International review for spatial planning and sustainable development, Vol.5 No.3 (2017), 100-115
ISSN: 2187-3666 (online)
DOI: http://dx.doi.org/10.14246/irspsd.5.3_100

Big data for intrametropolitan human movement studies

A case study of bus commuters based on smart card data

Jiangping Zhou[1] , Mingshu Wang[2*] and Ying Long[3]
1 Department of Urban Planning and Design, School of Architecture, University of Hong Kong
2 Department of Geography, University of Georgia
3 School of Architecture, Tsinghua University
** Corresponding Author, Email: mswang@uga.edu*

Key words: Big data, human movement, intra-metropolitan, bus

Abstract: Unlike the data from traditional sources, there have not been standard ways to validate the quality and reliability of information derived from big data. This article argues that the theory of urban formation can be used to do the validation. In addition, the information derived from big data can be used to verify and even extend existing theories or hypotheses of urban formation. It proposes a general framework regarding how the theory of urban formation can be employed to validate information derived from smart card data and how the validated information can supplement other data to reveal spatial patterns of economic agglomeration or human settlements. Through a case study of Beijing, it demonstrates the usefulness of the framework. Additionally, it utilizes smart card data to delineate characteristics of subcenters defined by bus commuters of Beijing.

1. INTRODUCTION

Human movements and related activity centers at the intrametropolitan level have been a topic of lasting interest to geographers, planners, modelers and the like (Cervero, 1998; Hanson & Giuliano, 2004; Salas-Olmedo & Nogués, 2012; de Dios Ortúzar & Willumsen, 1990). Data and information from traditional sources such as field trips, interviews, archives, surveys and censuses dominate related studies. Only in recent years have passive user-generated big data such as smart card data been introduced in those studies (Tao et al., 2014; Kim et al., 2014; Briand et al., 2017; Wang, M. et al., 2016). Existing studies based on smart card data have demonstrated that smart card data can be used to reveal the spatial-temporal dynamics of bus trips, to identify subway trip between stations and to detect zones that share trip origins or destinations in proximity. It is argued that smart card data could support evidence-based transit planning (Tao et al., 2014) and could facilitate the simultaneous discovery of zones and subway passenger movements between these zones (Kim et al., 2014).

Little has been done, however, on how smart card data can be used to verify the theory of urban formation, for instance, the Zipf's law or the power law in general and how the theory of urban formation can be used to

validate the quality and reliability of information derived from smart card data when they are employed to reveal spatial patterns of economic agglomeration or human settlements at the intrametropolitan level, that is, where people prefer to work or reside in a metropolis. In this study, we argue that the theory of urban formation can be used to validate and calibrate the quality and reliability of information derived from smart card data. We propose a general framework regarding how the theory of urban formation can be employed to validate information derived from smart card data and how the validated information can supplement other data to reveal spatial patterns of economic agglomeration or human settlements. Through a case study of Beijing, we demonstrate the usefulness of the framework. Specifically, we elucidate how the framework can guide us to (a) derive and calibrate bus commuters' residence and workplace based on smart card data and other data from traditional sources; (b) use the derived information to verify Zipf's law; (c) combine processed smart card data and other data to reveal spatial patterns of subcenters of employment and residence.

The reminder of the article is organized as follows. The next section (Section 2) is a review of relevant literature. Section 3 describes our proposed framework. Section 4 is our case study, which is used to demonstrate the usefulness of the framework. Section 5 concludes.

2. RELATED LITERATURE

2.1 Smart card data and human movement studies

Smart cards are not new technologies in the transit field. Transit professionals and administrators have used the data produced by smart cards to do jobs at three levels: (a) strategic (long-term planning); (b) tactical (services adjustments and network development); (c) operational (ridership statistics and performance indicators) (Pelletier, Trepanier, & Morency, 2011). Transit researchers have employed smart card data as new input to do more than the above, showing that smart card data have great potential. Bagchi and White (2005), for instance, demonstrate that smart card data can help estimate turnover rates, trip rates per card on issues and linked trips. Morency, Trepanier, and Agard (2007) successfully measure spatiotemporal variability of transit trips in Gatineau, Canada based on smart card data in that city. In Seoul, Park, Kim, and Lim (2008) describe the characteristics of public transit users, such as the number of transfers, boarding time, hourly trip distribution of the number of trips for different transit modes, and travel time distribution for all transit modes and user types by using both local smart card and survey data. They argue that smart card data have the potential to supplement and even replace survey data in those regards. Similar to Morency, Trepanier, and Agard (2007), Liu, L. et al. (2009) use the smart card of Shenzhen to characterize spatial and temporal mobility patterns at the city and individual levels. They argue that their methodologies are replicable and can be useful for transportation planning and management. Taking advantage of the individual-level subway movement data provided by "Oyster" card in London, Roth et al. (2011) show the structure and organization of that city in terms of intraurban movement, hierarchy and activity centers.

Using the smart card and household travel survey data from Singapore, Chakirov and Erath (2012) identify the number of work activities and their

locations in that city-state. They conclude that smart card data from public transport offer significant potential for studies of travel behavior and activity identification. Their work, however, shows that despite the fact that processed smart card data from the local public transit system can reasonably detect work places but are subject to biases. In their case study of Singapore, they admit that the number of work places based on smart card data from the local public transit system can be underestimated. In other words, smart card data are often not full-population data but data of a bigger sample than the traditional survey data. There are cases that we need to take this into account and find ways to correct possible biases in research results based on smart card data. This should not be a surprise to researchers, as smart cards' main function is collecting the fare in the transit field (Pelletier, Trepanier, & Morency, 2011) and thus smart card data could have their limitations, for instance, they do not collect information of interest to researchers such as trip length (Bagchi & White, 2005), trip destination (e.g., (Li et al., 2011)) and socio-demographics of trip makers (Long, Zhang, & Cui, 2012). Methodologies thus have to be developed and supplementary data have to be used for researchers to obtain relevant information based on smart card data. Li et al. (2011) and Munizaga and Palma (2012) are two cases in point, which show how smart card data and other data could be used together to derive origin-destination matrices of transit trips, which are necessary input for any serious transportation system analysis. More recently, authors have used smart card data to help complete extra studies of transit trips and activity centers. Zhong et al. (2014), for instance, have used smart card data of Singapore for multiple years to profile the polycentrism in that city and how it evolved over time. Tao et al. (2014) utilize the smart card data from the bus rapid transit (BRT) and regular buses in Brisbane, Australia to geo-visualize the spatiotemporal patterns of BRT and regular bus trips. They argue that similar work can enhance the evidence-based BRT planning. Kim et al. (2014) propose a new approach to using smart card data as input to identify zones and movements between zones simultaneously.

More recently, Wang, M. et al. (2016) apply smart card data to identify frequent visiting locations of college students in Beijing. Alsger et al. (2016) validate different origin-destination estimation algorithms. Briand et al. (2017) categorize public transit riders based on the temporal features of the smart card usage. Zhong et al. (2016) compare mobility patterns of smart card users in London, Singapore, and Beijing. Further, Ma et al. (2017) develop a data mining method to understand spatiotemporal commuting patterns of smart card users.

2.2 The theory of urban formation and smart card data

Researchers have always attempted to explain the universal driving forces such as economic agglomeration, economies of location or urbanization and to identify laws such as the gravity law, rank-size rule or Zipf's law that govern the formation, evolution and interaction of cities, including intra- and inter-metropolitan movements of people and cargo (e.g., Anas, Arnott, and Small (1998), Barthélemy (2011), Simini et al. (2012), Zipf (1946)). If we regard related knowledge and insights from the above explorations as "the theory of urban formation", then there have been numerous studies of the theory of urban formation. Existing studies of the theory of urban formation, however, reply heavily on data from traditional sources such as censuses and ad-hoc surveys. Giuliano and Small (1991), for instance, use the 1980 Census journey-to-work data to study employment subcenters in the Los

Angeles region. Anas, Arnott, and Small (1998) employ census data of multiple years and of different countries in their studies of urban spatial structure. Bento (2003) examine the impact of urban spatial structure on travel demand in the US based on the 1990 National Personal Transportation Survey data. It is only recently that a few researchers have started exploring how smart card data from transit can facilitate studies of the theory of urban formation. Roth et al. (2011) and Zhong et al. (2014) are two examples. Roth et al. (2011) are interested in characterizing intraurban movement, hierarchy and activity centers based on smart card data from London's Metro. Zhong et al. (2014) apply recent methods in network science and their generalization to spatial analysis to identify city hubs, centers, and borders in Singapore with the 2010, 2011 and 2012 smart card data of that city's transit system.

Few existing studies, however, have applied the theory of urban formation to verify reliability of smart card data or information derived from them. Roth et al. (2011), Eubank et al. (2004) and Gutiérrez and García-Palomares (2007), for instance, have all found that the movement patterns in large cities exhibit a heterogeneous organization of flows. But according to our knowledge, nobody has used this finding to verify reliability of smart card data, regardless such data cover a large or small sample.

In this article, we argue that on the one hand, smart card data can facilitate more studies of the theory of urban formation; on the other hand, the known theory of urban formation, for instance, the above finding about heterogeneous organization of flows in large cities, can be employed to verify representativeness and reliability of smart card data or information derived from them. Later in this article, we will use a case study to show we use smart card data from the bus system in Beijing for us to identify employment subcenters in the city and how we verify those identified subcenters are representative and reliable based on Zipf's law.

3. PROPOSED FRAMEWORK

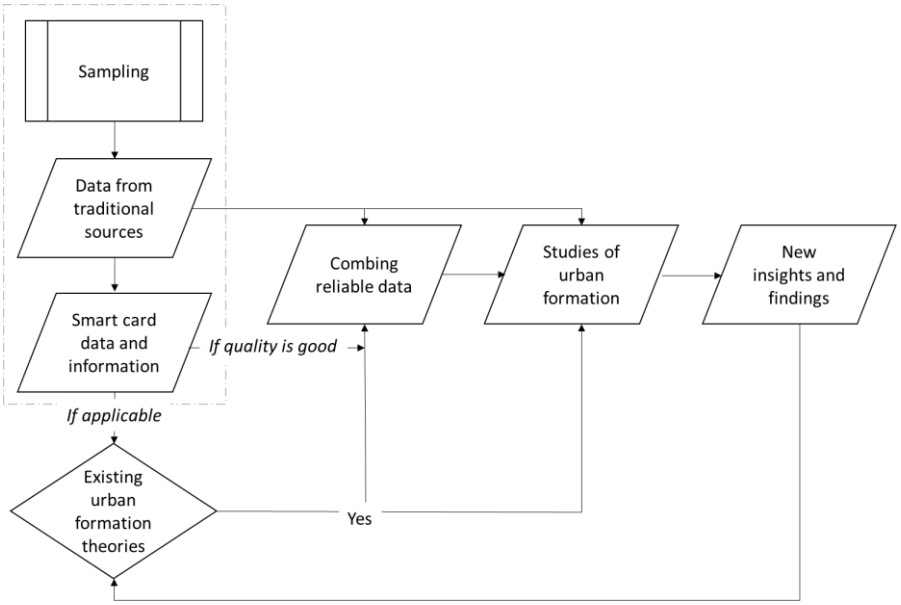

Figure 1. Proposed framework for better linking the theory of urban formation and smart card data

Considering the above literature review, we propose the following general framework regarding how we can have more meaningful linkages between the theory of urban formation and smart card data so that we could do a better job when we use smart card data to facilitate studies of the theory of urban formation and employ the theory of urban formation to verify the representativeness and reliability of smart card data and information derived from them.

In this framework, we argue that data from traditional sources (e.g., censuses, interviews and surveys), smart cards and the combination of traditional sources and smart cards can serve as input for studies of urban formation. There have been a notable number of publications on how we ensure the representativeness and reliability of data from traditional sources (e.g.,Box-Steffensmeier, Brady, and Collier (2008); Statistics Canada (1975); Groves (2009)). However, unlike data from traditional sources, there have been few documented mature and systematic procedures and methodologies to ensure their representativeness and reliability of data and derived information from smart cards. We thus propose that we could use both existing theories of urban formation and data from traditional sources to help us verify and calibrate representativeness and reliability of data and derived information from smart cards, if applicable, before they are fed into our studies of urban formation. We also believe that the introduction of smart card data into studies of urban formation would produce new theories (or hypotheses) of urban formation, which would enable us to more effectively verify and calibrate representativeness and reliability of data and derived information from smart cards.

4. CASE STUDY

To show the usefulness of the above framework, this section presents a case study, which shows how we use smart card data from Beijing to study bus commuters' employment and residential subcenters.

4.1 The Site

Beijing Metropolitan Area (BMA) is our site for case study. It covers an area of 16,410 km^2 and has a population of more than 22 million as of 2015. The BMA lies in northern China, to the east of the Shanxi altiplano and south of the Inner Mongolian altiplano. The southeastern part of the BMA is a flatland, extending east for 150 km to the coast of the Bohai Sea. BMA is the anchor city of the Beijing-Tianjin-Hebei polycentric city-region, which if one of the three most renown city-regions in China (Liu, X., Derudder, & Wang, 2017). Gaining momentum from China’s recent economic success, Beijing, as the capital city, is becoming one of the world’s most populous and fastest growing metropolises. The master city plan of Beijing has envisioned a polycentric urban form with one central city and ten subcenters. Detailed information about BMA can be found in Yang et al. (2013).

Beijing’s public transit system consists of buses and subways. The combined share of subway and bus trips in BMA was 38.9%, making Beijing the largest public transit system in terms of daily ridership in China (Beijing Transportation Research Center [BTRC], 2011). Bus trips still accounts for 29% of all trips and thus studies of bus travelers or commuters are still quite relevant in the context of BMA (BTRC, 2011).

4.2 Data

For the case study, we were granted access to a full week's historical data from the administrator of the smart card data of the Beijing transit system. The data contain 77,976,010 bus trips of 8,549,072 anonymized cardholders between April 7 and April 13, 2008. Data on subway trips were excluded by the data administrator due to security concerns. Given the fact that 95 per cent of bus users in Beijing are smart card holders, the one week sample is representative of all bus users in the city (Long, Zhang, & Cui, 2012). Thus, if we simply utilize the above data to study the general behaviors of bus users' in Beijing between bus stops, that is, similar to what Liu, L. et al. (2009) and Roth et al. (2011) do, there should not be any problems. However, if we manipulate the data to derive locations of residences and employment of bus commuters and then identify subcenters of residences and employment for bus commuters, we encounter the issue of representativeness and reliability of the derived information. Technical details regarding how we derive locations of residences and employment of commuters from the smart card data are elucidated in Long, Zhang, and Cui (2012). By and large, what Long, Zhang, and Cui (2012) does is (a) using data from traditional sources to establish rules for smart card data queries; (b) singling out the most probable locations of residences and employment from smart card data based on those rules. Long, Zhang, and Cui (2012) embodies the procedures in the dash-line rectangle in Figure 1. It is not unique, for instance, Chakirov and Erath (2012) has processed and queried the smart card data of Singapore in a similar fashion. In our case study here, we thus no longer detail how to derive probable locations of commuters' residences and employment from smart card data; instead, we focus on how we address representativeness and reliability of the derived information based on smart card data.

4.3 Representativeness and reliability of derived locations

Roth et al. (2011), Eubank et al. (2004), Gutiérrez and García-Palomares (2007) among others, find that trips between any two activity centers (e.g., a subway station) exhibit heterogeneous organization. In the log-log plot of the histogram format, the number of trips between any two activity centers follows the power law (Equation 1). Therefore, if we believe that bus commuting trips in Beijing are not exceptions to the above, the derived number of bus commuting trips, that is, flows between corresponding residence and employment based on smart card data should also follow the power law. By analyzing the number of trips (OD flows) for bus commuters and the corresponding histogram, we find home and employment centers for bus commuters in Beijing followed the power law (Figure 2).

$$P = a*N^b \qquad \text{(Equation 1)}$$

where
P denotes the frequency in the histogram distribution;
N is the number of trips between two traffic analysis zones (TAZs);
a and b are coefficients determined by the goodness of fit test.

In Figure 2, we find with a = 0.139 and b = -0.473, the goodness of fit test shows that R-square = 0.926, RMSE = 0.012. Figure 3 visualizes the

216,844 commuting trips between corresponding residence and employment locations for bus commuters. Again, the heterogeneous organization of the trips can be observed, which is in line with the pattern identified by Roth et al. (2011)for London's subway trips. Based on the above, we can at this point be more confident that locations of residence and employment for bus commuters derived from the smart card data in the case of Beijing are likely to be representative and reliable.

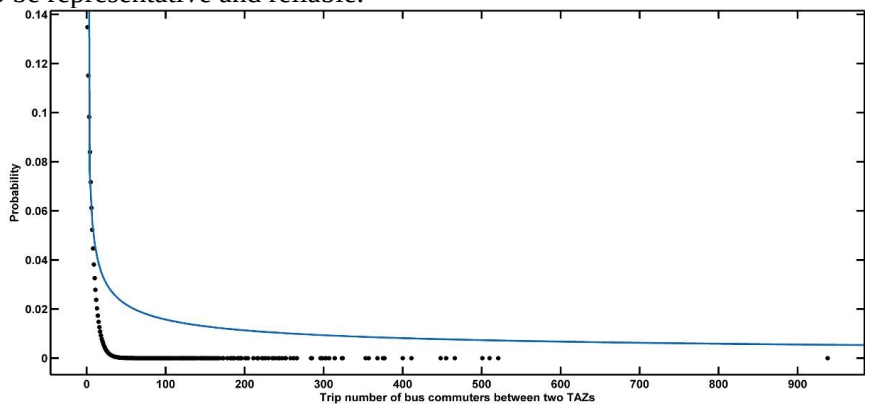

Figure 2. OD flow distribution. Plots of the histogram of the number of trips between two traffic analysis zones (TAZs). The black dots denote actual trip number; while the blue curve is a power law fit.

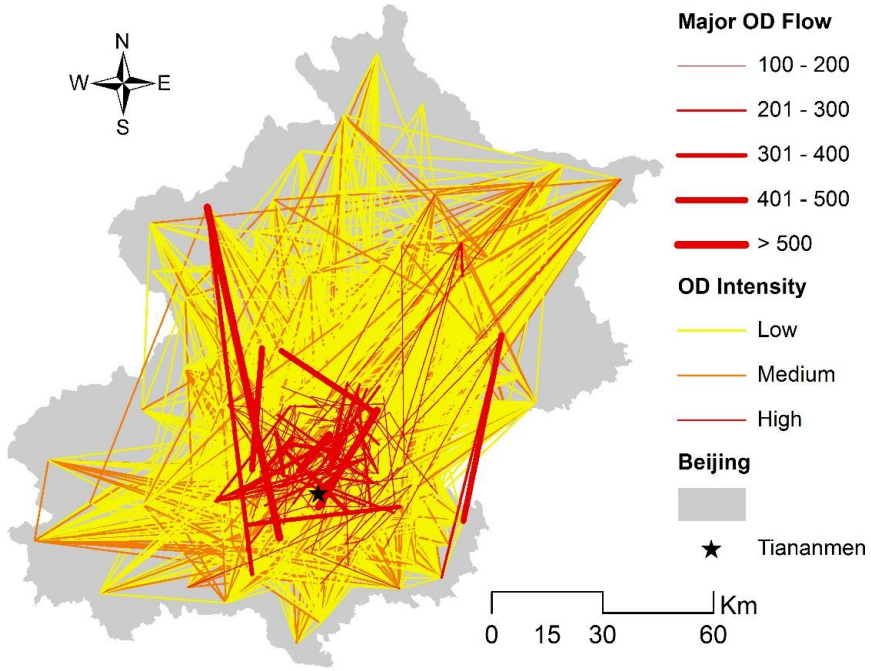

Figure 3. Visualization of OD flows for bus commuters between residence and employment locations. Major OD flows are categorized based on normalized trips.

4.4 Representativeness and reliability of derived subcenters

After verifying representativeness and reliability of the derived locations of bus commuters' residences and employment, we utilize spatial autocorrelation statistics to identify subcenters of bus commuters' residences and employment. We cannot replicate the approaches in existing studies such as Giuliano and Small (1991) or Anderson and Bogart (2001) to identifying those centers because those approaches deal with all workers.

Thus, their proposed thresholds for the total number of employment and density of employment would not be applicable to our case study.

Spatial autocorrelation analyzes the degree of dependency among observations in a geographic space. Positive spatial autocorrelation indicates the clustering of similar values across geographic space, while negative spatial autocorrelation indicates dissimilar values occur near one another. In other words, spatial autocorrelation can help us where there are concentrations of residences or employment of bus commuters in space.

Spatial autocorrelation statistics include Moran's I (Moran, 1950), Geary's C (Geary, 1954), Getis's G (Getis & Ord, 1992) and so forth. Among these statistics, Moran's I has the longest history and has been the most widely used (Lloyd, 2010; Wang, S. & Armstrong, 2009; Huang & Dennis Wei, 2014; Luo, 2014). For n observations on a variable x at location (i, j), Moran's I is calculated as:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad \text{(Equation 2)}$$

where

$\bar{x}$ is the mean of the x variable;

w_{ij} are the elements of the spatial weight matrix;

S_0 is the sum of the elements of the spatial weight matrix: $S_0 = \sum_i \sum_j w_{ij}$.

Spatial weights matrix reflects the intensity of the geographic relationship between observations in a neighborhood, such as the distances between neighbors. Moran' I allows us to testify whether there exist subcenters of bus commuters' residences and employment (I>0). However, the fact that the spatial heterogeneity of OD flows for bus commuters between residence and employment locations (Figure 3) suggests that the estimated degree of autocorrelation varies significantly across Beijing. Therefore, local version of Moran's I, as one of those well-established local spatial autocorrelation statistics (Anselin, 1995), is applied to provide estimates disaggregated to the TAZ level. In this case study, GeoDa software by Anselin, Syabri, and Kho (2006) is applied to test global and local spatial autocorrelation.

First, global Moran's I is used to determine if there exists any subcenter. Results of Moran' I show subcenters of both bus commuters' residences and employment exist ($p < 0.001$ for both cases). Second, local Moran's I is calculated for each TAZ to determine the residential and employment subcenters. We define a residential subcenter of bus commuters as a TAZ that is surrounded by TAZs with high bus commuters' residences or a TAZ that is surrounded by TAZs with significantly lower bus commuters' residences (i.e. the High-High and High-Low clusters derived from local Moran's I). Similarly, an employment subcenter of bus commuters is defined as a TAZ that is surrounded by TAZs with high bus commuters' employment or a TAZ that is surrounded by TAZs with significantly lower bus commuters' employment (i.e. the High-High and High-Low clusters derived from local Moran's I). Local Moran's I ($p = 0.05$) reveals that there are 35 subcenters of bus commuters' residences and 40 those of bus commuters' employment. Among these subcenters, there are 8 TAZs serving as both residential and employment subcenters (Figure 4).

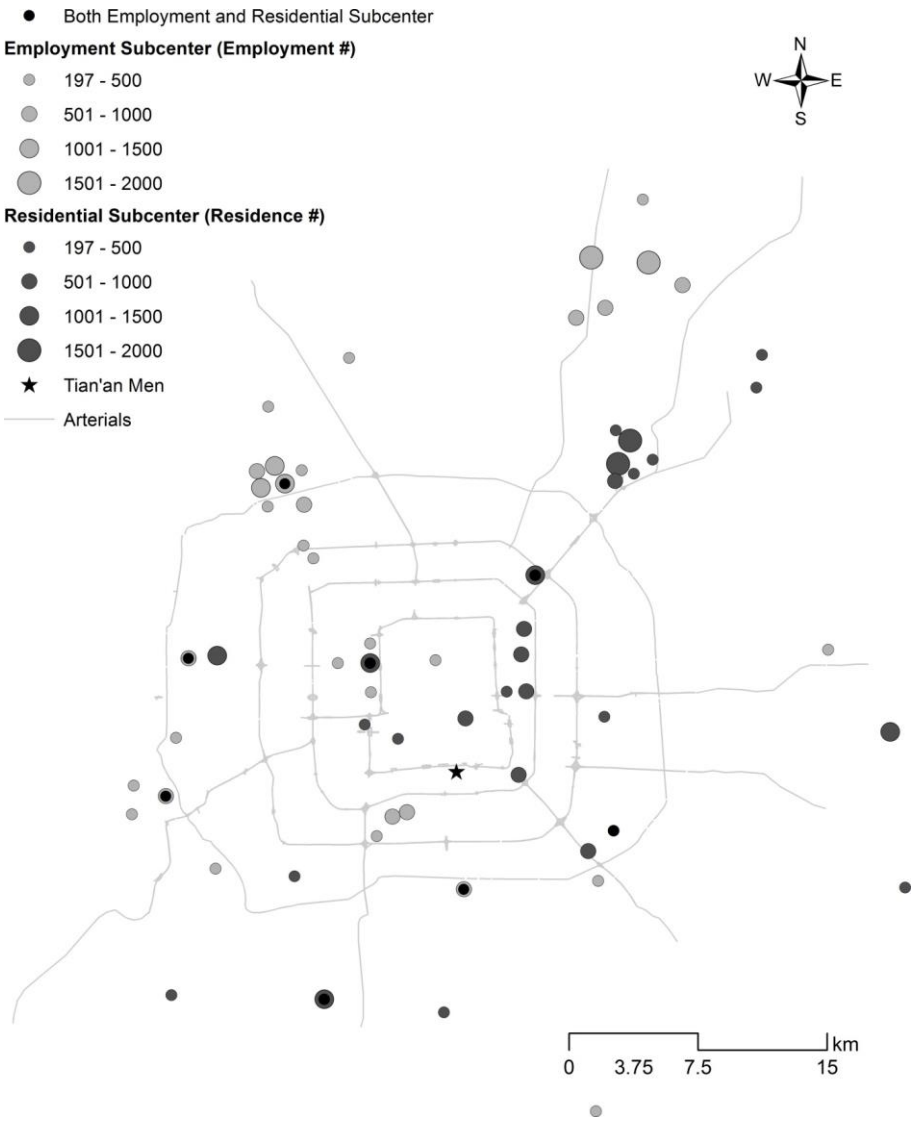

Figure 4. Subcenters of Bus Commuters in Beijing

Anas, Arnott, and Small (1998) contend that employment centers in a city are analogous to the system of cities in a larger regional or national economy and the former should therefore comply with Zipf's law as well. Based on this, one simple and further check we can do with the derived locations of residences and employment is to test whether the derived subcenters follow Zipf's law. In general, power-law distributions including Zipf's law are plotted on doubly logarithmic axes via cumulative distribution (Equation 3 and 4).

$$P(x)=P_r(X>x) \quad \text{(Equation 3)}$$

$$P(x)=P_r(X>x) = C\int_x^{\infty} p(X)dX = \frac{\alpha-1}{x_{min}^{-\alpha+1}} \int_x^{\infty} x^{-\alpha} dX = \left(\frac{x}{x_{min}}\right)^{-\alpha+1} \quad \text{(Equation 4)}$$

In Equations 3 and 4, x is the number of employment by subcenter, α is a constant to be calibrated.

In this case study, Zipf's law tests of derived subcenters of bus commuters are conducted (Figure 5), combining combination of maximum-likelihood fitting methods with goodness-of-fit test based on the Kolmogorov-Smirnov statistic and likelihood ratio based on (Clauset, Shalizi, & Newman, 2009). Results show both employment and residential subcenters of bus commuters follow Zip's law.

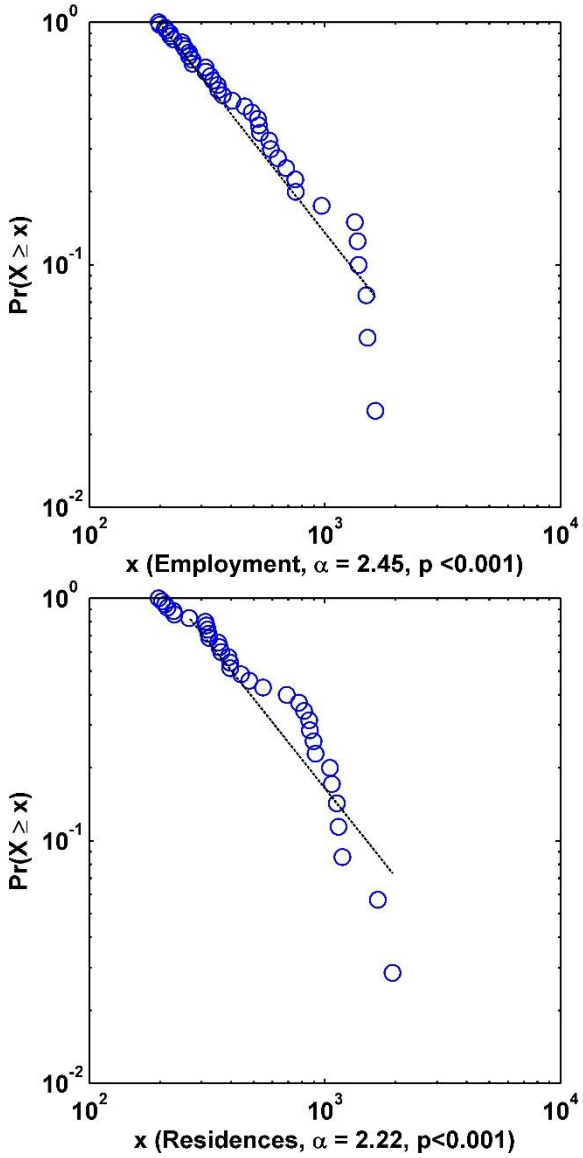

Figure 5. Power-distributions of subcenters

At this point, we have double-verified the representation and reliability of the derived locations of residences and employment based on the smart card data.

4.5 Characteristics of top subcenters

The smart card data alone do not tell us the land use and neighborhood characteristics of the identified subcenters. We thus have to rely on data from traditional sources such as land-use maps, satellite images, field trips and interviews if we want to find out those characteristics, which are of particular interest to geographers, planners and local policy analysts. They need to know those characteristics to better deal with issues such as economic agglomeration, traffic congestion and jobs-housing separation associated with subcenters. The characteristics, nevertheless, would also provide another opportunity for us to check the reasonableness of the identified subcenters. There have been a considerable number of existing studies of employment subcenters in metropolises (e.g., Agarwal, Giuliano, and Redfearn (2012); Cervero (1998); Giuliano and Small (1993)). So, when we check characteristics of subcenters we also focus on employment

subcenters so that we have more references to make comparisons. Table 1 summarizes characteristics of the employment subcenters for bus commuters we identified in Beijing and those by other researchers elsewhere.

Table 1. Characteristics of the Employment Subcenters: Beijing vs. Other Places

References	Geographical Focus	Land-use Characteristics	Neighborhood Characteristics
This study	40 employment subcenters for bus commuters Beijing	Mixed land use; All located in suburbs (outside the 5th ring road); University campuses; University employee apartment compounds; Gated communities; Suburb villages characterized by light industries and agriculture and related tourism	A large number of bus stops; Jobs≥1,300; Jobs/residences ≥1.19 (Max. 4.94); All have easy access to arterial roads; Except jobs at the universities, all jobs are recently emerging; ≤10% jobs are in the identified subcenters
Cervero (1998)	57 suburban employment centers (SECs) across American cities	Low density, single use and jobs-housing imbalance (He classified SECs into six groups: office parks, office centers and concentrations, large mixed-use developments, moderate-scale mixed use developments, subcities and large office corridors)	Free parking, low levels of transit services and lack of coordinated growth
Forstall and Greene (1997)	120 employment subcenters in Los Angeles	Most subcenters are recognized locally as separate activity centers and serve different functions	Jobs/workers≥1; At least one tract with jobs/workers≥1.25; Industrial profiles of the largest subcenters vary widely; Most subcenters had been in existence for more than 30 years
Giuliano and Small (1991)	35 employment subcenters in Los Angeles	Specialization in land use	Employment density ≥10 jobs/acre; Total employment≥10k; Subcenters are associated with agglomeration and industry mix)
Anderson and Bogart (2001)	Employment subcenters in Cleveland, Indianapolis, Portland, St. Louis	Specialization in land use	Subcenters follow a rank size distribution; ≤50% of metro employment is within the identified subcenters
Giuliano et al. (2007)	Employment subcenters in Los Angeles for three time points (1980, 1990 and 2000)	Subcenters remain stable over time; There are new subcenters emerging and growth at established subcenters at the same time; There is rapid growth of dispersed employment in outer suburbs.	The amount and density of employment have changed substantially. Employment and employment density has grown more rapidly in the suburban and exurban centers—but at an uneven rate among them.
Giuliano et al. (2012)	48 employment subcenters in Los Angeles	-	Subcenters have better road network and labor force accessibility; Subcenters follow a rank

References	Geographical Focus	Land-use Characteristics	Neighborhood Characteristics
			size distribution.
Agarwal, Giuliano, and Redfearn (2012)	48 employment subcenters in Los Angeles	Specialization in land use (e.g., the LA downtown is a specialized manufacturing/wholesale/ public administration center)	With at least 10k jobs; Jobs/population ≥1.62; Subcenters follow a rank size distribution; Subcenters have better road network accessibility

One thing should be noted is that all the existing studies cited in Table 1 do not separately consider employment subcenters for bus commuters, rather, they consider employment subcenters for all commuters. Thus, we cannot directly compare those subcenters with the subcenters for bus commuters in Beijing. But there are still some similarities between the two groups of subcenters. Most notably, like in Los Angeles, university campuses are subcenters in Beijng too. In addition, there tend to more local jobs than residences in the two groups of subcenters, indicating some degree of jobs-housing imbalance. There are also several notable differences between the two groups of subcenters. First, there tend to be more diverse land uses in employment subcenters in Beijing. Second, there may be more bus stops in or around employment subcenters in Beijing, even in two suburb villages characterized by light industries and agriculture (Table 2). Third, the subcenters other than university campuses in Beijing are recently emerging and have characteristics that are not found elsewhere, for instance, large-scale all-rounded university employee apartment compounds, high-end gated communities and villages characterized by light industries and agriculture and related tourism.

Table 2. Selected Characteristics of Top 5 Employment Subcenters in Beijing

TAZ ID	All jobs*	All population*	Incoming bus employees	Bus commuters' residences	Bus stops**	Land Use Characteristics***
291	2,826	14,355	1,342	396	25	University campus, concentration of star schools, university employee apartment compounds, a large number of bus stops
292	4,014	19,912	1,499	728	57	
294	3,427	17,418	1,377	1,156	17	
787	2,221	5,035	1,640	747	66	Suburb villages characterized by light industries and agriculture
788	1,427	3140	1,518	307	1	

Note: * Derived figures based on the 2008 local economic census data.
** Based on Google maps and Baidu maps.
***Based on local land-use maps, satellite images, field trips and interviews.

5. CONCLUSIONS AND DISCUSSION

The literature reviewed and the case study conducted in this study show that big data such as smart card data from transit operators have great

potential for us to better understand human settlement and movement patterns in metropolis. But as argued by Bagchi and White (2005) and Li et al. (2011), big data are often not designed to facilitate our studies of human settlement and movement patterns. As shown in the case study, big data have to be processed so as to derive useful information of relevance to those studies. But one challenge facing us is validation of the derive information based on big data. Unlike the data from traditional sources, there have not been mature and established ways of doing the validation. This study therefore proposes a framework regarding how we can validate derived information based on big data. It shows via a concrete case study of Beijing that the theory of urban formation can be used to validate the derived information from the smart card data. Combing the validated derived information from the smart card data with data from traditional sources, it can identify and profile land use and neighborhood characteristics of the employment subcenters in Beijing. This demonstrates that big data should be integrated to traditional data to best inform local researchers and decision-makers. The study also has the following generalizable implications for other researchers or users of big data:

First, asking the right and appropriate research questions is an important premise of putting big data to better and more usage. Pelletier, Trepanier, and Morency (2011), for instance, show that when transit professional and administrators look at or use big data, they legitimately focus on issues related to planning and operations of transit. But as demonstrated in other studies such as Zhong et al. (2014) and Roth et al. (2011) big data from transit companies can be used to answer questions beyond transit planning and operations. Given the above, we argue that asking the right and appropriate research questions is an important premise of putting big data to better and more usage. In addition, even we cannot answer those questions right away, those questions could inspire us improve our work of big data, for instance, why shouldn't we redesign our data collection mechanism in advance to capture more relevant information, as recommended by Pelletier, Trepanier, and Morency (2011).

Second, deriving and validating information from big data demands new protocols, methods and procedures. In our case study of Beijing, yes, 95 percent of bus commuters use a smart card when making a bus trip. But this does not mean that we can automatically and conveniently get bus commuters' locations of residence and workplace, which are of interest to geographers, planners and policy analysts. In the case study, we devise and implement an ad-hoc way to derive and validate the locations. But we should not always devise and implement an ad-hoc way to take advantage of big data each time. For certain big data such as the smart card data in Beijing, we should be able develop some routinized protocols, methods and procedures to increase our efficacy.

Third, linking big data and data from traditional sources (or simply "traditional data") is important to generate more relevant knowledge and insights. In our case study of Beijing, the smart card data can at most tell us where those bus commuters reside and work at the TAZ level. Knowing such information is good but to better inform local decision-makers and researchers, extra information such as land use and neighborhood characteristics is needed. Based on our experience of the case study, it can be more efficient for us to get the extra information based on traditional data. Finally, traditional data provide another opportunity to validate the derived information from big data.

Despite the above features and merits, this study can still be improved and enhanced in several aspects in the future. First, it can validate the locations of bus commuters using extra traditional data, for instance, the bus commuting flow matrices by the local transportation planning agency. Given the planning data hoarding issue in China, this would mean extra work for us to get access to those data (c.f., (Zhou & Wang, 2014)). But it is definitely worthwhile. Second, it can standardize and streamline procedures and methodologies for the work of deriving and validating residential and workplace locations of bus commuters from smart card data. Third, it can deepen the current studies of bus commuters by collecting extra socio-demographic information of bus commuters, for instance, conducting on-board survey of bus commuters and giving incentives to bus commuters who are willing to complete on-line surveys about their residential and mode choices. If the smart card data can help us identify the settlement and movement patterns of bus commuters, as described above, extra socio-demographic information of bus commuters would enable us to get insights into why there are those patterns and whether and how the patterns can be changed.

REFERENCES

Agarwal, A., Giuliano, G., & Redfearn, C. (2012). "Strangers in Our Midst: The Usefulness of Exploring Polycentricity". *The Annals of Regional Science*, *48*(2), 433-450. doi: http://dx.doi.org/10.1007/s00168-012-0497-1.

Alsger, A., Assemi, B., Mesbah, M., & Ferreira, L. (2016). "Validating and Improving Public Transport Origin–Destination Estimation Algorithm Using Smart Card Fare Data". *Transportation Research Part C: Emerging Technologies*, *68*, 490-506.

Anas, A., Arnott, R., & Small, K. A. (1998). "Urban Spatial Structure". *Journal of Economic Literature*, *36*(3), 1426-1464.

Anderson, N. B., & Bogart, W. T. (2001). "The Structure of Sprawl: Identifying and Characterizing Employment Centers in Polycentric Metropolitan Areas". *American Journal of Economics and Sociology*, *60*(1), 147-169.

Anselin, L. (1995). "Local Indicators of Spatial Association—Lisa". *Geographical analysis*, *27*(2), 93-115.

Anselin, L., Syabri, I., & Kho, Y. (2006). "Geoda: An Introduction to Spatial Data Analysis". *Geographical analysis*, *38*(1), 5-22.

Bagchi, M., & White, P. R. (2005). "The Potential of Public Transport Smart Card Data". *Transport Policy*, *12*(5), 464-474.

Barthélemy, M. (2011). "Spatial Networks". *Physics Reports-Review Section of Physics Letters*, *499*(1-3), 1-101.

Bento, A. M. (2003). *The Impact of Urban Spatial Structure on Travel Demand in the United States* (Vol. 3007). Washington, D.C.: World Bank, Development Research Group, Infrastructure and Environment.

Box-Steffensmeier, J. M., Brady, H. E., & Collier, D. (2008). *The Oxford Handbook of Political Methodology*. Oxford ; New York: Oxford University Press,. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199286546.001.0001/oxfordhb-9780199286546.

Briand, A.-S., Côme, E., Trépanier, M., & Oukhellou, L. (2017). "Analyzing Year-to-Year Changes in Public Transport Passenger Behaviour Using Smart Card Data". *Transportation Research Part C: Emerging Technologies*, *79*, 274-289.

Cervero, R. (1998). *The Transit Metropolis : A Global Inquiry*. Washington, DC: Island Press.

Chakirov, A., & Erath, A. (2012). "Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport". Zürich: Eidgenössische Technische Hochschule Zürich, IVT, Institute for Transport Planning and Systems.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). "Power-Law Distributions in Empirical Data". *SIAM review*, *51*(4), 661-703.

de Dios Ortúzar, J., & Willumsen, L. G. (1990). *Modelling Transport*. New Jersey: Wiley.

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). "Modelling Disease Outbreaks in Realistic Urban Social Networks". *Nature*, *429*(6988), 180-184. doi: http://dx.doi.org/10.1038/nature02541.

Forstall, R. L., & Greene, R. P. (1997). "Defining Job Concentrations: The Los Angeles Case". *Urban Geography*, *18*(8), 705-739.

Geary, R. C. (1954). "The Contiguity Ratio and Statistical Mapping". *The incorporated statistician*, 115-146.

Getis, A., & Ord, J. K. (1992). "The Analysis of Spatial Association by Use of Distance Statistics". *Geographical analysis*, *24*(3), 189-206.

Giuliano, G., Redfearn, C., Agarwal, A., & He, S. (2012). "Network Accessibility and Employment Centres". *Urban Studies*, *49*(1), 77-95.

Giuliano, G., Redfearn, C., Agarwal, A., Li, C., & Zhuang, D. (2007). "Employment Concentrations in Los Angeles, 1980–2000". *Environment and planning A*, *39*(12), 2935-2957.

Giuliano, G., & Small, K. A. (1991). "Subcenters in the Los Angeles Region". *Regional Science and Urban Economics*, *21*(2), 163-182.

Giuliano, G., & Small, K. A. (1993). "Is the Journey to Work Explained by Urban Structure?". *Urban Studies*, *30*(9), 1485-1500.

Groves, R. M. (2009). *Survey Methodology* (2nd ed.). Hoboken, N.J.: Wiley.

Gutiérrez, J., & García-Palomares, J. C. (2007). "New Spatial Patterns of Mobility within the Metropolitan Area of Madrid: Towards More Complex and Dispersed Flow Networks". *Journal of Transport Geography*, *15*(1), 18-30. doi: http://dx.doi.org/10.1016/j.jtrangeo.2006.01.002.

Hanson, S., & Giuliano, G. (Eds.). (2004). *The Geography of Urban Transportation* (3rd ed.). New York: Guilford Press.

Huang, H., & Dennis Wei, Y. (2014). "Intra-Metropolitan Location of Foreign Direct Investment in Wuhan, China: Institution, Urban Structure, and Accessibility". *Applied geography*, *47*, 78-88.

Kim, K., Oh, K., Lee, Y. K., Kim, S., & Jung, J.-Y. (2014). "An Analysis on Movement Patterns between Zones Using Smart Card Data in Subway Networks". *International Journal of Geographical Information Science*, *28*(9), 1781-1801.

Li, D., Lin, Y., Zhao, X., Song, H., & Zou, N. (2011). "Estimating a Transit Passenger Trip Origin-Destination Matrix Using Automatic Fare Collection System". Proceedings of the 16th International Conference on Database Systems for Advanced Applications, Hong Kong, pp. 502-513.

Liu, L., Hou, A., Biderman, A., Ratti, C., & Chen, J. (2009, 4-7 Oct. 2009). "Understanding Individual and Collective Mobility Patterns from Smart Card Records: A Case Study in Shenzhen". Proceedings of 12th Intelligent Transportation Systems Conference (ITSC '09), Hague, pp. 1-6.

Liu, X., Derudder, B., & Wang, M. (2017). "Polycentric Urban Development in China: A Multi-Scale Analysis". *Environment and Planning B: Urban Analytics and City Science*. doi: 10.1177/2399808317690155.

Lloyd, C. D. (2010). "Exploring Population Spatial Concentrations in Northern Ireland by Community Background and Other Characteristics: An Application of Geographically Weighted Spatial Statistics". *International Journal of Geographical Information Science*, *24*(8), 1193-1221.

Long, Y., Zhang, Y., & Cui, C. (2012). "Identifying Commuting Pattern of Beijing Using Bus Smart Card Data". *Journal of Geographical Sciences*, *67*(10), 1339-1352.

Luo, J. (2014). "Integrating the Huff Model and Floating Catchment Area Methods to Analyze Spatial Access to Healthcare Services". *Transactions in GIS*, *18*(3), 436-448. doi: http://dx.doi.org/10.1111/tgis.12096.

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.

Moran, P. A. (1950). "Notes on Continuous Stochastic Phenomena". *Biometrika*, 17-23.

Morency, C., Trepanier, M., & Agard, B. (2007). "Measuring Transit Use Variability with Smart-Card Data". *Transport Policy*, *14*(3), 193-203.

Munizaga, M. A., & Palma, C. (2012). "Estimation of a Disaggregate Multimodal Public Transport Origin–Destination Matrix from Passive Smartcard Data from Santiago, Chile". *Transportation Research Part C: Emerging Technologies*, *24*, 9-18.

Park, J., Kim, D.-J., & Lim, Y. (2008). "Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea". *Transportation Research Record: Journal of the Transportation Research Board*, (2063), 3-9.

Pelletier, M. P., Trepanier, M., & Morency, C. (2011). "Smart Card Data Use in Public Transit: A Literature Review". *Transportation Research Part C-Emerging Technologies*, *19*(4), 557-568.

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.

Salas-Olmedo, M. H., & Nogués, S. (2012). "Analysis of Commuting Needs Using Graph Theory and Census Data: A Comparison between Two Medium-Sized Cities in the Uk". *Applied geography*, *35*(1), 132-141.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). "A Universal Model for Mobility and Migration Patterns". *Nature*, *484*(7392), 96-100.

Statistics Canada. (1975). "Survey Methodology" (0714-0045). Ottawa: Household Surveys Development Division, Statistical Services Field, Social Survey Methods Division. Retrieved from https://sunsite2.berkeley.edu/NRLF_article?rec=b15226692.

Tao, S., Corcoran, J., Mateo-Babiano, I., & Rohde, D. (2014). "Exploring Bus Rapid Transit Passenger Travel Behaviour Using Big Data". *Applied geography*, *53*, 90-104.

Wang, M., Zhou, J., Long, Y., & Chen, F. (2016). "Outside the Ivory Tower: Visualizing University Students' Top Transit-Trip Destinations and Popular Corridors". *Regional Studies, Regional Science*, *3*(1), 202-206.

Wang, S., & Armstrong, M. P. (2009). "A Theoretical Approach to the Use of Cyberinfrastructure in Geographical Analysis". *International Journal of Geographical Information Science*, *23*(2), 169-193.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). "Detecting the Dynamics of Urban Structure through Spatial Network Analysis". *International Journal of Geographical Information Science*, 1-22. doi: http://dx.doi.org/10.1080/13658816.2014.914521.

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.

Zhou, J., & Wang, Y. (2014). "Mobile-Source Ghg Modeling Institutions and Capaci-Ties in China: Findings Based on Structured Interviews and on-Line Surveys". *China City Planning Review*, *23*(2), 14-23.

Zipf, G. K. (1946). "The P 1 P 2/D Hypothesis: On the Intercity Movement of Persons". *American sociological review*, *11*(6), 677-686.