Transportation Research Part C 46 (2014) 326–337

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

From traces to trajectories: How well can we guess activity locations from mobile phone traces?

Cynthia Chen [a,*], Ling Bian [b,1], Jingtao Ma [c,2]

[a] Department of Civil and Environmental Engineering, Box 352700, University of Washington, Seattle, WA 98195, United States
[b] Department of Geography, University at Buffalo, State University of New York, 105 Wilkeson Quad, Amherst, NY 14261, United States
[c] Traffic Technology Solutions, LLC, 17933 NW Evergreen Pkwy, Suite 121, Beaverton, OR 97006, United States

ARTICLE INFO

Article history:
Received 19 October 2013
Received in revised form 16 May 2014
Accepted 7 July 2014

Keywords:
Household travel survey
Mobile phone datasets
Activity locations
Clustering

ABSTRACT

Passively generated mobile phone dataset is emerging as a new data source for research in human mobility patterns. Information on individuals' trajectories is not directly available from such data; they must be inferred. Many questions remain in terms how well we can capture human mobility patterns from these datasets. Only one study has compared the results from a mobile phone dataset to those from the National Household Travel Survey (NHTS), though the comparison is on two different populations and samples. This study is a very first attempt that develops a procedure to generate a simulated mobile phone dataset containing the ground truth information. This procedure can be used by other researchers and practitioners who are interested in using mobile phone data and want to formally evaluate the effectiveness of an algorithm.

To identify activity locations from mobile phone traces, we develop an ensemble of methods: a model-based clustering method to identify clusters, a logistic regression model to distinguish between activity and travel clusters, and a set of behavior-based algorithms to detect types of locations visited. We show that the distribution of the activity locations identified from the simulated mobile phone dataset resembles the ground truth better than the existing studies. For home locations, *70%* and *97%* of identified homes are within *100* and *1000* m from the truth, respectively. For work places, *65%* and *86%* of the identified work places are within *100* and *1000* m from the true ones, respectively. These results point to the possibility of using these passively generated mobile phone datasets to supplement or even replace household travel surveys in transportation planning in the future.

1. Introduction

Human trajectories[3], sets of time-stamped locations describing individuals' movements in time (typically over a day), are potentially of great use for many applications. On their own, they reveal the fundamentals of human mobility patterns

* Corresponding author. Tel.: +1 206 543 8974.
E-mail addresses: qzchen@uw.edu (C. Chen), lbian@buffalo.edu (L. Bian), jingtao.ma@traffictechnologysolutions.com (J. Ma).
[1] Tel.: +1 716 645 0484.
[2] Tel.: +1 503 530 8487.

[3] The word, trajectory, has been used with different meanings in the literature. In this study, it specifically refers to a set of time-stamped locations describing individuals' movements in time (typically over a day). In the existing literature, it has been referred to drivers' route choices on a highway network (Schlaich et al., 2010).

http://dx.doi.org/10.1016/j.trc.2014.07.001
0968-090X/

(Gonzalez et al., 2008; Song et al., 2010). When overlaid with layers describing the built environment, they provide insights on the interactions between human and the environment and questions on how people use space and how space in turn affects people's behaviors can be answered (Ahas, 2010; Ahas et al., 2010a,b; Calabrese et al., 2013). Answering these questions is critical in unpacking the intricate, complex relationships between human and nature (Jianguo Liu et al., 2007) and paving our way to sustainability. When used purely in a forecasting context, the trajectory information is used to predict future travel demand for infrastructure investment and policy decisions (Kitamura et al., 1997; Pendyala et al., 1997; Kitamura et al., 2000). Most recently, the emerging location-based services utilize such data to predict individuals' next destinations and recommend location-based services (Cho et al., 2011).

Information on human trajectories has been traditionally collected by Metropolitan Planning Organizations (MPOs) (Stopher and Greaves, 2007; Federal Highway Administration, 2009) through self-reported household travel surveys conducted approximately every 10 years. It has long been recognized that self-reported travel survey data has been subject to a number of problems including missing activities and trips in particular non-motorized travels (Wolf et al., 2003; Pearson, 2004), declining sample sizes (Stopher and Greaves, 2007), increasing non-response rates (Wilson, 2004), non-representative samples (Stopher and Greaves, 2007; Murakami, 2008), increasing costs (Stopher and Greaves, 2007), and imprecise travel times (Stopher et al., 2005).

In early 2000s, concerns primarily on missing trips motivated the use of GPS in travel surveys (Stopher and Greaves, 2007). The role of GPS in travel surveys has ranged from primarily identifying missing trips to becoming a part in travel surveys (TMIP, 2013). Today 20 large MPOs have included a 5–10% GPS sub-sample in their regional household travel surveys (TMIP, 2013). In a GPS-assisted survey, participants are asked to carry GPS units for a period of time. GPS traces are then sent to a central server and displayed on the web so that the participants can enter activity and trip related information (Auld et al., 2009). Displaying trips on a web helps the respondents recall them and may also reduce the amount of information that need to be entered.[4] A few, recent studies have examined the feasibility of a completely passive travel survey by developing algorithms to automatically detect trip purposes (e.g., a home-based work trip) and modes of travel such that no information needs to be solicited from the respondents (Wolf et al., 2001). The initial results (on small samples) are promising (Stopher et al., 2008b; Chen et al., 2010a,b; Gong et al., 2011).

Both the traditional travel surveys and those using GPS are surveys of active solicitation by the surveyors. The rapid development of social media tools combined with location-aware sensors has made the collection of passive datasets possible. Such data results from either simply carrying and using mobile phones or using location-aware social media tools (e.g., Facebook, Flicker, Twitter), thus passively generated. Three distinct features characterize these passively collected datasets: first, since they are passively generated, there is little to no respondent burden and consequently the associated cost may be significantly reduced; second, these datasets usually encompass many people in a population.[5]; and third, the length of the period that these datasets cover is often much longer than household travel surveys.[6] There are obviously limitations: unlike the travel surveys that attempt to record every trip made by each respondent during a certain time period (typically 24 h) and thus, one's complete trajectory (information on places visited and when) is known, a passively generated dataset contains no information that directly tells one's trajectory. Rather, it contains traces that are generated whenever a device connects to the communications network (e.g., a phone call, or uses an application that requires internet). Thus these traces are often not continuous in time (unless a phone is being used nonstop). And because the cell towers are used to derive the position of a device, the traces recorded often have errors that range between 250 m and 900 m (AIRSAGE, 2013) in urban areas.

These new, passively generated datasets have offered high hopes for researchers and practitioners in their potential capability of replacing or supplementing traditional household travel survey data including GPS data. Many questions remain in terms how well such data can capture human mobility patterns. Despite that a number of studies using such datasets have been conducted (Eagle, 2008; Gonzalez et al., 2008; Calabrese et al., 2009; Phithakkitnukoon et al., 2010; Song et al., 2010; Kang et al., 2012a,b; Calabrese et al., 2013), only one study (Calabrese et al., 2013) has compared the attributes of a mobile phone dataset (passively generated) against those in a National Household Travel Survey (NHTS), though the comparison is done between a Boston sample and a national sample. Consequently, one cannot be sure how well the information on people's trajectories is captured in a mobile phone dataset or how well an algorithm or a model performs to identify mobility related information from a mobile phone dataset. Not a single study has systematically validated the methods used to derive activity locations from those passively generated mobile phone datasets. To a large extent, the main obstacle that prevents validation studies is that ground truth information on locations visited is often unavailable in a passively-generated dataset. In this paper, we illustrate a procedure through which a simulated mobile phone dataset is built. In this simulated dataset, the temporal and spatial information associated with the sightings[7] exhibit similar distributions to those in a real-world mobile phone dataset. Different from a real-world mobile phone dataset, this simulated mobile phone dataset contains ground truth information on the locations visited by each simulated individual and their associated location types. This procedure can

[4] For example, home can be automatically labeled on the map shown to a respondent.

[5] The sampling rate of a typical household travel survey is at most 1% for a small metropolitan area; large MPOs often have much lower sampling rates (TMIP, 2013).

[6] Most of the household travel survey datasets are one-day diaries. A few (e.g., the Seattle region) are of 2 days. Surveys of over 2 days are often much smaller in size.

[7] A sighting is generated when a mobile phone is observed on a phone network. Associated information generated is when (time information) and where (location information in *xy* coordinates) a mobile phone is observed.