ARTICLE IN PRESS

Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Big data and urban system model - Substitutes or complements? A case study of modelling commuting patterns in Beijing

Li Wan[a,*], Shuo Gao[b], Chen Wu[c], Ying Jin[a], Mingrui Mao[b], Lei Yang[c]

[a] Department of Architecture, University of Cambridge, United Kingdom
[b] Beijing City Quadrant Technology Co., Ltd., PR China
[c] Beijing Engineering Research Centre of Urban Design & Urban Renaissance, Beijing Institute of Architectural Design, PR China

ARTICLE INFO

Keywords:
Big data
Spatial equilibrium
Land-use and transport model
Commuting

ABSTRACT

The emergence of urban big data is transforming the existing research paradigms in urban studies. New theories and analytical methods are required to meet the methodological challenges. This paper empirically compares a data-driven approach and an urban-system-model approach through a case study of modelling the commuting patterns in Beijing. For the data-driven approach, the novel location-based-services (LBS) data are explored to identify the employment-residence location of the service users. For the modelling approach, a spatial equilibrium model is calibrated for base year 2010 and is used to simulate the commuting patterns for Beijing 2015 based on exogenous development projections. The results of the two approaches are then compared against the benchmark statistics for Beijing 2015. The comparison shows that the LBS data perform better in detecting residence locations than employment locations. The model prediction fits better with the benchmark, while the errors of the LBS data tend to vary significantly across space. For amplifying the LBS sample data to represent the full population, uniform scale factor thus should be avoided. In addition, the ineffectiveness of representing short-distance commuting for the LBS data is revealed by the comparison with the model predicted flows. In light of the strength and weakness of the respective approach, the prospect of a collaborative use of big data and urban system models is explored in the conclusion.

1. Introduction

The past twenty years have witnessed the rise of big data as both an academic topic and technology terminology. One of the most vivid definitions of big data is 'any data that cannot fit into an Excel spreadsheet' (Batty, 2013). This definition indicates the sheer size of the data and also suggests that new methods are required to process and understand the big data. In the sphere of urban studies, urban transport has been a fertile research field embracing the big data. Compared with conventional data sources such as travel survey and census, the transport big data are usually much finer at both spatial and temporal scale, which provides a new perspective to examine the relationship between locations and activities and the interaction with other urban systems.

Recent progresses of this research line include using smart card data to identify travel patterns (Kieu, Bhaskar, & Chung, 2014; Ma, Liu, Wen, Wang, & Wu, 2017; Seaborn, Attanucci, & Wilson, 2009) and the urban spatial structure (Roth, Kang, Batty, & Barthélemy, 2011). The smart card data are also applied to analyse the job-housing balance (Long & Thill, 2015), the travel behaviours of underprivileged residents (Long & Shen, 2015), long-distance commuters (Long, Liu, Zhou, & Chai, 2016), and the temporal mobility patterns (Zhong et al., 2016; Zhong, Manley, Arisona, Batty, & Schmitt, 2015). Shen and Chai (2012); Shen, Kwan, and Chai (2013) use GPS tracking data collected from voluntary samples to explore the spatial-temporal variations in commuting in Beijing. The use of multi-source geospatial data to identify urban spatial structure is reported by Cai, Huang, and Song (2017).

Contrast to the big-data approach, urban system modelling represents another long-standing methodology in urban land-use and transport studies. Among the wide spectrum of urban applied models, the land-use and transport integrated (LUTI) modelling framework has been the mainstream since the early spatial interaction models (Echenique et al., 1990; Lowry, 1964; Wegener, 1998). The incorporation of spatial equilibrium with the LUTI structure represents a significant advancement (Anas & Liu, 2007; Bröcker, 1998; Jin, Echenique, & Hargreaves, 2013). The spatial equilibrium theory provides a solid economic foundation for quantifying the impacts of urban land-use and transport policies within a circular causality. The LUTI models with equilibrium mechanisms have been widely applied for practical policy studies (Anas, 2013; Jin et al., 2017; Volterra & CBP,

* Corresponding author at: 1–5 Scroope Terrace, Cambridge CB2 1PX, England, United Kingdom.
E-mail addresses: lw423@cam.ac.uk (L. Wan), gaos@urbanxyz.com (S. Gao), wuchen@vip.sina.com (C. Wu), yj242@cam.ac.uk (Y. Jin), mmr@urbanxyz.com (M. Mao).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.10.004
Received 13 June 2017; Received in revised form 23 October 2017; Accepted 23 October 2017

2007; Wegener, Mackett, & Simmonds, 1991).

Given the prevalence of the big data, a new form of data-driven empiricism may declare "the end of theory" (Anderson, 2008) for urban studies. This proposition is often accompanied with the claim that "correlation is enough" given that the sheer size of the big data may have already depicted a near-complete answer to the question of interest, which is not possible with models based on aggregate data. This issue is critically discussed by Kitchin (2013, 2014) from an epistemological point of view. New theories and analytical methods are required to meet the methodological challenges posed by the emerging big data (Batty, 2013). Although most discussions have focused on either the data-driven approach or the theory-driven approach, few research compares the two methods in an empirical manner.

This paper aims to fill the gap by applying the two approaches on one case study of modelling the commuting patterns in Beijing. A novel data source, namely the location-based service (LBS) data collected from personal smart-device users, is explored to identify the employment-residence location of commuters in Beijing 2015. To represent the urban system modelling approach, a LUTI type urban system model is developed to predict the commuting pattern in Beijing 2015. The two derived commuting flows are then compared against the same benchmark for 2015. The comparison helps to explore the strength and weakness of the each approach in terms of its application in urban commuting analysis.

The paper is structured as follows. The next section introduces the LBS data and the processing method for commuting analysis. Section 3 presents the calibration of a LUTI model and how it is used to predict the commuting pattern in Beijing 2015. Section 4 compares the results of the two approaches against the benchmark. Section 5 concludes by way of considering the wider implications of the findings.

2. Location-based-service data processing

The location-based service (LBS) data cover a wide range of data sources, but in terms of location acquisition technology, Global Position System (GPS) and mobile positioning are most commonly used (Lu & Liu, 2012). Some LBS data are collected directly from user-end hardware, such as mobile phone or GPS receiver, while some LBS data do not require the positioning hardware on user end, such as the smart card system in public transit. The challenges of the LBS data to urban planning and public management are first brought up by Ahas and Mark (2005). Recent applications of LBS data in regional studies as well as the emerging technologies and challenges are discussed in Schintler and Chen (2017).

Studies on transport big data have focused on smart-card data (Long & Thill, 2015; Ma et al., 2017; Seaborn et al., 2009) and mobile phone data (Ahas, Silm, Järv, Saluveer, & Tiru, 2010; Gao, Liu, Wang, & Ma, 2013; Kung, Greco, Sobolevsky, & Ratti, 2014). Compared with the smart card data, the GPS-based LBS data has its own distinct features. First, the LBS data collected from individual smart devices is expected to be more accurate in terms of positioning than smart-card data, because it records the exact location of the user rather than the location of the bus/metro stations as in smart-card data. Secondly, as opposed to the smart-card data, the LBS data is not limited to public transit and includes travels of all modes and purposes. Nonetheless, the LBS data do not have explicit information on travel mode, route and duration. Thirdly, for smart-card data, the service record is usually collected as location pairs, i.e. the origin and destination of the travel. By contrast, the location records in LBS data tend to be single-ended. To investigate the travel pattern of LBS users, both the employment and residence location thus need to be derived. In addition, because the LBS data can only be collected from smart devices, the socio-demographic background of LBS users is likely to be biased towards younger population. The magnitude and spatial distribution of such bias in the LBS data needs to be investigated empirically.

The Location-based service data used in this paper is an exclusive dataset provided by 'TalkingData', a Chinese corporation that provides location-based services to thousands of applications on smart devices, e.g. smartphones and tablets. Established in 2011, TalkingData is currently the biggest third-party LBS provider in China. The location-based service provider collects geographic location from the embedded GPS module on smart devices when the service is requested. The prerequisites for collecting the data from users are 1) the LBS application is properly installed, and 2) permission to use the location service on smart device is granted by the user. The LBS data provided include location records of the users as well as a set of background attributes (see Table 1). Note that the background information is collected from users on a voluntary basis through the registration procedure. For disclosure control, all information are anonymously represented with a unique user code, and information that can be used to identify individuals are removed.

Table 1
Sample information of the LBS data.

Background attributes		Location records
User ID	aefeb5333	Location 1: timestamp
Age	0–110	longitude/latitude
Gender	Male/female/unknown	Location 2: timestamp
Marriage status	Single/married/ unknown	longitude/latitude
Car ownership	Yes/no/unknown	
If university student	Yes/no/unknown	
User birthplace	Beijing, Shanghai, etc.	
Device code	cbb0b5ad5a162	
Operation system on device	iOS, Android, etc.	
User language on device	Chinese, English, etc.	
Application name	xyz	

The core study area of this paper is the Beijing municipality, while the data collection area is expanded to cover the Greater Beijing city region, which consists of the Beijing Municipality, Tianjin Municipality and Hebei Province. The inclusion of the wider city region is to enable the modelling of cross-boundary commuting. The LBS raw data is collected between August and October in 2015, which has a total of 17,000 million records from 16 million devices in Greater Beijing, implying approximately 11.8 records per device per day on average. Fig. 1 presents the spatial distribution of the LBS record data at 9 am on a typical weekday in central Beijing.

2.1. Derive home and workplace from LBS data

LBS data provide detailed user location with specific timestamp. To utilize the data to extract commuting patterns, we first try to identify the "anchor points" of users in space. Anchor points are defined as locations which people tend to stay for a period of time, typically home and workplace. Anchor points reflect the key locations of people's daily routine, thus can be used to infer their residence location as well as workplace if the person of interest is deemed employed. To detect the anchor points of LBS users, the processing method needs to tackle two challenges. First, the LBS records can be transient in the sense that any single location record may be irrelevant to either the home or workplace of the user. Secondly, the observed spatial-temporal variations in commuting behaviour (see Shen et al., 2013 for the empirical evidence in Beijing) suggests that a relatively long period of observation is required to establish regular spatial patterns. Once regular locational patterns are detected, the user's employment-residence location pair may be inferred with certain behavioural assumptions.

To this end, the raw LBS data needs to be cleaned. To reduce the noises and unwanted variations, we define a LBS user to be valid if the following two criteria are met simultaneously, during the data collection period, 1) the number of days that records appear at both night