALSE               |
| G   | (E,F,G)     | 0           | FALSE               |
| E   | (E,F,G)     | 1           | FALSE               |
| G   | (E,F,G)     | 2           | TRUE                |
| F   | (E,F,G)     | 2           | TRUE                |
staying in a location and performing an activity if his or her cell phone remains within a radius of $R_d$ during a time limit of $T_d$. They ran a few sensitivity analyses and determined these thresholds in such a way that the trip-rates and number of stops matched the data from surveys. However, this does not sound like an appropriate procedure because the objective of using cell phone data is to derive traffic-related information without the need of surveys. Their method makes cell phone data dependent on survey data for calibration and practical use. Montoliu and Gatica-Perez (2010) employed the same method and added a third condition that restrains the time difference between two consecutive records in a stop location (Montoliu & Gatica-Perez, 2010).

The method proposed in this paper considers all the records of a user after applying the Ping-Pong pre-processing and decides if each record is a stop or not. Since our data is in five-minute-intervals, a subscriber may spend 1 s to 10 min in a given location, so it is not easy to tell if a location was recorded during a move or a stop. However, by combining other factors with the duration of a stop, it is possible to robustly detect a purposeful stop. Here are three conditions where, if any of them are met, the point under consideration is highly likely to be a stop.

5.2.1. If a cell phone connects to a tower (or remains within a cluster of Ping-Ponging towers) for more than 20 min
Odds are the user of a cell phone who meets condition one is purposefully staying near the tower. The configuration of the city and the BTSs inside it are in such a way that a moving cell phone is not likely to connect to one tower (or to remain within the cluster of the towers causing Ping-Pong handovers) for more than 20 min. In other words, four consecutive repetitions of the same location in the time-stamped sequence indicates a stop.

5.2.2. Points with two or three consecutive repetitions of the same location in dense presence of BTSs
If the cover range of the tower or towers representing the cell phone’s location is fairly small, it can be assumed that the user has stopped in that location to perform some activity.

Figure 2 illustrates how the second condition works. If the area that a BTS tower covers is small, or, that is to say, if there are several towers in close vicinity of it, a moving cell phone can easily leave the tower’s territory and connect to other BTSs. Hence, if the cell phone remains connected to one tower, or, in other words, its representative location repeats two or three times consecutively in the sequentially time-stamped records, this means that the subscriber holding this cell phone has intentionally stopped near that BTS and is not moving. By ‘small cover area’, we mean an area that a walking passenger can cross within the expected amount of time spent on that tower. For example, if the data of the $i_{th}$ user has recorded tower $j$ for three consecutive times, we expect that the user’s cell phone has been connected to that tower for 15 min (which can be between 10 and 20 min). Now, if we assume that the speed of an average passenger is 1.4 metres per second, a cell phone that is moving at least as fast as a pedestrian can move $1.4 \times 15 \times 60$, or 1260 metres, which is the diameter of a circle with an area of 1.25 km squared. If the area belonging to a certain tower is notably smaller than 1.25 km$^2$ and a cell phone is connected to this tower for 10 or 15 min, this indicates that the subscriber was not trying to move away from that
5.2.3. Changes in major directions

Significant changes in the overall trends and directions of a person are motivated by activities. People tend to use routes that can more or less directly take them from an origin to a destination. The paths on which subscribers move are normally smooth, especially when seen by a network that catches their location only every five minutes and the sharp U turns and left or right turns at intersections are neglected. Hence, if a person deviates considerably from the overall trend of his or her movements there is a reason behind it, and the reason is to make a stop to participate in an activity. Consider Figure 3. A tour guide leaves his home at Gasr dast (marked by 1), goes to pick up a tourist at the airport (marked by 2) and finally takes the tourist to the Hafez Tomb (marked by 3). Since the pick-up may take less than ten minutes, our first two conditions are highly likely to fail to detect a stop at the airport. The third condition however, notices that the traveller was initially moving south. At the airport, he changes his direction and goes north, which indicates that the traveller had something to do at the airport and he was not merely passing by it. In order to find out whether or not a point meets the third condition, the overall direction and moving trend before that point is compared to the trend after it, and if these trends have notable differences in direction, the point under investigation is considered to be a purposeful stop location. It is arguable that some direction changes might be due to the geometric structure of the road network. However, it should be noted that the cell phone network does not exactly follow a subscriber’s path and its movements are being observed by the network with an average accuracy of 350 metres. It seems
reasonable to assume that the sharp direction changes imposed by the transportation system on the subscriber’s path fade when seen from the cell phone network’s perspective. To add to this point, the movements of a subscriber are recorded every five minutes, which means that only a sense of the general trend of the movements is captured by the network. This extenuates the errors that might be imposed by the geometry of the transportation system.

In order to mathematically model the third condition, the trend before a point is defined as the unit vector that represents the direction of the line that best fits the two anterior points (recorded BTS locations) and the original one. Likewise, the trend after the point is the vector that represents the direction of the line that best fits the two posterior points and the original one. The angle between these two vectors is computed, and if it exceeds 110 degrees the point under investigation is probably a stop location.

**Figure 3.** How a significant change in the moving direction of a passenger indicates a stop.

**Figure 4.** Stop points and on move points.
If a point does not meet any of these three conditions, then it is assumed to be an on-move point and the cell phone holder was merely passing by this point while the periodic location-update recorded its location. Figure 4 shows a schematic output of this section.

All of the above conditions were applied to the records. For a given subscriber, there must be a trip between two consecutive stops. The number of trips for each subscriber has been counted, and an average of 1.85 trips per person was determined for the city of Shiraz. The trip rate determined by our method is consistent with that obtained from the survey.

5.3. Detecting types of activities

When a stop is detected, this means an activity is being performed during that stop. The next step would be to identify in which sort of activity the traveller is participating. Each type of activity has several characteristics that help its recognition. For example, the activity type of ‘work’ usually happens during working hours and in non-residential areas. In order to determine the type of activities a person performs in a day, each stop should be enriched with features that contribute to identifying the types of activities during that stop. Our method proposes 5 features for a given stop: the arrival time to the stop, the departure time from the stop, the duration of the stop, the frequency of visits and the land-use shares around the location of the stop. However, different types of activities are determined with different confidence and preciseness. For example, it is much easier to determine the types of activities such as ‘home’ or ‘work’ rather than ‘shopping’ or ‘health care’. Therefore, the two former types of activities are identified deterministically, whereas other types of activities are determined probabilistically. Figure 5 demonstrates what we mean by enriching activities with features. It also shows the end result of activity type detection.

5.4. Detecting ‘Home’ and ‘Work’

A home is an important anchor point in a traveller’s chain of trips. Most trips will either start or end at home. This paper defines ‘home’ as the place a person stops the longest during night (from 11.00 PM to 8 AM) in residential areas. However, sometimes there are more than one dominant stops during this period. Some people spend a significant part of their nights out doors and enter their homes quite late. In these cases, other features, such as frequency of visits, can help because normally leaving home will eventually be followed by coming back, and the home location is more frequently visited than other locations. Therefore, for these cases a ‘home’ is defined as a long stop during the night that can be seen as an anchor point in a person’s daily trips. To determine which of the stops in a person’s chain of trips is his or her ‘home’, we use the following equation. Formula 1 is computed for each stop and the one with the larger value is declared the home

\[ \frac{D_{n}(i)}{9h} \times \frac{F(i)}{1 + F(i)} \times LU(i) \]  

(1)

Where \( D_{n}(i) \) is the total duration spent in location \((i)\) during night, \( F(i) \) is the total
frequency of visit in the day of location \((i)\) and \(LU\) is the residential land-use share of location \((i)\).

Formula 1 multiplies three terms between 0 and 1, and the closer the resulting value is to number one, the higher is the probability of the stop being at ‘home’.

Work is detected in the same way, except that instead of using overnight for the expected time of activity, working hours are used, and a residential land-use share is replaced with a non-residential land-use share.

Formula 2 is computed for each stop, and the one with the larger value is declared as work. (Working hours are from 8 in the morning till 5 in the afternoon)

\[
\frac{Dw(i)}{9h} \times \frac{F(i)}{1 + F(i)} \times NLU(i)
\]

Where \(Dw(i)\) is the total duration spent in location \((i)\) during working hours, \(F(i)\) is the

\[\text{Figure 5. How different features in each stop is used to derive activity types and how the end result will be.}\]
total frequency of visit in the day of location \( (i) \) and NLU is the non-residential land-use share of location \( (i) \).

5.5. Detecting other types of activities

As mentioned earlier, other activities are not easy to identify. For example, ‘shopping’ may occur anytime during a day, and it may last across a wide range of time, from several minutes to several hours. Health-care, personal business, recreation and even sometimes school activities show the same complications, which complicates the feasibility of their exact identifications. As a result, a probabilistic method is employed. For each given stops that is not home or work, the areas of different land-use types in the vicinity of that stop are available from GIS. Each land-use type has an average rate of attraction that is provided by previous studies on the cities of Shiraz and Mashhad. These studies are similar to the trip generation list of ITE and report how many trips per hour a squared metre of a certain land-use type is expected to attract. These attraction rates are then used to find the probability of the purpose of a stop. The higher the attraction of a land-use type in the vicinity of a stop, the higher the probability of that land-use type to be the purpose of the stop. For example, if the dominant land-use share of an area is hospitals and treatment centres, and 90% of the trips attracted to that area are because of sanitary and health care, then a stop detected in that area that is not home or work has a 90% chance of being due to health care reasons. The remaining 10% will be distributed similarly between the other possible land-uses and types of activities. The possible types of activities are determined by filtering them all by using their arrival times and durations. For instance, an activity that lasts less than a class, say 1.5 h, cannot be for educational purposes. Likewise, since places of administration are usually not open after 5 pm, stops that occur after this time cannot be for administrative purposes. The same applies to education as well. Incorporating these filters enables us to use arrival time, departure time and duration to better estimate the types of activities preformed at each stop.

5.6. Estimating activity patterns

An activity pattern is a chain of trips that is comprised of several activities. A person who leaves home in the morning to go to work and returns home at noon to have lunch with his or her family follows a different activity pattern than a person who prefers to have lunch in his or her office. Identifying these sorts of activity patterns and finding the share that each pattern is used by the citizens is the main objective of this paper. Since the activities preformed in an individual chain of trips is not exactly known, all plausible patterns for each individual should be considered. Simultaneously, the probabilities of

| Activities | activity 1 | activity 2 | activity 2 | activity 3 | activity 4 | activity 4 |
|------------|------------|------------|------------|------------|------------|------------|
| Labels     | ‘H’        | ‘A’        | ‘HC’       | ‘W’        | ‘R’        | ‘Sh’       |
| Probabilities | 100%     | 65%        | 35%        | 100%       | 80%        | 20%        |
| Pattern    | ‘H-A-W-R’  | ‘H-A-W-Sh’ | ‘H-HC-W-R’ | ‘H-HC-W-Sh’| ‘H-HC-W-Sh’| ‘H-HC-W-Sh’|
| Probabilities | 52%      | 13%        | 28%        | 7%          |            |            |
occurrences of each pattern should be calculated. For example, if an individual conducts four activities in a day where two of them are home and work and the other two are probably administrative activities or health care and recreation or shopping, respectively, there will be four conceivable activity patterns for this individual. Table 2 shows these patterns and their corresponding probabilities.

When the patterns for all individuals have been extracted, the percentage that each pattern is used by the citizens can be easily computed. Our algorithm starts with a blank list and considers all the patterns of each individual one by one. Once it encounters a pattern that is not on the list, the pattern is added to the list along with its probability. However, if the pattern is already on the list, the corresponding probability of the pattern is added to the probabilities of previous similar patterns. After all the citizens have been studied, we have a list that shows all the patterns and the frequencies that were followed by the citizens.

### 6. Results

Among all of the 300,000 citizens, those who made moves in areas where the land-use information was available were separated, and the above-explained methodology was implemented on them for a 24-hour period. At first, outliers, noises and Ping-Pong handovers were omitted. Next, stops were detected and the process of assigning a trip between two consecutive stops yielded a trip rate of 1.86 per person, which is certified by the trip rate obtained from the survey conducted on the exact same day from 4% of the population. Table 3 shows the different patterns of movements and their relative frequencies. In this table, nodes are where subscribers engage in an activity, and stops are the number of times a subscriber has halted in any of those nodes. The most frequent combination in Table 3 is the two-node-three-stop pattern, which seems logically sound because it can be expected that home-based trips with a singular purpose occur more often than other patterns. As the patterns of movements become more complicated, the percentage of the people following the patterns decreases. This again corroborates our expectations.

After distinguishing the stops, the type of activity of each node is estimated. In order to validate the accuracy of our methodology, the resulting percentages of different trip

| Stops      | Percentage | Stops      | Percentage |
|------------|------------|------------|------------|
| Two-nodes  |            | Three-node |            |
| 2-stop     | 0.16239    | 3-stop     | 1.1        |
| 3-stop     | 0.68376    | 4-stop     | 0.2564     |
| 4-stop     | 2.4641     | 5-stop     | 1.96565    |
| 5-stop     | 2.7205     | 6-stop     | 0.94       |
| 6-stop     | 1.282      | 7-stop     | 0.8547     |
| 7-stop& more| 0.3547    | 8-stop     | 0.427      |
| Total      | 33.2643    | 9-stop& more| 0.2564    |
|            |            | Total      | 11.66      |

| Stops      | Percentage | Stops      | Percentage |
|------------|------------|------------|------------|
| Four-nodes |            | Five-node  |            |
| 4-stop     | 0.68376    | 5-stop     | 0.2564     |
| 5-stop     | 2.4641     | 6-stop     | 1.96565    |
| 6-stop     | 2.7205     | 7-stop     | 0.94       |
| 7-stop     | 1.282      | 8-stop     | 0.8547     |
| 8-stop     | 0.3547     | 9-stop     | 0.427      |
| 9-stop& more| 0.2564    | 10-stop    | 0.085      |
| Total      | 7.761      | Total      | 3.76       |
purposes and types of activities are compared with those of the trip survey conducted by the municipality of Shiraz in 2000 (Table 4). It is important to note that the regulations imposed by the ministry of education do not allow high school and middle school students to take their cell phones to class. As a result, the number of trips for educational purposes is severely underestimated, and such trips are hidden from the cell phone network. This qualification only exists in Iran and does not raise questions about the power of cell phone data in general. In order to compensate for this limitation during the comparison, the number of educational activities has been increased in a way that matches the percentage of educational trips obtained from the survey in 2000, which enables us to better compare the rest of the activities and trip purposes. By comparing the percentage of the purpose of each trip from 2000 and cell phone data, an overall congruency is observed. The closeness of activity shares suggests that the methodology presented in this paper is reliable. The reduction in the share of return-to-home trips indicates that citizens tend to participate in their daily activities in one multi-purpose trip chain rather than frequent departures from and returns to home, which will in turn add to none-home-based trips. The existence of such trip chains shows that trips are getting more and more dependent on each other, which emphasizes the need to use activity based models. One of the ways that cell phone data is superior to survey data is the accuracy of recording none-home-based trips. Respondents to questionnaires often forget to include minor none-home based trips in to their daily life because they consider them trivial and unimportant. However, such trips are observed and recorded by the cell phone network. The 0.05 difference in trip rates obtained by cell phone data and survey data are assumed to be due to the none-home-based trips not reported in the survey.

After implementing the last step of the methodology for citizens who left home to participate in at least one activity (59% of the subscribers), nearly 800 different activity patterns were detected. Many of these patterns are very improbable and infrequent. Patterns with more than 10 stops in a day were removed because a regular person normally cannot participate in more than 10 activities; usually only taxi or bus drivers can go to so many places in a city in just one day. Table 5 shows the top 15 most frequently used patterns in Shiraz. Please note that ‘H’ stands for home, ‘W’ is work, ‘C’ is commercial, ‘R’ is recreation, ‘A’ is administrative purposes, ‘E’ is education and ‘P’ is pick up or drop off. All the patterns in Table 5 are followed by at least one percent of the subscribers who left their

---

**Table 4. Trip shares categorized by purpose.**

|                      | Type | W  | E  | P  | C  | R  | S  | A  | H  |
|----------------------|------|----|----|----|----|----|----|----|----|
| Cellphone data in 2016 | Percentage | 15.35 | 13.5 | 4.45 | 9.5 | 6.45 | 3.98 | 3.85 | 42.92 |
| Survey data in 2000  | Percentage | 19.76 | 13.5 | 2.59 | 7.3 | 3.12 | 2.8 | 2.29 | 48.03 |

---

**Table 5. Most frequently used activity patterns in the city.**

| Pattern | Share (%) | Pattern | Share (%) | Pattern | Share (%) |
|---------|-----------|---------|-----------|---------|-----------|
| HWH     | 13.85     | HCH     | 10.06     | HPH     | 3.06      |
| HWHWH   | 2.88      | HWHCH   | 2.44      | HSH     | 3.58      |
| HWRH    | 2.34      | HRH     | 9.33      | HCHCH   | 1.55      |
| HWRHH   | 1.11      | HRRH    | 1.11      | HCHCH   | 1.33      |
| HWHCH   | 1.75      | HAH     | 4.29      | HEH     | 2.47      |
homes, which amounts to 61% in total. The rest of the 39% followed patterns that were very unique and infrequent, which justifies the rationale behind numerous alternatives and patterns in activity-based models.

From these 800 activity patterns, the percentage of citizens having jobs can easily be calculated. Since teleworking is not common in Iran, the employment rate can simply be obtained by adding the share of all the patterns that include work. In activity based models, other than the mere type of an activity, the time it starts and finishes also matters. Table 6 shows the distribution of the starting time of the ‘work’ activity, during the day, obtained from the cell phone data. In addition, Table 6 contains the distribution obtained from the survey conducted by the municipality of Shiraz in 2000. The table shows how different policies have made the starting time of work activities shift towards noon.

A more important result that is worth mentioning is the patterns that are finished outside of home. Nearly 0.1% of the citizens is this research leave home and don’t come back within a 24-hour period. They follow patterns such as ‘HW’ or ‘HR’. This is an important result because it shows the superiority of cell phone data over traditional methods. In surveys and home interviews, those who are filling out the questionnaires have already returned home. Hence, the patterns obtained from such interviews will always finish with a trip to home, while patterns from cell phone data show that a noticeable percentage of citizens tend to stay out even after 12 am. Another peculiar, yet interesting, pattern observed by the cell phone data is the ‘HEEH’. It is a bit unusual that a substantial number of subscribers leave home and visit two different places to study and then return home. However, after further pondering, it was revealed that Shiraz University, which is one of the most prestigious universities in the country and has thousands of students, is not located in one central location. Different departments in the university are scattered all over the city, and undergraduate students need to take busses provided by the university to participate in classes across different departments. However, in questionnaires students merely address the trips from home to the university and vice versa, and they often forget to mention the trips they made between different departments throughout the city.

7. Concluding remarks

Different people follow different patterns to perform their daily activities. Having sufficient knowledge of these patterns is very important for city planners. Different policies, such as road pricing in the CBD or incentives to use public transit, greatly affect urban mobility patterns. City planners need to anticipate these changes and be able to observe the patterns continuously. However, the current methods that are used to obtain activity patterns are either very time consuming or unreliable due to the small sizes of samples. This paper proposes a methodology that uses cell phone spatio-temporal data to observe activity patterns in a city. The method works better as the accuracy of locating subscribers improves or as the land-use shares of the city get more detailed and accurate.

| Table 6. Temporal distribution of the arrival time activity ‘Work’. |
|------------------|------------------|------------------|------------------|------------------|------------------|
| Cellphone data in 2016 | Arrival time to work | 6–9 | 9–12 | 12–15 | 15–18 | 18–21 | 21–24 |
| Percentage of all ‘W’ activity types | 56.67 | 10.69 | 12.76 | 18.09 | 1.2 | 0.6 |
| Survey data in 2000 | Arrival time to work | 6–9 | 9–12 | 12–15 | 15–18 | 18–21 | 21–24 |
| Percentage of all ‘W’ activity types | 62.78 | 6.57 | 7.66 | 17.44 | 1.48 | 0.19 |
If the service provider can omit the Ping-Pong handover while tracking the cell phones and provides a dataset that has less back and forth jumps of BTS connections, the method’s performance will improve dramatically.

Our result is merely a simple output that can be inferred from such data.

Sometimes more complicated and comprehensive activity patterns are required. For example, most activity based models incorporate time of day or mode of travel in their activity patterns. Some go even further and account for destination and route choices. The ideas proposed in this paper can be modified and built on to match the needs of modellers as they see fit.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**References**

Allahviranloo, M., & Will, R. (2015). Mining activity pattern trajectories and allocating activities in the network. *Transportation, 42.4*, 561–579.

Asakura, Y., & Hato, E. (2004). Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies, 12.3*, 273–291.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., … Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM, 56.1*, 74–82.

Cayford, R., & Johnson, T. (2003). *Operational parameters affecting the use of anonymous cell phone tracking for generating traffic information*. Institute of transportation studies for the 82th TRB Annual Meeting, 1.3.

Dash, M., Koo, K. K., Holleczek, T., Yap, G., Krishnaswamy, S. P., & Shi-Nash, A. (2015). *From mobile phone data to transport network-gaining Insight about human mobility*. 2015 16th IEEE International Conference on Mobile Data Management, Vol. 1. IEEE.

Gao, S. (2015). Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation, 15.2*, 86–114.

Hong, J., & Kim, H. (2009). *Refining Elusive user registration patterns for empirical mobility models*. *Computer Communications and Networks, 2009. ICCCN 2009. Proceedings of 18th International Conference on*, IEEE, 2009.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Estimating human trajectories and hotspots through mobile phone data. *Computer Networks, 64*, 296–307. http://www.airsage.com/

Jiang, S., Ferreira Jr, J., & González, M. C. (2015). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *International Workshop on Urban Computing, Kalatian, A., & Shafahi, Y. (2016). Travel mode detection exploiting cellular network Data*. MATEC Web of Conferences. Vol. 81, EDP Sciences, 2016.

Laasonen, K., Raento, M., & Toivonen, H. (2004). *Adaptive on-device location recognition*. International Conference on Pervasive Computing, Springer Berlin Heidelberg.

Lee, J.-K., & Hou, J. C. (2006). Modeling steady-state and transient behaviors of user mobility: *Formulation, analysis, and application*, Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing, ACM, 2006.

Montoliu, R., & Gatica-Perez, D. (2010). Discovering human places of interest from multimodal mobile phone data. Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia, ACM, 2010.

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018–1021.
Taghipour, H., & Shafahi, Y. (2016). Route choice estimation using cell phone Data. MATEC Web of Conferences. Vol. 81, EDP Sciences, 2016.

Tettamanti, T., Demeter, H., & Varga, I. (2012). Route choice estimation based on cellular signaling data. Acta Polytechnica Hungarica, 9.4, 207–220.

Vajakas, T., Vajakas, J., & Lillemets, R. (2015). Trajectory reconstruction from mobile positioning data using cell-to-cell travel time information. International Journal of Geographical Information Science, 29.11, 1941–1954.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. Transportation, 42.4, 597–623.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., & Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. Transportation, 42.4, 625–646.

Ygnace, J.-L., Drane, C., Yim, Y. B., & de Lacivier, R. (2000). Travel time estimation on the San Francisco bay area network using cellular phones as probes. California Partners for Advanced Transit and Highways (PATH).

Yoon, J., Noble, B. D., & Liu, M. (2006). Building realistic mobility models from coarse-grained traces. Proceedings of the 4th international conference on Mobile systems, applications and services. ACM, 2006.