Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records

Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, Carlo Ratti

Abstract—**Transportation mode inference is an important research direction and has many applications. Existing methods are usually based on fine-grained sampling -- collecting position data from mobile devices at high frequency. These methods can achieve high accuracy, but also incur cost and complexity in terms of the system implementation and computational resource requirements. Finally, fine-grained sampling is not always available, especially for large-scale deployment. This paper proposes a novel method to infer transportation mode based on coarse-grained call detail records. The method allows estimating the transportation mode share from a given origin to a given destination, looking also at how the share changes over time. The method can achieve acceptable accuracy with trivial cost and complexity and is suitable for the statistical analysis on transportation modes of a large population. The method can also be used as a complementary tool in situations where fine-grained sampling is unavailable or the balance between accuracy and complexity is critical. A case study using real call detail records data for the city of Boston shows the performance of the proposed method.**

I. Introduction

Mode of transportation specifies one of different kinds of transport facilities that are used to transport people, such as cars, buses, bicycles, and even walking. Transportation mode inference is a tool to determine the transportation mode of an individual traveler or a group of travelers, based on the speed, travel time or other information that can be collected from their trips. This tool has been used to provide traveling services, manage transportation and plan cities.

The research on transportation mode inference has a history of more than a decade. At early stage, the technology was studied in the field of pervasive or ubiquitous computing, where the computation needs to understand the context. The context includes human activities, such as walking or driving, when the computation is being done. Body worn sensors (sensors placed in one or more positions on the body) are the major data source; see for instance [1]. However, since body worn sensors are not widely available, many research efforts tried to adopt mobile sensors, such as mobile phone or GPS, as data collection devices. Therefore, existing methods for transportation mode inference can be divided into two categories according to the data collection devices.

A. *Mobile phone based methods*

Mobile phone based methods infer the transportation mode by estimating the mobile phone's speed. The speed can be estimated by measuring the low level signals from the GSM network, such as the variance of GSM signal strength or the switch rate of cells. Generally, the variance of signal strength and the number of connected cells are greater when the mobile phone is moving faster. When the speed is within a range, it is believed that the mobile phone user is in a specific transportation mode. Since these methods are based on the speed, they can hardly distinguish transportation modes with similar speeds, such as buses and cars. Experiments report that these methods can generally obtain 80% - 90% accuracy when inferring simple transportation modes -- remaining still, walking and driving. Intel's research in [2][3] is a relevant example. The GSM signals were measured at the frequency of one record per second, and the authors proposed a formula to calculate the GSM signal variance based on the concept of naïve Euclidean distance. They further proposed seven features and adopted ordinary data mining algorithms for transportation mode inference, reaching an average accuracy of 85%. Other similar approaches have been proposed in [4][5] introducing also a hidden Markov model for the mode inference.

In [6] the authors were able, just using WiFi signals, to infer whether the user is in motion. A hybrid approach was used in [7], which enhanced the Intel's work by integrating WiFi signals and was able to produce an average accuracy of 88%.

Transportation mode inference was also performed in real time for a massive mobile phone location dataset in Rome [8], using an average speed threshold. However, only inference during the period of the call was possible since called ID was reset every time a call ended.

B. *GPS-based methods*

GPS location data is more precise, and can be used to measure both speed and direction of an individual. Therefore, more features can be extracted from GPS data. GPS is promising to distinguish different transportation modes, even when two modes have similar speeds. Microsoft's research work [9]-[11] is a typical example. The GPS data is collected at a frequency of one record every two seconds. Firstly, the method determines walking segments – segments of path in which a user is only walking – based on the instantaneous speed and acceleration measured from GPS at each sampling point. Then, the method determines the

Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo and Carlo Ratti are with Senseable City Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. (e-mails: huayongw@mit.edu, fcalabre@mit.edu, giusy@mit.edu, ratti@mit.edu)

transportation mode in non-walking segments. The authors believe that it is generally difficult to change direction when people are driving or taking bus. So, the frequency of direction change is used as a feature to distinguish car/bus and walking/cycling. Authors also assume that people are likely to stop more times when on a bus than driving. Then, the frequency of stops is taken as another feature. Similarly, speed change rate and variance of speed are also features, for a total of nine. After feature extraction, some ordinary data mining algorithms, such as SVM and decision tree, are used to infer transportation mode. Similar research but using different features have also been developed in [12]-[17].

II. Problem Statement And Overview Of Our Method

In real world scenarios, large scale and frequent sampling from mobile devices may be unavailable. For example, mobile phones with GPS still cover a minor portion (less than 10% in 2011) of the market in China, although the market share growth is strong [18]. And users usually turn off the GPS modules to save power. On the other hand, large volume of data about the position of mobile phones can be collected from CDRs – Call Detail Records. This data is cheap because CDRs are already produced by the charging system of the telecom infrastructure when users make phone calls, send/receive messages/emails or browse web pages. To analyze CDRs and extract meaningful information brings no extra overhead for both mobile phone users and telecom operators.

We propose a method in which trip information (user id, origin, destination, start time, end time) is extracted from CDRs. Furthermore, based on a large amount of trip data that we can collect from the CDRs, we can measure how many travelers have moved from a same origin to a same destination, their starting time and travel times. This information inspires us to consider the possibility to infer travelers' transportation modes based on these travel times, and so evaluate what is the share of people who travel with a give transportation facility. The problem that we try to solve in this paper can be stated as following:

Transportation mode share inference problem: Given an origin and a destination, as well as the travel times of a group of travelers who move from the origin to the destination, infer the percentage of travelers using a given transportation facility.

This problem is of interest for transportation planning because it allows understanding the number and percentage of people moving between areas in the city using different transportation modes (and how this percentage changes over time). This could for instance complement Origin-Destination flows information (see, for instance [19]) with information about the percentage of people using a given transportation mode by moving from a given origin to a given destination.

Our approach to solve the problem can be summarized as follows. Fig. 1 shows the travel time distribution of an imaginary group of travelers. The travel times are not evenly distributed. The numbers of travelers in some travel time ranges are significantly more than other ranges. Therefore, the travelers can be clustered into subgroups according to the density. The three subgroups should correspond to different transportation modes (e.g. car, public transit, walking) in the real world.

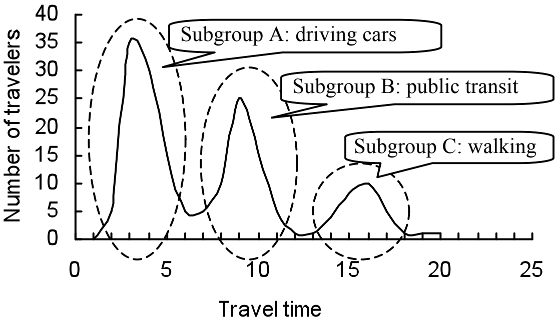


Fig. 1. Travel time distribution in an imaginary group of travelers. The travelers are clustered to three subgroups according to their travel times.

This method has several advantages:

1) It is novel in the sense that it does not rely on the frequent sampling on mobile devices. To the best of the authors' knowledge, no literature has provided research results about transportation mode inference only based on travel times.

2) It is cheap in the sense that it uses existing data generated by telecom infrastructures. No modification to mobile phones or telecom equipments is required.

3) It has good scalability so it is suitable for statistical analysis on a large population, which is desirable for city planning and transportation management.

To make a fair comparison, we admit that there are also some limitations for this method:

1) The CDRs are not a complete dataset since there is no CDR if traveler's mobile phone is turned off. But large volume of CDRs data can compensate the incompleteness.

2) This data collection method usually has low sampling frequency. So, it is quite coarse-grained if compared with the sampling frequency in existing research work.

III. Method Details

A. Data Set Description

The dataset used in this paper consists of anonymous cellular phone signaling data collected by AirSage (http://www.airsage.com), which turns this signaling data into anonymous locations. The dataset consists of 829 millions of anonymous location estimations – latitude and longitude – from close to 1 million devices in 1 month, which are generated each time the device connects to the cellular network. The location information is estimated through the AirSage's Wireless Signal Extraction (WiSE) technology, which aggregates, anonymizes and analyzes signaling data from cellular networks, and determines location information. A longer description of this data is available in [20].

B. Determining trips

To infer transportation mode for the travels, we first have to estimate the origins and destinations of trips that people make. A first approach for the trip estimation is to consider a trip as a path between user's positions at consecutive network connections and calculate the length as the distance between those points. This approach was used in [21] even if only the cell phone tower location for each network connection was available, so coarser grain spatial resolution. The drawback of this approach is that we can detect several very short trips due to localization errors and users making consecutive network connections in the same area. An example of this problem is shown in Fig. 2, where the user does not move in the time interval 0 - 3 and 3.2 - 5.5, but with this first approach we would detect more than 20 very short trips that are just due to the localization errors.

Since these fictitious trips could drastically modify the trip length distribution, we propose a second approach for which we manipulate the data applying the same methodology used for analyzing GPS [22][23]. In fact, we can consider our data as GPS traces due to the fact that Airsage provides us with the estimated position (and associated uncertainty) and not just the cell phone tower the phone is connected to.

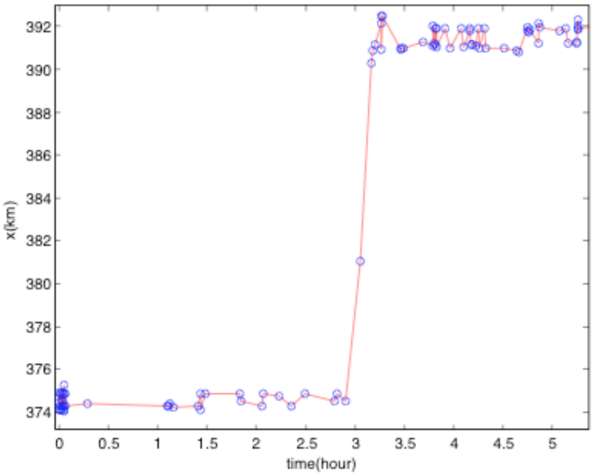


Fig. 2. Example of individual measurements and localization errors. The location changes even when the user remains still, which leads to fictitious trips.

For each user we have a location estimation measurement $m_i \in M$ each time his device connects to the cellular network. Each location measurement $m_i \in M$ is characterized by a position p_{m_i}, expressed in latitude and longitude, and a timestamp t_{m_i}. The locations measurements of each user are then connected into a sequence $M_n = \{m_1, m_2, \cdots, m_n\}$ according to their timestamp.

The methodology used to extract trips is composed of the following steps.
Given a sequence of locations measurements M_n,

1) We infer virtual locations by grouping consecutive locations measurements $\mathrm{M_S = m_q, m_{q+1}, \cdots, m_z \in M_n}$ where $\Delta \mathrm{T = t_{m_z} - t_{m_q} > 0}$ and

$\max \text{ distance} (\mathrm{p_{m_i}, p_{m_j}}) < \Delta \mathrm{S} \quad \forall\ \mathrm{q \le i, j \le z}.$

The spatial threshold ΔS has been defined as 1km, to take into account the localization errors estimated by Airsage.
2) The points $\mathrm{M_S = m_q, m_{q+1}, \cdots, m_z}$ are fused together so that a single geographic region:

$$\mathrm{p_s = (z - q)^{-1} \sum_{i=q}^{i=z} p_{m_i}}$$

This location becomes the origin or destination of a trip.
3) Once the virtual locations are detected, we can evaluate the trips as paths between user's positions at consecutive virtual locations.

C. Dataset subsampling

Given the massive dataset, we decided to subsample it by considering only mobile phone users living in the Middlesex County, and making calls with frequency greater than 1 per hour. The reason of this subsampling resides in the fact that we are interested in users making many calls in a day, so to be able to infer trips they make, and travel times. This results in 9,154,042 trip records for 56,715 mobile phone users. Each user has 6.2 trip records per day on average. Figure 3 and 4 show the spatial and length distribution of inferred trips. In particular, the trips length distribution was computed by counting the number of trips having a defined trip length.

Fig. 3. Spatial distribution of trips in the Boston downtown (green lines). Yellow lines show trips that happen at least 50 times. Blue lines show the subway system.

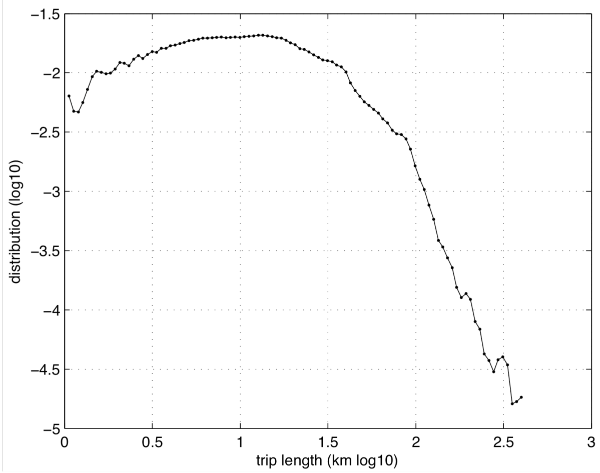


Fig. 4. Trip length distribution of the trip data.

Because of the coarse-granularity of the sampling over time (only when people make phone calls), the data set cannot be used to infer transportation modes for very short trips. For this reason, we concentrated on long trips above 3 km.

D. *Trip Data Grouping*

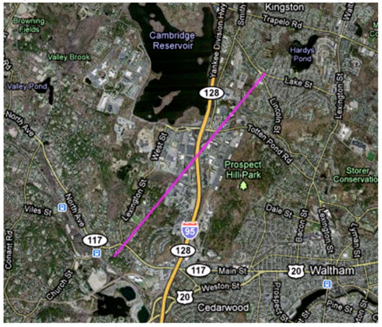
We grouped all trip records according to the same origin and destination. The same means two locations are in a same cell of a 500 x 500m cells grid. The choice of 500m comes from our localization error being 350m We remove all groups that contain less than ***k*** records (minimum group size), and denote with ***f***(***k***) the number of the associated groups. Table I shows the number of groups available for different group sizes.

TABLE I
NUMBER OF GROUPS WITH DIFFERENT GROUP SIZE

	k = 25	***k*** = 50	***k*** = 75	***k*** = 100
f(***k***)	860	107	30	7

For each group, we also label the records as weekday records or weekend records according to the date when the records are collected.

To give an example, Fig. 5 shows the origin and destination of one group (our running example). The position of the origin and the destination are (lat: 42.41, long: -71.25) and (lat: 42.38, long: -71.28) respectively.



Origin (86573)
lat: 42.40552199
long: -71.24900864

Destination (84168)
lat: 42.37862317
long: -71.27895031

Google Map reports:
by car: 9 mins
by public transit: 47 mins
by walking: 63 mins

Fig. 5. Origin and destination of a group of travelers, and the travel times reported from Google Maps.

Figs. 6, 7 and 8 show the travel time distribution in this group at weekdays, weekends and both. At weekdays, there are 20 travelers who spend 10 minutes traveling from the origin to the destination. While at weekends, there are only 7 travelers whose travel time is 10 minutes. Fig. 8 is the sum of the Fig 6 and Fig 7.

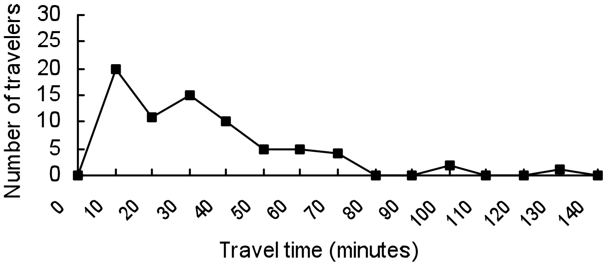


Fig. 6. Travel time distribution in a group (week day).

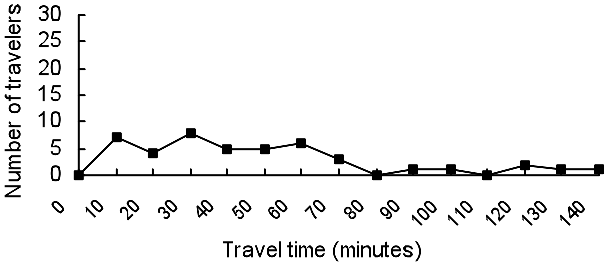


Fig. 7. Travel time distribution in a group (weekend).

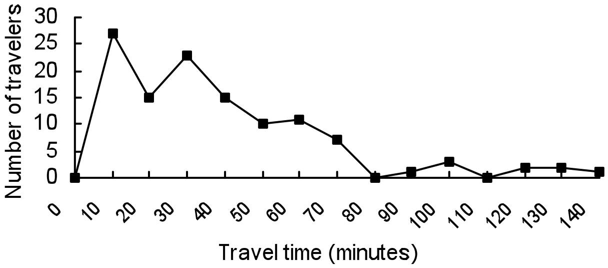


Fig. 8. Travel time distribution in a group (weekday + weekend). This figure is the sum of Figs. 6 and 7.

E. *Downloading Travel Time*

Since Google Maps (http://maps.google.com) provides reliable travel time information, we use it as reference to verify our inference results. It is also interesting to note that the available travel times for public transit change over time as function of the schedules. For the example in Fig. 5, Google Maps reports the travel time for driving is 9 minutes, for public transit is 47 minutes and for walking is 63 minutes (computed over a weekday 9am). Moreover, the travel length by driving is 5.5 km, by walking is 5.1 km.

F. *Noise Reduction*

Since it is a rare case for a traveler to walk more than 3 km (see for instance the area shown in Fig. 5), we hypothesize that the records whose travel times are larger than 63 minutes (walking travel time computed through Google Maps) are due to noise in the data (either localization or sampling errors). We then remove all records whose travel times are more than 63 minutes, shown in Fig. 9. After noise reduction, the data look like Fig. 10.

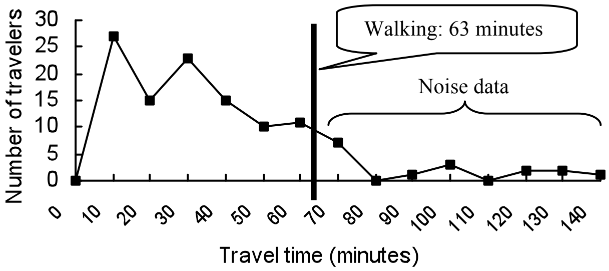


Fig. 9. Noise data in a group. Noise data refer to the trip data records that have travel times longer than the walking time.

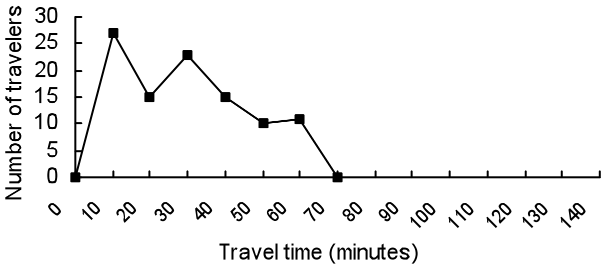


Fig. 10. Trip data after noise reduction. Compared with Fig. 9, trip data records that have travels time longer than 63 minutes are removed.

G. Data Clustering

We use the k-means unsupervised clustering algorithm to partition the records of a Group into two non-overlapping subgroups. This approach partitions data such that each observation is as much like its own group's members, and unlike other groups' members, as possible. We cluster the records in two groups that for us represent trip made by driving and using public transit.

For each subgroup we then compute the average travel time. Considering the running example, as shown in Fig. 11, the average travel time of the first subgroup is 13.4 minutes, while the average travel time of the second subgroup is 42.4 minutes. Fig. 11 also marks the travel times from Google Maps. The driving time is 9 minutes while the public transit travel time is 47 minutes. We define the error of transportation mode inference as the average of the differences between average travel times obtained from k-means and the travel times reported from Google Maps. In this case, the error is (| 13.4 – 9 | + | 42.4 – 47 |) / 2 = 4.5 minutes, which seems acceptable compared to the big variation of observed travel times.

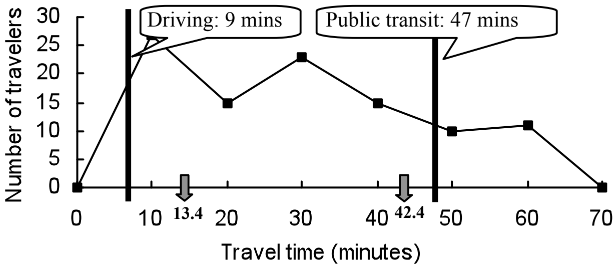


Fig. 11. Trip data clustered to two subgroups. The average travel times of the subgroups are 13.4 and 42.4 minutes. The travel times reported by Google Maps for driving and public transit are 9 and 47 minutes.

After data clustering, we can finally infer the percentage of travelers in each subgroup, which corresponds to the percentage of travelers using each transportation mode (driving and public transit).

IV. Performance Evaluation

We evaluate the performance of our method according to various metrics.

A. Comparison with Google Maps

In this section, we study the groups with minimum size 100 (k = 100). We have done the clustering experiments for the 7 groups. Considering the weekday and weekend may have different road conditions, the experiments are done separately for weekday records and weekend records. The results are shown in Table II and III.

TABLE II
K-means Error of Each Group in Weekdays (Seconds)

Group ID	Google Map			k-means		Error	Error / Walking Time
	Driving	Public Transit	Walking	Driving	Public Transit		
1	420	1740	3120	998	2323	580	18.6%
2	480	2100	3180	1044	2190	327	10.3%
3	420	2460	3780	690	2469	139	3.7%
4	600	2880	7740	1001	3155	338	4.4%
5	540	1740	3420	805	2206	366	10.7%
6	540	2820	3780	705	2358	314	8.3%
7	420	2280	3000	536	2178	109	3.6%
Average value						**310**	8.5%

TABLE III
K-means Error of Each Group in Weekends (Seconds)

Group ID	Google Map			k-means		Error	Error / Walking Time
	Driving	Public Transit	Walking	Driving	Public Transit		
1	420	2400	3180	1110	2412	351	11.0%
2	480	1500	3180	1049	2264	666	20.9%
3	360	2580	3780	671	2095	398	10.5%
4	600	2820	7620	843	3450	436	5.7%
5	540	1680	3480	688	1999	233	6.7%
6	540	2700	3780	978	2794	266	7.0%
7	420	2280	3060	363	1688	325	10.6%
Average value						**382**	10.3%

From these two tables, we can observe that the average errors of our method are about 5~6 minutes. The error of weekend records is higher than the weekday records. The reason is that the number of weekend records is smaller than the number of the weekday records in each group.

B. Validation by Sihouette

Another important aspect in all clustering techniques is to verify whether the elements of a cluster are well associated with the representative of the cluster. This is computed through the silhouette value [24]. Values close to 1 correspond to good associations. Values close to -1 instead are symptoms of bad clustering. Results for the 7 groups are shown in Table IV, and confirm the good performance of the k-means algorithm.

TABLE IV
Silhouette of K-means Clustering in Each Group

Group ID	Week days	Week ends	Average
1	0.666	0.654	0.658
2	0.557	0.593	0.584
3	0.651	0.583	0.624
4	0.687	0.728	0.691
5	0.631	0.636	0.626
6	0.595	0.617	0.595
7	0.675	0.749	0.674

C. Percentage of Travelers in Each Mode

In order to further verify the accuracy of our method, we compare the percentage of each transportation mode inferred by our method with other surveyed data for the city of Boston. While it is quite difficult to find local statistics on transportation mode choice (global statistics from the National Household Travel Survey (http://nhts.ornl.gov) are not adequate for comparison since sampled over the whole nation, and so do not consider the very good public transportation available in Boston), we have been able to find self-reports from people living in Boston, available on the online website Carpoolworld (http://www.carpoolworld.com). These reports tell us, for the first 3 months of 2010, that on average 45% of the trips of 110 surveyed individuals are made using public transportation. Our method instead predicts an average share of public transportation use of 38.1%. The two percentages are very close and the small difference could be explained by the slightly biased comparison dataset (people more conscious about efficiency and CO_2 saved, as mentioned in the website).

D. *Sensitivity to Group Size*

The number of groups increases if the group size k decreases. Fig. 12 shows the results of clustering experiment for 4 values of k. It is clear that the error of our method decreases when the group size increases. In order to keep the error less than 5~6 minutes, the group size k should be larger than 100.

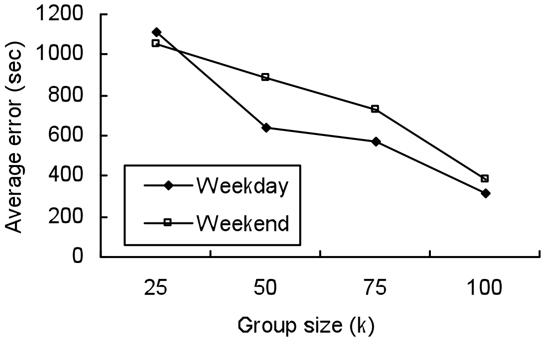


Fig. 12. Impact of group size to accuracy. The accuracy increases as the group size increases.

V. Conclusion

This paper proposed a method to infer transportation mode from CDR data. For a given origin and destination the method can determine the percentage of travelers using each transportation facility starting from their travel times. Experiments of the method using a real mobility dataset are performed and comparisons with travel times from Google Maps show promising results. The method can be easily implemented and applied in real world and for large populations, so could be a suitable candidate for augmenting existing transportation datasets used for city planning and transportation management.

Acknowledgment

The authors would like to thank Airsage for providing the data used for the experiments, and the colleagues at MIT Senseable City Lab for their useful comments.

References

[1] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In Proc. of the 2nd International Conference on Pervasive Computing. pp. 1-17, 2004.

[2] Evan Welbourne, Jonathan Lester, Anthony LaMarca, *et al.* Mobile context inference using low-cost sensors. In Proc. of the International Workshop on Location and Context-Awareness (LoCA). pp. 254-263, 2005.

[3] Timothy Sohn, Alex Varshavsky, Anthony LaMarca, *et al.* Mobility detection using everyday GSM traces. In Proc. of the 8th International Conference on Ubiquitous Computing. pp. 212-224, 2006.

[4] Ian Anderson and Henk Muller. Practical context awareness for GSM cell phones. In Proc. of the 10th International Symposium on Wearable Computers. pp. 127-128, 2006.

[5] Ian Anderson and Henk Muller. Context awareness via GSM signal strength fluctuation. In Adjunct Proc. of the 4th Conference on Pervasive Computing. pp. 27-31, 2006

[6] John Krumm and Eric Horvitz. Locadio: inferring motion and location from Wi-Fi signal strengths. In Proc. of the 1st Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services. pp. 4-13, 2004

[7] Mun Y. Mun, Deborah Estrin, Jeff Burke, *et al.* Parsimonious mobility classification using GSM and WiFi traces. In Proc. of the 5th Workshop on Embedded Networked Sensors. 2008.

[8] Francesco Calabrese and Carlo Ratti. Real time Rome. Networks and Communications Studies, 20(3-4), pp. 247-258, 2006.

[9] Yu Zheng, Quannan Li, Yukun Chen, *et al.* Understanding mobility based on GPS data. In Proc. of the 10th International Conference on Ubiquitous Computing. pp. 312-321, 2008.

[10] Yu Zheng, Like Liu, Longhao Wang, *et al.* Learning transportation mode from raw GPS data for geographic applications on the Web. In Proc. of the 17th International Conference on World Wide Web. pp. 247-256, 2008.

[11] Yu Zheng, Yukun Chen, Quannan Li, *et al.* Understanding transportation modes based on GPS data for web applications. ACM Transactions on the Web. 4(1), 2010.

[12] Donald J. Patterson, Lin Liao, Dieter Fox, et al. Inferring high-level behavior from low-level sensors. In Proc. of the 5th International Conference on Ubiquitous Computing. 2003.

[13] Young-Ji Byon, Amer Shalaby and Baher Abdulhai. Travel time collection and traffic monitoring via GPS technologies. In Proc. of IEEE Intelligent Transportation Systems Conference. pp. 677-682, 2006.

[14] Sasank Reddy, Jeff Burke, Deborah Estrin, *et al.* Determining transportation mode on mobile phones. In Proc. of International Symposium on Wearable Computers. pp. 25-28, 2008.

[15] Sasank Reddy, Min Mun, Jeff Burke, *et al.* Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN), 6(2), 2010.

[16] Cenceme project. http://cenceme.org/index.html

[17] MetroSense project. http://metrosense.cs.dartmouth.edu

[18] Prediction on the volume of mobile phones with GPS in Chinese market. http://it.sohu.com/20070708/n250952491.shtml

[19] US department of transportation. Census transportation planning products (CTPP). http://www.fhwa.dot.gov/ctpp, 2010.

[20] Francesco Calabrese, Francisco C. Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In Proc. of International Conference on Pervasive Computing, 2010.

[21] Marta C. Gonzalez, Cesar A. Hidalgo, Albert-Laszlo Barabasi. Understanding individual human mobility patterns. Nature 453 (7196), pp. 779-782, 2008.

[22] John Krumm. Real time destination prediction based on efficient routes. In: Society of Automotive Engineers (SAE) 2006 World Congress, 2006.

[23] Quannan Li, Yu Zheng, *et al.* Mining user similarity based on location history. In Proc. of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2008

[24] Peter Rousseuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Computational and Applied Mathematics, 20(1), pp. 53-65, 1987