Transportation Research Part C 33 (2013) 37–49

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

Urban link travel time estimation using large-scale taxi data with partial information

Xianyuan Zhan [a], Samiul Hasan [b], Satish V. Ukkusuri [a,*], Camille Kamga [c]

[a] School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907, USA
[b] School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907-2051, USA
[c] Civil Engineering, Marshak Hall, Suite 910, 160 Convent Avenue, The City College of New York, New York, NY 10031, USA

ARTICLE INFO

Article history:
Received 16 August 2012
Received in revised form 1 April 2013
Accepted 1 April 2013

Keywords:
Traffic state estimation
Path inference
Large scale data analysis
GPS-enabled taxicab
Probe vehicle
Urban networks

ABSTRACT

Taxicabs equipped with Global Positioning System (GPS) devices can serve as useful probes for monitoring the traffic state in an urban area. This paper presents a new descriptive model for estimating hourly average of urban link travel times using taxicab origin–destination (OD) trip data. The focus of this study is to develop a methodology to estimate link travel times from OD trip data and demonstrate the feasibility of estimating network condition using large-scale geo-location data with partial information. The data, collected from the taxicabs in New York City, provides the locations of origins and destinations, travel times, fares and other information of taxi trips. The new model infers the possible paths for each trip and then estimates the link travel times by minimizing the error between the expected path travel times and the observed path travel times. The model is evaluated using a test network from Midtown Manhattan. Results indicate that the proposed method can efficiently estimate hourly average link travel times. This research provides new possibilities for fully utilizing the partial information obtained from urban taxicab data for estimating network condition, which is not only very useful but also is inexpensive and has much better coverage than traditional sensor data.

1. Introduction

Accurate estimation and prediction of urban link travel times are important for improving urban traffic operations and identifying key bottlenecks in the traffic network. They can also benefit users by providing accurate travel time information, thereby allowing better route choice in the network and minimizing overall trip travel time. However, to accurately assess link travel times, it is important to have good real-time information from either in-road sensors such as loop detectors, microwave sensors, or roadside cameras, or mobile sensors (e.g. floating cars) or Global Positioning System (GPS) devices (e.g. cell phones). However, in most of these cases, only limited information is available related to speed or location, hence, one has to develop appropriate methodologies to accurately estimate the performance metric of interest at the link, path or network level.

In the last few years, there has been a growing trend of implementing GPS installed taxicabs in urban areas. While GPS-equipped taxicabs have many advantages, including the ability to locate taxis and track lost packages, they also serve as useful real-time probes in the traffic network. Taxis equipped with GPS units provide a significant amount of data over days and months thereby providing a rich source of data for estimating network wide performance metrics. However, currently there

* Corresponding author.
E-mail addresses: zhanxianyuan@gmail.com (X. Zhan), samiul.hasan@gmail.com (S. Hasan), sukkusur@purdue.edu (S.V. Ukkusuri), ckamga@utrc2.org (C. Kamga).

0968-090X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2013.04.001

are limited methodologies making use of this new source of data to estimate link or path travel times in the urban network. Within this context, this paper proposes a new method for estimating hourly urban link travel times using large-scale taxicab data with partial information. The taxicab data used in this research provides limited trip information, which only contains the origin and destination location coordinates, travel time and distance of a trip. However, the extensive amount of data records compensates for the incompleteness of the data and makes the link travel time estimation possible. A novel algorithm for estimating the link travel times will be presented and tested in this paper using a test network in New York City.

1.1. Related work

Previous research on urban link travel time estimation and prediction has largely relied on various data sources, including: loop detectors (Coifman, 2002; Zhang and Rice, 2003; Oh et al., 2003; Wu et al., 2004), automated vehicle identification (AVI) (Park and Rilett, 1998; Li and Rose, 2011; Sherali et al., 2006), video camera, Remote Traffic Microwave Sensors (RTMS) (Yeon et al., 2008), and automated number plate recognition (Hasan et al., 2011). All of these data collection methods require installing corresponding sensors to retrieve data. Therefore a large number of sensors are required to achieve a reasonable accuracy level based on these data sources. The cost of installing and maintaining such a large number of sensors is prohibitive. Hence predicting link travel times with reasonable accuracy and network coverage based on sensor data could be expensive.

On the other hand, there is a significant potential to use emerging large-scale data sources to estimate dynamic demand and dynamic network conditions in urban areas. For instance, GPS devices in dedicated fleets of vehicles or in users' mobile phones can be viable sources of data for monitoring traffic in large cities (Herrera et al., 2010). Industry models, such as Inrix,[1] have also gained popularity in recent years where private entities install, collect, utilize and sell "large-scale" historical traffic data from GPS-equipped vehicles or mobile phones. With an increasing amount of GPS data available from taxi, transit, and mobile phones, a new option of using such large-scale decentralized data for link travel time estimation becomes realistic. Herring et al. (2010) used GPS traces data from a fleet of 500 taxis in San Francisco, CA. to estimate and predict traffic conditions. However, in this work, instead of link travel times, discrete traffic states were predicted. Zheng and Van Zuylen (in press) also proposed an ANN model to estimate urban link travel times based on sparse probe vehicle data (e.g., GPS traces from GPS-equipped vehicles or smartphones). Hunter et al. (2009) proposed a statistical approach for path and travel time inference using GPS probe vehicle trajectory data. The GPS data used in their study has been recorded each minute, where the inferred path consists of at most five link segments. This method is not applicable if the GPS data has a longer recording interval or only has the starting and ending coordinates. Estimating link travel times from GPS data provides a much cheaper and a larger coverage area in the urban network compared with approaches using fixed sensor data. However, all of the above mentioned approaches are only applicable for GPS trace data, in which the trajectories of vehicles are available. To the best of our knowledge, there is no study found in literature that used OD level GPS data for urban link travel time estimation, even though extensive amount of such less detailed data (e.g. taxicab data) is generated and recorded every day.

1.2. Background and objectives

In New York City, GPS devices are installed in each taxicab. The taxicab data is collected and archived by the New York City Taxi and Limousine Commission (NYTLC), an agency that is responsible for all taxi related issues in New York City. The New York City has the largest market for taxis in North America with 12,779 (in 2006) yellow medallion taxicabs serving about 240 million passengers a year. The taxi service transports 25% of all fare-paying bus, subway, taxi and for-hire vehicle passengers that are traveling within Manhattan (Schaller Consulting, 2006; King et al., 2012).

In this paper, data collected from New York City taxicabs is used to estimate the link travel times. The dataset provides an extensive amount of taxi trip data, which records the trip starting and ending geo-location, along with information about trip distance, time and fare. Unlike the detailed GPS trajectory data used in previous studies, the dataset only provides the trip origin and destination information (i.e. starting, ending location and time) without the exact trajectory of the taxicab; only path travel time and distance are known. However, the advantage of the massive amount of data (the number of observations recorded within a day range between 450,000 and 550,000) makes it possible to infer the possible routes that the taxicab is taking and further, to estimate the link travel times in the New York City network. There is potential bias associated with measuring network link travel times from taxis, as taxi drivers are just one particular group of all drivers in the network. However, given the high penetration rate of taxicabs, it is reasonable to assume that taxis are good probe vehicles and therefore taxi travel times are a good representation of the actual network condition.

In this research we propose a methodology to estimate urban link travel times based on taxi GPS data that includes only the information about the origin and destination of the trip and total travel time to reach the destination. The goal of this study is to show the potential of using taxicab data as a complimentary data source in urban transportation operation and management. The link travel times estimated from taxicabs provide an hourly aggregate measure of the urban network condition, which can be fused with the information from other existing data sources such as fixed sensors in the future.

[1] Inrix, Inc. http://www.inrix.com.