Estimating Activity Patterns Using Spatio-temporal Data of Cellphone Networks

Seyedmostafa Zahedi and Yousef Shafahi

Sharif University of technology, Civil and environmental engineering department, Azadi Ave., Tehran, Iran

Abstract. The tendency towards using activity-based models to predict trip demand has increased dramatically over recent years, but these models have suffered insufficient data for calibration. This paper discusses ways to process the cellphone 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. Movements of each subscriber is described by a sequence of stays and trips and each stay is labeled by an activity. The type of activities are estimated using features such as land use, duration of stay, frequency of visit, arrival time to that activity and its distance from home. Finally, the chains of trips are identified and different patterns that citizens follow to participate in activities are determined. The data comprises 144 million records of the location of 300,000 citizens of Shiraz at five-minute intervals.

1 Introduction

Transportation planners need to understand human movements to better design networks that suit citizens' needs and their demand 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 getting 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 engineer's analyses. Moreover, not all the citizens are comfortable with such detailed questions on 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 the society understands the importance of these surveys and cares to participate in the procedure, which leave us with 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 the models. Between the two prevalent models, the trip-based ones have long been used and calibrated by the 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 the 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, cellphone networks stand out as a more promising way of collecting data. Cellphone networks have a built-in capability of recording the location of their subscriber's cellphone 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 the readiness for a fast connection. There are four events that result into registering a cellphone's location. When a cellphone leaves a location area, during a call, when a cellphone is turned on and a periodic location update. Cellphone networks are capable of tracking a large portion of citizens for as long a period as required and for as many times as needed. The location updates are passive and the subscribers do not take any actions to make it happen. When we use the data obtained by cellphone networks to represent citizens' movements the participant 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 cellphones spatio-temporal data. Location accuracy might be insufficient. The accuracy by which networks estimate a subscriber may vary between 50 meters to several kilometers. In addition, it is assumed that the cellphone network records the location of the nearest BTS (Base Transceiver Station) to the subscriber's cellphone which is not necessarily always true. A cellphone which is holding still may connect to several nearby BTSs and have the network record different locations for it while it has been stationary the whole time. A phenomenon known as the Ping-Pong handover. Scholars have tried to overcome these issues and provide frame works for practical use of these spatio-temporal data. Next section presents a review of the literature on this topic.

2 Related work

Scholars have made efforts to develop methods to use cellphone data to understand and estimate urban mobility and trip patterns. Studies were initially simple and were mainly addressing the feasibility of using mobile phones to measure traffic variables. Ygnace et al. (2000) tried to find out if it is possible to use cellphones as a probe to estimate travel times [1]. Cayford and Johnson (2003) studied the effective parameters on practical use of cellphone traces for generating traffic information [2]. One of the obvious outcomes was spatial accuracy. Afterwards Asukara and Hato undertook a more complicated research to elicit travel behavior from cellphone data [3]. They tried to actually use this data for traffic interpretations. They enumerated some advantages for using cellphone data; advantages such as higher accuracy in comparison to traditional methods, all-day-long coverage of cellphone networks, reachability of data in all sorts of weather conditions and etc. However their main contribution was developing a label setting algorithm that distinguishes stay locations from on-move locations. Recent studies are more precise in their objectives. Yang Xu et al. (2015) tried to understand aggregate human mobility patters using passive mobile phone location dataset from Shenzhen, China [4]. Their study presented a home-based approach to find human movements patterns that considered 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 show similar mobility patterns. Mahdieh Allahviranloo et al. (2015) conducted a research to mine activity pattern trajectories and allocate activities in different parts of the network [5]. 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 duration of stop and distance from home. Later they used a Markov chains with conditional random fields to find the relations between individual's socio-economic attributes and activity sequencing and spatio-temporal trajectory of activities. Again in the same year Peter Widhalm et al. made an effort to discover urban activity patterns in cell phone data [6]. In order to do so they developed a two-staged method. In the first stage, they detected stays and extracted geocoded time stamps that form trip chains. In the second stage they combine stays with land use data to cluster activities. They model 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 of the Boston city and claimed that the results agreed with the city surveys. Shan Jiang et al. used data of Singapore to infer activity-based human mobility pattern from mobile phone data [7]. The patterns are supposed to be suitable for the activity-based approach. By parsing trajectories to extract stays and monitoring the most frequently communicated tower during night they were able to detect homes of the individuals. They used the concept of "motifs" as a representative of the complex activity pattern and chains of trips which can go into numerous combinations 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 Data

Previous scholarly works have been more or less with Call Detailed Records. 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 database 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 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 stays and activity locations that remain hidden because there has been no record for them. Our data however, are records of periodic update. Three hundred thousands of citizens of Shiraz have been tracked every five minutes and their location have been recorded for 40 hours which is nearly one fifth of the city's population for almost two days and nights and comprises 144 million records. These records are anonymous so that the privacy of the citizens are not violated. The precision by which the location is recorded in the network differs regarding the density of BTSs in various parts of the city, however, in crowded zones with dense presence of BTSs the network can locate its subscribers with an error of less than 50 meters.. The 300,000 participants' cellphones were scanned the night before to make sure that they reside within the urban areas of the city, however, some of them made trips outside the city in the 40 hour period. We used ArcGIS to filter those who had spent a fair share of their time outside the city. Other than cellphone data, 4 percent of the citizens were surveyed in a day during the 40 hour period. Twenty thousand questionnaires were handed to random household and their daily chains of trips were gathered by these questionnaires and can be used to verify the methods presented here in this paper.

4 Methodology

The data is intractable and quite difficult to handle. Different phenomena such as Rayleigh fading effect and Ping-Pong hand over distort the image seen by the network of the subscriber's movements. Therefore, some preprocessing needs to be done in order to convert the data into meaningful trips and stay locations that show movements and activity participation. Peter Widhelm et al [6] used a low-pass filter to eliminate the outliers and smooth the movements with a velocity higher than an acceptable range. Then they use an incremental clustering algorithm to detect stays and to convert raw cellphone track into a sequence of visited places. Jiang et al [7] applied an outlier detection algorithm based on time intervals and distances between consecutive points to eliminate the two types of noises that exist in the raw data. Other than eliminating signal jumps and outliers they agglomerated points that are spatially close but not necessarily adjacent in temporal sequence to obtain one unique location that represents an activity location. 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. That 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 hand over 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. In the following sections we describe how to distinguish Ping-Pong handover from real displacements and then provide stepwise instructions to discern stays that were motivated by an activity from passing-by points.

4.1 Distinguishing Ping-Pong handover from real movements

Cellphones connect to the BTS from which they receive the strongest signal. Since signal strength decays exponentially as the distance between cellphone and tower increases, it seems a logical assumption that cellphones connect to the nearest tower which 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 strengthens and weakens randomly. These random fluctuations in strength change the BTS that the cellphone connects to and therefore, changes the location that represents the cellphone in the network database. As a result, within consecutive records a cellphone may be represented by different locations while its actual location has not changed. Sometimes presence of tall buildings block the way and the cellphone connects to a more accessible tower. And after a slight displacement the obstacle is removed and cellphone changes the connecting BTS to a nearer one. Sometimes a tower is fully loaded and a nearby cellphone is not welcome and has to connect to another BTS While shortly afterwards load decreases and the former tower is suddenly able to accommodate the cellphone. Sometimes the topology of the region is in a way that cellphones tend to connect to farther towers that are in similar altitude as the cellphone device. All of these examples cause frequent back and forth movements of cellphones between towers that resembles a game of Ping-Pong and therefore, 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 cellphone users, even those that are not making any moves. There is no discernable pattern for Ping-Pong handover. It can happen between any numbers of towers for a single cellphone. It can happen every now and then at unknown intervals. Sometimes towers hand over a cellphone every minutes or so and sometime a cellphone lingers on a 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 make a trip and try to minimize the trips necessary to achieve these purposes. Lasonen et al. [8] used these oscillations and came up with three conditions that detects 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 and that is if 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 cellphone is oscillating among three towers say A, B and C and then moves to another point and starts Ping-Ponging among different cells that has an intersection with the previous ones, for example C, D and E this algorithm will combine all of the towers and outputs a single location for the cellphone with the meaning that no actual movement has occurred, which is not true. Hong and Kim notice that the same Ping-Pong handover happens in WLAN traces and that 90% of transitions are irrelevant to actual user movements [9]. They try to filter false transitions. 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 transition among these l BTSs. Readers can refer to [9] for a more complete explanation. We have employed the method provided by Hong and Kim to distinguish ping pong hand over from real movements. We chose l to be three and p_{nu} to one. However, their method had some deficiencies. First, their conditions detect Ping-Pong hand over with a delay. It takes at least three hand overs for their method to detect Ping-Pong. Second, there were finite number of situations where their method couldn't detect a Ping-Pong hand over. We completed their method, added those undetected conditions, removed the delay and inserted some modification to better distinguish Ping-Pong hand over from real movements.

4.2 Distinguishing stays 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 was moving to get to the places he or she wanted to visit. Correctly detecting stays is of paramount importance when it comes to extracting activity patterns; it is crucial to other transportation uses of cellphone data such as mode choice or route assignment. Failing to detect a stay will 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. Besides, by missing a stay point the path that the traveler moves along is unusually long. Hardly exists a logical explanation as to why the traveler uses such path. As discussed earlier Widhelm et al. used an incremental clustering algorithm to detect such stays. In the time-stamped location sequence, Shang et al. clustered the points which are spatially close (within the threshold of Δd) and temporally adjacent. But, as mentioned earlier these methods are not efficacious for PLU data. Dash et al. [10] used two thresholds to detect stays. They postulated that a person is staying in a location and performing an activity if its cellphone remains within a radius of Rd during a time limit of Td. They ran some sensitivity analyses and determined these thresholds in such a way that trips rates and number of stays matched the data from surveys. However this does not sound like an appropriate procedure because the objective of using cellphone data is to derive traffic-related without the need of surveys. Their method makes cellphone data dependent to survey data for calibration and practical use.

The method proposed in this paper considers all the records of a user after applying the Ping-Pong preprocessing and decides if each record is a stay or not. Since our data is in five-minute-intervals a subscriber may spend 1 second to 10 minutes in a given locations so it is not easy to tell if a location was recorded during a move or during a stay. By considering factors other than duration of stop it is possible to robustly detect a stay. Here are three conditions that if any of them is met, the point under consideration is highly likely to be a stay.

1- If a cellphone connects to a tower (or remains within a cluster of Ping-Ponging towers) for more than 20 minutes. It can surely be stated that the user of that cellphone is purposefully staying near that tower. Because the configuration of the city and the BTSs inside it are in a way that a moving cellphone is not likely to connect to one tower (or remains within the cluster of the towers with Ping-Pong handover) for more than 20 minutes. In other words four consecutive repetition of the same location in the time-stamped sequence indicates a stay

2- For the points with two or three consecutive repetition of the same location. If the cover range of the tower or towers representing that location is fairly small it can be assumed that the user has stopped in that location to perform some activity.

Fig.1 illustrates how the second condition works. If the area that a BTS tower covers is small or if there are several towers in close vicinity of it; a moving cellphone can easily leave the tower's territory and connect to other BTSs. Hence, if the cellphone remains connected to one tower or in other words its representative location repeats two or three times consecutively in the sequentially time-stamped records, it means that the subscriber holding this cellphone has intentionally stopped near that BTS and is not moving. By small cover area, we mean an area that a walking passenger can cross it 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 cellphone has been connected to that tower for 15 minutes (which can be between 10 to 20 minutes). Now, if we assume that the speed of an average passenger is 2 meters per second, a cellphone that is moving at least as fast as a walking passenger can move 2*15*60= 1800 meters. Which is the diameter of a circle with an area of 2.54 km squared. If the area belonging to a certain tower is notably smaller than 2.54 km^2 and a cellphone is connected to this tower for 10 or 15 minutes, it indicates that the subscriber was not trying to move away from that tower. The cover range of each BTS in the city is obtained from a Thiessen polygon.

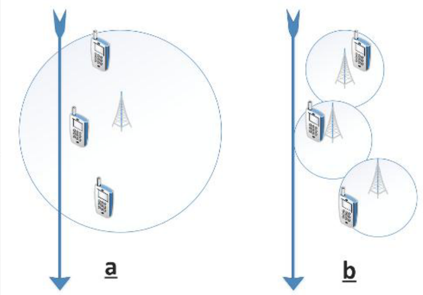

Figure 1. - a) if a cellphone is passing by a tower with a large coverage area, only one location may be recorded for the cellphone in consecutive time stamps b) if the cellphone is moving in a zone that BTSs are placed closely to each other, different locations will be recorded in consecutive time stamps for a moving cellphone as it connects to different BTSs.

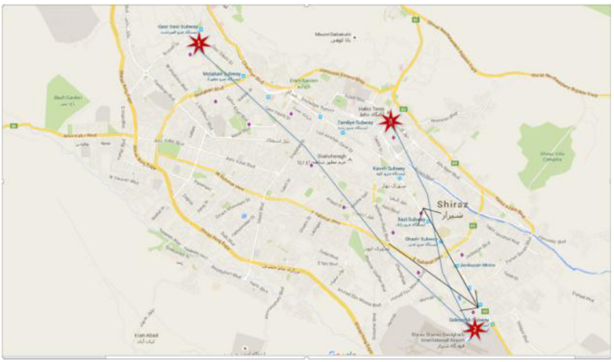

Figure 2. Significant change in the moving direction of a passenger and how our method detects this change and finds the stay location

3- Major direction change

Significant changes in overall trend and direction of a person is motivated by activities. People tend to use the 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 turns on 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 Fig.2. 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 him 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 traveler was initially moving south and at the airport he changes his direction and goes north which indicates that the traveler 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 understudy is considered to be a stay location. 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 current one. Like-wise the trend after the point is the vector that represents the direction of the line that best fits the two posterior points and the current one. The angle between these two vectors are computed and if it exceeds 130 degrees the point under study is probably a stay location.

4.3 Feature extraction to detect activity types

When a stay is detected, it means an activity is being performed during that stay. The next step would be to identify in which sort of activity is the traveler participating. Each type of activity has several characteristics that help to its recognition. For example the activity type of “work” usually happens in working hours. A pick-up/drop off activity usually takes no longer than several minutes. In order to determine the type of activities a person performs in a day, each stay should be enriched with features that contribute to identifying the type of activities during that stay. Our method proposes 8 features for a given stay. The arrival time to the stay, the departure time from the stay, duration of the stop, sequence of stay in the chain of trips, frequency of visits to that stay in the trip chain, distance of the stay to home, distance from previous stay and the dominant land use type of the location of the stay.

4.4 Assigning activity types to stays

When a stay is detected it means that the traveler has made a stop to participate in some activity. The problem here is to identify what sort of activity was the traveler doing during that stop. [5] Used adaptive boosting algorithm to infer activity types. The adaptive boosting algorithm is a set of classifiers that needs train data which we do not. We used K-mean clustering algorithm to classify activities to six groups and then manually label them to “Work”, “School”, “Shop”, “Personal business”, “Recreation” and “drop off or pick up”. All different stays of all the people in the dataset has been aggregated and each stay was assigned features that were previously extracted except the sequence of stay in the chain of trips which will later be used to find the patterns of activity participation. Features have been normalized in a way so that they can be compared to each other. After running extensive sensitivity analyses 11 seeds were inserted in the model that resulted in to 11 clusters. One for work, one for school, two for personal business, two for shopping, two for recreation, and three for pick up/ drop off.

5 Concluding remarks

The methodology presented in this paper gets the raw PLU data from cellular phone network as input and gives a series of activities that the subscriber has performed during a day as output. After finding the types and timings of activities that each citizen has participated in, the next step would be to extract patterns that are similar in the citizens’ chains of activities. This step is still understudy and will be presented in the future publication of the authors.

References

1. J.-L. Ygnace, C. Drane, Y. Yim, and R. De Lacvivier, PATH*,* *(*2000)
2. R. Cayford and T. Johnson, *Geography compass,* **8,** 49-62, (2003)
3. Y. Asakura and E. Hato,*Transportation Research Part C: Emerging Technologies,* **12**, 273-291, (2004).
4. Xu, Yang, Shih-Lung Shaw, Ziliang Zhao, Ling Yin, Zhixiang Fang, and Qingquan Li. *Transportation* ,**42**, 625-646, (2015)
5. Allahviranloo, Mahdieh, and Will Recker. *Transportation*, **42**, 561-579 ,(2015)
6. Widhalm, Peter, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C. González. ,*Transportation*, **42**, 597-623, (2015)
7. J. Shan, J. Ferreira Jr, and M.C. González. In *Int. Workshop on Urban Computing,* (2015)
8. Laasonen, Kari, Mika Raento, and Hannu Toivonen. In Pervasive Computing. Springer Berlin Heidelberg, 287-304., (2004)
9. J. Hong, H. Kim, Computer Communications and Networks, *Proceedings of 18th International Conference on. IEEE*, (2009)
10. M. Dash, K.K. Koo, T. Holleczec, G. Yap, et.al.*(MDM), 2015 16th IEEE International Conference* ,**1**, 243-250, (2015)