GeoJournal
https://doi.org/10.1007/s10708-018-9885-0

Uncovering patterns of ties among regions within metropolitan areas using data from mobile phones and online mass media

Hui Peng · Yunyan Du · Zhang Liu · Jiawei Yi · Yuhao Kang · Teng Fei

© Springer Science+Business Media B.V., part of Springer Nature 2018

Abstract Given the abundant quantities of big spatiotemporal geographic data that are available, interactions among spatial entities can now be extracted from various perspectives. This research investigates the spatial interactions within the metropolis of Beijing quantitatively. Two methods of quantifying the interactions are proposed. These interactions can be calculated from either individual trajectories extracted from mobile phone records or the co-occurrence of the toponyms of administrative units mentioned in online news items. By fitting these two types of data with a gravity model and comparing the results, we determine that the distance decay effect exists in both data sets, and this effect is more obvious in the interactions computed from the human trajectories. The spatial interactions and connections quantified from the two data sources display greater numbers of mutual patterns in the central urban areas, whereas more diversity is observed in the suburban areas. We conclude that the choice of assumptions as to which data can adequately represent spatial interactions significantly affects the results; therefore, rigorous examination of specific problems is needed to redefine the problems in a more specific way.

Keywords Spatial interactions · Mobile phone data · Mass media data · Co-occurrence · Beijing

Introduction

New approaches to the planning and management of urban regions, such as sustainable development and smart growth, require information on both the apparent properties of urbanized areas, such as their sizes, shapes, and magnitudes, and their implicit properties, such as the intensity of connectivity (ties) among the regions that make up a city.

Urban development promotes an intensification of traffic and subsequently fosters stronger interactions between districts in terms of flows of people, materials and information. Interactions among regions usually result in significant changes inside cities in terms of their spatial extents, structures and functions; thus, they have received considerable attention in the scientific community (Herold et al. 2003).

Traditionally, remote sensing provides spatially consistent data sets that cover large areas with both

H. Peng · Y. Du (✉) · Z. Liu · J. Yi
State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
e-mail: duyy@lreis.ac.cn

H. Peng · Z. Liu · J. Yi
University of Chinese Academy of Sciences, Beijing 100049, China

Y. Kang · T. Fei
School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, China

Published online: 14 May 2018

high spatial detail and high temporal frequency. Through the analysis of remotely sensed data, urban growth and land-use change can be monitored and modelled (Herold et al. 2003; Luo and Wei 2009; Vermeiren et al. 2012). However, remote sensing data do not represent the short-term human activities that contribute to urban processes (Jendryke et al. 2017). The wide use of mobile electronic devices equipped with location awareness technologies produces abundant spatiotemporal geographical big data, such as mobile phone records, taxi GPS trajectories (Kang et al. 2013; Liu et al. 2015; Veloso et al. 2011), bus integrated circuit (IC) card records (Long et al. 2012), and social network check-in data (Van Zanten et al. 2016). These new geospatial data make it possible to obtain information on cities from the perspective of individuals and to chart different flows at fine spatial and temporal resolutions accurately (Lu and Liu 2012). Data, such as mobile-based trajectory data and online mass media data, combined with innovative techniques even offer the potential to significantly improve the analysis, understanding, representation and modelling of urban dynamics.

From a bottom-up perspective, individual trajectories link places together through place-human interactions. Greater numbers of trajectories linking two regions indicate greater flows of people and stronger interactions among the places considered.

From a top-down perspective, mass media data can capture the social, economic, cultural, and political aspects of a society synergistically at an aggregate scale, rather than the individual scale. For example, it is possible to build a semantic linkage between areas of interest from the frequency of co-occurrence of place names in news archives. If the names of two places appear in the same web news document, the two geographical entities may be linked. The number of co-occurrences of pairs of place names or the probability of their co-occurrence can be used as a quantitative index, with higher probabilities indicating tighter connections. From an aggregate perspective, web-based toponym co-occurrence analysis of mass media data can reveal the macroscopic connections between regions and urban structures, in particular when the relevant socioeconomic data are not available.

This study represents a contribution and provides a tool for use by urban planners and designers, as well as scientists who use spatial information from various sources to observe urban processes. We examine both the movement trajectories of individuals derived from mobile phone data and the toponym co-occurrence data extracted from mass media articles to characterize the intensity of the interactions among administrative units in Beijing, China. In this study, the intensities of ties between pairs of administrative units are decomposed into the intensities of interactions and connections, which can be calculated separately using the mobile phone and online news data.

A brief review on quantifying spatial interaction

The study of spatial interactions, which investigates the movements of humans, materials and information, began long ago (Haggett et al. 1978). These interactions usually result in significant changes in cities in terms of their spatial extents, structures and even functions. Therefore have received substantial attention in the scientific community (Lathia et al. 2012; Smith et al. 2013). Researchers have used different methods and datasets to examine the flows of people, materials, and information. Abel and Sander (2014) visually quantified the international migration flows globally from 1990 to 2020 using migration data for 196 countries. Using data on 4,000,000 commuters, Nelson and Rae (2016) identified patterns of economic interconnection in the US and divided the states into labour markets. Shen (2004) used a simplified algebraic method to examine passenger transport data obtained from an airline passenger database to study inter-city spatial interactions across the US. Chen et al. (2013) studied the connections of city functions in the Pearl River delta using data describing inter-city passenger flows. Transportation data are widely used to examine material flows. Xu et al. (2015) studied the spatial interactions among global shipping networks using container shipping data. Djankov and Freund (2002) studied the interactions among the countries of the former Soviet Union based on the trade flows among them. The flow data used in the aforementioned studies are mainly obtained from census statistics or questionnaire surveys. These data have coarse spatial resolutions and lack real-time dynamical information. Furthermore, these data also have difficulty in accurately reflecting human activities, due to their lack of precision.

The strength of the interaction between two places is controlled by the distance decay effect; i.e., it declines as the distance between the places increases (Zipf 1949; Liu et al. 2014a). The distance decay effect can be described by the gravity model, in which the numerator can be any specified characteristic of the places, such as their masses, populations, or economic values (Matsumoto 2004; Fuellhart 2003; Xiao et al. 2013; Kang et al. 2013). These interactions can also be measured using other indexes from social science or even virtual networks, such as check-ins and following numbers (Liu et al. 2014a, b, c). However, gravity models only work well at certain spatial scales because the decay of the interaction intensity is sensitive to the distance, which is not always true (Simini et al. 2012; Stefanouli and Polyzos 2017; Masucci et al. 2013).

Significant progress has been achieved in studying spatial interactions over the past several decades. However, previous studies focus mainly on countries or urban agglomerations and treat individual cities as nodes; insufficient attention has been devoted to examining the spatial interactions within cities. In the big data era, large amounts of geographic information have been acquired that reveal human behaviour in urban spaces and provide new opportunities to perceive the flows of various elements among regions and the interactions within cities quantitatively. Birkin and Malleson (2012) deduced the behaviours and locations of users from 1 year of Twitter data that covered Leeds to construct an individual-level model of city structure and dynamics. Hollenstein and Purves (2010) sketched out the core areas of London and Chicago using 8 million Flickr images and their locations posted over the course of 1 month. Liu et al. (2012) used GPS data collected by taxis in Shenzhen, China, to quantify the flow of resident commuters and explored the relationship between intra-urban trip patterns and land use. Shi et al. (2015) used mobile phone data to study the relationships among individual users and to categorize these users based on their mobility patterns. Niu et al. (2014) used mobile phone data collected in Shanghai to investigate the dynamic density of users at different times during workdays and weekends; they identified the functions of urban public centres to better understand the spatial structure of central cities. Wang and Zhen (2016), Li et al. (2013), and Wang and Deng (2016) analysed the interactions among cities in China using website data from the SINA micro-blog service. As interdisciplinary urban computing has become more prevalent, the field of urban sensing has shifted its focus from a pure "city" perspective to the agglomeration of "people, environments, and cities" (Feng and Xiao 2014).

Another facet of the study of spatial connections lies in the field of natural language processing, by which the connection intensity between two places can be inferred from large amounts of text. In fact, the study of word co-occurrence data can be traced back to the 1970s or even earlier (Van Rijsbergen 1977). Today, it is much easier to obtain and calculate the probability of co-occurring place names in digital mass media data. Though many studies examine toponym co-occurrence in mass media outlets (Roark and Charniak 1998; Peat and Willett 1991; He 1999; Bhattacharya and Basu 1998; Zhong et al. 2017), only a few of these studies focus on the analysis of spatial interactions. Liu et al. (2014b) analysed the spatial structure of China based on the co-occurrence of the toponyms of provincial units in the mass media and identified four kinds of relationships among the provinces: part-whole relations, spatial proximity, interactions, and similarities. The latter two relationships mainly reflect the interactions between spatial entities and are commonly reported in the news, whereas the former two relationships typically receive less coverage.

Materials and methods

Mobile phone data

The mobile phone dataset used in this research was obtained from the China Mobile Communication Company (CMCC), which is currently the largest communication provider on the market. The dataset includes 16 million records of CMCC mobile phone users collected on Dec. 3, 2015. A record is generated when a user makes a phone call, sends a text message, or connects to the Internet. Another record is generated when a user moves and connects to a new base station. Each record contains information including the ID of the user, the time when the record was made, the ID of the base station, and the location where the record was generated. The dataset, which contains information on 16 million users, covers all age ranges (12–90); for comparison, the number of permanent

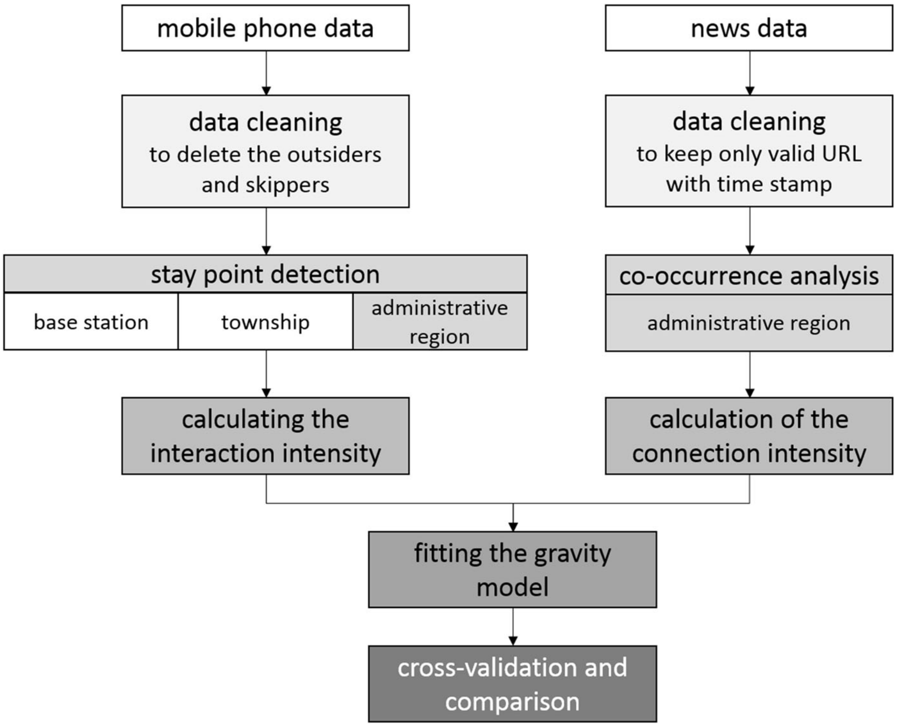

mobile phone data
news data
data cleaning
to delete the outsiders and skippers
data cleaning
to keep only valid URL with time stamp
stay point detection
base station
township
administrative region
co-occurrence analysis
administrative region
calculating the interaction intensity
calculation of the connection intensity
fitting the gravity model
cross-validation and comparison

Fig. 1 The framework of the methodology

residents in Beijing was approximately 21.7 million in 2016. The entire area of the city, from urban to rural areas, is covered by these base stations.

Online news data

The online news data used in this study were crawled from Baidu News (baidu.com), the largest online news aggregation website in China. We found 71,163 news items that contain the Chinese name of any of the 16 administrative units of Beijing with dates between January 1st, 2014 and February 22nd, 2017. None of the articles are repeated news re-distributed by different websites, as duplicate news items are filtered by the Baidu news aggregation service; therefore, all the news items obtained are unique. We save the URL of each piece of news that includes the Chinese name of at least one administrative unit within Beijing. All of the saved URLs are further searched to determine whether the Chinese names of the other 15 administrative units also appear in the same news item. We count the number of news items that include the Chinese names of at least two of the 16 administrative units and prepare a 16-by-16 co-occurrence matrix. Each row and column in the matrix represents one of the administrative units in Beijing, and the values in the matrix represent the co-occurrence probabilities of the names of any two administrative units that coexist in a news item.

Methodology

This study examines the mobile phone and news datasets in parallel, although they are obtained at different spatial scales (Fig. 1). The two datasets are first preprocessed to remove unreasonable records. Different methods are then applied to examine the two datasets to extract the ties among the regions at different spatial scales. Finally, both datasets are aggregated to extract the ties among the regions at the administrative unit scale. The ties among the regions are then fitted using the gravity model, and the modelling performance is evaluated by examining the goodness of fit (GOF).

Mobile phone data preprocessing

The mobile phone data is first preprocessed to detect and eliminate outsiders and skippers. In this study, “outsiders” are defined as the points that are located outside of the study area of Beijing. All of the records that were generated beyond our study area are deleted from the database and are not considered in this study. "Skippers" are defined as users who move impossibly fast; that is, they "jump" from one base station to another at an impossibly high speed. A record may also be generated when a user is within a crossing area at the border between two or more base stations. In such circumstances, mobile phone connections sometimes switch back and forth between these base stations, even if the user makes no significant movement at all. A speed threshold (200 km/h) is used to detect the pseudo-movements and delete all of the "skipper" locations.

Stay point detection from mobile phone data

Previous studies have shown that the interactions within a city can be characterized by examining the patterns of a large number of individual movements (Liu et al. 2012). However, mobile phone records are not generated repeatedly on a periodic basis. As noted above, a record is generated when a user makes a phone call, send a text message, or connects to the Internet. As a result, active users tend to trigger more records; thus, their trajectories are more accurately depicted. Therefore, trajectories with different numbers of records must be standardized before they are used to reflect the interactions among different places. In this study, we simplify the trajectories through extracting the important places users visit or stay, such as their homes, workplaces, and restaurants. In other words, the places along the trajectories of city residents where they spend most of their time are highlighted. In this study, we apply a stay point algorithm (Ye et al. 2009; Zheng and Zhou 2011) to extract the stay points along the movement trajectories of the users. A stay point represents a geographic region where a user remains for a pre-defined time period. In this study, we define a stay point as a specific site where a user spends more than 30 min within a range of 500 m. Once all of the stay points have been identified, the movement trajectories of the users can be simplified into a sequence of stay points, from which we can infer individual activities once the time and duration information in the records is further considered.

Figure 2a exemplifies the daily trajectory of one individual with six stay points. This person leaves home at 7:00 am, as indicated by the point 1 record. He/she makes two phone calls on the way to work at 7:30 and 8:30, which are recorded at points 2 and 3. Point 4 is recorded after the person arrives at his/her office and makes another call at 9:00. After work, the person decides to go shopping. He/she arrives at the shopping mall at 19:00 and generates point 5 when he/she makes a call. Finally, the person makes another call after he/she goes home at 21:00, thus generating point 6. Of these records, only points 1, 4, 5, and 6 are detected as stay points because the user stays at these places for a sufficiently long time, rather than simply passing by. Note that there are two kinds of stay points, as shown in Fig. 2b. Stay point A represents a place where a person stays for a long time, such as a shopping mall. In contrast, stay point B represents the geometric centre of several densely clustered records. However, stay point B does not lie exactly on the original trajectory.

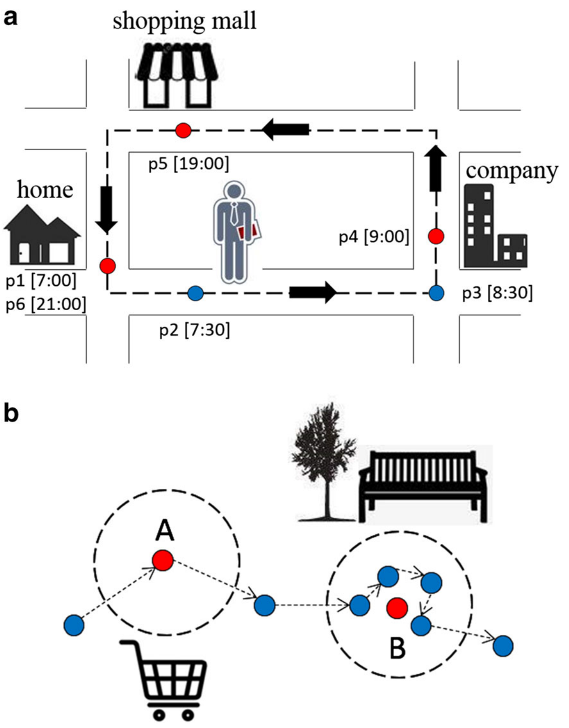
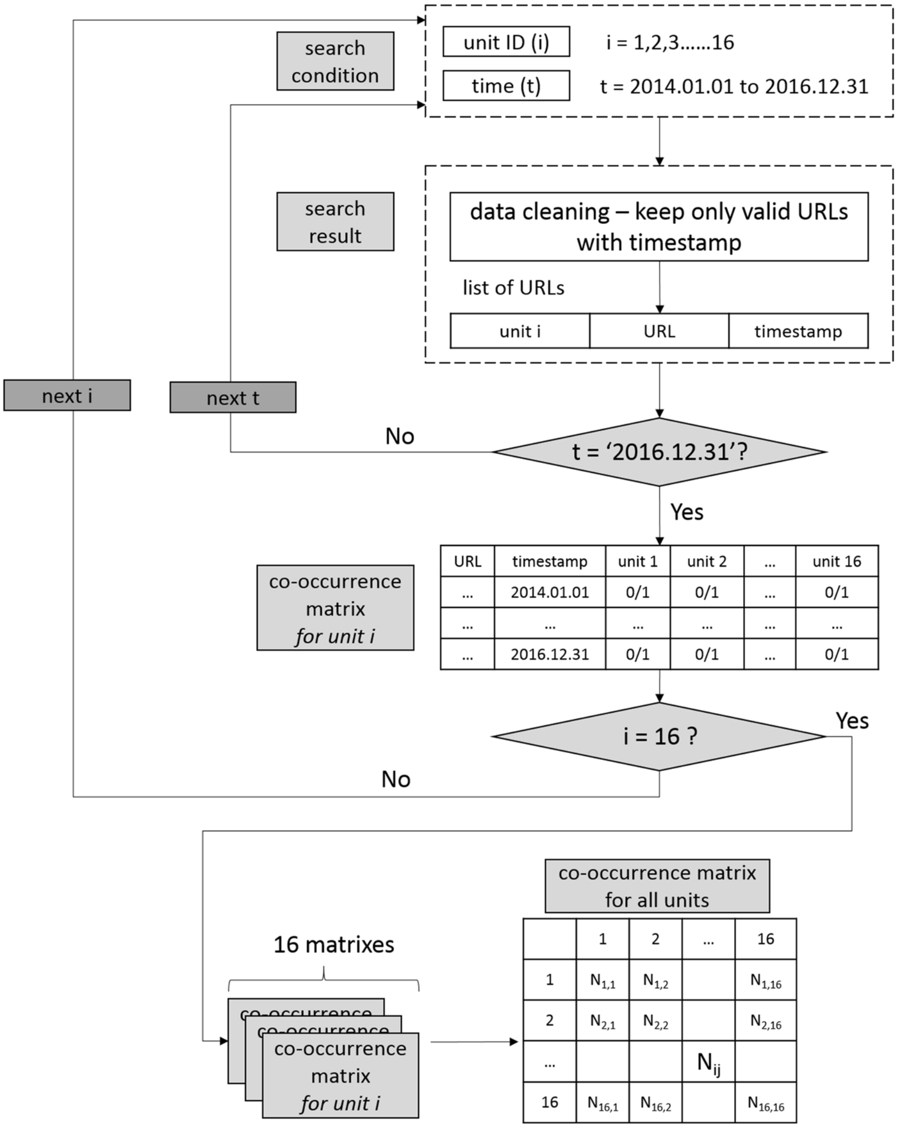


Fig. 2 Stay point detection. **a** An example of simplifying the daily trajectory of an individual into a stay point sequence; **b** two kinds of stay points: a stay point that represents a fixed point and a stay point that represents a set of points



unit i	URL	timestamp

URL	timestamp	unit 1	unit 2	…	unit 16
…	2014.01.01	0/1	0/1	…	0/1
…	…	…	…	…	…
…	2016.12.31	0/1	0/1	…	0/1

	1	2	…	16
1	$N_{1,1}$	$N_{1,2}$		$N_{1,16}$
2	$N_{2,1}$	$N_{2,2}$		$N_{2,16}$
…			N_{ij}	
16	$N_{16,1}$	$N_{16,2}$		$N_{16,16}$

Fig. 3 The flow diagram of extracting the co-occurrence statistics of toponyms in news items

Co-occurrence matrix of toponyms in the news

The news dataset is also preprocessed (Fig. 3) before analysis. Only news items with a date stamp are considered valid in this study, whereas all of the URLs without a time stamp are deleted. We then build a co-occurrence matrix by counting the number of valid news items that include the names of at least two of the administrative units of Beijing.

We then construct a graph structure to model the toponym co-occurrence network. As an undirected graph, the vertices in the network are expressed by the names of the 16 administrative units. The co-occurrence relationships among the 16 administrative units form the edges of the graph. Let G = (V, E) be the graph, which can be denoted by a co-occurrence matrix. In the corresponding matrix A, $\forall v_i, v_j \in V$; if $e_{ij} \in E$, then a_{ij}= 1; otherwise a_{ij}= 0. A is a symmetric matrix; i.e., $\forall v_i, v_j \in V, e_{ij} \in E, e_{ij} = e_{ji}$.

By aggregating the matrix A to the edges of graph G, the co-occurrence network can be transformed to a weighted undirected graph G = (V, E, W), which is characterized by a set of vertices V, a set of edges E, and a set of weights W. The values of W form a 16-by-16 matrix in which each element corresponds to the relative frequency of occurrence. w_{ij} is calculated as follows:

$$w_{ij} = \sum_{v_j \in B(v_j)}^{n} c_j$$

where B(v_j) is the set of toponym vertices that have a co-occurrence relation with v_j, and c_j is the frequency of v_j on the online news pages D_j. In short, the values of w_{ij} measure the strength of the interactions among the administrative units.

Spatial interaction modelling

In this study, we fit the gravity model to the interaction intensities that are derived from the mobile phone data and the news data. Although the Poisson model may be more appropriate than the log-normal model in fitting a gravity model to the data of counting type (Flowerdew and Aitkin 1982), in order to compare the distance attenuation coefficient with most of other researches that use popularized log-normal model with ordinary least square fitting, the same method were used. As used in the study of spatial interactions, the gravity model can be stated as:

$$I_{ij} = K \frac{P_i P_j}{D_{ij}^{\beta}}$$

where I_{ij} and D_{ij} are the intensity of a spatial interaction and the distance between units i and j, respectively. K is a constant, and β is the distance decay coefficient, which usually ranges from − 2 to − 1 (Kang et al. 2012; Gao et al. 2013; Xiao et al. 2013).

The variable P_i represents the number of phone call records made within the administrative units, whereas P_j represents the total number of news items that include the names of the administrative units. The GOF represented by the R^2 is calculated to determine whether the interaction intensity complies with Tobler's law of geography. The GOF (that is, the R^2) is calculated at the scale of base stations, townships, and administrative units to evaluate how the interaction intensity varies with spatial scale. The GOF of the interaction intensities that are derived from the news data is calculated only at the administrative unit level, and the results are compared to the GOF derived from the mobile phone data at the same spatial scale.

Results

The inner urban interactions based on the mobile phone data

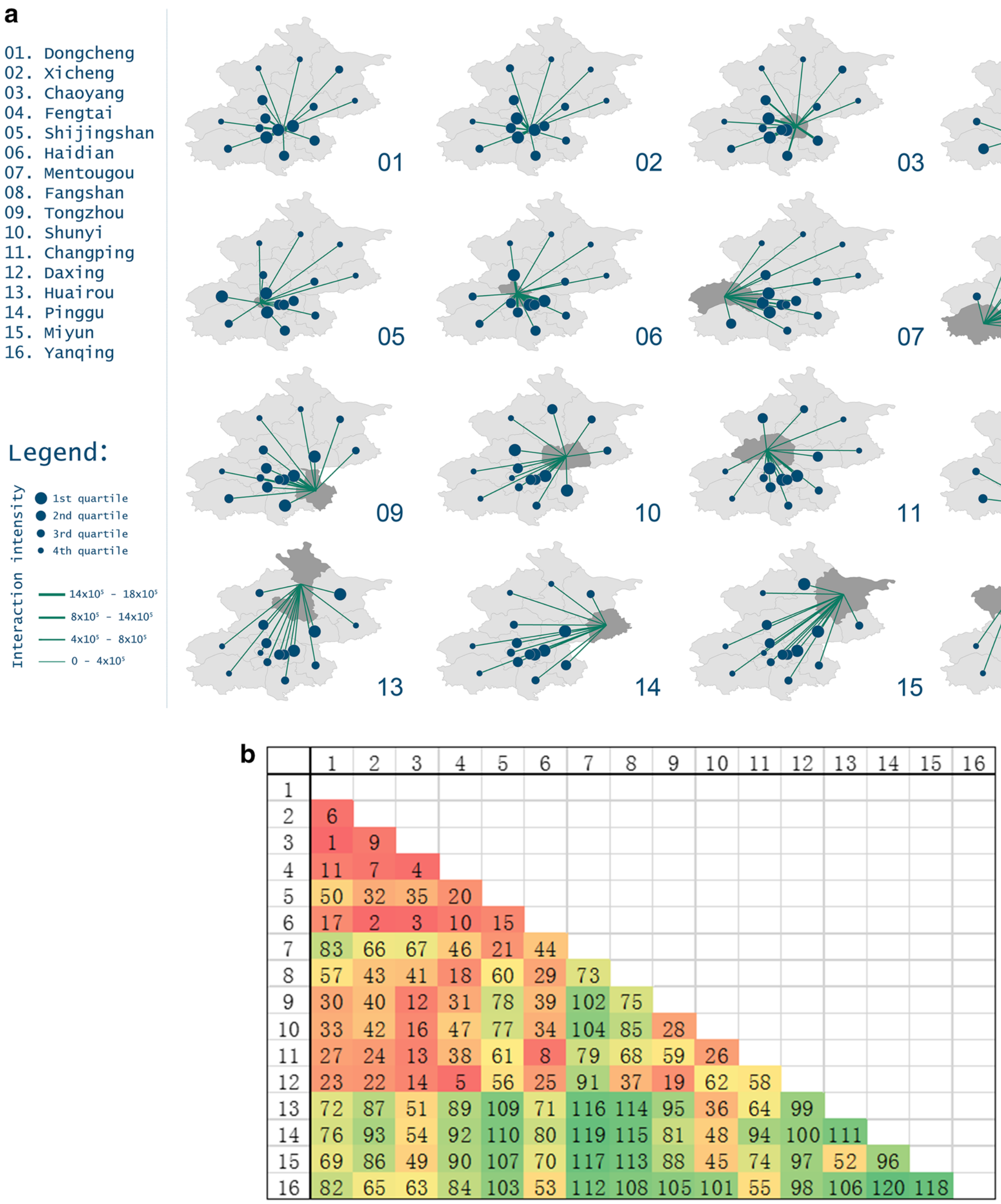
In total 12,279,121 trajectories are reconstructed within our study area from the mobile phone records collected on December 6th, 2015 (Table 1). The average length of the trajectories is 17.7 km, and the standard deviation is 29.45 km. On average, each trajectory includes records that were generated at 28 base stations.

The 16 undirected weighted sub-networks show the interactions between each administrative unit and the 15 other units. The sizes of the disks and the thicknesses of the lines in each sub-network show the strength of the interaction intensity, which is determined by the number of trajectories that link specific pairs of administrative units. As shown in Fig. 4, the five strongest interaction intensities are those between Dongcheng and Chaoyang; Xicheng and Haidian; Chaoyang and Haidian; Chaoyang and Fengtai; and Fengtai and Daxing. In contrast, the five weakest interaction intensities are those between Pinggu and Yanqing; Mentougou and Pinggu; Miyun and Yanqing; Mentougou and Miyun; and Mentougou and Huairou.

A multi-scale analysis is shown in Fig. 5, above. The left three panels present the interactions among

Table 1 Summary statistics for the original trajectories and the stay point sequences

	Number of trajectories	Average number of points	Average trajectory length (km)	Standard deviation (km)
Original trajectories	12,270,000	28	17.7	29.5
Stay point sequences		2	3.7	2.5



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1																
2	6															
3	1	9														
4	11	7	4													
5	50	32	35	20												
6	17	2	3	10	15											
7	83	66	67	46	21	44										
8	57	43	41	18	60	29	73									
9	30	40	12	31	78	39	102	75								
10	33	42	16	47	77	34	104	85	28							
11	27	24	13	38	61	8	79	68	59	26						
12	23	22	14	5	56	25	91	37	19	62	58					
13	72	87	51	89	109	71	116	114	95	36	64	99				
14	76	93	54	92	110	80	119	115	81	48	94	100	111			
15	69	86	49	90	107	70	117	113	88	45	74	97	52	96		
16	82	65	63	84	103	53	112	108	105	101	55	98	106	120	118	

Fig. 4 The inter-unit interactions based on the mobile phone data. **a** Maps showing the interaction intensities among the administrative units; **b** ranked interaction intensity matrix

regions at different scales in Beijing, whereas the right three panels reveal the relationships between interaction intensity and distance.

Over these three scales, the overall pattern is that, as the colour changes from dark yellow to light yellow, the interaction intensity decreases from the centre of Beijing to its edges. The existence of the distance decay effect is further confirmed by the scatter diagrams, which reveal that the interaction intensity decreases as the distance increases. However, measurements of the interactions made using different scales lead to different patterns in some local areas.

We constructed 320 and 20,978 sub-networks at the township and base station levels. Each sub-network has its own centre node, which is connected with all of the other nodes. We then use the gravity model to fit

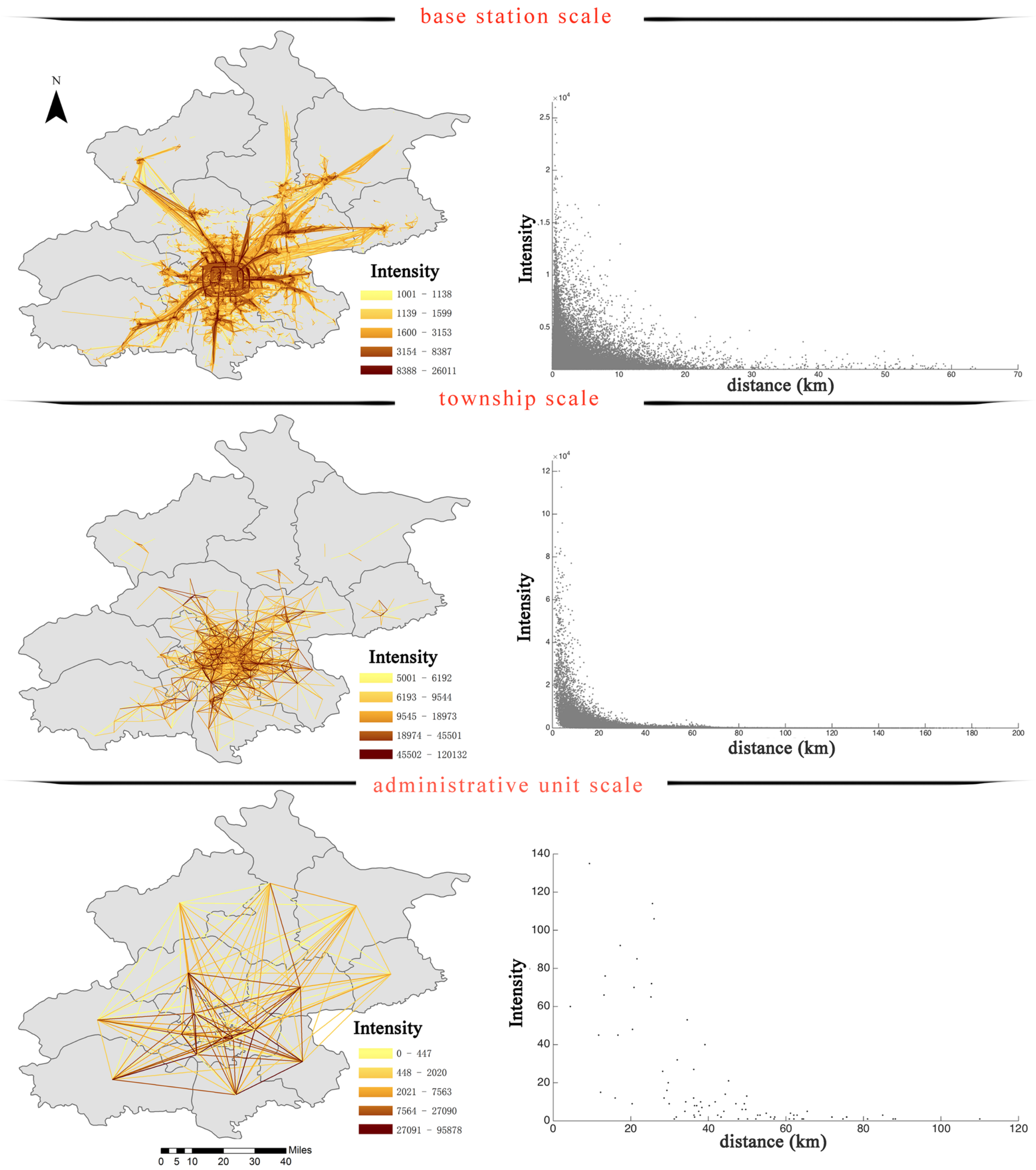


Fig. 5 Maps of inter-unit interactions and the distance decay effect on multiple scales

each of the sub-networks. Table 2 shows the relationship between the distance coefficient β of the gravity model and the intensity at different scales. The results conform to the law of distance decay and show that the interaction intensity decreases as the distance between administrative units increases. The GOF, as represented by R^2, increases from − 0.31 to 0.92 as the interactions are examined from the base station to the administrative levels.

A significant "centre radiation" pattern is present in Beijing at all three levels; that is, the central administrative units, including Dongcheng, Xicheng,

Table 2 Statistics and fitting coefficients at each spatial scale

	Base station	Township	Administrative unit
Number of areas	20,978	320	16
Average distance of the nearest neighbour	480 m	4 km	20 km
Coefficient of distance decay β	0.34	1.12	1.13
Goodness of fit R^2	− 0.31	0.63	0.92

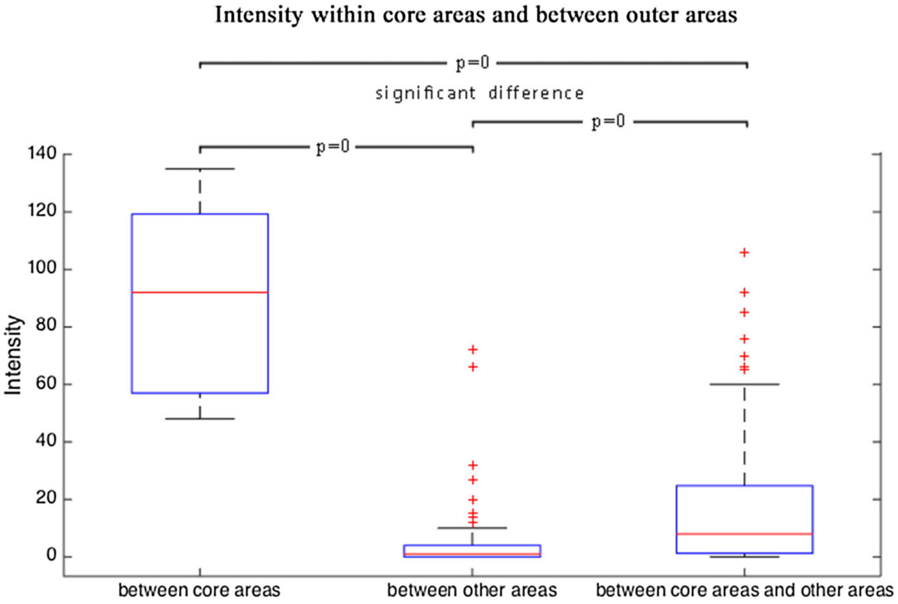


Fig. 6 *T* test for the intensity within the core areas and between the outer areas

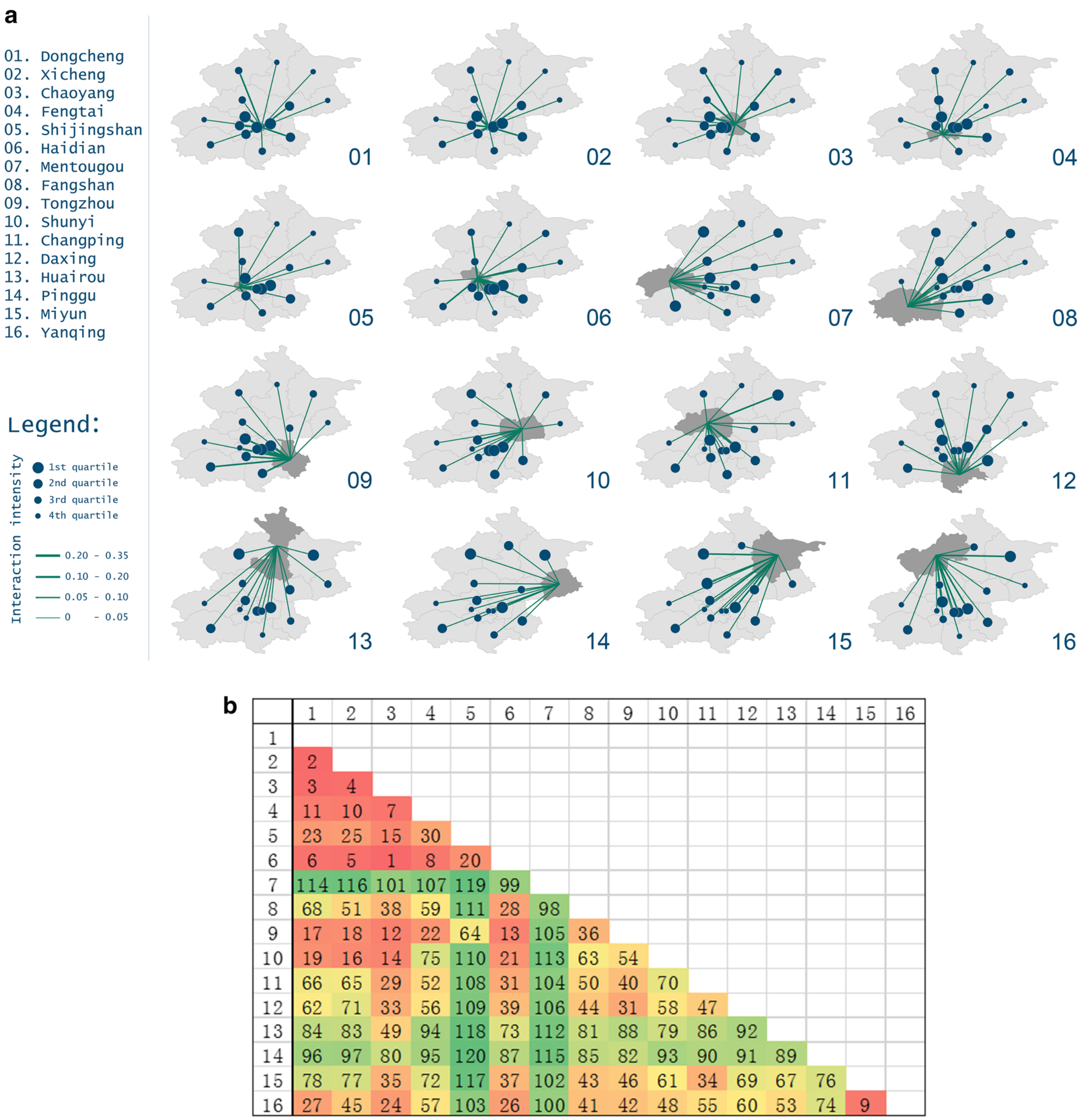
Chaoyang and Haidian, have significantly stronger interactions with all of the other units than do the suburban administrative units (Fig. 6).

The interaction intensity also varies at different scales. Strong interactions at the administrative unit scale may become weak at the township and base station scales. At the administrative unit scale, units serving the same function within the city tend to have fewer connections. In contrast, a weak interaction at the administrative unit scale is likely to be strong at the township scale. Some residential areas around the 5th Ring Road have strong interactions with the central urban area, which is obvious at the base station or township scales. However, these intense interactions do not bring two administrative areas significantly closer. For example, the largest residential area in Beijing, Tiantongyuan, has a high population density and strong interactions with the central urban area inside the 5th Ring Road because of daily commuting. These interactions are obvious in the first two panels in Fig. 5. However, as shown in the bottom panel, the interaction between Changping and Chaoyang is relatively weak at the administrative unit scale.

Here, we note that the administrative boundaries are defined artificially; therefore, when different boundary systems are used in this analysis, the resulting correlation coefficients may vary substantially, reflecting inconsistent results. This behaviour is called the modifiable areal unit problem (MAUP) (Fotheringham and Wong 1991). The MAUP affects the results when point-based measures of spatial phenomena are aggregated into districts. Nevertheless, when the results of two data sources are compared using the same set of administrative boundaries, as in this study, this problem can be minimized (Fotheringham and Wong 1991).

The inter-unit connections based on the news data

Figure 7 shows the patterns of interaction intensity between administrative units that are calculated based on the news data. The values in the matrix represent the number of news items that include two administrative units. As shown in Fig. 7, the 5 strongest interactions are identified between Chaoyang and Haidian; Dongcheng and Xicheng; Dongcheng and Chaoyang; Xicheng and Chaoyang; and Xicheng and Haidian. The 5 weakest interactions are those between Shijingshan and Pinggu; Shijingshan and Mentougou; Shijingshan and Huairou; Shijingshan and Miyun; and Xicheng and Mentougou.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1																
2	2															
3	3	4														
4	11	10	7													
5	23	25	15	30												
6	6	5	1	8	20											
7	114	116	101	107	119	99										
8	68	51	38	59	111	28	98									
9	17	18	12	22	64	13	105	36								
10	19	16	14	75	110	21	113	63	54							
11	66	65	29	52	108	31	104	50	40	70						
12	62	71	33	56	109	39	106	44	31	58	47					
13	84	83	49	94	118	73	112	81	88	79	86	92				
14	96	97	80	95	120	87	115	85	82	93	90	91	89			
15	78	77	35	72	117	37	102	43	46	61	34	69	67	76		
16	27	45	24	57	103	26	100	41	42	48	55	60	53	74	9	

Fig. 7 The inter-unit connections based on the news data. **a** Maps of connection intensity between administrative units; **b** ranked interaction intensity matrix

Discussion

Consistency between the results derived from the mobile phone and online news data

The distance decay effect exists in the ties among the regions at different spatial scales in Beijing. Closer regions usually have stronger bonds. The two data sources reveal a mutual regional bonding pattern of concentric circles, in which the inner circles display closer bonds.

At the administrative unit scale, the mobile phone and online news data show very similar inter-unit interactions within our study area (Fig. 8). The correlation coefficients between the interactions based on the mobile phone data and the connections based on the news data decay gradually from the central urban area to the suburbs. The correlation coefficients range from 0.79 to 0.91, and the correlations are statistically significant for the districts of Dongcheng, Xicheng, Chaoyang, Fengtai, and Haidian. These 5 regions are clustered in the centre of Beijing. A lower correlation coefficient of less than 0.5 ($p > 0.01$) is found for Mentougou, Fangshan, Daxing, Huairou, Pinggu,

Fig. 9 The differences between the interactions based on the mobile phone data and the connections based on the news data. **a** Rank difference matrix; **b** map of interaction intensity based on the mobile phone data and the connections based on the news data; **c** map of the intensity differences

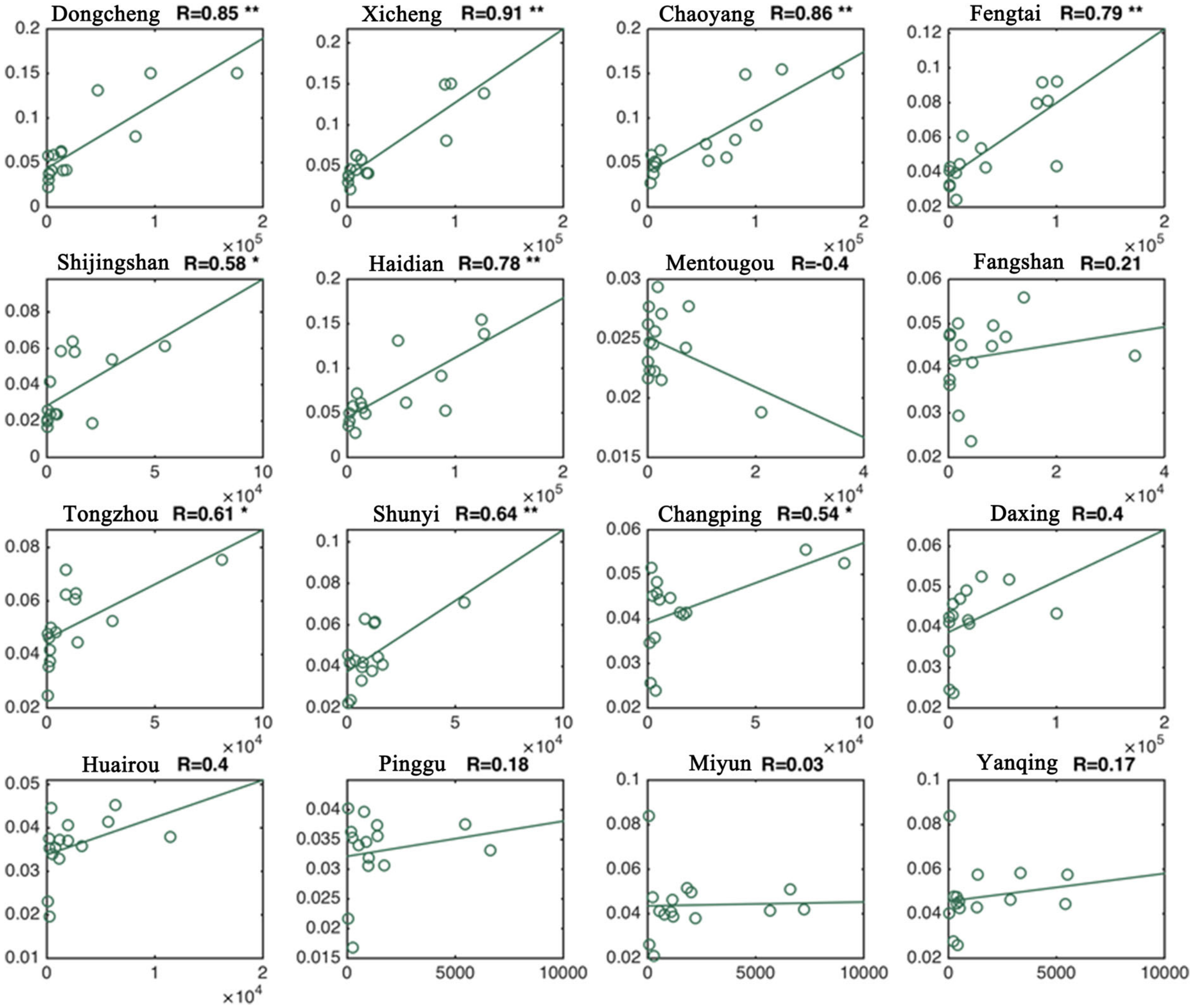
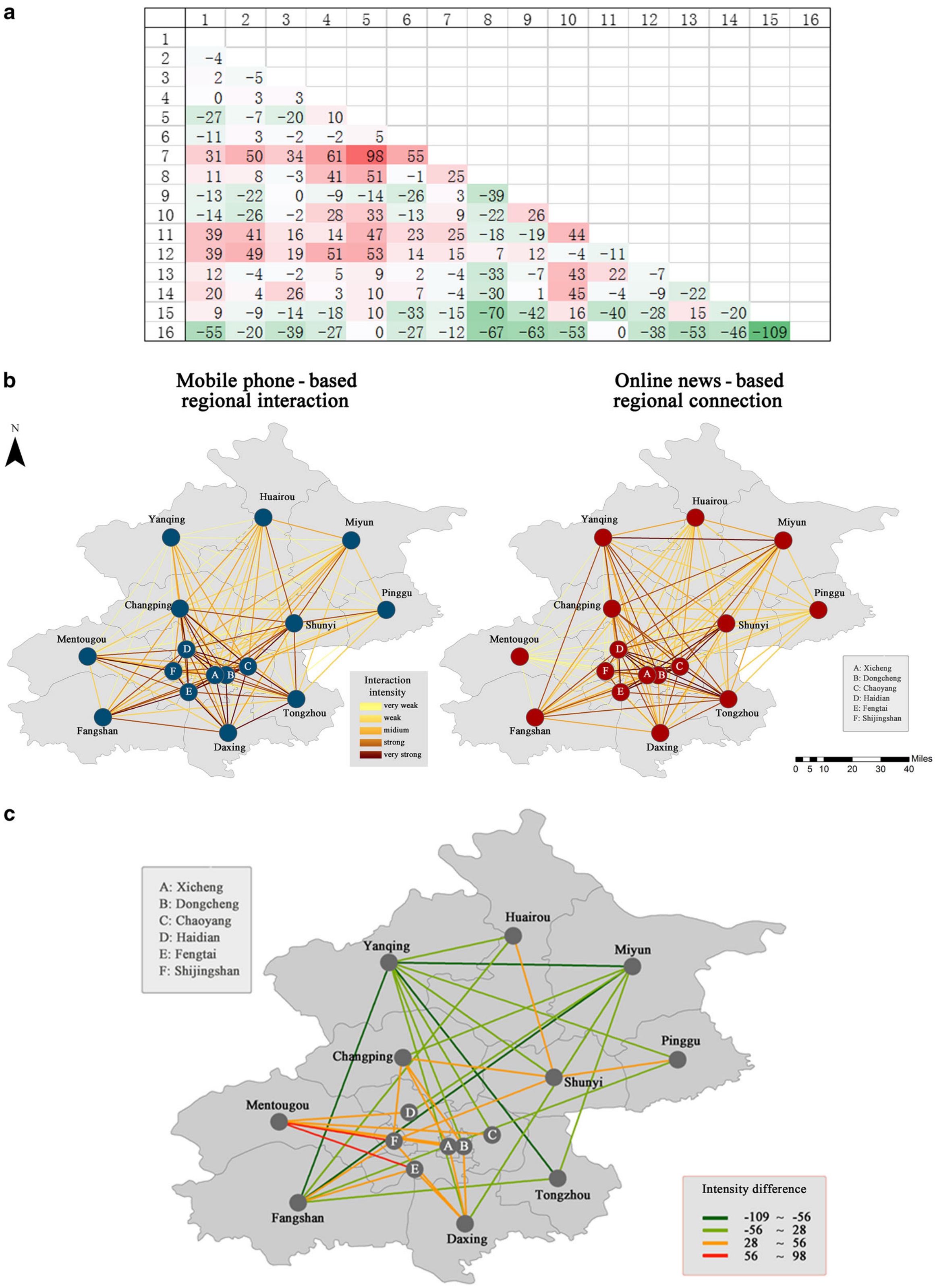


Fig. 8 The correlations between the interactions determined using the mobile phone data and the connections based on the news data



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1																
2	-4															
3	2	-5														
4	0	3	3													
5	-27	-7	-20	10												
6	-11	3	-2	-2	5											
7	31	50	34	61	98	55										
8	11	8	-3	41	51	-1	25									
9	-13	-22	0	-9	-14	-26	3	-39								
10	-14	-26	-2	28	33	-13	9	-22	26							
11	39	41	16	14	47	23	25	-18	-19	44						
12	39	49	19	51	53	14	15	7	12	-4	-11					
13	12	-4	-2	5	9	2	-4	-33	-7	43	22	-7				
14	20	4	26	3	10	7	-4	-30	1	45	-4	-9	-22			
15	9	-9	-14	-18	10	-33	-15	-70	-42	16	-40	-28	15	-20		
16	-55	-20	-39	-27	0	-27	-12	-67	-63	-53	0	-38	-53	-46	-109	

Miyun, and Yanqing. These units are located in the suburban areas. The transitional zone between the central and suburban areas includes Shijingshan, Tongzhou, Shunyi ($p < 0.01$), and Changping, which show correlation coefficient values ranging between 0.5 and 0.75. The correlations are not statistically significant for most of the administrative units in this zone, except Shunyi.

Divergence between the results derived from the mobile phone data and the online news data

Distance affects the bond strength between administrative units to a greater degree when the strength is calculated using the human trajectories, rather than the online news data.

Although the administrative units are at the same political level, their importance in urban functions may vary, and this study may show that administrative units serving similar functions tend to show stronger interactions. For example, the two central units of Dongcheng and Xicheng share very similar urban functions; thus, they are often mentioned together in news data, leading to a relatively strong connection between these units. The two suburban units of Miyun and Yanqing also show a relatively tight connection, as reflected by their co-occurrence probability, despite the long distance between them. This phenomenon cannot be simply explained by the gravity model and does not show a clear distance decay effect (Fig. 9b).

A difference matrix is created to illustrate the divergence between the intensity of the mobile phone-based interactions and the online news-based connections. The numbers in the matrix represent the differences in rank between the results derived from the news data and the cell phone data. Positive values indicate that the rankings based on the news data are higher than those obtained using the cell phone data, whereas negative values reflect the opposite difference. On the whole, the units with positive values (darker red), which indicate cases in which stronger interactions are inferred from the cell phone data than the news data, are mainly distributed in the middle of this difference matrix. On the other hand, the units with negative values (darker green) lie at the bottom of this matrix.

Projecting the difference matrix (Fig. 9a) into geographical space (Fig. 9c) shows the spatial patterns revealed by the differences in ranks. The results show that the differences between the regional interactions and connections are spatially heterogeneous. Following the colour scheme used in the matrix, the interactions represented by darker red lines are distributed in the central portion of the study area and concentrate in the southwest. The units connected with darker green lines describe a large circle around the boundary of the city. Most parts of the administrative units located to the north and east, including Yanqing, Huairou, Miyun, Pinggu, Shunyi and Tongzhou, display an obvious pattern; specifically, the interactions with other areas inferred from the mobile phone data are almost always stronger than the connections inferred from the news data. However, in the southwestern part of the study area, including Changping, Mentougou, Shijingshan, Fengtai, and Daxing but excluding Fangshan, the connections between these areas inferred from the news data are generally stronger than their interactions based on the mobile phone data. The positive values of the connections between Fangshan and Fengtai and between Fangshan and Shijingshan may result from the very good transportation links between these units, though there are no significant inter-unit differences in economic conditions and resources. The news data indicate tight connections among the four suburban districts, Yanqing, Fangshan, Miyun and Tongzhou. The mobile phone data do not reflect strong interactions among these districts because very few individuals travel between these units within 1 day. However, because these units are located along the outermost circle of Beijing, they have similar city functions and thus tend to be mentioned together in the news. For example, one news item reported "five ecological conservation areas are now the first choice for leisure tourism in Beijing" (http://travel.cnr.cn/list/20170221/t20170221_523611976.shtml); this news item mentioned the five administrative areas of Daxing, Changping, Miyun, Mentougou and Huairou all at once. Ecological conservation zones are written into the master plan of Beijing as "one zone" in the "one core, one main, one assistance, two axes, multiple points, one zone" principle. This term refers to the ecological conservation zones of the districts of Mentougou, Pinggu, Huairou, Miyun, Yanqing, Changping and Fangshan, which are important parts of the ecological conservation area in the northwestern portion of Beijing–Tianjin–Hebei.

The characteristics of the two types of data

The use of mobile phone data to collect and characterize the movements of individuals has the following advantages. The widespread use of mobile electronic devices enables the assessment of overall trends through the recording of the movements of individuals; the coverage of the signal is wide, and the accuracy of positioning is high; and the records can be accessed in real time. Given the distribution of base stations within the city, the data are available on a sub-kilometre scale. However, when the granularity of the spatial analysis has the same order of magnitude as the service radius of the base stations, the results are sometimes unstable because of the imperfect nature of mobile phone data. However, as the scale increases, studies on variable scales lead to more comprehensive analyses.

Given its free access and availability, mass media-based toponym co-occurrence analysis can assist in measuring the degree of overall connection between regions; however, the spatial scale must be correctly selected to ensure that sufficient data fall into each region to produce stable co-occurrence probabilities. In this study, interactions are defined solely in terms of human flows, whereas the connections among regions are defined more broadly. We believe that the mobile phone data can help us to uncover the inter-regional interactions, whereas the news data primarily reflects the connections between regions.

This paper notes that, besides the inherent information bias that exists in big data, including mobile phone signals, it is also important to consider another category of bias, cognitive bias, which involves assumptions as to what the data represent. For example, it seems that both human flows and the co-occurrence probabilities of toponyms in news items can reasonably represent the connection intensity between regions inside a city. However, these datasets actually differ strongly; the data on daily individual trajectories primarily provides a snapshot of regional interactions, whereas the news data can be considered to represent a running account of regional connections. In short, the use of mobile phone data focuses on short-term interactions in space, whereas the news data reflect the functional interactions over longer periods. Researchers must be wary of the cognitive biases that they bring to the interpretation of data.

One of the contributions of this paper is that we identify the cognitive bias in analyses of big data, as exemplified by an urban case study that uses two sources of big data. We conclude that it is of importance to make assumptions that are suitable for specific problems.

Conclusions

In this paper, we chart the interactions and connections between the administrative units of Beijing, China using individual trajectories derived from mobile phone data and a toponym co-occurrence analysis performed using mass media data. Given the above-mentioned results and analysis, conclusions can be drawn for the study area and the methodology we follow.

Within the study area of Beijing, the overall pattern of interactions and connections among administrative districts have the shape of concentric circles. Dongcheng and Chaoyang share the strongest bond among the pairs of regions. Compared with the mass media-based regional connections, the interactions among regions based on human flows is better represented by the gravity model.

Both the mobile phone-based trajectories and the mass media-based toponym co-occurrence analysis can measure the connections between regions within a city. However, the assumptions as to which data are appropriate in representing spatial interactions can significantly affect the results. Thus, a rigorous examination of specific problems is needed to redefine each problem in a more specific way.

These conclusions only hold for our study area. To test the generality of these conclusions, more studies must be carried out in different cities in the future. For the previously mentioned scale effect in measurements of interactions using mobile phone data, further research using grids with various resolutions could be performed to quantitatively identify the resolution at which the gravity model provides the best fit to the data. In the analysis of toponym co-occurrence within news data, the topic of each news item is worthy of consideration. Different kinds of connections could then be tagged according to the co-occurrence of toponyms in news items with different topics,.

Acknowledgements This research is funded by NSF-4159840011, 088RA500PA, and QYZDY-SSW-DQC007-2. Funding was provided by the Major Program of the National Natural Science Foundation of China, Innovation Program of State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, and Science Foundation of the Chinese Academy of Sciences.

References

Abel, G. J., & Sander, N. (2014). Quantifying global international migration flows. *Science, 343*(6178), 1520–1522.

Bhattacharya, S., & Basu, P. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics, 43*(3), 359–372.

Birkin, M., & Malleson, N. (2012). *Investigating the behaviour of twitter users to construct an individual-level model of metropolitan dynamics*. Working paper.

Chen, W., Ma, X., Cai, L., Luan, X., & Li, G. (2013). Characteristics of regional city connection's spatial pattern based on intercity passenger traffic flow in Pearl River Delta. *Economic Geography, 33*(4), 008.

Djankov, S., & Freund, C. (2002). Trade flows in the former Soviet Union, 1987 to 1996. *Journal of Comparative Economics, 30*(1), 76–90.

Feng, Z., & Xiao, Q. (2014). The application of big data in smart city research and planning. *Urban Planning International, 29*(6), 44–50.

Flowerdew, R., & Aitkin, M. (1982). A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science, 22*(2), 191–202.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025–1044.

Fuellhart, K. (2003). Inter-metropolitan airport substitution by consumers in an asymmetrical airfare environment: Harrisburg, Philadelphia and Baltimore. *Journal of Transport Geography, 11*(4), 285–296.

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environment and Planning B: Planning and Design, 40*(1), 135–153.

Haggett, P., Cliff, A. D., & Frey, A. (1978). Locational analysis in human geography: Volume 1: Locational models. *Journal of the Royal Statistical Society*, *141*(4), 554.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133.

Herold, M., Goldstein, N. C., & Clarke, K. C. (2003). The spatiotemporal form of urban growth: Measurement, analysis and modeling. *Remote Sensing of Environment, 86*(3), 286–302.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science, 2010*(1), 21–48.

Jendryke, M., Balz, T., McClure, S. C., & Liao, M. (2017). Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. *Computers, Environment and Urban Systems, 62*(1), 99–112.

Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1702–1717.

Kang, C., Zhang, Y., Ma, X., & Liu, Y. (2013). Inferring properties and revealing geographical impacts of intercity mobile communication network of China using a subnet data set. *International Journal of Geographical Information Science, 27*(3), 431–448. https://doi.org/10.1080/13658816.2012.689838.

Lathia, N., Quercia, D., & Crowcroft, J. (2012). The hidden image of the city: Sensing community well-being from urban mobility. In *International conference on pervasive computing, 2012* (pp. 91–98).

Li, Q., Chang, X., Shaw, S., Yan, K., Yue, Y., & Chen, B. (2013). Characteristics of micro-blog inter-city social interactions in China. *Journal of Shenzhen University Science and Engineering, 30*(5), 441–449.

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography, 43*(1), 78–90. https://doi.org/10.1016/j.jtrangeo.2015.01.016.

Liu, Y., Gong, L., & Tong, Q. (2014a). Quantifying the distance effect in spatial interactions. *Acta Scientiarum Naturalium Universitatis Pekinensis, 50*(3), 526–534.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems, 14*(4), 463–483.

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014b). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS ONE, 9*(1), e86026. https://doi.org/10.1371/journal.pone.0086026.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014c). Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS, 18*(1), 89–107.

Long, Y., Zhang, Y., & Cui, C. (2012). Identifying commuting pattern of Beijing using bus smart card data. *Acta Geographica Sinica, 67*(10), 1339–1352.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems, 36*(2), 105–108.

Luo, J., & Wei, Y. D. (2009). Modeling spatial variations of urban growth patterns in Chinese cities: The case of Nanjing. *Landscape and Urban Planning, 91*(2), 51–64.

Masucci, A. P., Serras, J., Johansson, A., & Batty, M. (2013). Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E, 88*(2), 22812.

Matsumoto, H. (2004). International urban systems and air passenger and cargo flows: Some calculations. *Journal of Air Transport Management, 10*(4), 239–247.

Nelson, G., & Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE, 11*(11), e0166083.

Niu, X., Ding, L., & Song, X. (2014). Understanding urban spatial structure of shanghai central city based on mobile phone data. In *Urban planning forum, 2014* (Vol. 6, pp. 61–67).

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*(5), 378.

Roark, B., & Charniak, E. (1998). Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In *Proceedings of the 17th international conference on Computational linguistics, 1998* (Vol. 2, pp. 1110–1116). Association for Computational Linguistics.

Shen, G. (2004). Reverse-fitting the gravity model to inter-city airline passenger flows by an algebraic simplification. *Journal of Transport Geography, 12*(3), 219–234.

Shi, L., Chi, G., Liu, X., & Liu, Y. (2015). Human mobility patterns in different communities: A mobile phone data-based social network approach. *Annals of GIS, 21*(1), 15–26.

Simini, F., Gonzalez, M. C., Maritan, A., & Barabasi, A. L. (2012). A universal model for mobility and migration patterns. *Nature, 484*(7392), 96–100. https://doi.org/10.1038/nature10856.

Smith, C., Quercia, D., & Capra, L. (2013). Finger on the pulse: Identifying deprivation using transit flow analysis. In *Conference on computer supported cooperative work, 2013* (pp. 683–692).

Stefanouli, M., & Polyzos, S. (2017). Gravity vs radiation model: Two approaches on commuting in Greece. *Transportation Research Procedia, 24*(1), 65–72. https://doi.org/10.1016/j.trpro.2017.05.069.

Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation, 33*(2), 106–119. https://doi.org/10.1108/eb026637.

Van Zanten, B. T., Van Berkel, D. B., Meentemeyer, R. K., Smith, J. W., Tieskens, K. F., & Verburg, P. H. (2016). Continental-scale quantification of landscape values using social media data. *Proceedings of the National Academy of Sciences, 113*(46), 12974–12979. https://doi.org/10.1073/pnas.1614158113.

Veloso, M., Phithakkitnukoon, S., & Bento, C. (2011). Urban mobility study using taxi traces. In *Proceedings of the 2011 international workshop on Trajectory data mining and analysis, 2011* (pp. 23–30). New York: ACM.

Vermeiren, K., Van Rompaey, A., Loopmans, M., Serwajja, E., & Mukwaya, P. (2012). Urban growth of Kampala, Uganda: Pattern analysis and scenario development. *Landscape and Urban Planning, 106*(2), 199–206.

Wang, K., & Deng, Y. (2016). Identification of spatial connection intensity of Zhongyuan urban agglomeration based on microblogging. *Journal of University of Chinese Academy of Sciences, 33*(6), 775–782.

Wang, B., & Zhen, F. (2016). The role of distance in online social networks: A case study of Sina micro-blog. *Progress in Geography, 35*(8), 983–989.

Xiao, Y., Wang, F., Liu, Y., & Wang, J. (2013). Reconstructing gravitational attractions of major cities in china from air passenger flow data, 2001–2008: A particle swarm optimization approach. *The Professional Geographer, 65*(2), 265–282. https://doi.org/10.1080/00330124.2012.679445.

Xu, M., Li, Z., Shi, Y., Zhang, X., & Jiang, S. (2015). Spatial linkage of global container shipping network. *Journal of Shanghai Maritime University, 36*(3), 6–12.

Ye, Y., Zheng, Y., Chen, Y., Feng, J., & Xie, X. (2009). Mining individual life pattern based on location history. In *2009 tenth international conference on mobile data management: Systems, services and middleware, 2009* (pp. 1–10). IEEE.

Zheng, Y., & Zhou, X. (2011). *Computing with spatial trajectories*. Berlin: Springer.

Zhong, X., Liu, J., Gao, Y., & Wu, L. (2017). Analysis of co-occurrence toponyms in web pages based on complex networks. *Physica A: Statistical Mechanics and its Applications, 466*(C), 462–475. https://doi.org/10.1016/j.physa.2016.09.024.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison Wesley.