AN OPTIMIZATION FRAMEWORK FOR TRAVEL PATTERN INTERPRETATION OF CELLULAR DATA

Sarit Freund *

Department of Industrial Engineering and Management,
Ben-Gurion University, Be'er-Sheva, Israel

Hillel Bar-Gera

Department of Industrial Engineering and Management,
Ben-Gurion University, Be'er-Sheva, Israel

***** Email: sarit.freund@gmail.com

Collection of travel data by traditional survey methods is costly and time consuming, thus limiting the amount of data being collected, as well as collection frequency and coverage. For many years the main method for collecting data on travel patterns has been household travel surveys. A major challenge in travel surveys, especially in recent years, is non- responsiveness [1, 2, 3]. In the last U.S. NHTS (National Household Travel Survey), conducted from March 2008 through May 2009, response rate was 19.8% [4].

Advanced technologies offer new types of data collection options. In particular, cellular systems generate substantial amounts of data, including records regarding the connection between handsets (phones) and base stations (antenna) [5, 6, 7, 8]. These records, collected by cellular service providers for various internal purposes, may provide an excellent source of information regarding travel, with several critical advantages relative to traditional travel surveys: low cost, large sample, long duration, and high response rate. Cellular data has some limitations too, particularly with respect to accuracy and traveler identity; therefore, such data cannot provide a complete replacement for traditional surveys, but it can complement and enhance them.

The dataset used in this study includes information regarding 9454 user cellular handsets, containing 3.7 M records, collected over 132 hours (nearly one week). For privacy reasons, the dataset includes an arbitrary handset ID, and not the real cellular number. By analyzing this database, one can know the location of the base station that communicated with the handset at the stated time, which in turn gives a general idea on the user's location. Base stations are located several hundred meters from each other in densely populated areas, and few kilometers apart in rural areas.

In a weekly perspective it is possible to observe substantial movements, when the handset user was probably travelling, and periods of minor base station location changes, when the handset user was probably stationary, participating in a certain activity.

The main challenge in this research is to distinguish between these two, periods when the user is staying in one place (activity) and periods when the user is moving (travel). To differentiate between activity intervals and travel intervals, various approaches can be utilized. A typical procedure is to rely on measurement of geographical distance between two consecutive observations compared to a predefined threshold.

The current research aims to develop a more holistic approach, which is based on a definition of an objective function that is used for grading and comparing alternative interpretations of the raw data. An interpretation of a single handset-user raw data is a specification for each time interval between two records [T1, T2] whether the user was in activity or travelling during that interval. The grading function for each interpretation examines the entire interpretation as a whole. The aim is to define an objective (grading) function in such a way that more plausible interpretations will get higher scores. We focused our attention on objective functions that grade sequences of activity or travel intervals, where the overall score is the sum of these components.

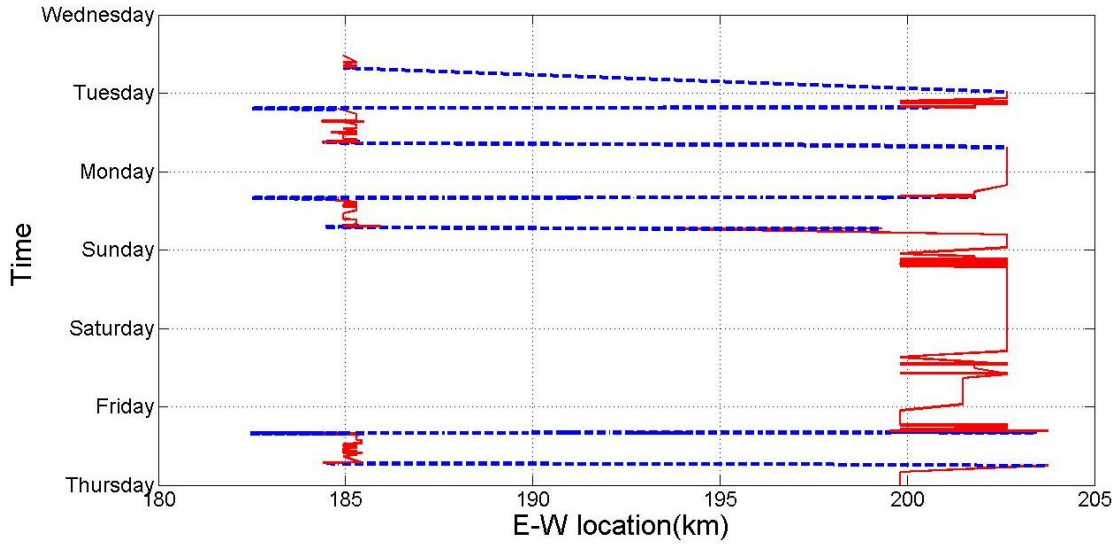
Activity is defined here as staying in the same location over a period of time. Considering this definition, two parameters are being checked to identify activities, location and duration. The grading function relies on these two parameters, but taking into consideration special attention to short activities; Identifying short periods of time as activities can lead to misunderstanding of the user's behavior. Such sequences can stem from slow travel or waiting in a traffic jam. If a handset is nearly stationary for a relatively long period of time, the probability that the user was indeed performing a real activity is higher. Therefore short activities are questioned, and require thorough inspection.

Travel is defined as a movement from an origin point to a geographically separated destination point. The travel grading relies on the distance between origin and destination points, calculated as the distance between the first and the last records (as the crow flies distance).

Once grading function definitions have been determined, the best-value interpretation of the data from a specific user is sought, for example by a genetic algorithm. The genetic algorithm process starts with creating a chromosome population in which each member describes a possible interpretation for the user's behavior. After grading all current interpretations (i.e. "chromosomes"), those individuals with highest score procreate and generate a new and improved generation, and so on, until a nearly-optimal solution is achieved.

The algorithm classification for travel or activity was compared with a manual classification for a 780 intervals chromosome. Figure 1 describes the handset location throughout the sampling period (E-W coordinates only), solid thin red sections describe intervals interpreted as activity and dashed thick blue sections describe intervals interpreted as travel.

(A)



(B)

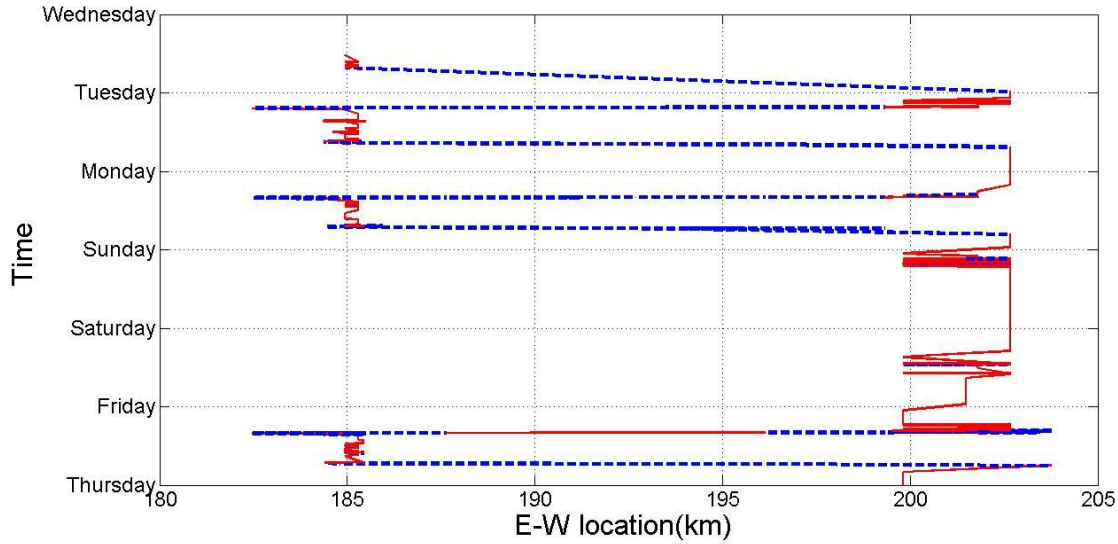


FIGURE 1 - Comparing manual (A) and genetic algorithm (B) classification of travel and activity intervals.

The particular handset user presented in Figure 1 had two main activity locations: in the first she stayed during mornings and afternoons on Thursday, Sunday, Monday and Tuesday; and in the second she stayed during nighttime and over the weekend.

The algorithm correctly identified the main travel intervals, and most of the activities. The difference between the two classifications presented in Figure 5 includes nine sequences with a total duration of 303 minutes, i.e. 3.8% of the total time covered by the data.

Comparison between the two highest scoring interpretations produced by the algorithm for a different handset shows that discrepancies between the two interpretations occur in 18% of 830 intervals, representing 215 minutes (3%) of the total recorded time.

To get a broader picture of the data, the algorithm was applied to all handsets with less than 1500 samples (totaling 9272 handsets), and for each one the highest scoring interpretation was selected.

The application of the algorithm to all handsets suggests average number of activities per handset of 11.7, i.e. about two activities per day. Figures 2&3 demonstrate the cumulative distribution of travel distances and activity durations (logarithmic vertical axis). 20% of the activities are shorter than 15 minutes, and 24% are between 30 minutes to one hour. Intervals identified as activities with average distance between base-stations larger than the set threshold (1500), representing 1.26% of the total sample time, are not displayed in the figure.

On average 13.14% of the intervals have different classification between the two highest scoring interpretations, representing 392 minutes per handset or 4.31% of the total recorded time.

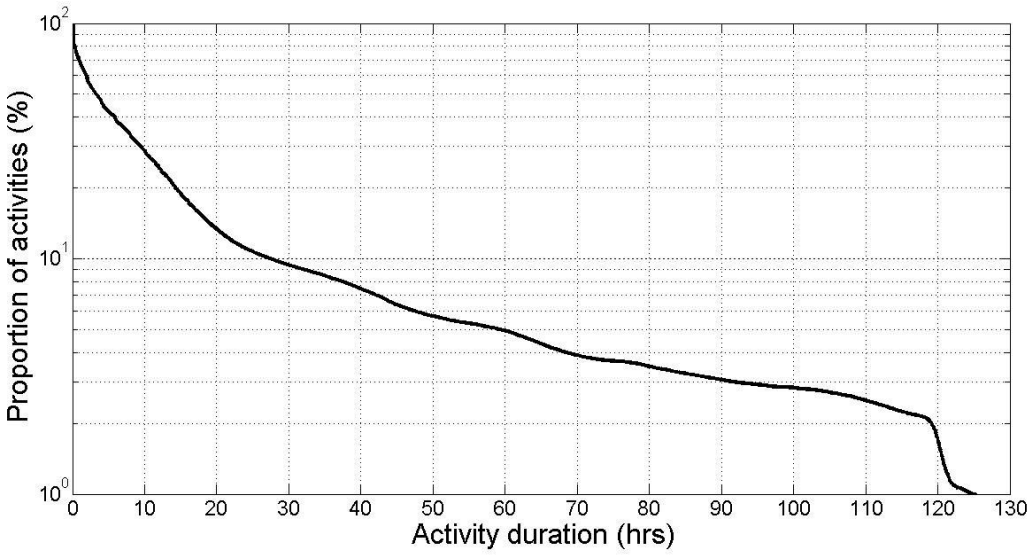


FIGURE 2 - Cumulative distribution of activity duration

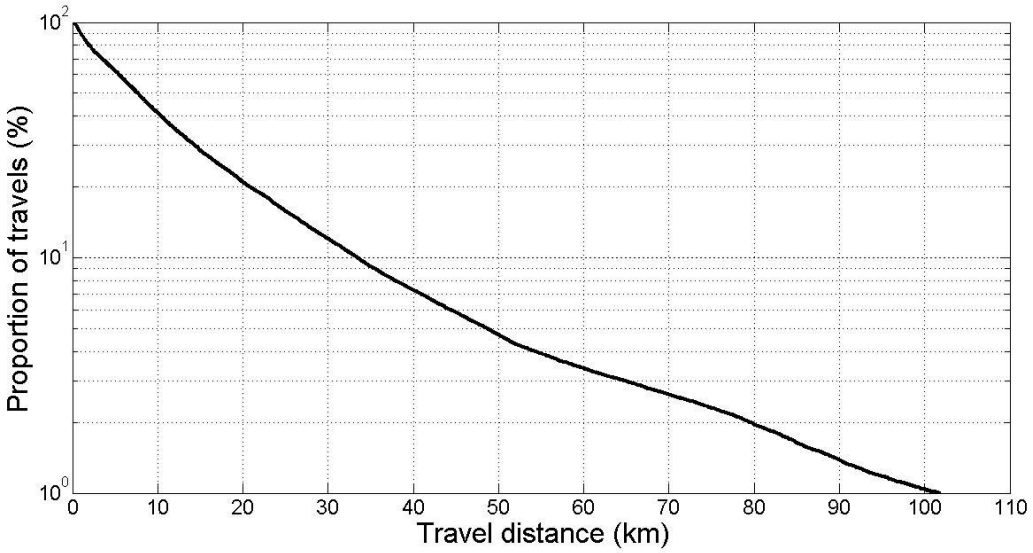


FIGURE 3 - Cumulative distribution of travel distance

References

[1]. Keeter, S., C. Miller, A. Kohut, R. M. Groves, and S. Presser. Consequences of Reducing Nonresponse in a National Telephone Survey*. *Public Opinion Quarterly*, Vol. 64, No. 2, 2000, pp. 125-148.

[2]. Moritz, G., and W. Brög. Redesign of the Dutch Travel Survey: Response Improvement. *Voorburg Heerlen*, 1999, pp. 7.

[3]. Kalfs, N., and H. van Evert. Nonresponse in Travel Surveys. *Transport Survey Quality and Innovation*, 2003, pp. 567-585.

[4]. Hu, P. S., T. Reuscher, R. L. Schmoyer, and S. Chin. 2009 National Household Travel Survey. *U.S. Department of Transportation Federal Highway Administration, Washington, DC.*, 2010.

[5]. Ahas, R., A. Aasa, A. Roose, Ü. Mark, and S. Silm. Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study. *Tourism Management*, Vol. 29, No. 3, 2008, pp. 469-486.

[6]. Ahas, R., A. Aasa, S. Silm, and M. Tiru. Daily Rhythms of Suburban Commuters' Movements in the Tallinn Metropolitan Area: Case Study with Mobile Positioning Data. *Transportation Research Part C*, 2009.

[7].Friedrich, M., P. Jehlicka, T. Otterstätter, and J. Schlaich. Monitoring Travel Behaviour and Service Quality in Networks with Floating Phone Data. In *International Conference on Transportation Decision Making: Issues, Tools, Models and Case Studies, Venice, Italy*, 2008.

8. Gur, Y. J., S. Bekhor, C. Solomon, and L. Kheifits. Intercity Person Trip Tables for Nationwide Transportation Planning in Israel obtained from Massive Cell Phone Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2121, No. -1, 2009, pp. 145-151.