MATEC Web of Conferences **81**, 03009 (2016) DOI: 10.1051/matecconf/20168103009
ICTTE 2016

Route Choice Estimation Using Cell Phone Data

Homa Taghipour and Yousef Shafahi

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

Abstract. Nowadays development of cell phone network provides huge and ubiquitous data, with wide application in transportation science. One of the most important advantages of these kinds of data is enabling the process of collecting information without any active users' interference. A big data set consisting of 300,000 cell phone users' information in Shiraz are studied. This data set includes spatiotemporal information of travelers for every 5 minutes in a time span of 40 hours in two consecutive days. The spatial part of each user's information contains the position of the BTS (Base Transceiver Station) to which his cell phone is currently connected. Due to the existence of outliers, it is necessary to smooth the data initially. One of the main reasons of errors in the data set is ping pong handover, which leads to false transitions and must be eliminated. After the data preparation, stay locations are determined for each user and a trajectory for each pair of origin and destination is estimated. At this step based on network information of the city, a method to match trajectories with the network map is applied. Finally the obtained results indicate whether travelers choose the shortest path or other possible alternatives.

1 Introduction

Nowadays the advancement of technologies and cell phone network expansion can provide a massive and ubiquitous database. This comprehensive data set that belongs to the majority of people in society, yields a wide application in transportation science. Knowing the trajectories that are chosen by travelers, is really important and useful for practice in certain transportation planning steps, such as traffic assignment. According to the Wardrop's first principle of equilibrium, the travel times in all routes, which are used, are equal and also less than those routes that are not used. Utilizing this principal is completely common in traffic assignment step. However, based on observations and recent researches, this equilibrium does not necessarily occur and people sometimes decide to choose longer trajectories because of a couple of reasons such as the number of stops in a route, the quality of route roads and even the route aesthetic.

There are lots of recent researches conducted in this field, but the major deficiency existed in most previous researches is the lack of sufficient data. The main goal of this research is human trajectory estimation between a pair of origin and destination by using numerous cell phone data information and the result shows that how travelers choose their routes.

There are some kinds of methods to collect data in order to estimate travelers' trajectories. Although one of the most popular methods is using GPS database, collecting these kinds of data encounters some problems. The need for direct cooperation of travelers is one of the main problems, furthermore using of cell phones equipped with GPS is not popular among people and also GPS consumes battery charge of cell phones severely. Therefore, GPS data utilization cannot provide a data set that covers the majority of people travel information. The other common method to collect data is gathering questionnaire which has some imperfections. The high cost of collecting data, the possibility of information inaccuracies and the extremely time-consuming process of preparing data are some of the imperfections in questionnaire gathering method.

In this research frequent spatiotemporal cell phone data set is employed. One of the most important advantages of these kinds of data is enabling the process of collecting information to be accomplished passively and without any active interference of the users. Moreover, a large amount of data, which are attainable, lead to high accuracy. It is also easily possible to collect these appropriate data in a short duration. On the other side, cell phone data include some inaccuracies and errors that must be corrected.

After the data preparation, stay locations are determined for each user and a trajectory for each pair of origin and destination is estimated. At this step based on network information of the city, a method to match trajectories with the network map is applied. Finally the obtained results indicate whether travelers choose the shortest path or other possible alternatives.

2 Literature review

Researchers believe that it is possible to extract travel information from cell phone data, but accuracy and quality of the data depends on the network conditions

such as density, situation and power of BTSs (Base Transceiver Station) in the network. There are several researches which study on estimating travel information using cell phone data. In this section some of these related researches are explained briefly.

The appropriate data dramatically affect the outcome of these kinds of researches. Most researchers had access to the event-based cellular data. These kinds of data contain the special information of mobile users at the moments that they are connected to the cellular network, by making a call, sending a SMS or using internet and sometimes when their BTS zone is changed. Authors in [1] used the event-based cellular data collected by Airsage [2]. In this research the spatial information was calculated by triangulation method by means of the Airsage's Wireless Signal Extraction technology to reach more accurate and precise data. The researchers selected the information of those users who had more than 1000 connections during a given day. This selection is biased because of getting limited. In this fashion, only information of special group of society who are more active on using cell phone is acquired. The event-based data are also used in [3]-[7], etc.

Removing errors from data plays a significant role in gaining better results. Three methods to detect outliers from a set of cellular data in order to improve traffic modeling, are discussed by authors in [5]. In this research, the transition between two points is defined as an outlier, if the transition speed is more than 250 Km/h. Using the recursive naive filter, the look-ahead filter and the Kalman filter are the mentioned three methods, each of which is advisable for a special purpose.

The most challenging part of cellular data preparation is to eliminate the false result caused by ping pong handover, which is discussed in many researches. Authors in [8] introduced an algorithm to detect and remove ping pong handover by defining cell graph and cell cluster. Researchers in [9] used a moving average method to smooth out high frequency ping pong handover. Authors in [10] defined the repetitive transitions between two or three antennas as ping pong handover, put those antennas in a cluster and did not count the internal transitions. Researchers in [11] introduced a method by defining a set of recent antennas and a ping pong counter, to filter false transitions generated by ping pong phenomena. Authors in [12] applied an almost similar method to remove ping pong phenomena.

In order to determine the stay locations, each research use its own method. As an instance, authors in [6] determined the stay locations by defining the spatiotemporal threshold. In other words, in a sequence of data records, a point which is closer than one kilometer to its last point and its time stamp is more than 1 hour after its last point, is a stay location.

Author in [13] at map matching step used Hidden Markov Model to match a sequence of points to map, instead of matching them one by one, in order to increase the accuracy. Authors in [1] divided users into sedentary and commuter categories and calculated the error between real trajectory and mobile data trajectory using an interpolation method. Authors in [6] after an initial data smoothing, at map matching step used a method to detect some probable trajectories between each pair of origin and destination based on local transportation network. Afterward, authors selected the closest one to the cellular data trajectory as the correct trajectory. Authors in [4] proposed a method to estimate trajectories, using the information of high probability intermediate areas and weighting streets based on their coverage. They showed that travelers do not choose the shortest path consistently. In order to match trajectory to transportation map, authors in [14] applied a multi-path route generation algorithm to find the set of plausible trajectories and then they chose the best one by using Needleman-Wunsch algorithm. The goal of researchers in [7] was not gaining the maximum likelihood by finding the most likely trajectory. They generated set of possible trajectories based on the right probability distribution. Authors in [3] to map match the trajectories, made a set of probable transportation network trajectories by incremental assignment between each pair of origin and destination, then chose the best trajectory which had the shortest squared deviations between each point of mobile data trajectory to that network trajectory.

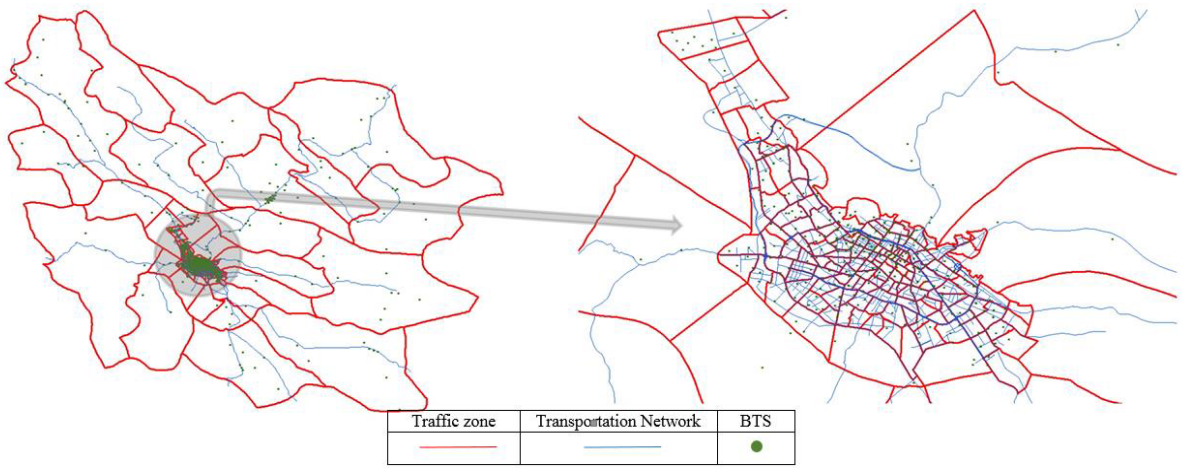

Figure 1. Shiraz traffic analysis zones, transportation network and BTSs position.

3 Data

A big data set consisting of 300,000 cell phone users' information in Shiraz city of Iran are studied to achieve travelers' trajectories between each pair of origin and destination. Shiraz is the fifth most populated city of Iran, with the population of about 1.5 million. The city divided into 183 traffic zones and there are 484 BTSs throughout the city, which are denser in the central zones. The figure 1 depict the study zone. In the most related researches usually cell phone data are recorded at the moment that a person uses his cell phone, in the other word cell phone data consist of phone events, such as phone calls, short messages and connecting to the internet. However, the data set that have been used in this research include frequent periodic spatiotemporal information of travelers for every five minutes in a time span of 40 hours in two consecutive days. The spatial part of each user's information contains the position of the BTS to which his cell phone is currently connected. Due to the existence of outliers, it is necessary to smooth the data initially. One of the main reasons of errors in the data set is caused by ping pong handover, which leads to false transitions and must be eliminated.

4 Methodology

In this section, the successive steps of estimating travelers' trajectories are discussed. The first step of this procedure is to prepare the data. After eliminating all the errors existed in the data, the stay locations are determined. Eventually, by using the BTSs trajectories, it is possible to achieve the main goal of this research. Final step is to map match the BTSs trajectories and reach the correct transportation trajectories. In the following steps, three fundamental assumptions have been applied:

1. The signaling power of all BTSs are the same as each other.
2. Each cell phone connect to the nearest BTS around it.
3. The disruptive effect of spatial objects like tall buildings on power of signaling is neglected.

4.1 Data preparation

These kinds of data are usually error prone and need to get prepared before utilization. Initially, some records which have favorable characteristics are selected. The selected data contain records of the travelers which were entirely in Shiraz city zones in the studied day and their spatiotemporal information in every five minutes of the day are available. Afterward, the raw primary data are converted to the spatial information of the selected travelers in addition to the stay duration of each location.

Commonly, cell phones are consecutively connected to some different BTSs, even if the user of that cell phone is immobile, it is called ping pong handover and this phenomena leads to false movements. At this step it seems necessary to look for a solution in order to modify the ping pong handover. It is possible to detect ping pong transitions by using suggested algorithm of Hong and Kim [11]. Based on mentioned algorithm, a transition to j is a ping pong handover if conditions (1) and (2) meet:

$$j \in Bn_u \quad (1)$$

$$pn_u = pn_{max} \quad (2)$$

When Bn_u is the set of three recent BTSs to which user "u" is connected at its n_uth transition and pn_u is ping pong counter which takes measure in what follows:

0	if	$1B_n(j) = 0$	and	$pn_{u-1} = 0$
pn_{u-1} -1	if	$1B_n(j) = 0$	and	$0 < pn_{u-1}$
pn_{u-1} +1	if	$1B_n(j) = 1$	and	$pn_{u-1} < pn_{max}$
pn_{max}	if	$1B_n(j) = 1$	and	$pn_{u-1} = pn_{max}$

The pn_{max} is the maximum value of pn_u which is equal to 2 based on this research conditions, and the indicator function of $1A_x$ is 0 if $x \notin A$ and 1 if $x \in A$.

After ping pong handover detection, those locations which the ping pong transition occurs between them, are replaced with an expected new location. The expected new location is the weighted average of initial locations, and can be calculated by weighting the locations by the stay duration of each one. Finally, at the end of this step, the BTSs trajectories of the travelers are extracted. A BTSs trajectory is the total of straight lines between each two BTSs that a traveler is connected to them.

4.2 Stay location extraction

Stay locations are the places which travelers have a stop there for a special purpose. These purposes might be going to home, work, shopping, entertainment, etc. In this section the process of determining the stay locations is discussed. One location is defined as a stay location if a traveler's cell phone remain connected to a BTS more than 20 minutes and this duration has been chosen due to the Shiraz local network conditions. Sometimes a cell phone is connected to a special BTS in two consecutive records and the distance between that BTS to its adjacent BTSs is short enough that if the traveler is moving even in walking speed, his cell phone can be connected to the other BTS easily. In these situations the stays are defined as stay locations too.

4.3 Estimationing the trajectory between a pair of origin and destination

After determining the stay locations, for each traveler, there are some stay locations as origins and destinations of his travels and some intermediate points between them. Now it is the time to estimate the network trajectory that each traveler choose to transfer from an origin to a destination.

4.4 Map matching

Conspicuously, cellular data are inaccurate and one of the most important steps to make them qualified enough to use is map matching. Some kinds of data like GPS data approximately give the information of cell phone exact location on the network, but cellular data can give just the position of the BTS to which the cell phone is connected.

So converting BTSs locations to actual locations is really necessary. There are lots of approaches in order to meet this goal. Some of them are categorized into simple search techniques such as point-to-point matching and point-to-curve matching and some of them are the complex techniques and use the theories like probability theory and fuzzy logic theory. These methods are discussed in [15]. In this paper the result of map matching process is to extract network trajectory from BTSs trajectory. The approach is to find some trajectory options initially and then choose the best one as the final transportation trajectory.

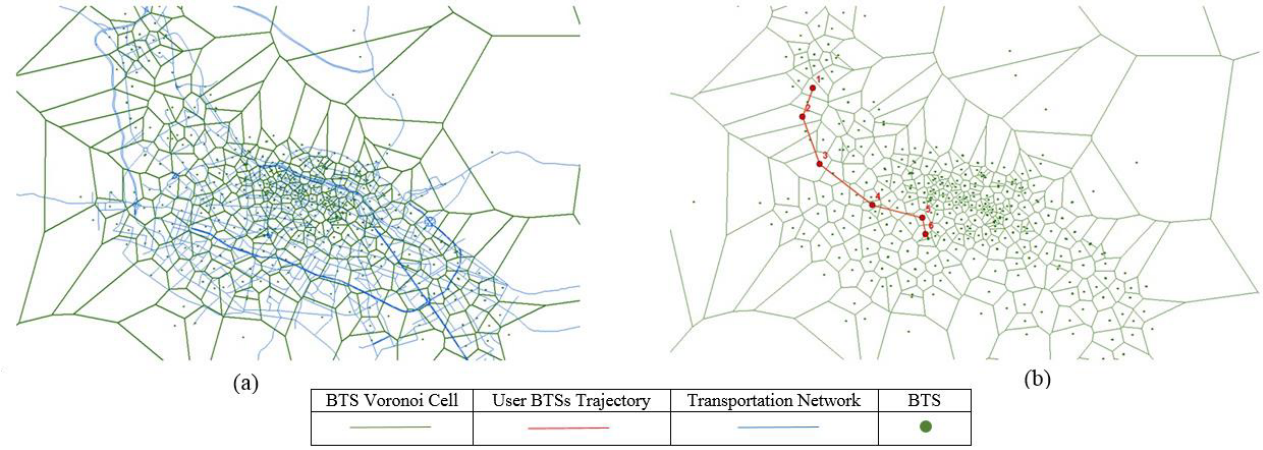

Figure 2. (a) BTSs Voronoi cells, (b) a sample user BTSs trajectory.

In order to determine some trajectory options for each pair of origin and destination, a method almost like the one discussing in [6] is applied. Firstly, it is essential to divide the whole study zone into voronoi cells using BTSs as seeds and keep these cell's characteristics in a set named "V". Voronoi cell is the region consisting of all points which are closer to its seed than the other seeds. At the next step all transportation network links of the study zone are kept in another set named "t". After preparation of these sets, it is possible to make a new set for each transportation network link, named "Vt" and put the information of all cells intersected by that link in the set. Then perform the same process for each BTSs trajectory, in other words, for each BTSs trajectory create a set named "Vb" and store the information of all cells intersected by the BTSs trajectory in that set. Hence, for each BTSs trajectory with the determined Vb set, the selected links to estimate the correct transportation trajectory, are all the t links which there is at least one common element in their Vt set and the Vb set, it means $Vt \cap Vb \neq \emptyset$.

At last, the correct transportation trajectory must be chosen from the set of candidate trajectories for each pair of origin and destination. To reach this goal, total distance between the BTSs trajectory and each trajectory options is calculated and the trajectory with the shortest distance is the correct final transportation trajectory. The total distance is the sum of distances between the intermediate points of BTSs trajectory and each transportation trajectory.

5 Result

This section illustrates the above mentioned methodology that is used to determine the transportation trajectory for one trip of a random user. After preparing data and removing all the aforesaid errors, the methodology is performed by utilizing GIS software. The divided study zone into voronoi cells is stored in the V set. The Shiraz transportation network is also kept in the t set. At this step it is possible to estimate the Vt sets for each transportation network link, using the characteristics of V and t sets, as illustrated in Figure 2.a. For a random user, a trip between two stay locations is considered, this trip comprises six points: one origin, four intermediate points and a destination. The points are obtained after the process of error eliminating, such as ping pong handover. The BTSs trajectory is indicated in Figure 2.b. The Vb set is created by the BTSs trajectory and V set.

Therefore, those transportation links which are candidate to make the set of probable trajectories, based on the methodology, are selected as shown in Figure 3.a. Finally the most acceptable transportation trajectory that is the closest one to the BTSs trajectory is determined. Figure 3.b. indicates the ultimate trajectory for the mentioned trip.

6 Conclusion

In this paper some algorithms have exploited to estimate the transportation trajectory of cellphone users. At first process of data preparation which is really essential to get a better result is discussed. After that, stay locations are determined. Then, map matching is applied to reach the correct transportation trajectory. This paper explicates the fundamental part of research and result can be used in other related researches. Finding trajectories for a group of users between a pair of origin and destination using periodic cellular data might be the subject of next researches.

An appropriate data plays a significant role to reach the goal. A repetitive, periodic and large dataset can easily increase the accuracy and preciseness of the result. Some practical applications of this research is in traffic controlling, urban planning, etc.

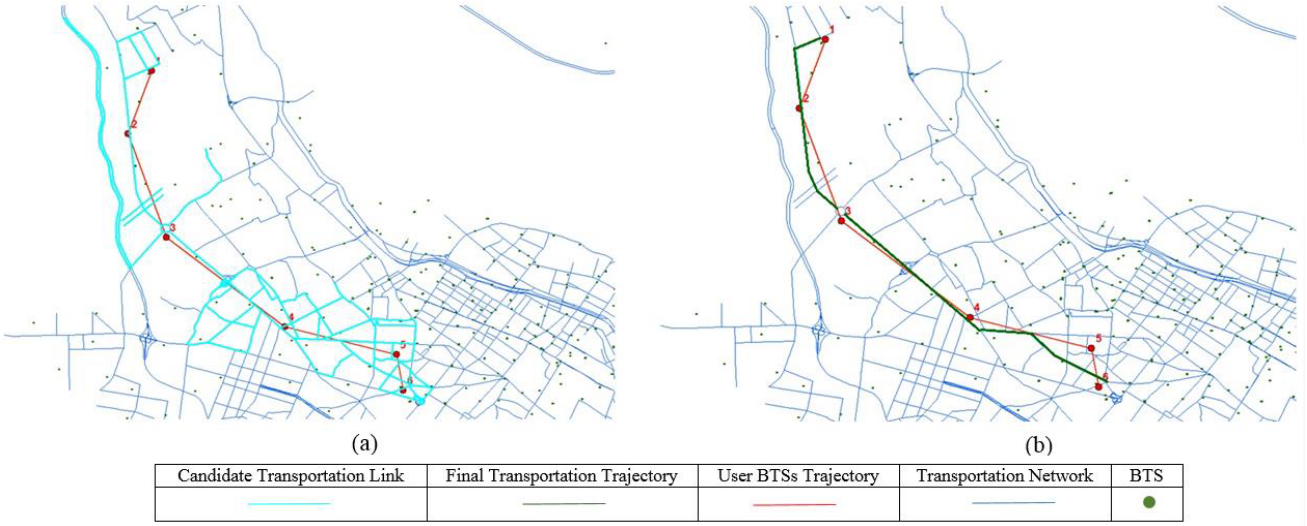

Figure 3. (a) Candidate transportation links and (b) final transportation trajectory (for sample BTSs trajectory).

References

1. S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, G. Pujolle, Comput. Netw. **64,** 296-307 (2014)
2. http://www.airsage.com
3. T. Tettamanti, H. Demeter, I. Varga, Acta polytech. Hung. **9**, 207-220 (2012)
4. I. Leontiadis, A. Lima, H. Kwak, R. Stanojevic, D. Wetherall, K. Papagiannaki, *Emerging Networking Experiments and Technologies*, 121-132 (2014)
5. C. Horn, S. Klampfl, M. Cik, T. Reiter, TRR. **2405**, 49-56 (2014)
6. M. Dash, K. Kiat Koo, T. Holleczek, G.Yap, Sh.P. Krishnaswamy, A. Shi-Nash, *Mobile Data Management*, **1**, 243-250 (2015)
7. T. Vajakas, J. Vajakas, R. Lillemets, Int. J. Geogr. Inf. Sci. **29**, 1941-1954 (2015)
8. K. Laasonen, M. Raento, H. Toivonen, Pervasive. Comput., 287-304 (2004)
9. J. Yoon, B.D. Noble, M. Liu, M. Kim, *Mobile systems, applications and services*, 177-190 (2006)
10. J. Lee, J.C. Hou. *Mobile ad hoc networking and computing*, 85-96 (2006)
11. J. Hong, H. Kim, *Computer Communications and Networks,* 1-6 (2009)
12. B. Nunes, K. Obraczka, *World of Wireless, Mobile and Multimedia Networks*, 1-6 (2011)
13. A. Thiagarajan, *PhD diss., Massachusetts Institute of Technology,* (2011)
14. J. Schlaich, T. Otterstätter, M. Friedrich, *transportation research board of the national academies*, (2010)
15. M.A. Quddus, W.Y. Ochieng, R.B. Noland, Transp. Res. Part C. Emerg. Technol. **15**, 312-328 (2007)