Transportation Research Part C 79 (2017) 136–155

Contents lists available at ScienceDirect

Transportation Research Part C

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

An adaptive hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data

Wangsu Hu, Peter J. Jin *

Rutgers, The State University of New Jersey, CoRE Building, 96 Frelinghuysen Road, Piscataway, NJ 08854-8018, USA

ARTICLE INFO

Article history:
Received 2 March 2016
Received in revised form 11 December 2016
Accepted 2 February 2017

Keywords:
Location-based social networking
Dynamic trip arrival estimation
Human activity patterns
Hawkes process
State space model

ABSTRACT

Location-Based Social Networking (LBSN) services, such as Foursquare, Facebook check-ins, and Geo-tagged Twitter tweets, have emerged as new secondary data sources for studying individual travel mobility patterns at a fine-grained level. However, the differences between human social behavioral and travel patterns can cause significant sampling bias for travel demand estimation. This paper presents a dynamic model to estimate time-of-day zonal trip arrival patterns. In the proposed model, the state propagation is formulated by the Hawkes process; the observation model implements LBSN sampling. The proposed model is applied to Foursquare check-in data collected from Austin, Texas in 2010 and calibrated with Origin-Destination (OD) data and time of day factor from Capital Area Metropolitan Planning Organization (CAMPO). The proposed model is compared with a simple aggregation model of trip purposes and time of day based on a prior daily OD estimation model using the LBSN data. The results illustrate the promising benefits of applying stochastic point process models and state-space modeling in time-of-day zonal arrival pattern estimation with the LBSN data. The proposed model can significantly reduce the number of parameters to calibrate in order to reduce the sampling bias of LBSN data for trip arrival estimation.

1. Introduction

The increasing popularity of social networking services (SNS) and location-based services (LBS) has offered new opportunities for urban mobility patterns analysis. The combination of SNS and LBS leads to a new type of social networking service, Location Based Social Networking (LBSN) service. In LBSN, users can "check-in" with their LBS-enabled mobile devices to a nearby "venue," or point of interest (POI) to declare their arrivals. Such information can be shared with friends and family, as well as with business owners for potential discounts and promotions. Given the pre-registered location and POI type information of venues, travelers' trip arrivals are recorded with accurate location and trip purpose information. When aggregated, such data can provide a new secondary data source for the estimation of urban travel demand.

LBSN data is one of the emerging technology-based travel demand data collection methods of recent years. Table 1 summarizes the characteristics of the latest technology-based primary and secondary data collection methods for travel demand estimation. Primary data collection methods, including GPS and smartphone based travel survey (Barceló et al., 2010; Bohte

* Corresponding author.
E-mail address: jinjing.wisc@gmail.com (P.J. Jin).

http://dx.doi.org/10.1016/j.trc.2017.02.002

Nomenclature

$x_{i,p}$, $x_{i,p}(t)$	the check-in counts at location i for trip purpose p overall and at time t, respectively
$\hat{A}_{i,p}, \hat{A}_{i,p}(t)$	the estimated trip arrivals at location i for trip purpose p overall and at time t, respectively
I	the number of Traffic Analysis Zones in the study area
t	time step $t \in T$
s	the points at time occurring prior to (check-in arrivals) time t
$\lambda(t)$	the trip arrival rate with respect to t
μ	the background trip arrival rate
T	the duration of Hawkes estimation process
τ	the time interval between two (trip) arrivals

and Maat, 2009; Caceres et al., 2007; Wolf et al., 2001). Secondary data sources include Bluetooth (Barceló et al., 2010), cellphone location (Liu et al., 2008) and location-based social networking (LBSN) data (Gao et al., 2012b). Table 1 compares the pros and cons of the emerging data sources with conventional travel demand data collection methods.

The advantage of the LBSN data can be classified into four categories. First, the LBSN data has the advantage of relatively low-cost secondary planning data sources. LBSN services are tightly integrated with personal smartphones and tablets through mobile applications. The only cost incurred is that of a data subscription fee. Second, unlike other secondary data sources such as cell phones and Bluetooth, the LBSN check-in data comes with the confirmed trip purposes. Each check-in is linked to a Foursquare venue whose category is defined by venue owners with a three-level Foursquare venue classification system (Foursquare). Table 2 lists the trip purposes determined based on the Foursquare venue types. Third, the LBSN services are self-sustained twenty-four hours a day, seven days a week. They contain users' interests in exploring new points of interest and business owners' interests in attracting and maintaining their customer base. Finally, a comprehensive privacy protection mechanism has been implemented in LBSN services that combine general anonymization (e.g. only counts at business are posted) and user consents public information sharing (the Foursquare – Twitter bridge).

Despite the above advantages, LBSN data is not without its limitations for estimating urban travel demand, as well as dynamic travel demand estimation for proactive urban congestion mitigation and operations (Neudorff and McCabe, 2015; Zheng et al., 2012). First, LBSN can have a systematic temporal error for estimating travel demand. The LBSN activity does not always mimic travel activities throughout the day. LBSN check-in activities tend to be more intensive during afternoons and evenings at social recreational places, as opposed to during morning peak hours when commuters are rushing against time to get to workplaces. Second, LBSN data includes a sampling bias for different population groups and venue types. Third, the stochastic nature of human activities, especially the POI arriving patterns are critical for travel demand estimation. In previous studies (Jin et al., 2013, 2014; Hu and Jin, 2015; Yang et al., 2015), it was observed that the accurate estimation of zonal departures and arrivals are critical in reducing the errors in the subsequent OD estimation.

To address the above issues in LBSN-based travel demand estimation, this paper focuses on introducing a time-of-day-variant and trip-purpose-specific dynamic estimation model with respect to zonal LBSN arrival data. The time-of-day (TOD) variations allow the model to adapt to different sampling patterns during the day; trip-purpose-specific modeling provides the flexibility to capture the different sampling characteristics for different venue types. A Hawkes process based state propagation model and a state-space framework are introduced to model the stochastic arrival patterns with sampling error

Table 1
Emerging versus Traditional Travel Demand Data Collection Methods.

Characteristics	Trad. Survey Method	GPS	Bluetooth	Smart Phone Survey	Cell Phone Signals	Social Media	LBSN Check-in
Spatial Resolution	Low	Low	Low	High	High	High	High
Temporal Resolution	Low	High	High	High	High	High	High
Large-scale Deployment	*Yes*	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*
Survey/Data Cost	*High*	*Medium*	*Medium*	*Medium*	*Low*	*Low*	*Low*
Survey Needs	*Yes*	*Yes*	*No*	*Yes*	*No*	*No*	*No*
Social Demographic Data	*Yes*	*No*	*No*	*Inferred*	*No*	*Yes*	*Inferred*
Origin-Destination Data	Yes	Yes	Yes	Yes	Yes	Yes	Inferred
Trip Chain	Yes	Yes	Yes	Yes	Yes	Inferred	Inferred
Trip Purpose Confirmation	*Yes*	*Limited*	*Limited*	*Yes*	*Limited*	*Inferred*	*Yes*
Mode Share	*Yes*	*Inferred*	*Inferred*	*Yes*	*Inferred*	*Inferred*	*Inferred*
Arrival Time Resolution	*Low*	*High*	*High*	*High*	*High*	*High*	*High*
Arrival Location Resolution	*Low*	*High*	*Low*	*Low*	*High*	*Medium*	*High*
Sampling bias	Low	Medium	Medium	Medium	Medium	Yes	Yes
Privacy Concern	No	Medium	No	No	No	Medium	Medium

Non-italicized characteristics are based on NCHRP Report 735 Table D.2 (Schiffer, 2012), and previous papers (Hu and Jin, 2015; Yang et al., 2014).