Travel Mode Detection Exploiting Cellular Network Data

Arash Kalatian and Yousef Shafahi

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

Abstract. There has been growing interest in exploiting cellular network data for transportation planning purposes in recent years. In this paper, we utilize these data for determining mode of travel in the city of Shiraz, Iran. Cellular data records -including location updates in 5minute time intervals- of 300,000 users from the city of Shiraz has been collected for 40 hours in three consecutive days in a cooperation with the major telecommunications service provider of the country. Depending on the density of mobile BTS's in different zones of the city, the user location can be located within an average of 200 meters. Considering data filtering and smoothing, data preparation and converting them to comprehensible traces is a large portion of the work. A novel approach to identify stay locations is proposed and implemented in this paper. Origin-Destination matrices are then created based on trips detected, which shows acceptable consistency with current O-D matrices. Finally, Travel times for all trips of a user is estimated as the main attribute for clustering. Trips between same origin and destination zones are combined together in a group. Using K-means algorithm, records within each group are the portioned in two or three clusters, based on their travel speeds. Each cluster represents a certain mode of travel; walking, public transportation or driving a private car.

1 Introduction

Studying users' travel behaviors have always been of paramount importance for transportation planners and policy makers. Transportation modal share, which is defined as the percentage of users using a given mode of transportation between a given origin and destination, is of interest as it gives insight into utilities of different transportation facilities and their changes and trends over time. For this purpose, Surveys (including questionnaires and interviews) have conventionally been conducted to gather sample data required. Being time consuming and expensive are concomitant problems of such methods [1].

Development of location-aware technologies such as body-worn sensors, Bluetooth transceivers, satellite positioning systems and mobile phones in recent decades has led to the introduction of novel alternative approaches for gathering users' data and extracting survey parameters. Determination of Origin-destination matrices, mode of travel, travel time, route choice and activity patterns are some of the main issues that have been discussed in recent years.

Among these new methods, body worn sensors and Bluetooth transceivers need additional facilities and therefore have the lowest penetration rate. GPS-based methods, although precise, requires GPS modules to be turned on, which leads to the reduction in devices' battery usage. Hence, collecting large-scale data for transportation planning purposes passively is not possible [2].

Mobile phone based methods employ GSM signal strength, fluctuation or connections to extract users' positions and movements. No infrastructure is required to obtain records as mobile phone operators already acquire such information, often as Call Detailed Records, for their own purposes. Because of the inherent inexpensiveness and high penetration rate of these methods, this paper aims to exploit cellular network data to estimate modal split. In order to achieve this goal, origin-destination matrices are also developed and travel time between different zones is estimated using the same data.

This paper is organized as follows. Section 2 is a review of previous works on the subject. The data set and the preparation process which involves handling ping-pong handovers and smoothing data are described in Section 3 and 4 respectively, while the process to distinguish stay locations from moving locations and development of origin-destination matrices is discussed in Section 5. In Section 6, travel times and speeds between different origins and destinations are estimated and by clustering, modes of travel for number of trips is identified. Conclusions and future research plans are outlined in Section 7.

2 Related works

Most of the existing works have used GPS tracking data or GSM data along with external data obtained from additional modalities such as body sensors or accelerometers to acquire users' travel information, which can only be applied to a limited number of users [3]-[5].

Among those who have utilized cellular data information, the main attitude toward solving the problem consists of clustering/classification of users' trips based on features like travel time and speed along with additional infrastructural information to infer transportation mode of travel. In [6] for instance, GSM data were merged with real-time location of buses and taxis in the city of Rome and developed a software to monitor real-time urban mobility. The addition of location of buses and taxis helps to understand their trajectory which can be used as an additional feature to detect the mode of transportation.

Another relevant example is [1], in which authors coupled anonymized call detailed records with spatial coverage of mobile network cells and route map information of main transportation modes and inferred users' trajectory and mode of transport. For validation purposes, Kernel density path generation was used.

K-means clustering algorithm and hidden Markov model were implemented to infer transportation mode in [7]. Users' activities were smoothed out by knowledge of "normal" behavior. By using five prior states and implementing Hidden Markov Model, the most likely state of the users is predicted.

While fine-grained sampling may not always be available, in [8], the authors used coarse-grained mobile phone call detailed records to infer transportation mode. Travel times between pairs of defined origins and destinations were estimated -for weekdays and weekends separately- and travelers were clustered into three subgroups using k-means algorithm: walking, public transit and driving cars. However, the variation of travel times in different times of the day has not been considered by the authors. In addition, since clustering requires sufficient data to be meaningful, trips within origins and destinations with a small number of users traveling in between, have been neglected.

3 Dataset description

Anonymized spatiotemporal data of 300,000 mobile phone users in the city of Shiraz, Iran have been collected for approximately 40 hours in three consecutive weekdays, in a cooperation with Hamrah-Aval, Iran's leading telecommunication service provider. The data set covers about 1/3 of Shiraz's mobile phone users.

Unlike Call Detailed Records Data, in which location of each user is updated when the user makes a call, sends an SMS, connects to internet or changes cell area, records for each individual user is updated in 5-minute time intervals. Thus, the data set consists of about 144 million records, each containing anonymous user ID, time of the day and coordination (Latitude and Longitude) of the transceiver the user is connected to.

For this research, 24-hour records of the second day were extracted as the main data. Other records can be used for validation purposes. In order to have a strong focus on trips within the urban area zones, users whose cell phones were never turned off or dropped signal from 6 a.m. to 9 p.m., and spent the whole period in urban areas of Shiraz were extracted. About 50,000 users fulfilled the criteria.

4 Data preparation process

After the extraction of users meeting the criteria defined previously, anomalies were detected and removed or replaced. In the second step in smoothing out data, bouncing of cells back and force between two or three base stations while the user is stationary, was handled. This phenomenon, known as ping-pong handover effect, occurs because of fluctuations in the received signal strength and causes changing in cells IDs while the user is not actually moving [9]. An algorithm introduced in [10] was modified slightly and implemented to distinguish movements due to ping-pong H/O effect from actual mobilities. The slightly modified algorithm is as follows:

Step 1- A cell graph is defined for each user with vertices of all observed GSM cells. There is an edge between two vertices if and only if a transition occurs between them.

Step 2- Any two or three vertices set a cell cluster if and only if:

A) They form a connected subgraph. This makes sure all cells in a cluster are adjacent to each other.

B) The average time spent at the cluster is larger than the sum of the individual average times spent at each cell within the cluster. This condition is met only if there is a cell fluctuation in the cluster.

Step 3- The clusters with cell(s) in common are merged together.

Step 4- Location of each cluster is represented by the weighted average of its consisting cells coordinations. The weight of each cell coordination is set according to the time spent by the user at that cell, divided by total cluster time.

The above algorithm is at first implemented for two-way ping-pong handovers, meaning that only clusters containing two cells are detected. The algorithm is repeated for the output data of the previous implementation until no significant change occurs in the cell graphs. The next stage would be the detection of three-way ping-pong handovers, which takes into account clusters including three cells. About 65% of movements (not trips) were detected as a result of ping-pong H/O effect. Which shows the necessity of replacing them with reasonable records, weighted average of corresponding cells here. Fig. 1 shows a three-way ping pong H/O effect detected by the algorithm for sample user: the user is bouncing back and force between points A, B, and C. these points are clustered together and replaced with point S, as indicated by a star sign.

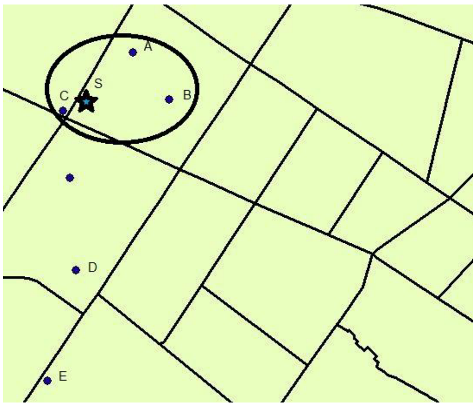

Figure 1. Ping-pong detection of a sample user.

5 Origin-destination matrices

5.1 Stay location detection

In order to recognize trips between different origins and destinations for a user, it is required to know whether a record is an activity point (or stay location, including home, work, school, shopping, etc.) or an on-the-path point (a point user passed by during her trip). For this purpose, a method is proposed as follows:

Step 1- The geographic coverage area of all Shiraz urban area cells is estimated by Voronoi tessellation method. The locations of transceivers are considered as the generating points for the tessellation. There are 544 transceiver antennas in Shiraz urban area, partitioning the urban area into 544 distinct zones. Fig. 2 depicts part of Shiraz traffic zones (thick lines) along with BTS [a] antenna's cover area (thin lines), created with ArcMap Software.

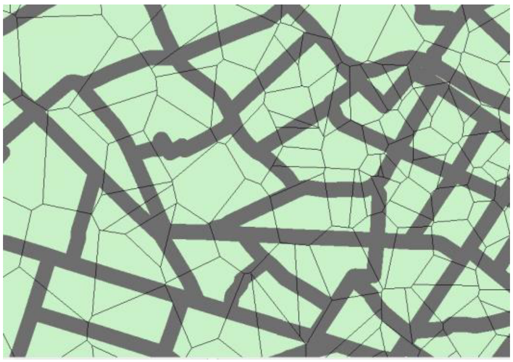

Figure 2. Part of Shiraz traffic zones (thick lines) and transceivers cover area (thin lines).

Step 2- If the largest dimension of the corresponding cell's cover area is smaller than the length the user can cover when she walks during time *t*, it can be inferred that the user has been stationary during that time. In other words, for a given user *u* that spends *t* minutes consecutively at location *l* in cover area *v*, if the following condition is true, *l* is marked as a stay location:

$$t * S_{walk} > D_v \tag{1}$$

In which S_{walk} is the walking speed, approximately 80 meters/minute according to HCM[b] and D_v is the largest diameter of cover area *v*. When a user meets this condition, even if she walks, the cover area would change. Not changing means that the user has remained still.

It must be noted that (1) is a sufficient but not necessary condition for a record to be marked as stay location. Other methods must be taken into account to identify other records moving/stationary status

Step 3- Trips are investigated with origins and destinations extracted from Step 2, trips with similar origin and destination locations are spotted and converted to two trips, with the furthest point in trajectory marked as destination and origin for each trip respectively. This will make sure that no user makes a trip without a destination, and separate origin and destination pairs are assigned to all trips.

Implementing the methods mentioned above, users' statuses can be identified. The average number of trips per day for all users is computed to be 1.81 trips/day/person (neglecting intrazonal trips), which is acceptable compared to Shiraz City Transportation Studies statistics [11].

5.2 Studying trips

Trips between different traffic analysis zones of Shiraz were aggregated to build the origin-destination matrices for different hours of the day. For each user, a stay location that the user spent at least 3 hours during night hours is marked as home. Each stay location acts as the origin for one trip and the destination for another one, except for those recognized as home.

Total number of inter-zonal trips for each hour during the day is depicted in Fig. 3. As shown in the figure, the A.M. peak hour in Shiraz occurs between 11 A.M. and 12 P.M. and the P.M peak hour occurs between 5 P.M. and 6 P.M., both consistent with previous transportation studies of the city. A relative peak can be observed between 7 A.M and 8 A.M., which is also similar to the one in studies.

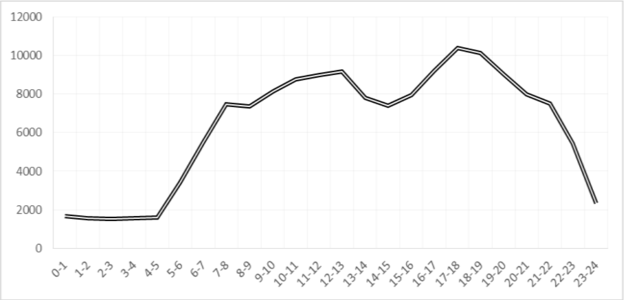

Figure 3. Total number of inter-zonal trips per hour.

6 Determining mode of travel

Our goal is to infer transportation mode of each trip by estimating speed. A first approach is to combine trips between origin *O* and destination *D* in hour *h* in one

[a] Base Transceiver Station

[b] Highway Capacity Manual, 2010

group and cluster this group into three subgroups: walking, driving a private car and public transportation (mainly buses in Shiraz). There will be a group for every origin-destination pairs in every hour. All conditions of trips within a group are similar together that travel speed can be used to determine the mode of the travel. The drawback of this approach is that not enough trips occur between any origin and destination in every hour to partition the records into clusters. The maximum number of trips between an origin and destination in the peak hour appears to be less than 100. More than 95% of origin-destination-hours do not include sufficient records to be clustered.

A second approach to solving the problem is to neglect the effect of time of the day on traffic speed and combine all the records of trips between origin *O* and destination *D* for all 24 hours together. Records from different weekdays can also be added as done in [8]. The problem with this approach is the fact that travel speed during different hours of the day varies significantly, especially in central congested zones. This approach may be applicable to trips on not congested roads.

To overcome these issues, we combined trips occurring at similar hours of the day together. Travel speeds for different modes of transport in these similar hours are assumed to be comparable. Take time period of 4 P.M to 9 P.M for instance; all trips from TAZ[c] 88 to TAZ 64 of the city of Shiraz occurring in this time period are merged together in a group. We then partition this group into clusters each representing a certain mode of travel. There is a total of 128 trips in this group.

Using K-means unsupervised learning algorithm, this group is divided based on travel speed into two clusters. The distance between these two zones is less than 2 km at the furthest points, which makes it possible for users to walk between these zones. The centers of clusters are 5km/h and 17 km/h and 29 km/h, representing walking and public transport and private car respectively. The size for each mode of transport is estimated to be 6, 29 and 93, which means the modal share of 4.5% for walking, 22.4% for public transit and 72.1% for driving. The method can be applied to several origin-destination pairs. Although there still remains a number of trips undecided, due to the insufficient number of similar trips.

7 Conclusions and future work

In this paper, we tried to convert big cellular network data to comprehensible transportation trajectories of users. After handling the inherent anomalies of real GSM data, stay locations were detected and origin-destination matrices for different hours of the day were developed. We also proposed a method to estimate modal share. Although we reached sufficient records for clustering in some origin-destination pairs, the majority of trips does not belong to these pairs and therefore their mode of travel cannot be estimated. As this paper presents some parts of an ongoing research, mode detection methods will be extended in several aspects in future works.

A proposed method being conducted for future works is to group trips similar to each other in some defined features together. These defined features are adding a feature to distinguish trips within CBD from other trips and trips between zones that public transit covers from other trips, adding economic level of origin zone, traffic congestion among available routes, etc. these features are assumed to be factors affecting mode choice and travel speed, thus, travel times and speeds of trips with similar features would be more comparable. By taking these features into consideration, a major number of trips would be covered. In the final stage, results would be compared to available transportation statistics.

Another problem that may lead to a decrease in accuracy is assigning each stay location to one TAZ. Some zones, especially in central areas of the city, are fairly small that assigning a BTS Antenna to them may be a reasonable choice. For other zones, this assumption would cause errors due to the large area. To solve this issue, some novel approaches are proposed and being worked on, defining membership functions for TAZs.

As travel speeds for public transit and private cars do not differ significantly especially in congested zones, to better distinguish between driving by vehicle and by bus, an android application is being developed for bus drivers of different bus lines of the city of Shiraz to capture the traces to identify their trajectory. The proximity of the user to these trajectories in different hours can be used as an attribute for detecting trips through public transit. Transportation mode share between different origins and destinations can be inferred thereafter using a large amount of data that can be acquired from cellular network data. Validation of the method is done by comparing the results with surveyed data collected from the city.

Acknowledgement

The authors would like to gratefully acknowledge the support of Hamrah-Avval for providing the dataset used in this research. Colleagues from Shiraz Municipality are also gratefully acknowledged for providing geographical region polygons.

References

1. J. Doyle, P. Hung, D. Kelly, S. McLoone, R. Farrell, *Irish Signals and Systems Conference*, (2011)
2. S. Timothy, A. Varshavsky, A. LaMarca, M.Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, E. De Lara. UbiComp 2006: Ubiquitous Computing, pp. 212-224, (2006)
3. S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, M. Srivastava. ACM TOSN, **6**, no. 2 (2010)
4. L. Stenneth, O. Wolfson, P. S. Yu, B. Xu. *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 54-63 (2011)
5. H. Xia, Y. Qiao, J. Jian, Y. Chang. Sensors. **14**, no. 11 (2014)

[c] Traffic Analysis Zone

6. F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, C. Ratti. IEEE ITS Transactions, **12**, no. 1, pp. 141-151 (2011)
7. I. Anderson, H. Muller, *ISWC*, pp. 127-128 (2006)
8. H. Wang, F. Calabrese, G. Di Lorenzo, C. Ratti. *IEEE ITSC,* **13**, pp. 318-323 (2010)
9. T. Tettamanti, Z.A. Milacski, A. Lorincz, I. Varga, Periodica Polytechnica. Transportation Engineering, **43**, no. 2 (2015)
10. K. Laasonen, M. Raento, H. Toivonen. Pervasive Computing, pp. 287-304 (2004)