Computers, Environment and Urban Systems 68 (2018) 64–77

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
Available online 06 November 2017
0198-9715/

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.

Table 1
Sample information of the LBS data.

Background attributes		Location records
User ID	aefeb5333	Location 1: timestamp longitude/latitude
Age	0–110	
Gender	Male/female/unknown	Location 2: timestamp longitude/latitude
Marriage status	Single/married/unknown	
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 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.

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

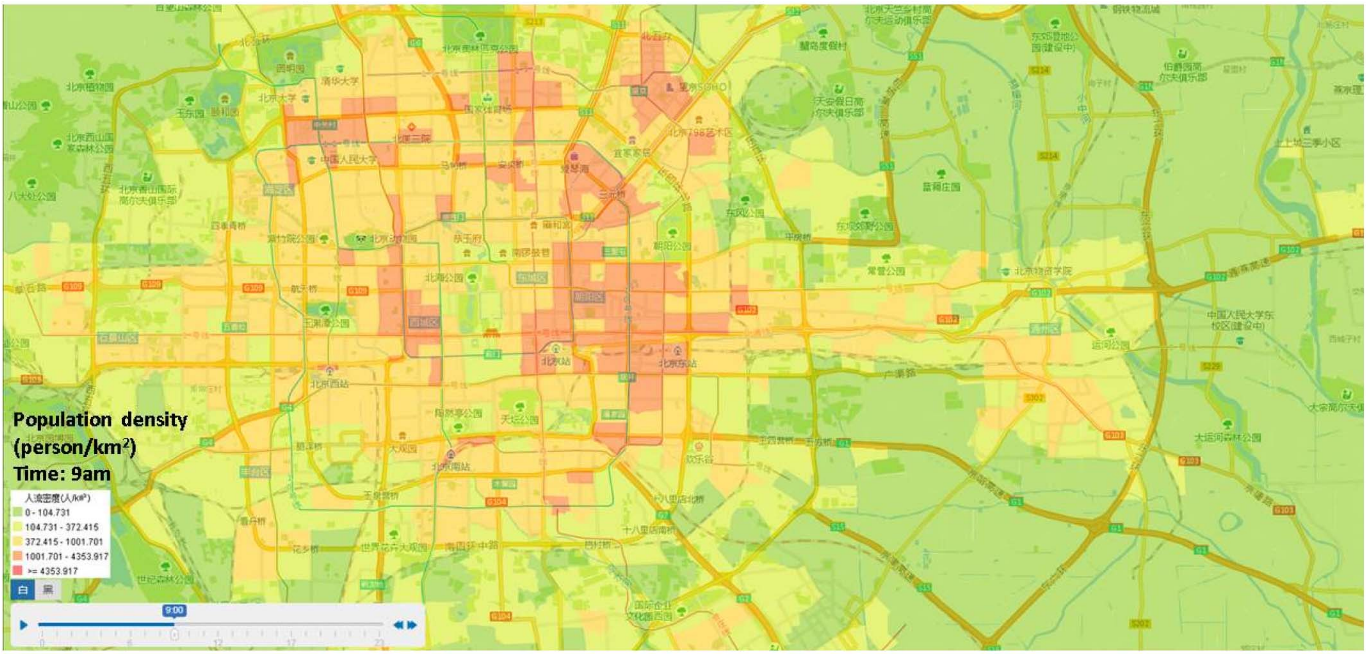

Fig. 1. Mapping of LBS data records at 9 am on a typical workday in central Beijing 2015.

time (9 pm–6 am) and daytime (6 am–9 pm) is longer than *N* days, and 2) the number of record occurrence (i.e. frequency) at both night time and daytime is more than *F* times per day. Only records of the valid users are selected and used in this research. For valid users, due to the regularity in their night-time and day locations, we further assume that all valid users are employed, and the night-time and daytime records reflect their potential home and workplace, respectively. We report that the selected valid user data has a total of 11,520 million records from about 8.1 million users in Greater Beijing, implying approximately 16 records per device per day. Because the employment-residence location is processed on a location-pair basis, the derived total number of employed workers thus equals the total number of employed residents.

To detect the employment/residence anchor points, Long and Thill (2015) presents a rule-based approach for processing bus smart-card data for Beijing, where an exclusive parcel-level land use data is applied to enhance the algorithm. Compared with the smart-card data, our LBS data does not include explicit information about the origin and destination of the journey. As a result, the rule-based approach based on origin-destination travel data is difficult for processing the LBS data. As opposed to the rule-based approach, more general methods have been proposed for detecting the anchor points (also named clustering) in big data mining. A survey of clustering methods is provided by Berkhin (2006). Two main methods include the standard k-means method (Lloyd, 1982) and the Density-Based Scanning Algorithm with Noise (DBSCAN) (Ester, Kriegel, Sander, & Xu, 1996). Both methods aim to partition observations into distinguishable clusters. The key difference lies in that the k-means method requires an a priori number of clusters (k value), while the DBSCAN method does not require exogenous cluster number, which improves its computing efficiency significantly. In addition, the DBSCAN method is able to detect noise in data provided that a threshold parameter is set for noise recognition. Due to the advancement in dealing with noise and the outstanding computing efficiency, the DBSCAN method has been widely applied for clustering analysis with spatial big data (Kieu et al., 2014; Ma et al., 2017).

For processing the LBS data using the DBSCAN method, the threshold parameter *D* is defined as the maximum radius within which records can be regarded as one cluster. The parameter *D* is determined by two factors. First, the locational accuracy of the embedded GPS on smart devices. Zandbergen and Barbeau (2011) reports a maximum horizontal error of 30 m and 100 m for outdoor and indoor positioning respectively using GPS-enabled mobile phones. The second factor is the spatial variation of user movement at the anchor point, i.e. users may not use the LBS service at exactly their home or workplace, but a place nearby. The authors deem that such spatial variation around anchor point is plausible when gated communities and large-scale building complexes are considered in the metropolitan context of Beijing.

The workflow of LBS data processing is summarized in Fig. 2. The result validation involves aggregating the processed LBS users and comparing the totals with the observed data, typically the total number of employed population at residence and work place. Such comparison is conducted in Section 4. To understand the broad coverage of the processed data, Table 2 presents the distribution of the LBS users by age

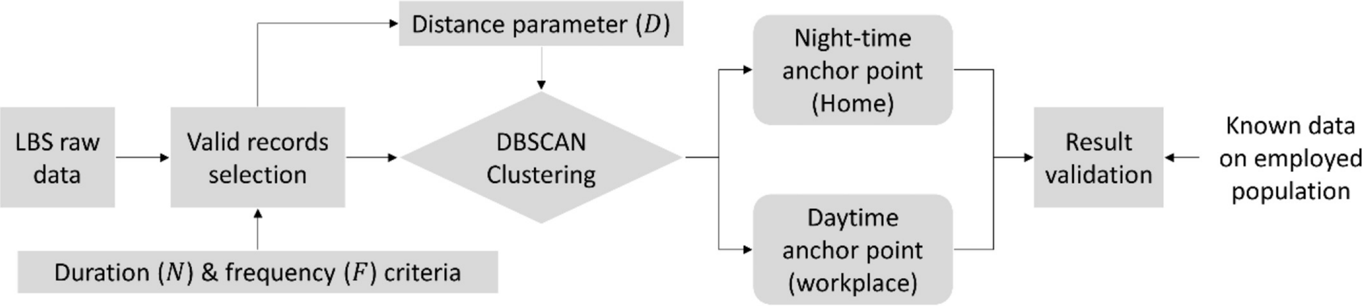
LBS raw data
Valid records selection
Distance parameter (D)
DBSCAN Clustering
Duration (N) & frequency (F) criteria
Night-time anchor point (Home)
Daytime anchor point (workplace)
Result validation
Known data on employed population

Fig. 2. Diagram for processing the anchor points from the LBS data.

Table 2
User profile of LBS data by age and gender - Greater Beijing 2015.

Distribution by age			Distribution by gender		
Age	Number of users	% Share	Gender	Number of users	% Share
< 19	28,309	0.3%	Male	3,413,296	42.1%
19–25	644,427	8.0%	Female	3,087,805	38.1%
26–35	2,631,225	32.5%	Unknown	1,600,491	19.8%
36–45	9,36,582	11.6%			
46–55	179,370	2.2%			
> 55	105,981	1.3%			
Unknown	3,575,698	44.1%			
Total	8,101,592	100.0%	Total	8,101,592	100.0%

and gender. Note that the data only cover users whose derived employment and residence location are both within Greater Beijing. People commuting to/from outside Greater Beijing are not included. The results in Table 2 confirm that among the age-known users, the majority are under 46 years old, implying a sampling bias towards the younger population. In terms of gender distribution, the processed data include slightly more male users than the female.

Given the 8.1 million LBS users identified and the 111.4 million total population in Greater Beijing in 2015 (National Bureau of Statistics PRC, 2016), the derived sample rate of the LBS data is approximately 7.3% for Greater Beijing. In big data processing, a scaling procedure is often considered for amplifying the sample data to represent the whole population. However, in this paper we do not apply a priori data scaling method and the processed LBS data are used without scaling for two considerations. 1) The LBS data is a quite new source of data, and the possible biases are yet fully investigated in literature. The scaling method tends to be case-specific and needs to be explored empirically. 2) It is difficult to scale the LBS data without using observed statistics for 2015. In this paper, however, the 2015 observed data are preserved as the benchmark for comparing the derived flows. The use of 2015 data is intentionally avoided in flow estimation, such that the benchmark is independent from the derived flows. The proposed comparison is expected to provide useful insights on how the LBS data should be scaled and corrected for bias.

3. Modelling commuting flows

Modelling commuting flows using urban system model requires the model to be first calibrated for a base year before used for prediction. In model calibration, key model parameters are estimated using observed base-year data; the calibrated model is expected to reproduce the observed commuting patter at base year. In the case of Beijing, the census year 2010 is defined as the base year and the year 2015 is set as the forecast year, which is in line with the LBS data. For the base-year calibration, the origin-destination (OD) commuting flows have to be estimated first because observed flow data are not available from open sources. This section first introduces how the base-year commuting flows are estimated using a discrete choice model with log-linear travel cost function. This is followed by modelling the location choice in a spatial equilibrium model. The scenario design for model forecast is presented in the end of the section.

3.1. Estimating base-year commuting flows

To model the commuting pattern using urban system model, we start from estimating the base year (2010) commuting flow. Two estimation methods can be found in the literature. One is the Fratar method (Fratar, 1954) and its variations (see a review by Horowitz, 2005). The Fratar method estimates the travel flow based on a seed matrix, taking the origin- and/or destination-end number of the trips as constraint. Specifically the Fratar method multiplies the rows and columns of the seed matrix in an iterative manner until the row and column totals agree with the exogenous constraints. Murchland (1966) and Evans (1970) have proven the existence and the uniqueness of the estimated matrix using the Fratar method, provided that the travel volume between all zone pairs is larger than zero. The second method for flow estimation is to use a distribution model, such as the gravity model (Wilson, 1967) or discrete choice model based on random utility (McFadden, 1973). It is assumed that the number of trips between pairs of zones for a particular purpose is proportional to the size of the zone pair and inversely proportional to the travel cost between them.

The key difference between the Fratar method and the model-based approach is that the effectiveness of the Fratar method relies on a good-quality seed matrix to be provided; while the model-based approach does not require the seed matrix input. For the case of Beijing, because no observed flow data are available to serve as the seed matrix, a multinomial discrete choice model is applied to estimate the base-year commuting flows.

The estimation process starts with defining the travel cost function. City regions with reasonably self-contained commuting catchment today tend to have a radius of 50 km or more. At this metropolitan scale, extensive analyses of travel choices data show that a travel disutility function that is linear to travel costs and times will have great difficulties in representing realistic demand elasticity throughout (Jin et al., 2013). To this end, a log-linear function (Daly & Zachary, 1978) is proposed in order to reflect the realistic travel demand elasticity at city-region level.

$$d_{fij} = a_f \chi_{fij} + (1 - a_f) \ln \chi_{fij} - a_f$$

where d_{fij} is the travel disutility of commuting from zone i to zone j for commuter type f; χ_{fij} is the annual commuting time (twice a day, 250 days per year) between zone i and zone j for commuter type f; $a_f = 0.003$ is a log-linear transformation parameter. The network travel distances, times and costs in Beijing are obtained from the associated transport model, which is calibrated and validated to the AM-peak road congestion times using a taxi GPS vehicle trace sample for 2008 (Deng, Denman, Zachariadis, & Jin, 2015).

To demonstrate the non-linear feature of the above function, the log-linear travel disutility is plotted versus the linear counterpart in Fig. 3. It shows that the elasticity of the log-linear function varies for different distance ranges. Specifically, the elasticity of disutility with regard to distance is higher for short-distance range (approx. 0–15 km), and becomes lower for long-distance range (approx. > 15 km).

This travel disutility term enters a multinomial logit model:

$$\widetilde{P}_{fij} = \frac{S_{ij} \exp(\lambda_f(-d_{fij} + E_{fij}))}{\sum_{k,l} S_{kl} \exp(\lambda_f(-d_{fkl} + E_{fkl}))}$$

$$\textit{subject to } \sum_j H_f \widetilde{P}_{fij} = H_{fi|Obs} \textit{ and } \sum_i J_f \widetilde{P}_{fij} = J_{fj|Obs}$$

where $\widetilde{P}_{fij}$ is the probability of commuter type f choosing residence zone i and employment zone j; $S_{ij} = \Sigma_m b_{mi} \Sigma_k B_{kj}$ is the size term that corrects for the bias introduced by the uneven sizes of zones (M. E. Ben-Akiva & Lerman, 1985), where b_{mi} and B_{kj} is the type-m housing and type-k business building stock (m^2) at zone i and j respectively; λ_f is the dispersion parameter and E_{fij} is the residual attractiveness term for residence-employment pair choice (i,j), which is to be endogenously solved.

In the estimation process, the observed zonal number of employed residents ($H_{fi|Obs}$) and workers ($J_{fj|Obs}$) are exogenous constraints. For any given λ_f, the E_{fij} is solved subject to the double constraints of $H_{fi|Obs}$ and $J_{fj|Obs}$. Note that solving E_{fij} requires i^*j functions while there are only $i + j$ constraints available from $H_{fi|Obs}$ and $J_{fj|Obs}$. Thus we need to reduce the dimension of E_{fij} by defining $E_{fij} = E_{fi} + E_{fj}$, where E_{fi} and E_{fj} are zonal residual attractiveness for residence and employment, respectively. This treatment reduces the number of variables from i^*j to $i + j$, thus speeds up the solving algorithm.

The λ_f can be solved through an optimization problem, which is

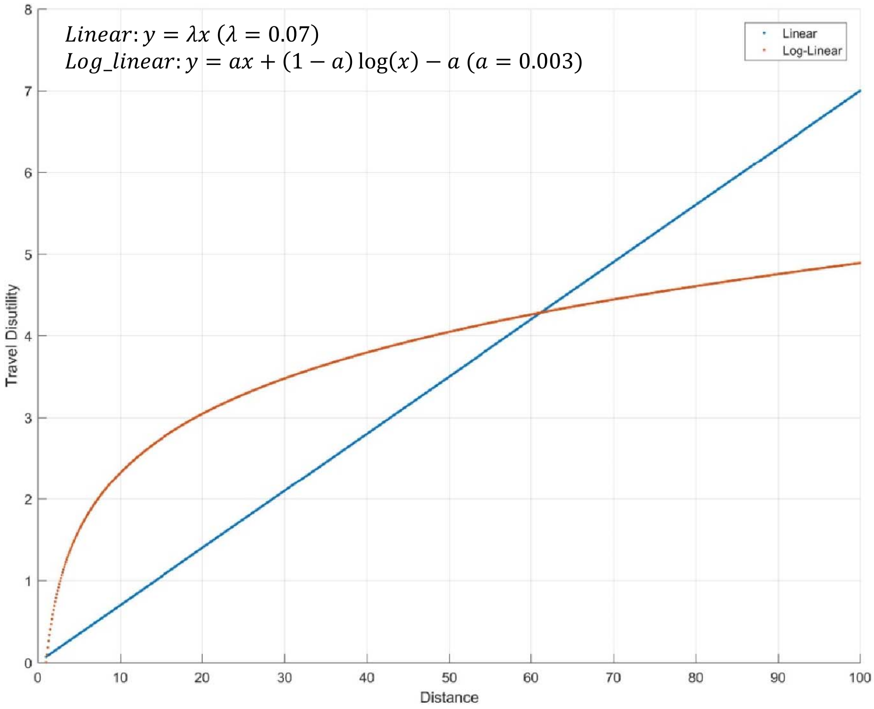

Fig. 3. Comparison of travel disutility - log-linear vs linear function.

defined as:

$$\min_{\lambda_f}\left(\left|\sum_{i,j}\widetilde{P}_{fij}Dist_{fij} - ACD_{f|Obs}\right|\right)$$

where $Dist_{fij}$ is the estimated network distance between zone i and j for commuter type f; and $ACD_{f|Obs}$ is the observed average commuting distance for commuter type f in Beijing (BCT & BTRC, 2010).

The estimated commuting matrices $\widetilde{P}_{fij}$ can reproduce the average commuting distance that is compatible with the observed travel data in Beijing for 2010. The proposed method provides reasonable approximations when observed OD matrices are not available. The derived commuting matrices are manually verified with modellers' local knowledge before being used for calibration purpose. The statistics of the estimated commuting matrices of each commuter type for base year 2010 are summarized in Appendix A.

3.2. Location choice within spatial equilibrium

The estimated base year commuting matrices are then used to calibrate a spatial equilibrium model for Greater Beijing. Within the spatial equilibrium framework, producers and consumers are assumed to be profit-maximizing and utility-maximizing, respectively. Employment and residents are both mobile in space, subject to the simultaneous equilibrium in the product, labour and building floorspace market. In this section, we focus on the employment-residence joint choice for employed residents. Other model equations as well as the equilibrium conditions of the spatial equilibrium model are provided in Appendix B.

In terms of model zoning, the Greater Beijing city region is represented with a total of 209 model zones (see Fig. 4). We define the 130 zones of Beijing Municipality as *core* zones of study and the rest 79 zones of Tianjin and Hebei as *peripheral* zones. The inclusion of the wider city region enables the modelling of cross-boundary commuting and within-region labour migration.

For modelling the employment-residence joint choice, a nested logit model (M. Ben-Akiva, 1974) is developed. This extends the existing spatial equilibrium models such as Anas and Liu (2007) and Jin et al. (2013). Specifically, we partition the employment-residence alternative pairs (total number $\mathbb{N} \times \mathbb{N}$) into $\mathbb{N}$ nests according to the workplace, i.e. each workplace j represents a nest that consists of $\mathbb{N}$ number of residence location alternatives. The underlying assumption is that, for any given employment location, the residence location alternatives are correlated. The existence of correlations among alternatives can be verified empirically. If the correlation among alternatives is significant, the multinomial logit model is no longer appropriate due to its strict limitation on the univariate structure. More flexible model specifications, such as the nested logit model, should be considered (Train, 2003).

Following the nested-logit structure, the joint probability of employed resident type f choosing to live in zone i and work in zone j is defined as the product of a marginal probability and a conditional probability:

$$P_{fij} = P_{fj}P_{fi|j}$$

where P_{fj} is the marginal probability for consumer f choosing to work in zone j, regardless of the residence location, and $P_{fi\,|\,j}$ is the conditional probability for consumer f choosing to live in zone i, given the chosen employment location j. The marginal and the conditional probability are defined as:

$$P_{fj} = \frac{S_j\, exp(\lambda_{f|J}(\upsilon_{fj} + V_{f|j}))}{\sum_k S_k\, exp(\lambda_{f|J}(\upsilon_{fk} + V_{f|k}))}$$

$$P_{fi|j} = \frac{S_i\, exp(\lambda_{f|I}\upsilon_{fi|j})}{\sum_m S_m\, exp(\lambda_{f|I}\upsilon_{fm|j})}$$

where

$$\upsilon_{fj} = ln\, M_{fj} + \psi_{fj} + E_{fj} + e_{fj}$$

$$\upsilon_{fi|j} = \widetilde{U}_{fi|j} - d_{fi|j} + \psi_{fi|j} + E_{fi|j} + e_{fi|j}$$

$$V_{f|j} = \frac{1}{\lambda_{f|I}}\, ln \sum_{i\in C_j} S_i\, exp(\lambda_{f|I}(\widetilde{U}_{fi|j} - d_{ij} + \psi_{fi|j} + E_{fi|j}))$$

υ_{fj} is the employment location utility of zone j for labour type f; $\upsilon_{fi\,|\,j}$ is the residence location utility of zone i for resident type f, given the chosen workplace j; $\lambda_{f\,|\,I}$ and $\lambda_{f\,|\,J}$ $(\lambda_{f\,|\,J}/\lambda_{f\,|\,I}\in(0,1])$ are dispersion parameters. $V_{f\,|\,j}$ term is called the log-*sum* or *inclusive utility* in the

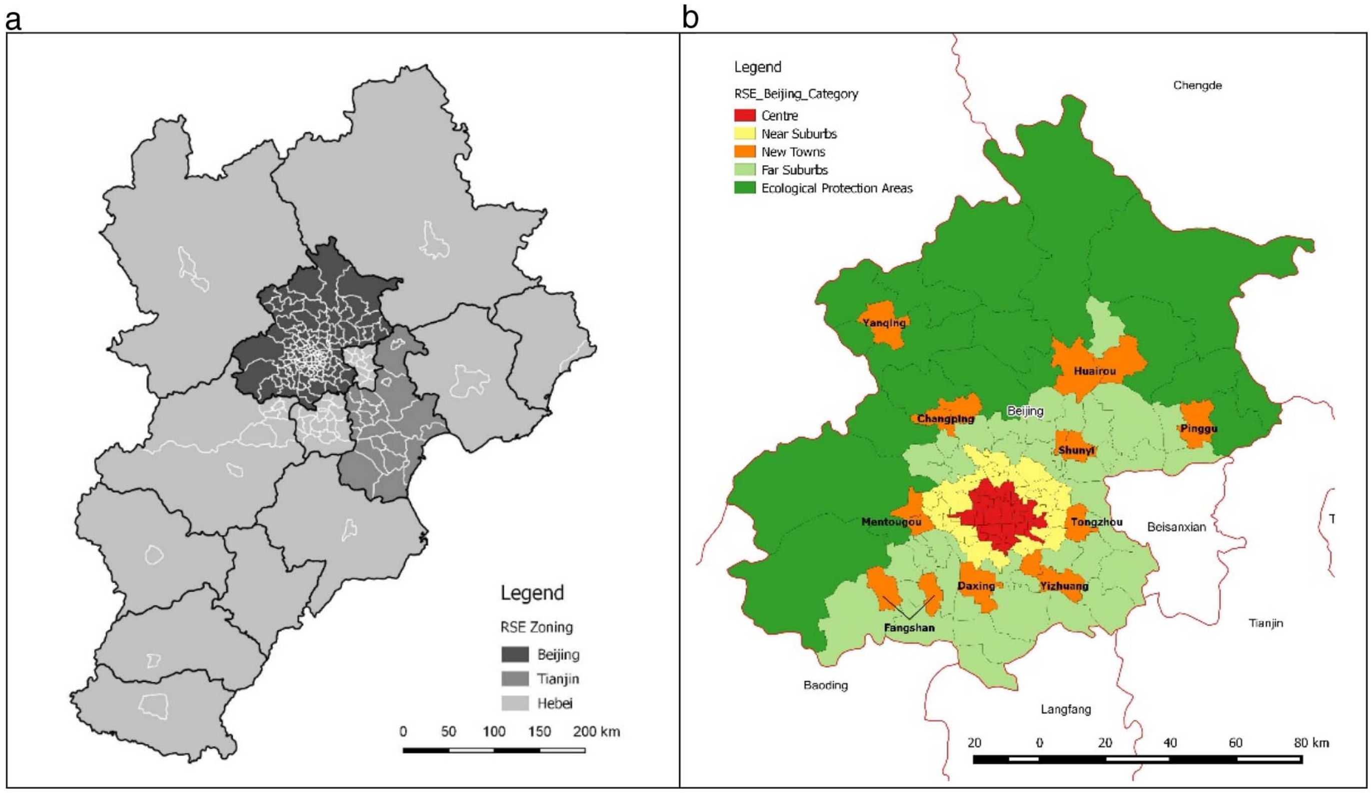

Fig. 4. a Zoning Map of Greater Beijing (Left). b Zoning Classification in Beijing (Right).

literature (Ben-Akiva & Bierlaire, 2003; Train, 2003), which represents the expected utility that employed worker type *f* in zone *j* would receive from all residence location choices, no matter which alternative is chosen. This expected utility enters as an explanatory variable into the upper-level employment location model. E_{fj} and $E_{fi\,|\,j}$ are zonal residual attractiveness terms.

To calibrate the nested logit model, we make use of the estimated commuting flows for Beijing 2010. The dispersion parameters (λ) are first estimated using maximum log-likelihood method. The residual terms are then estimated using standard least squares method, given the estimated dispersion parameters. We report the calibrated dispersion parameters and the goodness of fit in Table 3.

The estimated values of $\lambda_{f\,|\,I}$ are larger than one for all commuter types, which proves the existence of correlation among residence location alternatives. This finding justifies the use of the nested-logit model, instead of the multinomial logit model. In addition, we also find significantly lower R-squared values for the Low-SEG commuters in 2010. The underlying reason is that the residence locations of the Low group are geographically much more dispersed than the other groups, albeit their average commuting catchment is generally shorter. The Low-SEG commuters include agricultural workers whose residence location choices are subject to various historical and institutional factors.

Once the base year spatial equilibrium model is calibrated, parameter values are then retained for future model forecast. In forecast mode, the model requires exogenous development projections, such as the total number of employed residents, total building floorspace growth and transport supply. The model then predicts the spatial distribution of employment and residents.

Table 3
Dispersion parameter $\lambda_{f\,|\,I}$ of the location choice model for Beijing 2010.

	Maximum likelihood estimation		Regression analysis	
Consumer type *f*	Estimation value $\lambda_{f\,|\,I}$	Maximized log likelihood	R-squared	Sum of standard error (SSE)
2010				
$f=1$ High-SEG	3.0562	− 1.1980e + 07	0.8824	0.0013
$f=2$ Middle-SEG	3.6920	− 2.2592e + 08	0.8192	0.0054
$f=3$ Low-SEG	3.3725	− 2.0080e + 07	0.4643	0.0118

Note: we follow the tradition of nested-logit model calibration (Ben-Akiva & Bierlaire, 2003) by normalizing $\lambda_{f\,|\,J}=1$ for employment location choices of all employed residents. Thus only the $\lambda_{f\,|\,I}$ is estimated.

3.3. Scenario design for model forecast

The model forecast is based on the exogenous projections on total population, employment and land-use and transport development. The strategic demographic and floorspace growth assumptions for Greater Beijing 2015 are summarized in Table 4. Note that the assumptions are derived from a priori growth estimations, which include both official and third-party sources. The authors deem that the strategic assumption

Table 4
Strategic demographic and floorspace growth assumptions (Greater Beijing 2010–2015).

Item	2010 (Base year)	2015 (Forecast year)
Demographic growth		
Total population	104.55	118.13
Annualised growth rate	–	2.5%
Employed residents (million)	57.74	63.85
Annualised growth rate	–	2.0%
Employed-to-total-population ratio	0.55	0.54
Building floorspace stock		
Housing floorspace (million m^2)	3558.4	4659.7
Annualised growth rate	–	5.5%
Business floorspace (million m^2)	1154.9	1277.0
Annualised growth rate	–	2.0%

Table 5
Summary of the land-use and transport scenario (Greater Beijing 2010–2015).

Policy setting	Beijing					Rest of Greater Beijing
	Central districts	Near suburbs	New towns	Far suburbs	EPA[a]	
Land use development - natural growth						
Housing floorspace	✓	✓	✓	✓	✓	✓
Business floorspace	✓	✓	✓	✓	✓	✓
Land use development - discretionary growth						
Housing floorspace	×	✓	✓	✓	×	✓
Business floorspace	×	✓	✓	✓	×	✓
Transport development						
The access time between zones in Greater Beijing remains the same as in 2010, which implies a proportional transport supply growth with the population growth.						

[a] EPA: Ecological Protection Area.

Table 6
District-level number of residents and employment in Beijing 2015 (Benchmark).

District of Beijing	Permanent residents (10,000 persons)	Employed residents (ER) (10,000 persons)	Employed workers (EW) (10,000 persons)	EW/ER ratio
Dongcheng	90.5	46.7	100.2	2.1
Xicheng	129.8	65.8	147.6	2.2
Chaoyang	395.5	217.6	231.7	1.1
Fengtai	232.4	127.1	96.9	0.8
Shijingshan	65.2	36.0	30.3	0.8
Haidian	369.4	208.1	259.5	1.2
Fangshan	104.6	55.8	23.3	0.4
Tongzhou	137.8	77.5	34.5	0.4
Shunyi	102.0	56.5	71.3	1.3
Changping	196.3	108.1	44.9	0.4
Daxing	156.2	86.3	77.1	0.9
Mentougou	30.8	16.0	8.6	0.5
Huairou	38.4	20.6	14.6	0.7
Pinggu	42.3	22.2	19.0	0.9
Miyun	47.9	25.2	15.9	0.6
Yanqing	31.4	16.7	10.8	0.6
Total of Beijing	2170.5	1186.1	1186.1	1.0

represents a relatively fast growth trend in Greater Beijing.

Based on the above assumptions, a *Trend* scenario is then developed, which reflects the observed policy trend in Beijing from 2010. The scenario features an enhanced concentric urban growth pattern with moderate control of floorspace growth in central Beijing. Based on the five zoning categories of Beijing (see Fig. 6b), Table 5 summarizes the land-use and transport configurations of the *Trend* scenario.

In terms of land use development, we differentiate two types of floorspace growth as follows. *Natural growth* represents the spontaneous expansions of housing estates and business premises through infilling and densification in existing built-up areas. We follow Jin et al. (2017), assuming that the amount of natural growth is in proportion to existing zonal building stocks and applies to all locations. We further assume that natural growth accounts for 50% of the total projected floorspace growth. The second type is *discretionary growth* which addresses policy-oriented land development and subject to zoning regulations. In terms of transport development, the same level of network accessibility as in 2010 is assumed for 2015.

To distribute the projected aggregate floorspace growth to zones, a logit-type probability model is developed, which refers to not only the endogenous market variables from the base-year equilibrium, but also the durability of development and one-off policy interventions. The modelled zonal floorspace stock for 2015 are then used as exogenous inputs to predict the spatial location of employment, residents and the commuting flows across Greater Beijing in 2015.

4. Results and discussion

4.1. Benchmark statistics

To set up a benchmark for comparing the model prediction and the LBS data, we make use of the conventional statistics for Beijing 2015 (Beijing Bureau of Staistics, 2015), which is the core study area. Because no commuting flow data is available from existing data sources, we use the district-level total number of employed residents (ER) and employed workers (EW) as the benchmark values. For employment, total number of employment is only available at the municipal level (11.86 million in Beijing 2015), while the district-level statistics only cover legal-entity and non-agricultural workers, and do not include the self-employed and agricultural workers. To derive the district-level totals, we assume a fixed share of self-employed and agricultural workers (34.5%) in total employment for all districts in Beijing 2015. This share constant is calculated from the municipal-level statistics, and can be refined when detailed data become available in future.

For employed residents, we first calculate the percentage share of employed residents among 15–65 year-old permanent residents at municipal level (68.6%) in Beijing 2015. This share is then used to estimate the district-level total number of employed residents. Table 6 presents the estimated district-level employed residents and employed workers in Beijing in 2015. In terms of the EW/ER ratio, the central districts tend to have larger-than-one ratio, implying more jobs than local employed residents. For the whole of Beijing, it is assumed that the total number employed residents equals the number of employed workers (11.86 million), thus the EW/ER ratio equals one. Note that no observed data for Beijing 2015 are used in model prediction.

4.2. Comparison: district-level totals

Given the district-level benchmark, the employed residents (ER) and employed workers (EW) derived from the model and the LBS data are compared in Figs. 5 and 6. Note that the LBS data is essentially a sampled data in nature thus the district-level totals are quantitatively smaller than the benchmark values. By contrast, the model predicted flows represent the full population. In order to facilitate the comparison, a secondary vertical axis is introduced in the figures for plotting the LBS data.

Comparing the model prediction with the benchmark, a high level of goodness of fit is shown for both ER and EW. Nonetheless, the model overestimated the total number of ER (14.03 million) and EW (14.00 million) for Beijing 2015, compared with the actual total (11.86 million). In particular, relatively large overestimation is seen at Chaoyang district for both ER and EW, and also at Tongzhou and Changping district for EW. The error is likely to be attributed to the exogenous projection of floorspace growth. The *Trend* scenario features moderate control of floorspace growth in the centre, while the actual level of planning control in Chaoyang may well be stricter than the scenario. Nonetheless no housing or business floorspace data is available in Beijing 2015 for verification purpose. In fact acquiring floorspace data for non-census years is very difficult in this particular context.

By contrast, a larger discrepancy is shown between the LBS data and the benchmark ER and EW. To investigate the varying errors among district locations, Table 7 presents the district-level scale factors that may potentially correct the sampling errors of the LBS data. We define the scale factor as the ratio between the benchmark and the processed ER and EW from the LBS data. We also divide the districts into three categories, which are in line with the masterplan of Beijing.

First, the scale factors vary significantly across locations, which indicates that using one uniform scale factor for the whole study area is

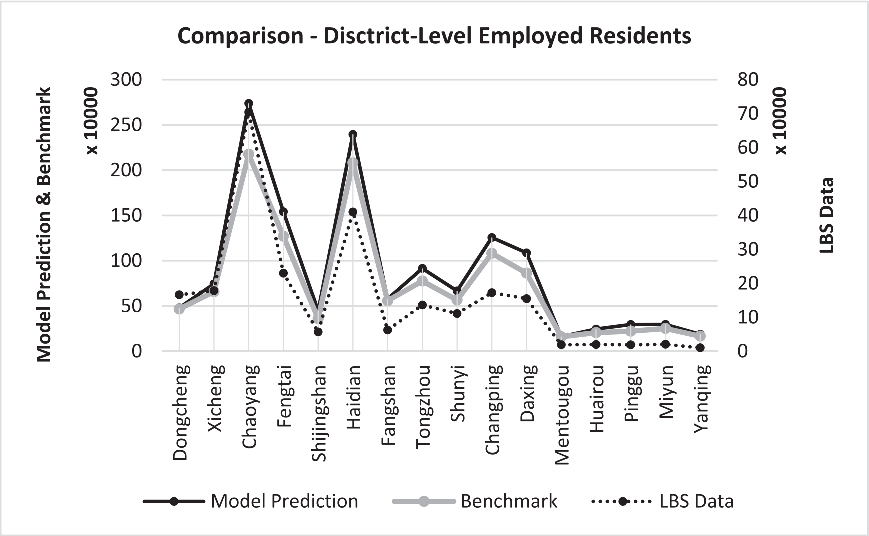

Fig. 5. Comparison of district-level employed residents in Beijing 2015.

not a feasible scaling method. Varying scale factors are thus suggested for scaling purpose, which need to be examined not only statistically but also geographically. Secondly, for employed residents, the scale factors tend to be smaller in the Central Districts than those in Suburban and Ecological Protection Districts. The authors deem that this factor gradient is reasonable because the sampling rate of LBS data is likely to be higher in relatively more developed central districts, where more people as well as smart devices are concentrated. Thirdly, for employment workers, however, the factor gradient is less significant. In particular, the scale factor for Dongcheng (8.03) and Xicheng (9.96) is much higher than the municipal average (4.76), albeit they are both central districts. It implies that the proposed data processing method for detecting employment locations may be ineffective for some locations. Because such inconsistency occurs even within the same district category, the varying sampling rate is unlikely to be the main cause. One possible reason is that people's daytime movements tend to be more dynamic and variable than those at night, when people generally stay at home. The variability of people's daytime locations increases the difficulty of detecting a stable workplace for LBS data.

The comparison results seem to imply that the system model approach performs better than the LBS data in terms of goodness of fit. However it should be noted that the outperformance of the system model approach does not indicate its superiority over the LBS data approach. This is because the outperformance may be attributed to the model calibration procedure, in which the historical spatial pattern of residents and employment is reproduced by the model at the base year (2010). Urban development processes tend to be inertia-prone and path-dependent. Drastic changes may happen at local level but the aggregate spatial pattern would remain relatively stable over a short period of time in the absence of political or economic shocks. The use of historical data in model calibration may contribute to the good performance of the model. Therefore the comparison in this section is for verifying the results of each approach, rather than comparing the analytical capability of the two methods.

4.3. Comparison: origin-destination analysis

The comparison of district-level ER and EW totals provides an overview of the quality of the derived flows for Beijing 2015. The next step is to extend the discussion by comparing the derived commuting pattern in terms of origin-destination pairs. As no observed commuting flows are available from existing statistical sources, thus the model-predicted flows and the LBS-data flows are compared horizontally. Fig. 7 presents the scatter plot of the model-predicted flows and the LBS data at district-pair level (16 districts in Beijing thus 256 district pairs including intra-district flows). To prevent biased fitting due to skewed

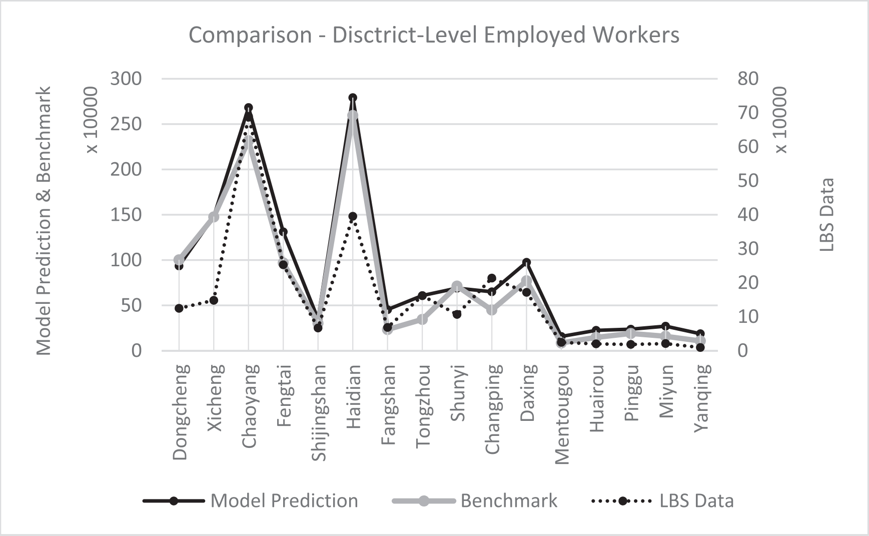

Fig. 6. Comparison of district-level employed workers in Beijing 2015.

Table 7
District-level scale factors of the LBS data - Beijing 2015.

	Scale factor: benchmark/LBS Data	
	Employed residents	Employed workers
Central districts		
Dongcheng	2.80	8.03
Xicheng	3.69	9.96
Chaoyang	3.09	3.36
Fengtai	5.53	3.83
Shijingshan	6.33	4.53
Haidian	5.07	6.56
Suburban districts		
Fangshan	8.92	3.41
Tongzhou	5.69	2.13
Shunyi	5.09	6.67
Changping	6.27	2.10
Daxing	5.56	4.49
Ecological protection districts		
Mentougou	8.31	3.49
Huairou	10.29	7.25
Pinggu	11.62	10.27
Miyun	12.29	7.57
Yanqing	16.22	11.37
Total of Beijing	4.80	4.76

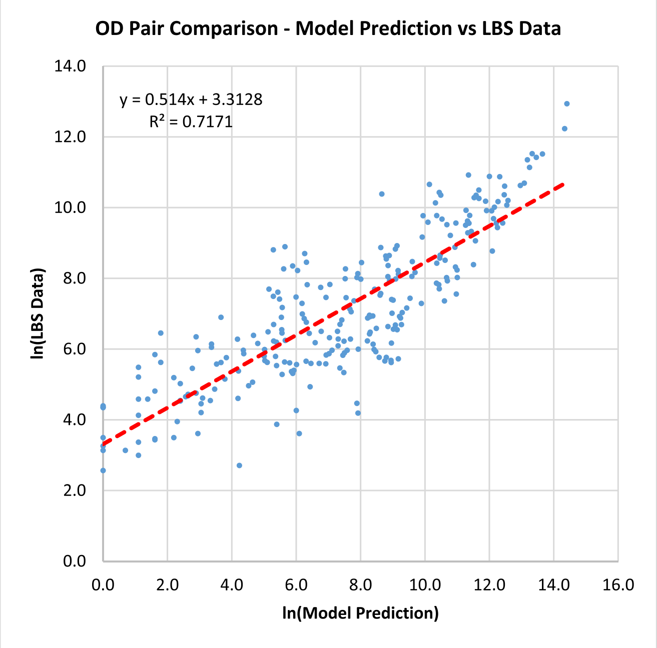

Fig. 7. District-level OD pair comparison - model prediction vs LBS data.

Table 8
OD flows comparison for selected districts in Beijing 2015 – model prediction vs LBS data.

Destination	Model prediction			LBS data		
	Origin			Origin		
	Xicheng	Tongzhou	Pinggu	Xicheng	Tongzhou	Pinggu
Dongcheng	10.4%	5.0%	0.4%	7.7%	2.1%	1.4%
Xicheng	64.7%	3.9%	0.4%	25.1%	1.7%	1.8%
Chaoyang	4.8%	19.9%	0.8%	17.9%	15.5%	9.3%
Fengtai	4.1%	1.1%	0.1%	14.4%	2.6%	2.8%
Shijingshan	0.4%	0.1%	0.0%	2.7%	0.4%	0.7%
Haidian	14.5%	1.8%	0.3%	17.8%	2.7%	3.4%
Fangshan	0.1%	0.0%	0.0%	1.4%	0.5%	0.6%
Tongzhou	0.0%	60.5%	0.1%	2.2%	65.5%	2.5%
Shunyi	0.0%	0.9%	4.7%	1.2%	1.2%	4.3%
Changping	0.0%	0.0%	0.0%	4.2%	1.4%	2.5%
Daxing	0.7%	6.4%	0.0%	4.1%	5.5%	1.9%
Mentougou	0.1%	0.0%	0.0%	0.5%	0.2%	0.2%
Huairou	0.0%	0.0%	0.2%	0.2%	0.2%	0.4%
Pinggu	0.0%	0.0%	91.8%	0.2%	0.2%	67.4%
Miyun	0.0%	0.0%	1.2%	0.2%	0.2%	0.5%
Yanqing	0.0%	0.0%	0.0%	0.1%	0.1%	0.2%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 9
Comparison of commuting statistics in Beijing - model prediction vs LBS data.

	Average commuting distance (km)	Average commuting time (min)
Model Prediction		
Beijing	10.7	45.2
Xicheng	5.2	34.8
Tongzhou	13.6	47.6
Pinggu	14.4	40.7
LBS data		
Beijing	15.3	56.5
Xicheng	16.9	61.4
Tongzhou	20.5	61.6
Pinggu	35.6	79.0

Average commuting distance and time is measured from resident locations and is the weighted average of all commuters to all employment destinations.

data distribution, natural logarithm is applied to both data. A significant linear correlation is shown between the two derived flows with an R-square value 0.72.

To further investigate the difference between the two derived flows, three districts in Beijing are selected as residence locations (i.e. origins) and Table 8 presents the percentage share of commuters by destination including the origin district itself. The three selected districts cover different location categories in Beijing, namely Xicheng representing the urban core, Tongzhou representing the suburban new town and Pinggu representing the ecological protection area. Due to the sampling nature of the LBS data, we focus on the percentage share, rather than the actual flow volume. As a further exercise, we make use of the network distance/time matrices to calculate the average one-way commuting distance/time for both derived flows, which are presented in Table 9. Note that the derived average commuting distance/time are based on the same transport network and congestion level as applied in the model.

The OD comparison (Table 8) reveals three findings. First, both the model and the LBS data show that the majority of employed residents commute within their residence districts, albeit the percentage share of intra-district commuting varies among locations. For the model predicted flows, the intra-district commuting in both central and rural district is higher than that in suburban new town. For central district where jobs concentrate, local employed residents are likely to work in the centre rather than commuting outward to the suburbs. For far suburbs, although a certain share of residents would still commute to the centre, the majority tends to work locally, particularly the agricultural workers. However, for new towns at an intermediate stage of development, local jobs may not be sufficient to support the local labour force, employed residents thus may have to commute to the centre. Meanwhile the new towns also accommodate the workers fled from the centre due to the high living cost in the centre, which together explain the lower share of intra-district commuting for new towns.

Second, the LBS data show much longer average distance and time travelled than the model predicted flows. For the whole Beijing, the average commuting distance (time) of the model predicted flows and the LBS data is 10.7 km (45.2 min) and 15.3 km (56.5 min), respectively. To articulate the discrepancy, we refer to the municipal annual transport survey, which provides aggregate transport statistics of

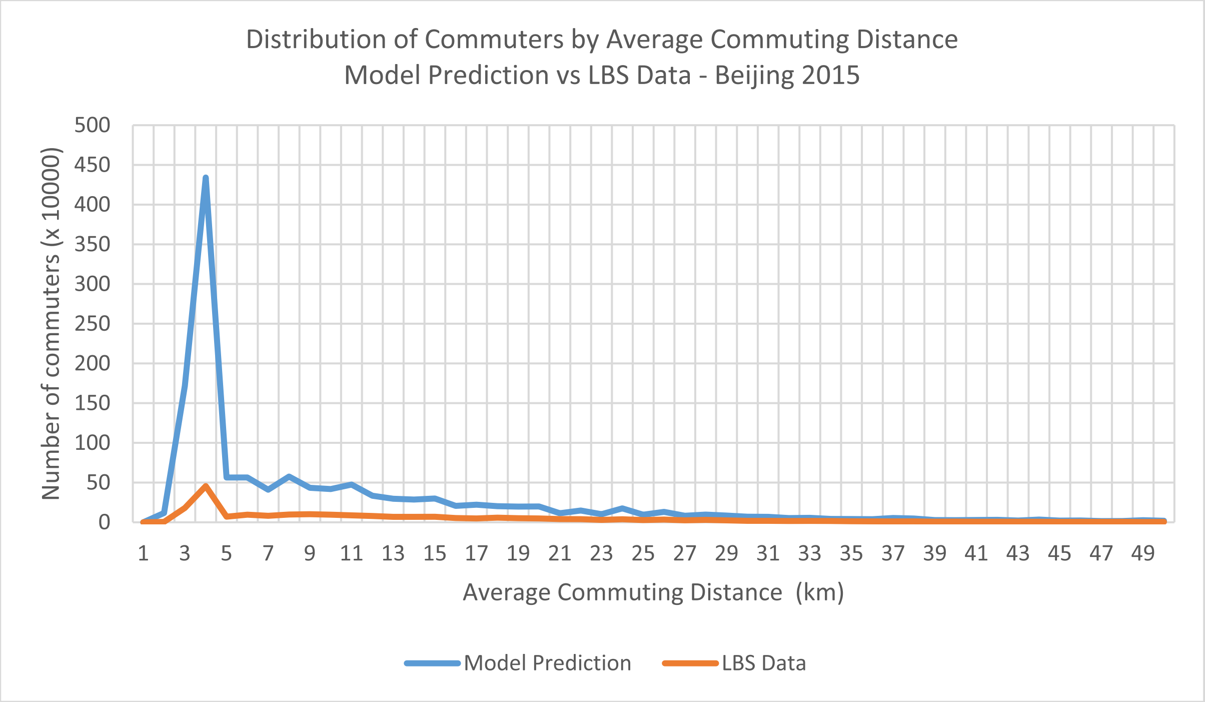

Fig. 8. Commuter distribution by average commuting distance - Model prediction vs LBS data.

Beijing. According to the latest published survey (BCT & BTRC, 2015), the average commuting distance and time in Beijing is 10.1 km and 48 min respectively. Given this data as the benchmark, the LBS data seem to overestimate the commuting catchment. The authors deem that the overestimation is mainly caused by the underestimated share of intra-district commuting.

To confirm the hypothesis, Fig. 8 presents the distribution of commuters by distance travelled (one-way) of the two derived flows. It shows that the LBS data have significantly lower share of short-distance (0–5 km) commuting (29.5%) than the model predicted flows (48.5%). Given the average size of districts in Beijing, these short-distance travels are typically intra-district. In terms of the reason why the LBS data would underestimate the short-distance commuting, one plausible explanation is that the spatial proximity of home and workplace implies that the commuter may know the area very well, which in turn reduces the need for using location-based services. Nonetheless the less usage of LBS for short-distance commuters is yet to be confirmed with further empirical evidence.

It should be noted that the underestimation of intra-district commuting in the LBS data appears in both central (e.g. Dongcheng) and rural (e.g. Yanqing) locations (see Fig. 9), but the source of errors may be different. For central locations, errors may be caused by the ineffectiveness of detecting employment locations; while for rural locations, underestimation may be attributed to the low LBS device usage due to the prevalence of lower-class workers. Interestingly, for some districts like Tongzhou, the share of intra-district commuting given by the LBS data (65.5%) is similar to the model-predicted fig. (65.6%). The inconsistency is likely to be case-specific but again it implies the uneven locational bias of the LBS data.

The third finding is that both the model and LBS data show shorter commuting distance/time in central districts than suburban districts. This is in line with the monocentric employment pattern in Beijing, where many suburban workers have to commute inwards to the centre on a daily basis. According to the model simulation, the current employment agglomeration in central Beijing is likely to hold in the near future, while strong decentralization policy is underway. The decentralization scheme would foster employment development in suburbs, which may eventually mitigate the super-commuting issue.

It should be pointed out that the geographical granularity of the LBS data is much finer than the model. In this paper, the LBS data are aggregated according to the model zoning system as a compromise, such that the results can be readily compared. The LBS data record accurate user location in a nearly continuous space, while the proposed model is operated on a discrete zoning system. Such zoning system tends to be geographically aggregate particularly for spatial equilibrium models in order to reduce the burden on data and computing. The LBS data thus provides a much finer perspective to examine the spatial and behavioural heterogeneity than conventional system models.

5. Conclusions

This paper has two purposes, 1) empirically comparing a data-driven approach and a system modelling approach in terms of modelling the commuting pattern for Beijing 2015, and 2) identifying the strength and weakness of each approach, including the prospect of a joint approach. The LBS data used in this paper are collected from personal smart-device users. For the Greater Beijing city region, the processed data include a total of 11,520 million location records from about 8.1 million LBS users. The DBSCAN method is applied to infer the employment and residence location of the LBS users using the daytime and night-time records respectively.

Parallel to the data-driven approach, a LUTI type urban spatial equilibrium model is developed to predict the commuting flows for Beijing 2015. To estimate the base-year commuting flows, a nested-logit discrete choice model with log-linear travel cost function is applied, which represents a useful alternative when observed commuting flows are not available. The model prediction for 2015 is based on exogenous projections on demographic, employment, land-use and transport development for Greater Beijing.

To establish a benchmark for comparison, district-level totals of employed residents and workers are processed from conventional statistics of Beijing 2015. The comparison reveals that the proposed LBS data processing method performs better in detecting residence locations than employment locations. This finding may also be applicable to

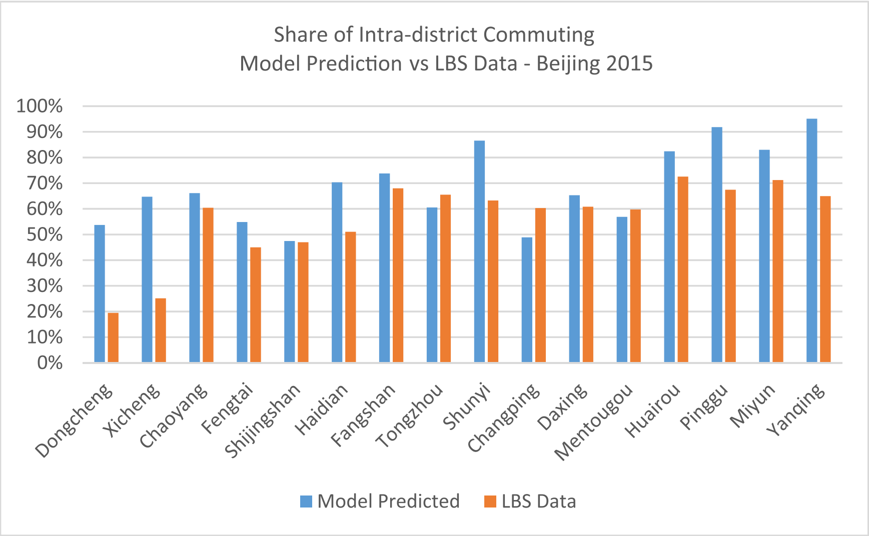

Fig. 9. Share of intra-district commuting - modelled predicted vs LBS data.

other types of LBS data, given that people's movement pattern at nigh time is likely to be more stable thus recognisable than that at daytime. The comparison also shows that the sampling errors of the LBS data vary significantly among district locations. For scaling the sample data to represent the full population, location-specific scale factors should thus be considered.

Based on the observed aggregate travel statistics in Beijing 2015, the LBS data is found to overestimate the average commuting distance and time for Beijing. By comparing the LBS data with the model-predicted flows, the authors find that the overestimation is caused by the ineffectiveness of representing short-distance (0–5 km) commuting in the LBS data. One plausible yet not verified hypothesis is that the spatial proximity of home and workplace may reduce the use of location-based services. More empirical evidence is required in order to understand such implicit biases of the LBS data.

In light of the findings, the collaborative use of big data and urban system models seems to be a win-win strategy that would enable improvements that neither approach could achieve in isolation. On the one hand, the LBS data, once well processed, can serve as a novel data source for calibrating and validating the urban system model, which would enhance the model's empirical foundation as well as its predictability in scenario analysis. The LBS data are collected much more frequently than conventional statistics, thus provide useful insights on how the urban stocks and flows change over time, which would inspire new model design to address temporal dynamics in urban development. The fine granularity of the big data would also help urban system models to embrace the spatial heterogeneity. On the other hand, theoretically sound models can provide good reference for investigating the possible errors and biases of the big data, and help big data to overcome its descriptive nature by probing into the derived yet articulated correlations and exploring the underlying causalities. The findings would in turn shed new light on how big data can be better processed and used. This paper is but a very small and experimental step towards the joint approach.

Acknowledgement

For Li Wan and Ying Jin, this research is supported by special fund of Key Laboratory of Eco Planning & Green Building, Ministry of Education (Tsinghua University), China.

Appendix A. Statistics of estimated commuting matrices - Greater Beijing 2010

Commuting statistics - Greater Beijing 2010	Beijing	Tianjin	Hebei	Great Beijing Average
High-socio-economic commuters				
Average commuting distance (km)	11.0	N/A[a]	N/A	11.0
Average monetary cost (fen = 0.01 yuan)	306.0	N/A	N/A	306.0
Average travel time (min)	40.0	N/A	N/A	40.0
Middle-socio-economic commuters				
Average commuting distance (km)	8.7	7.8	7.4	7.6
Average monetary cost (fen = 0.01 yuan)	149.2	93.2	92.7	100.0
Average travel time (min)	35.0	26.6	26.1	27.3
Low-socio-economic commuters				
Average commuting distance (km)	6.1	N/A	N/A	6.1
Average monetary cost (fen = 0.01 yuan)	81.0	N/A	N/A	81.0
Average travel time (min)	29.5	N/A	N/A	29.5
All commuters				
Average commuting distance (km)	8.3	7.8	7.4	7.7
Average monetary cost (fen = 0.01 yuan)	155.2	93.2	92.7	105.6
Average travel time (min)	34.2	26.6	26.1	27.8

[a] For lack of data, we model all employed residents in Tianjin and Hebei as one composite group. We further assume that the composite group share the same socio-economic

characteristics as the Middle group in Beijing.

Appendix B

Producers. We follow Anas and Liu (2007) and Jin et al. (2013), and define the production function as a variant of the CD-CES specification.

$$X_{rj} = E_{rj}A_{rj}(K_r)^{\nu_r}\left(\sum\nolimits_f \kappa_{rfj}L_{fj}^{\theta_r}\right)^{\frac{\delta_r}{\theta_r}}\left(\sum\nolimits_k \chi_{rkj}B_{kj}^{\zeta_r}\right)^{\frac{\mu_r}{\zeta_r}}$$

where

X_{rj}: production output of industry r in zone j;
E_{rj}: a constant scalar representing any additional zonal effects on total factor productivity;
A_{rj}: a function of the economic mass that represents the total factor productivity;
Productivity (TFP) effects;
K_r, L_{fj}, B_{kj}: capital, labour and business floorspace input;
ν_r, δ_r, μ_r: cost share parameters for the respective input group ($\nu_r + \delta_r + \mu_r = 1$);
κ_{rfj}, χ_{rkj}: input-specific constants for labour and business floorspace varieties;
θ_r, ζ_r: elasticity of substitution between any two labour and building floorspace varieties.

The *conditional input demand* (given target output X_{rj}) of each input factor can be derived as follows:

$$K_r = \frac{1}{\rho}\nu_r p_{rj}X_{rj}$$

$$L_{rfj} = \frac{\kappa_{rfj}^{\frac{1}{1-\theta_r}} w_{fj}^{\frac{1}{\theta_r-1}}}{\sum\nolimits_s \kappa_{rsj}^{\frac{1}{1-\theta_r}} w_{sj}^{\frac{\theta_f}{\theta_r-1}}}\delta_r p_{rj}X_{rj}$$

$$B_{rkj} = \frac{\chi_{rkj}^{\frac{1}{1-\zeta_r}} R_{kj}^{\frac{1}{\zeta_r-1}}}{\sum\nolimits_s \chi_{rsj}^{\frac{1}{1-\zeta_r}} R_{sj}^{\frac{\zeta_r}{\zeta_r-1}}}\mu_r p_{rj}X_{rj}$$

where

p_{rj}: unit production price of industry r in zone j;
ρ: exogenous price of business capital (i.e. the real interest rate);
w_{fj}: hourly wage of labour type f;
R_{kj}: average rent for business floorspace type k;
p_{rj}: production price.

Final consumers. Final consumers are categorized into $f = 1, \ldots, F$ types according to their employment status and socio-economic level. Following the random utility framework (McFadden, 1973), the observable utility U_{fij} is given by:

$$U_{fij} = \alpha_f \ln\left(\sum\nolimits_r \left(\sum\nolimits_z Z_{rz|fij}\right)^{\eta_f}\right)^{\frac{1}{\eta_f}} + \beta_f \ln\left(\sum\nolimits_m \iota_{mfi}(b_{m|fij})^{\sigma_f}\right)^{\frac{1}{\sigma_f}} + \gamma_f \ln l_{fij}$$

subject to budget constraint: $\sum\nolimits_{r,z}(p_{rz} + c_f 2g_{fiz})Z_{rz|fij} + \sum\nolimits_m r_{mi}b_{m|fij} + \Delta_f 2Dg_{fij} = \Delta_f w_{fj}\left(N - 2DG_{fij} - \sum\nolimits_{r,z} c_f Z_{rz|fij}2G_{fiz} - l_{fij}\right) + \mathcal{M}_{fi}$

and time constraint: $N - \sum\nolimits_{r,z} c_f Z_{rz|fij}2G_{fiz} - \Delta_f(l_{fij} + 2DG_{fij}) \geq 0$

We assume Cobb-Douglas preference between goods & services $Z_{rz|fij}$, housing $b_{m|fij}$ and leisure time l_{fij}. $\alpha_f + \beta_f + \gamma_f = 1$ are the expenditure coefficients for each consumption bundle. The varieties of goods & services and housing are assumed to be imperfect substitutes (Dixit & Stiglitz, 1977), and the elasticity of substitution is governed by η_f and σ_f for goods & services and housing, respectively. $\iota_{mfi} > 0$ is the input-specific constant measuring the inherent attractiveness of the housing type m for consumers type f living in zone i, which is calibrated empirically.

For the budget constraint, the right-hand side of the function is the total income and the left-hand side is the total expenditure. Specifically, p_{rz} is the mill price for goods & services type r produced in zone z; g_{fiz} and G_{fiz} is the expected one-way monetary cost and travel time from i to z for customers type f, respectively; c_f is an exogenous coefficient that measures the cost for delivering a unit of goods & services as percentage of the normal trip cost. r_{mi} is the housing rent of type m in zone i; w_{fj} is the hourly wage rate for labour type f working in zone j. Δ_f is the employment status of the consumer type f. For all employed consumers $\Delta_f = 1$; otherwise $\Delta_f = 0$. $\mathcal{M}_{fi}$ is the nonwage income of consumer type f in zone i. It consists of normal investment returns on real estate in the city region (endogenous in the model) as well as the individual share of social welfare transfer and amenity gains. As for the time constraint, D is the exogenous number of working days per annum; $N = 24D$ is the exogenous total annual time endowment. For the non-employed consumers ($\Delta_f = 0$), the model only accounts for the time for shopping, as they do not commute and have zero value of time for leisure time.

Under the above budget and time constraint, the *Marshallian* demand for goods & services, housing and leisure time can be derived as the following.

$$\overline{Z}_{r|fij} = \frac{\overline{p}_{r|fij}{}^{\frac{1}{\eta_f-1}}}{\sum\nolimits_s \overline{p}_{s|fij}{}^{\frac{\eta_f}{\eta_f-1}}}\alpha_f \Omega_{fij}$$

$$b_{m|fij} = \frac{\iota_{mfi}^{\frac{1}{1-\sigma_f}} r_{mi}^{\frac{1}{\sigma_f-1}}}{\sum_s \iota_{si}^{\frac{1}{1-\sigma_f}} r_{si}^{\frac{\sigma_f}{\sigma_f-1}}} \beta_f \Omega_{fij}$$

$$l_{fij} = \frac{\gamma_f \Omega_{fij}}{w_{fj}}$$

where

$\overline{Z}_{r|fij}$: aggregate demand for product type r for consumer type (fij);

$\overline{p}_{r|fij}$: probability-weighted average price of product type r faced by consumer type (fij).

Equilibrium Conditions. The general equilibrium structure of the RSE model requires four sets of equilibrium conditions to be satisfied simultaneously, conditional on the transport conditions **g** and **G**.

1) All consumers maximize utility subject to budget and time constraint.
2) All producers minimize cost subject to supply constraint of input factors and technology. Producers are competitive and operate under constant returns to scale. The minimized production price equals the average and marginal cost, implying zero economic profit.
3) The endogenous part of nonwage incomes must be consistent with the regional building stocks and the associated rental levels. Other sources of nonwage income are specified as exogenous input.
4) All markets clear with zero excess demands. This applies to: a) the residential and business floorspace markets; b) the labour market for each socio-economic group at each production zone; c) the product market of each product type at each production zone.

References

Ahas, R., & Mark, Ü. (2005). Location based services - New challenges for planning and public administration? *Futures, 37*(6), 547–561. http://dx.doi.org/10.1016/j.futures.2004.10.012.

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology, 17*(1), 3–27.

Anas, A. (2013). A summary of the applications to date of RELU-TRAN, a microeconomic urban computable general equilibrium model. *Environment and Planning B: Planning and Design, 40*(6), 959–970.

Anas, A., & Liu, Y. (2007). A regional economy, land use, and transportation model (RELU-TRAN??): Formulation, algorithm design, and testing. *Journal of Regional Science, 47*(3), 415–455. http://dx.doi.org/10.1111/j.1467-9787.2007.00515.x.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine, 16*(7), 7–16.

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography, 3*(3), 274–279. http://dx.doi.org/10.1177/2043820613513390.

BCT, & BTRC (2010). *Beijing travel survey.* China: Beijing. Retrieved from http://www.bjtrc.org.cn.

BCT, & BTRC (2015). *Beijing travel survey.* China: Beijing. Retrieved from http://www.bjtrc.org.cn.

Beijing Bureau of Statistics (2015). *Beijing statistical yearbook 2015*.

Ben-Akiva, M. (1974). *The Structure of Travel Demand Models.* PhD thesis. MIT Ben-Akiva.

Ben-Akiva, M., & Bierlaire, M. (2003). Discrete choice methods and their applications to short-term travel decisions. *Handbook of transportation science*. *1985*. *Handbook of transportation science* (pp. 7–37). . http://dx.doi.org/10.1007/0-306-48058-1_2.

Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand. Vol. 9.* MIT Press.

Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping multidimensional data* (pp. 25–71). Springer.

Bröcker, J. (1998). Operational spatial computable general equilibrium modeling. *The Annals of Regional Science, 32*(3), 367–387.

Cai, J., Huang, B., & Song, Y. (2017). Using multi-source geospatial big data to identify the structure of polycentric cities. *Remote Sensing of Environment*0034-4257http://dx.doi.org/10.1016/j.rse.2017.06.039http://www.sciencedirect.com/science/article/pii/S0034425717302985.

Daly, A., & Zachary, S. (1978). Improved multiple choice models. *Determinants of travel choice. 335. Determinants of travel choice* (pp. 357–). .

Deng, B., Denman, S., Zachariadis, V., & Jin, Y. (2015). Estimating traffic delays and network speeds from low-frequency GPS taxis traces for urban transport modelling. *EJTIR, 15*(4), 639–661.

Dixit, A. K., & Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *American Economic Review, 67*(3), 297–308. http://dx.doi.org/10.1016/S0167-7187(98)00007-1.

Echenique, M. H., Flowerdew, A. D. J., Hunt, J. D., Mayo, T. R., Skidmore, I. J., & Simmonds, D. C. (1990). The MEPLAN models of Bilbao, Leeds and Dortmund. *Transport Reviews, 10*(4), 309–322.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd. Vol. 96. Kdd* (pp. 226–231).

Evans, A. W. (1970). Some properties of trip distribution methods. *Transportation Research, 4*(1), 19–36. http://dx.doi.org/10.1016/0041-1647(70)90072-9.

Fratar, T. J. (1954). Vehicular trip distribution by successive approximations. *Traffic Quarterly, 8*(1).

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS, 17*(3), 463–481.

Horowitz, A. J. (2005). *Tests of a family of trip table refinements for long-range, quick-response travel forecasting, (1921)*. 19–26. http://dx.doi.org/10.3141/1921-03.

Jin, Y., Denman, S., Deng, D., Rong, X., Ma, M., Wan, L., ... Long, Y. (2017). Environmental impacts of transformative land use and transport developments in the Greater Beijing Region: Insights from a new dynamic spatial equilibrium model. *Transportation research part D: Transport and environment*.

Jin, Y., Echenique, M., & Hargreaves, A. (2013). A recursive spatial equilibrium model for planning large-scale urban change. *Environment and Planning B: Planning and Design, 40*(6), 1027–1050. http://dx.doi.org/10.1068/b39134.

Kieu, L. M., Bhaskar, A., & Chung, E. (2014). *Transit passenger segmentation using travel regularity mined from smart card transactions data*.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography, 3*(3), 262–267. http://dx.doi.org/10.1177/2043820613513388.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), http://dx.doi.org/10.1177/2053951714528481.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS One, 9*(6), e96180.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Long, Y., Liu, X., Zhou, J., & Chai, Y. (2016). Early birds, night owls, and tireless/recurring itinerants: An exploratory analysis of extreme transit behaviors in Beijing, China. *Habitat International, 57*, 223–232.

Long, Y., & Shen, Z. (2015). Profiling underprivileged residents with mid-term public transit smartcard data of Beijing. *Geospatial analysis to support urban planning in Beijing* (pp. 169–192). Springer.

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. *Computers, Environment and Urban Systems, 53*, 19–35.

Lowry, I. S. (1964). *A model of metropolis.* Santa Monica, CA, USA: Rand Corporation.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems, 36*(2), 105–108. http://dx.doi.org/10.1016/j.compenvurbsys.2012.02.002.

Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography, 58*, 135–145. http://dx.doi.org/10.1016/j.jtrangeo.2016.12.001.

McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*.

Murchland, J. D. (1966). *Some remarks on the gravity model of traffic distribution, and an equivalent maximization formulation.* London School of Economics and Political Science, Transport Network Theory Unit.

National Bureau of Statistics PRC (2016). *China statistical yearbook 2016.* Beijing, China: China Statistics Press.

Roth, C., Kang, S. M., Batty, M., & Barthélemy, M. (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS One, 6*(1), e15923.

Schintler, L. A., & Chen, Z. (2017). *Big data for regional science.* Routledge.

Seaborn, C., Attanucci, J., & Wilson, N. (2009). Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record: Journal of the Transportation Research Board, 2121*, 55–62.

Shen, Y., & Chai, Y. W. (2012). Study on commuting flexibility of residents based on GPS data: A case study of suburban mega-communities in Beijing. *Acta Geographica Sinica, 67*(6), 733–744.

Shen, Y., Kwan, M. P., & Chai, Y. (2013). Investigating commuting flexibility with GPS data and 3D geovisualization: A case study of Beijing, China. *Journal of Transport Geography, 32*, 1–11. http://dx.doi.org/10.1016/j.jtrangeo.2013.07.007.

Train, K. E. (2003). Discrete choice methods with simulation. *Discrete choice methods with simulation. Vol. iii–viii*http://dx.doi.org/10.1017/CBO9780511753930.

Volterra, & CBP (2007). *Modelling transport and the economy in London*.

Wegener, M. (1998). *The IRPUD model: Overview.* Website of the Institute of Spatial Planning, University of Dortmund.

Wegener, M., Mackett, R. L., & Simmonds, D. C. (1991). One city, three models: Comparison of land-use/transport policy simulation models for Dortmund. *Transport Reviews, 11*(2), 107–129. http://dx.doi.org/10.1080/01441649108716778.

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research, 1*(3), 253–269.

Zandbergen, P. A., & Barbeau, S. J. (2011). Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. *Journal of Navigation, 64*(3), 381–399.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., & Schmitt, G. (2016). Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PLoS One, 11*(2), e0149222.

Zhong, C., Manley, E., Arisona, S. M., Batty, M., & Schmitt, G. (2015). Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science, 9*, 125–130.