Transportation Research Part C 58 (2015) 240–250

Contents lists available at ScienceDirect

Transportation Research Part C

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

Origin–destination trips by purpose and time of day inferred from mobile phone data

Lauren Alexander [a,*], Shan Jiang [b], Mikel Murga [a], Marta C. González [a]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States

ARTICLE INFO

Article history:
Received 1 June 2014
Received in revised form 18 January 2015
Accepted 17 February 2015
Available online 9 March 2015

Keywords:
Mobile phone data
Data mining
Human mobility
Trip production and attraction
Trip distribution
Travel surveys

ABSTRACT

In this work, we present methods to estimate average daily origin–destination trips from triangulated mobile phone records of millions of anonymized users. These records are first converted into clustered locations at which users engage in activities for an observed duration. These locations are inferred to be *home*, *work*, or *other* depending on observation frequency, day of week, and time of day, and represent a user's origins and destinations. Since the arrival time and duration at these locations reflect the *observed* (based on phone usage) rather than *true* arrival time and duration of a user, we probabilistically infer departure time using survey data on trips in major US cities. Trips are then constructed for each user between two consecutive observations in a day. These trips are multiplied by expansion factors based on the population of a user's *home* Census Tract and divided by the number of days on which we observed the user, distilling average daily trips. Aggregating individuals' daily trips by Census Tract pair, hour of the day, and trip purpose results in trip matrices that form the basis for much of the analysis and modeling that inform transportation planning and investments. The applicability of the proposed methodology is supported by validation against the temporal and spatial distributions of trips reported in local and national surveys.

1. Introduction

The ubiquity of cell phones, along with rapid advancement in mobile technology, has made them increasingly effective sensors of our daily whereabouts (Lane et al., 2010). Call detail records (CDRs) from mobile phones contain time-stamped coordinates of anonymized customers, thereby providing rich spatiotemporal information about human mobility patterns. Since CDRs are automatically collected by cell phone carriers for billing purposes, this data can be gathered more frequently and economically than travel survey data collected once (or twice) a decade for transportation planning purposes. Additionally, mobile phone data offers digital footprints at a scale and resolution that may not be captured by surveys that typically record one day of travel diaries per household.

Despite these advantages, mobile phone data lacks information typically available from travel surveys about a respondent (e.g. age or income) or his/her trip (e.g. purpose or mode) (Richardson et al., 1995; Stopher and Greaves, 2007; Hu and Reuscher, 2004). Furthermore, CDRs contain traces of a user at approximated locations when his/her phone communicates

* Corresponding author.

http://dx.doi.org/10.1016/j.trc.2015.02.018

with a cell phone tower, providing an inexact and incomplete picture of daily trip-making. Accordingly, much research has focused on developing methods to extract meaningful information about human mobility from mobile phone traces as well as understanding its limitations.

It has been demonstrated that CDR data can be used to infer origin–destination (OD) trips using microsimulation and limited traffic count data (Iqbal et al., 2014). At the level of the individual, daily trip chains/trajectories constructed from mobile phone data are consistent with household surveys (Jiang et al., 2013; Schneider et al., 2013). Further, road usage inferred from the CDR data has been validated against GPS speed data (Wang et al., 2012) and highway assignment results from a travel demand model (Huntsinger and Donnelly, 2014).

There is still work to be done to explore the usage of phone data to generate trip distributions of different modes, purposes, and times of day. As a step in that direction, this research proposes a methodology to extract OD trips by purpose and time of day from CDR data. This segmentation captures distinct trip-making patterns pertinent for transportation planning applications. Moreover, other than CDR data, the techniques presented in this paper rely only upon nationally-available survey data to allow transferability of the methodology to other study areas in the US.

Extensive research has been conducted into OD estimation, as these trips provide the basis for transportation feasibility and impact studies. Conventional OD estimation approaches rely on surveys and/or travel demand models to provide trip matrices. Often, such trip matrices are calibrated or updated using traffic counts and estimation techniques such as maximum likelihood, generalized least squares, and optimization (Spiess, 1987; Cascetta, 1984; Bell, 1991; Yang et al., 1992). This research provides a realistic, cost-effective alternative to these traditional OD data sources and estimation approaches. By presenting a systematic and replicable procedure to extract data relevant to the transportation community, we hope this work will help to facilitate the use of mobile phone data in practice.

In this paper, we demonstrate methods to analyze mobile phone records for the Boston metropolitan area. In Section 2, we present an overview of the data and the methods developed to produce OD trips by purpose and time of day. In Section 3, we summarize and validate our results against independent data sources for the study area, including the US Census and household travel surveys. Based on these findings, we conclude with a discussion of the limitations and applications of CDR data in the context of transportation planning and modeling.

2. Data and methods

2.1. CDR data

The studied dataset contains more than 8 billion anonymized mobile phone records (from several carriers) from roughly 2 million users in the Boston metropolitan area over a period of two months in the Spring of 2010. Although the CDR data spans 60 days, the data provider reindexed the anonymous user IDs for most of the users after the 17th day of the dataset. Effectively, we observe some users for at most 17 days, some users for at most 43 days, and still others for up to 60 days.

Each record contains an anonymous user ID, longitude, latitude, and timestamp at the instance of a phone call or other types of phone communication (such as sending SMS, etc.). The coordinates of the records are estimated by service providers based on a standard triangulation algorithm, with an accuracy of about 200–300 m. In typical mobile phone data sets, locations are represented by cell towers rather than triangulated coordinates and therefore have a lower spatial resolution; however, the method proposed here is expected to hold for such cases (Song et al., 2010a; Wang et al., 2012).

2.2. Stay extraction

The first step to reliably infer activities and trips from CDR data is to filter out noise resulting from (1) tower-to-tower call balancing performed by the mobile service provider, creating the appearance of false movements, and (2) inexact signal triangulation. Furthermore, we wish to distinguish users' stationary stay locations (when/where users engage in an activity) from their moving pass-by locations (when/where users are en-route to activities). To do so, we develop a method based in the work of Hariharan and Toyama (2004) for processing GPS traces. The spatial and temporal filtering methods are discussed below and illustrated in Fig. 1.

Let sequence $D_i = (d_i(1), d_i(2), d_i(3), \ldots, d_i(n_i))$ be the observed data for a given anonymous user i, where $d_i(k) = (t(k), x(k), y(k))'$ for $k = 1, \ldots, n_i$, and $t(k)$, $x(k)$, and $y(k)$ are the time, longitude, and latitude of the k-th observation of user i. First, we extract points $d_i(k)$ that are spatially close (i.e. within roaming distance of 300 m) to their subsequent observations, say, $d_i(k+1), d_i(k+2), \ldots, d_i(k+m)$. To reduce the jumps in the location sequence of the mobile phone data, we assume that $d_i(k), \ldots, d_i(k+m)$ are observed when user i is at a specific location, i.e., the medoid of the set of locations $(x_i(k), y_i(k))'$, $\ldots, (x_i(k+m), y_i(k+m))'$, which is denoted by

$$Med((x_i(k), y_i(k))', \ldots, (x_i(k+m), y_i(k+m))').$$