Computers, Environment and Urban Systems 74 (2019) 50–61

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling

Kang Liu[a,b], Song Gao[c], Feng Lu[b,d,e,f,*]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[c] Department of Geography, University of Wisconsin-Madison, Wisconsin-Madison, USA
[d] Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou, China
[e] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China
[f] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Spatial interaction
Vehicle movement
Stroke
Road network
Topic model

ABSTRACT

The development of mobile positioning technologies makes massive individual trajectory data easily accessible, which facilitates the revisit of spatial interaction issue in recent years. Researchers have proposed many methods to investigate the spatial interactions derived from human movements, such as the gravity model and radiation model. However, these studies have mainly focused on the interactions among areal units at an aggregated level, neglecting that in most cases, human movements are carried by vehicles and constrained by the underlying road network, which causes the interactions among roads. To fill this gap, we propose a novel approach to identify spatial interaction patterns of vehicle movements on urban road network. As the topic model originating from the domain of natural language processing has powerful advantages in extracting semantic relations of words from corpus, we utilize it to extract interaction relations of urban roads from massive vehicle trajectories. First, "strokes" (i.e., natural streets) are chosen as geographical units to represent the vehicle moving paths. Then, an analogy between geographical elements (i.e., stroke, moving path) and textual elements (i.e., word, document) is established, and a topic model is applied to the moving paths to identify the spatial interaction patterns on road network. From a mass of trajectory data collected by GNSS-equipped taxis in Beijing, the "topic patterns", which can be viewed as clusters of spatially interacted strokes, are identified, visualized and verified. It is argued that our proposed approach is effective in identifying spatial interaction patterns, which provides a new perspective for spatial interaction modelling and enriches the current spatial interaction studies.

1. Introduction

Spatial interaction is a traditional issue in geographical research. Modelling spatial interactions between geographical units based on population immigration, commodity flow, or information communication can help us understand spatial structure of a region and plan an efficient spatial configuration for cities and regions (Fotheringham & O'Kelly, 1989; Gao, Liu, Wang, & Ma, 2013; Liu, Sui, Kang, & Gao, 2014; Wan et al., 2018; Wilson, 1967).

In recent years, with the development of ICT (Information and Communications Technology), especially mobile positioning technology, massive individual trajectory data are becoming easily accessible. Mobility information gathered at the individual level can be aggregated to study the human movements from one region to another. Hence it facilitates emerging studies on spatial interactions collected from ICT big data, which get much attention from the fields of transportation, physics, epidemiology, urban planning, and so on (Balcan et al., 2009; Lenormand, Huet, Gargiulo, & Deffuant, 2012; Tizzoni et al., 2014; Yan, Zhao, Fan, Di, & Wang, 2014).

Many researchers have devoted to predicting human moving fluxes between any two locations using a limited set of “static” attributes of the locations using various spatial interaction models (Barbosa et al., 2018). The differences of those models are mainly in the choice of attributes, and the specific functional forms. There are two related but diverging school of thoughts in classical studies. The gravity models initially introduced by George K. Zipf in 1946 assume that the number of trips between two locations is a decreasing function of their distance (Zipf, 1946). Instead of highlighting the importance of the distance, the

* Corresponding author at: 11A, Datun Road, Chaoyang District, Beijing 100101, China.
E-mail address: luf@lreis.ac.cn (F. Lu).

https://doi.org/10.1016/j.compenvurbsys.2018.12.001
Received 27 June 2018; Received in revised form 9 November 2018; Accepted 3 December 2018

intervening opportunities models derive from Stouffer's conceptual framework assume that the number (or cumulative number) of intervening opportunities between the origin and the destination plays a key role in determining migration (Stouffer, 1940). In recent years, with the support of massive geo-locating data, and on the basis of the classical thoughts, some new models have been proposed, including the radiation model (Simini, González, Maritan, & Barabási, 2012), rank-distance model (Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012), single-parameter model (Lenormand et al., 2012), population-weighted opportunity model (Yan et al., 2014), etc. Compared to previous models, they can predict the actual human movements among sub-regions more accurately with fewer parameters. Differing from the abovementioned mainstream models, Zhu, Huang, Shi, Wu, and Liu (2018) recently proposed to infer the spatial interaction patterns from sequential snapshots of spatial distributions, which provides a new viewpoint and a new framework to model spatial interactions.

In addition, with the rise of complex network science, some geographers have regarded the interactions among different regions as a spatially embedded graph, where the areal units are represented as nodes, and the interacting flows are represented as weighted edges (Gao et al., 2013; Hawelka et al., 2014; Liu et al., 2014; Liu, Gong, Gong, & Liu, 2015; Peng et al., 2018). Based on such transformation, the community detection algorithms, which can divide a network into sub-networks that have stronger inner connectivity, are applied to discover hidden patterns in the spatial interaction network. For instance, by partitioning both the network of call interaction and the network of movements extracted from mobile phone data, Gao et al. (2013) found that people tend to communicate within a spatially-proximate community, and the phone-users' movements in physical space and the phone-call interaction in cyberspace are highly correlated. Liu et al. (2014) found that the detected communities are spatially cohesive and roughly consistent with province boundaries when they investigated the inter-city trips and spatial interactions using social media check-in data.

In sum, existing studies have mainly focused on the spatial interactions among areal units, such as countries (Hawelka et al., 2014), cities (Liu et al., 2014; Noulas et al., 2012), and parcels inside a city (e.g., administrative districts, traffic analysis zones, grid cells, and Voronoi cells) (Gao et al., 2013; Liu et al., 2015; Yan et al., 2014), neglecting that the majority of human movements are constrained by the underlying road network, causing spatial interactions among the linear elements, i.e., roads. Recent studies also indicated that road unit is a promising data assembly and analysis unit for quantitative urban studies (Zhu, Wang, Wu, & Liu, 2017). Investigating spatial interactions on road network at geographical unit of road is meaningful as more detailed human moving processes can be revealed.

Therefore, in this paper, we propose an innovative approach to extracting the spatial interaction patterns on urban road networks using a topic model borrowed from natural language processing (NLP) domain.

Topic models can discover the abstract “topics” and hidden semantic structures from vast textual documents, hence they are widely used in document clustering, information retrieval, and feature selection in NLP (Blei, Ng, & Jordan, 2003). In recent years, topic models are also introduced to geographical and urban studies. For instance, Gao, Janowicz, and Couclelis (2017) and Yuan, Zheng, and Xie (2012) used topic models to identify urban functional regions from POIs (Points of Interests). Liu and Cheng (2018) as well as Mohamed, Côme, Oukhellou, and Verleysen (2017) applied topic models using smart card data to identify passengers' travel behavior according to their boarding/alighting time in public transit. Hu et al. (2017) extracted and analyzed the semantic relatedness between cities with the help of topic model and news articles. To improve the targeted outdoor advertising, Lai, Cheng, and Lansley (2017) applied a topic model to geotagged Tweets to identify the popular interests around given places (e.g., London Underground stations). Chu et al. (2014) and Zhang et al. (2016) applied topic models to vehicle trajectory data, but their approaches are different in both analogy ways and research purposes compared to our current study. Regarding named roads as words and taxi trips as documents, Chu et al. (2014) visualized the hidden themes of taxi movements in a city. While with each taxi's trip destinations (each of which is attached to a road segment ID) in a time period (e.g., a specific day) being analogous to the words in a document, Zhang et al. (2016) analyzed urban human mobility patterns by exploring the hot spots of taxis' trip destinations.

In this study, taking the topic model's advantage in extracting semantic relations of words from corpus, we utilize it to extract the interaction relations of urban roads from massive vehicle movement data, i.e., identify spatial interaction patterns of vehicle movements on urban road networks. Firstly, in order to better reflect human driving behavior in the real world, we choose “stroke” as geographic unit to express urban road, and represent each vehicle moving path as a series of strokes. Secondly, we establish an analogy between geographical elements (i.e., stroke, moving path) and textual elements (i.e., word, document), and then apply a topic model to a mass of moving paths to identify the spatial interaction patterns of vehicle movements on road network. This study can help traffic managers and urban planners better understand the traffic patterns and spatial structures of the city in the road network space, and plan an efficient traffic control or spatial configuration for the city.

The following of this paper is structured as follows: Section 2 introduces the methodology of this paper, including the representation of vehicle moving paths and the identification of spatial interaction patterns. Section 3 conducts a case study using massive taxi trajectory data collected in Beijing. Section 4 is devoted to discussions, and Section 5 concludes this work.

2. Methodology

Essentially, spatial interactions among roads originate from numerous vehicle movements, which are constrained and influenced by the underlying road networks (Hillier & Iida, 2005; Jiang, 2009; Jiang & Jia, 2011). Based on this fact, first, we choose "stroke" as geographical unit to express road, and propose to represent vehicle moving path by strokes, in order to better reflect human driving behavior. Then, we creatively propose to apply a topic model borrowed from NLP domain to the vehicle moving paths, in order to identify spatial interaction patterns on road network.

2.1. Representation of vehicle moving paths by strokes

2.1.1. Why choose "stroke" as geographical unit to express road?

A "stroke" (also in terms of "natural street") is defined as naturally merged road segments (each of which is a section between two adjacent intersections) following the principle of good continuity (Jiang, Zhao, & Yin, 2008; Thomson & Richardson, 1999; Yang, Luan, & Li, 2011), e.g., minor direction change of two adjacent road segments. Fig. 1(a) illustrates several strokes, where nodes are intersections, edges between nodes are road segments, and the chains rendered with same colors and letter are strokes generated from the road segments. Fig. 1(b) illustrates the spatial interactions among strokes, where nodes represent strokes, edges represent the associations between strokes, and widths of the edges reflect the moving flows between strokes. The target of this paper is to identify spatial interaction patterns where we can know which strokes are spatially interacted with each other in a strong way.

The concept of "stroke" originates from the fields of space syntax, spatial cognition, and the emerging network science, which emphasize on examining the relationship between human cognition and urban space (Jiang & Liu, 2009; Thomson & Richardson, 1999). Strokes can reflect two important aspects of human cognition on road network space (i.e., human driving behavior), that is, visual perception and hierarchical cognition.

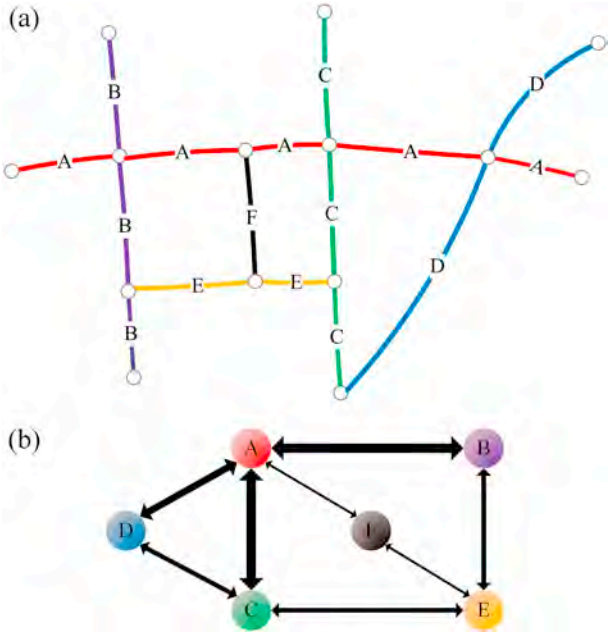

Fig. 1. Illustration of strokes and the spatial interactions among strokes.

Visual perception is a concept at small-scale space that can be perceivable from a single vantage point. Studies indicate that people tend to follow the longest line of sight that approximates their heading in order to reduce the complexity of environment, hence reduce the cognition burden in way-finding process (Conroy, 2001; Dalton, 2003; Golledge, 1995). In space syntax, such perceptive unit is represented by the longest visibility line, which is called axial line. In recent years, strokes are often used to replace axial lines for road network analysis (Thomson, 2004).

Hierarchical cognition is a concept at large-scale space (i.e., continuous open space). It refers to that anchors, particularly salient features in urban space, shape people's memory of a city, around which subjective knowledge is hierarchically organized (Couclelis, Golledge, Gale, & Tobler, 1987; Golledge & Spector, 1978; Passini, 1984). Actually, the road network itself is hierarchically organized at the granularity of stroke. Strokes demonstrate a scaling law (Zipf, 1946) that a majority of strokes are less connected, while a minority of strokes are well connected (Jiang, 2007). Whereas the minority of strokes accounts for a majority of traffic flow (Jiang, 2009), as they are salient features in individuals' spatial cognition, and chosen more frequently by the drivers in their way-finding process. Precisely due to their hierarchical nature, strokes have been widely used in map generalization, road selection, and road network analysis (Jiang et al., 2008; Thomson, 2004; Yang et al., 2011).

Based on the specific properties of strokes described above, we use "stroke" as geographical unit to express road in our spatial interaction study.

2.1.2. Representation of vehicle moving paths

Choosing strokes as geographical units, we further present a novel method to represent vehicle moving paths along road network in a meaningful way.

First, a moving path is represented by a series of consecutive road segments as follows:

$$route = [road\ segment_1, road\ segment_2, ..., road\ segment_N],$$

where N is the number of road segments included in the path.

Second, the road segment IDs are replaced by the stroke IDs they belonging to, and the moving path is finally represented as:

$$route = [stroke_1, stroke_2, ..., stroke_N].$$

Fig. 2 shows a representation example of a moving path. If we represent the path by road segments, then the path is formed as:

[23, 14, 7, 4, 15, 46, 117];

while if we represent the path by strokes, then the path is formed as:

[44, 12, 12, 7, 7, 7, 7].

Such representation implies the drivers' moving behavior by the frequency of stroke IDs appearing in a moving path, denoting which stroke(s) is (are) preferred in the trip. Take Fig. 2(2) for instance, "stroke 7" occurs frequently in the path, implying that the linear element is consistent with the driver's visual perception, and salient in the her/his hierarchical cognition along the road network.

2.2. Identification of spatial interaction patterns by topic modelling

Topic models in Natural Language Processing (NLP) domain are very popular and widely used in extracting implicit semantic relations of words from corpus. Therefore, in this section, we take advantage of the model to extract the associations of urban roads (i.e., identify the spatial interaction patterns) from massive vehicle moving paths.

2.2.1. Topic models and LDA

Topic models are unsupervised machine learning models, which are widely used in NLP for discovering the abstract "topics" and hidden semantic structures from vast textual documents. Intuitively, given that a document is about a particular topic, for instance, football, then some particular words such as "football", "team" and "league" would co-occur in the document more frequently. Topic models can automatically analyze the documents in the corpus and extract potential topics according to the co-occurrence of words in documents. As the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model is the most popular and classical one in the topic model family, we will utilize it in our study.

LDA introduces sparse Dirichlet prior distributions over document-topic and topic-word distributions, encoding the intuition that each document covers a small number of topics and each topic can be described by a small number of semantically related words. The graphical representation of latent Dirichlet allocation is shown in Fig. 3. The latent variables in the generative process are represented as single-circle nodes, while the observed variable is represented as a double-circle node. The rectangles are “plate” notation that denotes replication. K is the number of topics; N is the total number of words in the vocabulary; M is the number of documents in the corpus. $\vec{\alpha}$ and $\vec{\beta}$ are the prior parameters for the Dirichlet document-topic and topic-word distribution, respectively. $\vec{\vartheta}_m$ represents the topic distribution for document m, which is a K-dimension vector; while $\vec{\varphi}_k$ represents the word distribution for topic k, which is a N-dimension vector. $z_{m,n}$ is the topic of the n-th word in document m, while $w_{m,n}$ represents the specific word.

As shown in Fig. 3, the generative process of the LDA model can be described as follows.

---

for topic $k \in [1, K]$:
  draw $\vec{\varphi}_k \sim \text{Dirichlet}(\vec{\beta})$
for document $m \in [1, M]$:
  draw $\vec{\vartheta}_m \sim \text{Dirichlet}(\vec{\alpha})$
  for word $n \in [1, N_m]$:
    draw the topic of the n^{th} word: $z_{m,n} \sim \text{Multinomial}(\vec{\vartheta}_m)$
    draw the n^{th} word $w_{m,n} \sim \text{Multinomial}(\vec{\varphi}_{z_{m,n}})$

The variational Bayes (VB) algorithm is used in our study for parameter estimation, and more details along with the parameter initialization can be found in Hoffman, Bach, & Blei, 2010. After finishing

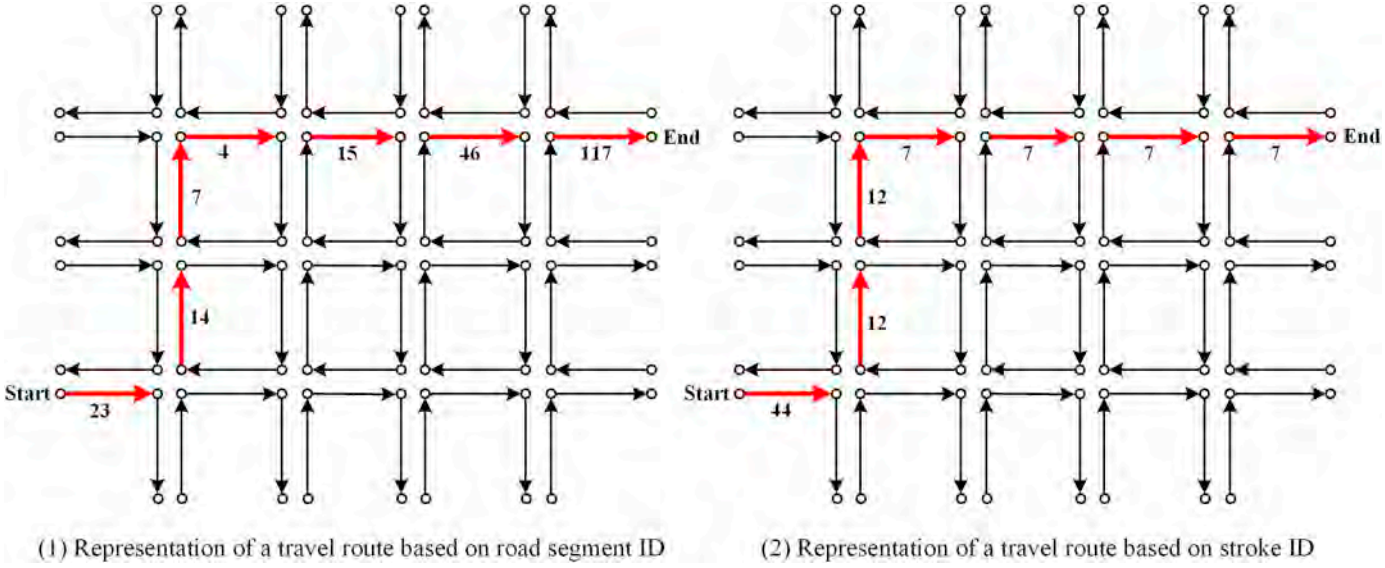

Fig. 2. Representation of a moving path.

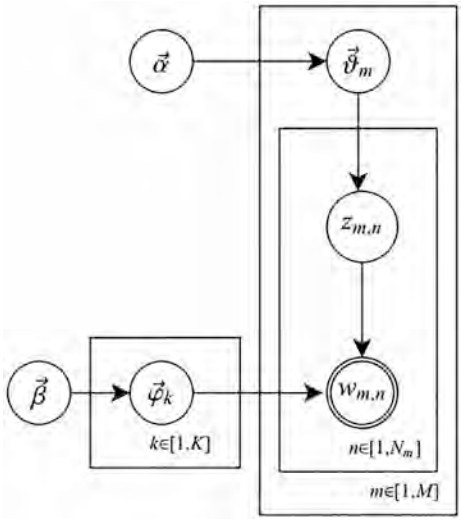

Fig. 3. The graphical representation of latent Dirichlet allocation (Blei et al., 2003).

the inference, $\vec{\alpha}$ and $\vec{\beta}$ associated with topic proportions and assignments are generated. Then each topic can be represented as a *N*-dimension vector, and words with high probabilities are semantically related or similar, which can well describe the topic; while each document can be represented as a *K*-dimension vector, and topics with high probabilities indicate the themes that the document talks about.

2.2.2. Combination of topic model with massive moving paths

Taking the advantage of topic models in extracting implicit relations of words, we propose to apply the LDA model to massive vehicle moving paths - by regarding each stroke as a word and each moving path as a document - in order to extract the interaction relations of strokes and identify the spatial interaction patterns on road network.

Here we explain why the combination of LDA model and moving paths is meaningful. According to the principle of LDA model, words with high probabilities in the topic-word distribution are semantically related or similar with each other (e.g., "pop" and "rock"), as they co-occur frequently in a considerable number of documents. Similarly, strokes with high probabilities in the topic-stroke distribution (an example is shown in Table 1) co-occur frequently in a considerable number of moving paths, which means that in the physical world, massive vehicles passed through these strokes from one (directly or indirectly) to another, making these strokes spatially interacted with each other. Therefore, we can extract the interaction relations of strokes and identify the spatial interaction patterns on road network by the probabilities of strokes to each topic.

Table 1
Example of topic-stroke distribution.

	stroke ID								
	28	51	211	79	466	267	1217	358	...
Topic 1	0.03	0.01	0.03	**0.14**	0.03	**0.39**	0.01	0.01	...
Topic 2	**0.25**	0.02	**0.24**	0.01	**0.15**	0.05	0.01	0.01	...
Topic 3	0.01	**0.39**	0.01	0.01	**0.17**	0.01	**0.12**	0.01	...
...	...	...	...	...	...	...	...	...	...

3. Case study

3.1. Data

We conduct a case study using the data of downtown Beijing, China.

The road network of the city contains 26,621 road segments, including expressways, arterial streets, side streets and collector streets,

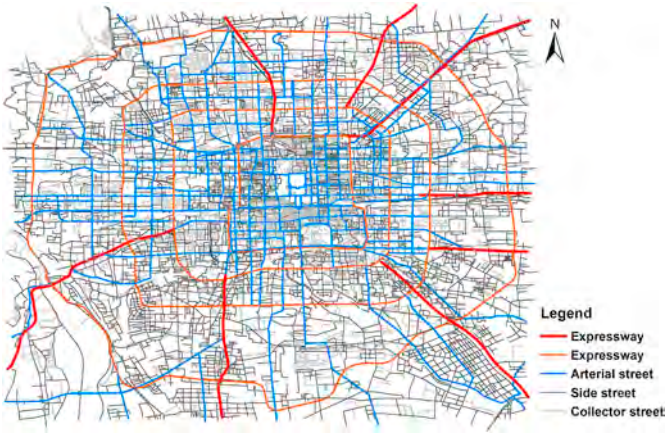

Fig. 4. The road network of downtown Beijing.

Table 2
Data schema of taxi GNSS records.

Taxi identification	Data	Time	Longitude	Latitude	Velocity (km/h)	Occupancy state
300,984	20,120,503	09:25:45	116.392	39.929	29	0
300,984	20,120,503	09:25:53	116.381	39.930	46	1
…	…	…	…	…	…	…

as shown in Fig. 4.

As strokes are chosen as geographical units in our study, we generate strokes from the road segments by the restriction of both their deflection angles and road names, which are two main elements used for generating strokes according to the continuity principle of perceptual grouping into networks (Yang et al., 2011). Finally, we derive 2762 strokes from the 26,621 road segments.

The moving paths are derived from the floating car data (FCD) collected by the GNSS (Global Navigation Satellite System) equipped taxis operating in Beijing during May 2012, with a sampling frequency of about 50 s. As shown in Table 2, a GNSS record consists of taxi's identification, coordinates (longitude and latitude), timestamp, velocity, and occupancy state (i.e., loading passengers or not), etc. As the taxi drivers are experienced drivers who are familiar with the city, they can be regarded as the samples of one type of local citizens living in the city.

To generate the moving paths, we first map the GNSS trajectories between passengers' pick-up and drop-off locations to the road network using a map matching algorithm called ST-CRF, which can process low-frequency (i.e., sparse) FCD (Liu, Liu, Li, & Lu, 2017). Then we proceed as follows:

- if more than one consecutive GNSS points are mapped into the same road segment, then the road segment is only counted once in the moving path;
- if two consecutive GNSS points are mapped onto different road segments that are not topologically adjacent in a road network, then we use the shortest path between the two road segments to fill in the moving path.

After the above processing, we obtain moving paths represented by sequential road segments, and then transform them into the format represented by sequential strokes introduced in Section 2.1.

By analyzing the frequencies of strokes in the moving-path datasets, we find that the rank-frequency distribution shows a long tail (Fig. 5), indicating that a minority of strokes are used highly frequently, while a majority of other strokes are much less used. This statistic is consistent with that of words in natural language texts, where a small number of words are used highly frequently while most other words are only used once a while (Manning & Schütze, 1999). This indicates that the geographical units, i.e., strokes, can hold the underlying statistical assumption in NLP, proving the rationality of representing vehicle moving paths by strokes, and applying the topic model to moving-path datasets.

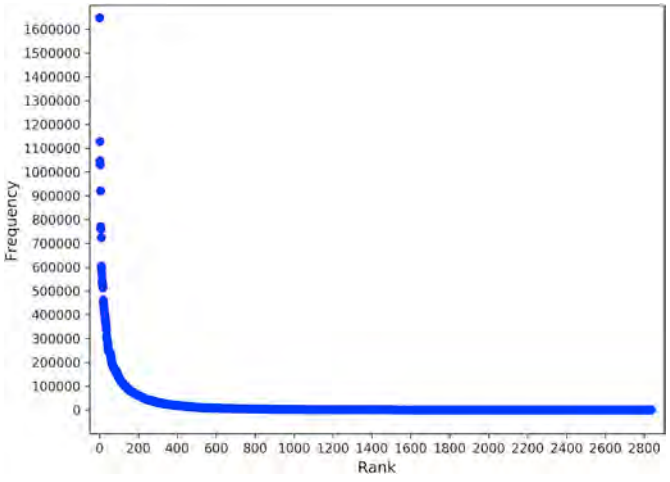

Fig. 5. Rank-frequency of strokes in the moving path datasets.

Furthermore, considering the dynamic characteristic of human moving behavior, we divide the moving paths into five datasets according to the departure time of the routes: workday morning rush hours (7:00–9:00), workday evening rush hours (17:00–19:00), workday non-rush hours (0:00–7:00; 9:00–17:00; 19:00–24:00), weekends and holidays.

3.2. Topic number selection

Setting an appropriate number of topics is important but still a challenging problem in LDA modelling. Several metrics have been developed to find optimal topic numbers, including perplexity (Blei et al., 2003), $\log P(w|T)$ (Griffiths & Steyvers, 2004), topic coherence (Mimno, Wallach, Talley, Leenders, & McCallum, 2011), topic correlation (Cao, Xia, Li, Zhang, & Tang, 2009), etc. In this paper, we use both the topic correlation and topic coherence to assess topic quality and determine an appropriate topic number.

(1) Topic correlation

Cao et al. (2009) studied the connection between LDA performance and topic correlation, and demonstrated that the LDA model performs best when the average cosine similarity of topics reaches the minimum. The calculating process of topic correlation is as follows.

Mark the topic-word distribution of topic i in N-dimension space as vector z_i where N is the number of words in the vocabulary; while the value of its t^{th} dimension is z_i^t. The correlation between topic i and j is measured by cosine similarity:

$$cor(i,j) = \frac{\sum_{t=1}^{V} z_i^t z_j^t}{\sqrt{\sum_{t=1}^{V} (z_i^t)^2}\sqrt{\sum_{t=1}^{V} (z_j^t)^2}} \quad (1)$$

where smaller value indicates that topic i and j are more independent.

The topic correlation of K topics is calculated as:

$$cor = \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} corr(z_i, z_j)}{K \times \frac{K-1}{2}} \quad (2)$$

where a smaller value indicates a better performance of the topic model.

(2) Topic coherence

Topic coherence is a metric based on the co-occurrence of words that word pairs belonging to the same topic should co-occur more frequently in documents.

Letting $D(v)$ be the document frequency of word v (i.e., the number of documents containing at least one v) and $D(v,v')$ be co-document frequency of word v and v' (i.e., the number of documents containing at least one v and at least one v'), the topic coherence of topic t is defined

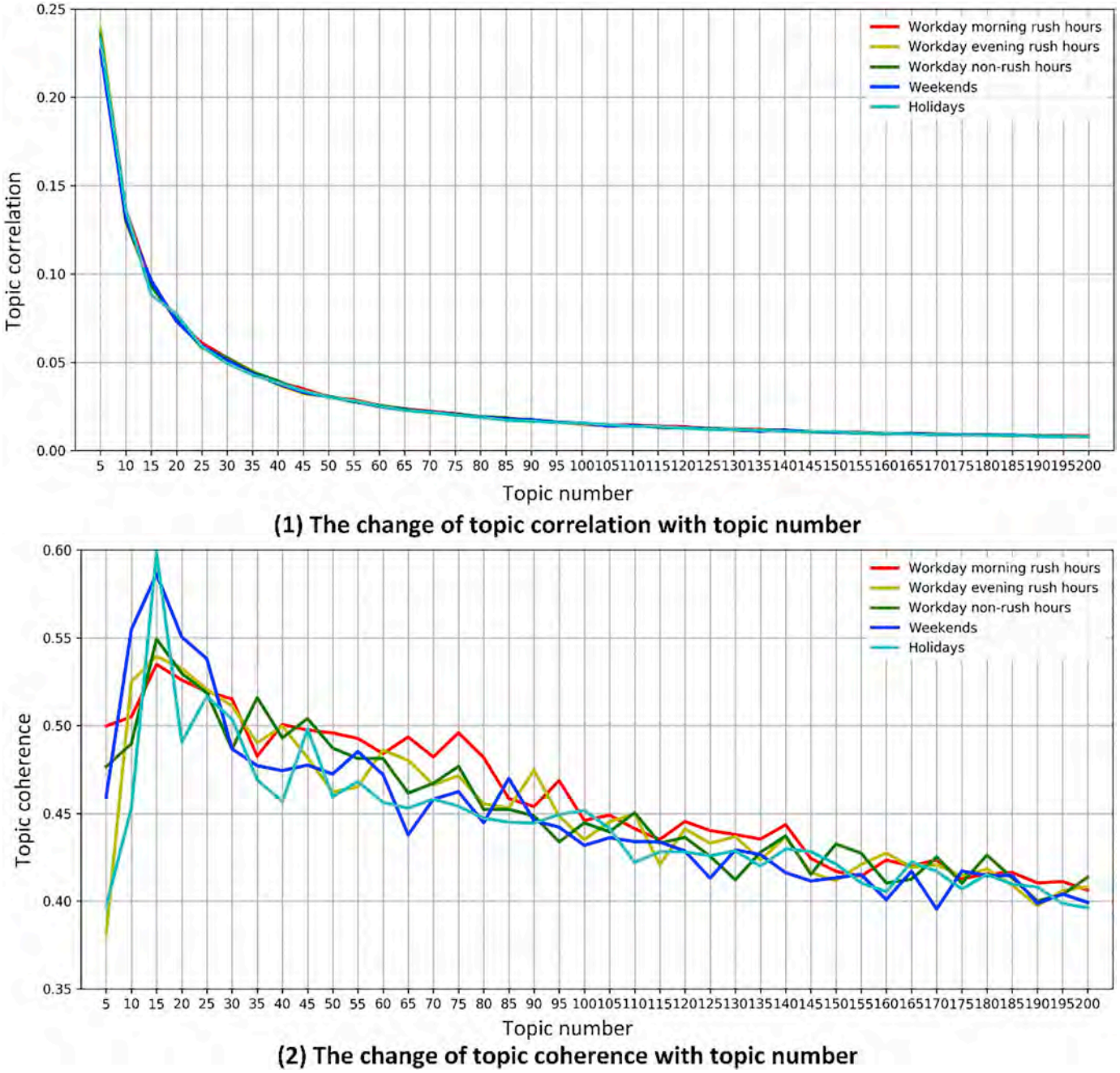

Fig. 6. Selection of appropriate topic number using topic correlation and topic coherence metrics.

as:

$$coh(t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \quad (3)$$

where $(v_1^{(t)}, \ldots, v_M^{(t)})$ is a list of the M most probable words in topic t. A smoothing count of 1 is included to avoid taking the logarithm of zero.

The topic coherence of K topics is calculated as:

$$coh = \frac{\sum_{t=1}^{K} coh(t)}{K} \quad (4)$$

where a greater value indicates a higher quality of topics.

In order to choose an appropriate topic number in our case study, we run the LDA model on our moving path datasets by setting the topic number from 5 to 200 with an interval of 5, and calculate the corresponding topic coherence and topic correlation metrics. As shown in Fig. 6, with the increasing of topic number, the topic correlations decrease first, and then keep steady at a low level; while the topic coherences increase first until reaching the peak at the topic number of 15, and then decrease slowly. Considering both the two metrics, we choose 15 as the optimal topic number for our datasets, even though the topic correlations not yet drop to a steady and low level - the topics do not need to be strictly independent with each other, i.e., some overlapping is allowed and one stroke may belong to more than one topic.

3.3. Topic pattern extraction

The LDA model is a statistical model, and the results run in different rounds are usually slightly different, even if apparent topic patterns did exist in the dataset. As only stable and reliable topics are expected, in this section, we use the following steps to identify significant topic patterns in each of the five datasets of different time periods.

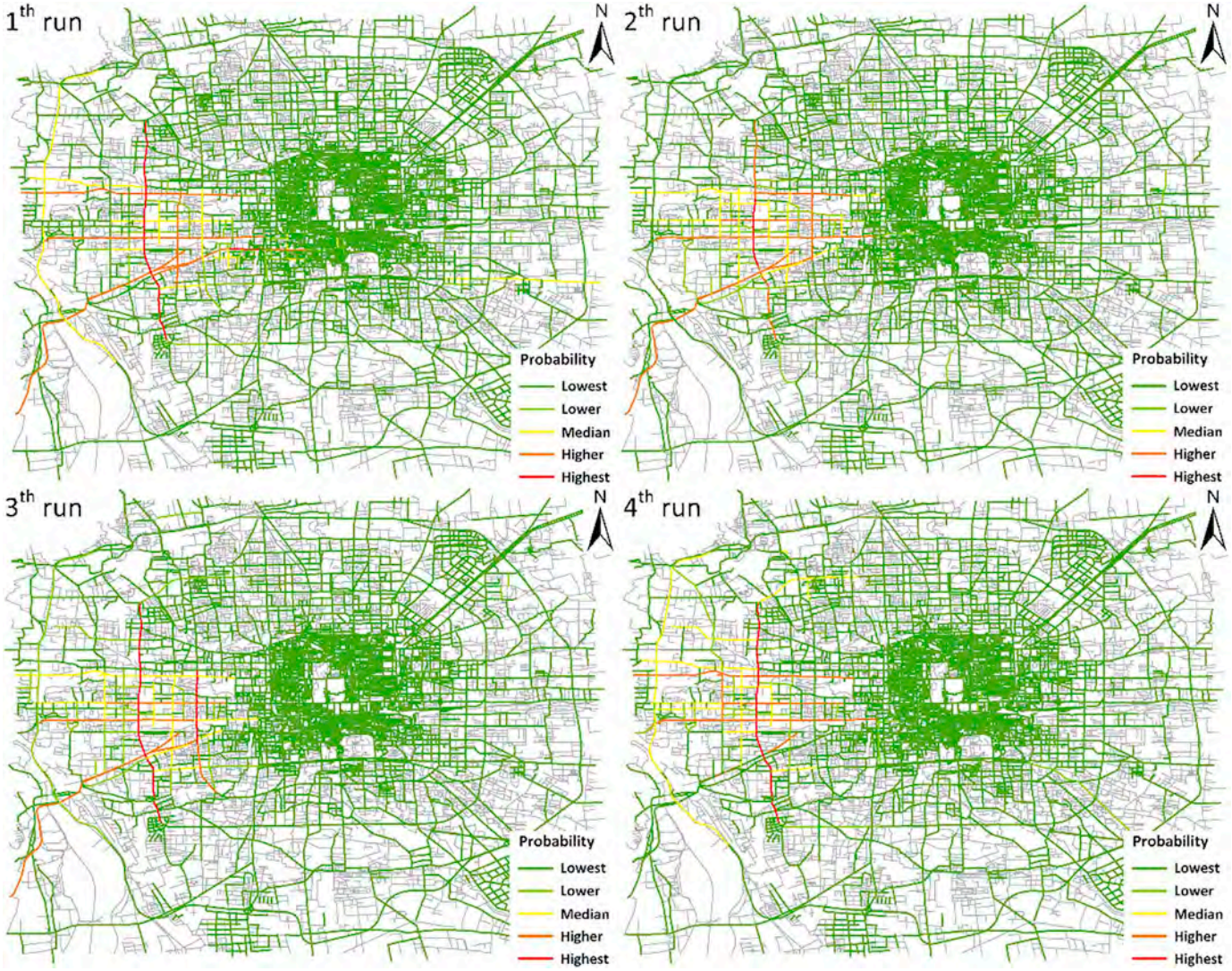

Fig. 7. Topics of the same pattern run in four different rounds.

Table 3
Cosine similarities of the topic-stroke distributions run in four different rounds.

	1	2	3	4
1	1			
2	0.889	1		
3	0.829	0.809	1	
4	0.893	0.809	0.912	1

(1) Run the LDA model five times;
(2) For each topic derived from each running round:

- Join the topic-stroke distribution with the shapefile of road network in ArcMap software through their common attribute - stroke ID;
- Classify the strokes into five classes by their probabilities using the Jenks natural breaks classification method, which seeks to reduce the variance within classes and maximize the variance between classes (Jenks, 1967);
- Render the five clusters of strokes with red, orange, yellow, light green, and green, reflecting relatively highest, higher, median, lower and lowest probabilities to the topic respectively.

(3) Based on the visualized maps, we manually select topics that appear more than three of the five times as the significant topic patterns.

Fig. 7 shows one topic pattern that appears in four of the five running rounds in the dataset of workday evening rush hours; while Table 3 shows the cosine similarities of topic-stroke distributions run in the four rounds. It shows that the topics run in the four different rounds are very similar, reflected in both the visualized maps and the similarity measurements, indicating that the topic pattern is meaningful and significant.

3.4. Result analysis

Using the topic pattern extraction method introduced above, we identify the topic patterns from the datasets of workday morning rush hours, workday evening rush hours, workday non-rush hours, weekends and holidays, respectively, and find that the majority of the topic patterns in different time periods are almost the same, so we focus on these common patterns only. The topic number is initially set as 15, and we finally obtain 11 topic patterns after topic pattern extraction.

From Table 3 and Fig. 7, we can see that after topic pattern extraction, the same patterns derived from different runs of the algorithm are very similar, so we just pick one of them for demonstration. Table 4 shows the cumulative numbers of strokes at different probability levels in each pattern, and Fig. 8 shows the visualized maps of the patterns, which distribute at various important positions across the road network. Polylines rendered by red, orange, yellow, and light green in

Table 4
Cumulative numbers of strokes at different probability levels in each topic pattern.

Prabability level	Topic pattern										
	1	2	3	4	5	6	7	8	9	10	11
Highest	5	3	5	4	3	6	3	11	2	2	10
Higher	7	10	15	11	12	20	11	27	7	7	23
Median	17	25	29	31	24	44	31	67	22	14	39
Lower	51	52	58	63	52	85	80	136	47	36	73
Lowest	2762	2762	2762	2762	2762	2762	2762	2762	2762	2762	2762

Fig. 8 can be regarded as clusters of strongly spatially interacted strokes, among which a mass of vehicles have passed through frequently.

- Pattern 1 mainly includes the North 4th Ring freeway, the North 5th Ring freeway, and some important roads connecting them (e.g., the Beijing-Tibet freeway) or surrounding them in the north of the city.
- Pattern 2 mainly includes some high-grade roads around Wangjing (a relatively independent commercial and residential sub-district), such as Beijing-Chengde freeway, the East of the North 4th Ring freeway, and the Beijing Capital Airport Expressway, and some main roads inside Wangjing.
- Pattern 3 mainly includes the West 4th Ring freeway and some main roads surrounding and crossing it in the west of the city.
- Pattern 4 mainly includes the West 3th Ring freeway, and some roads surrounding and crossing it, such as the Zizhuyuan Road.
- Pattern 5 mainly includes the North 3th Ring freeway and the surrounding main roads.
- Pattern 6 mainly includes some main roads jointed by Xizhimen flyover, such as the West 2th Ring freeway, the North 2th Ring freeway, etc.
- Pattern 7 mainly includes the main roads jointed by Dongzhimen flyover, such as the Capital Airport Expressway, the North 2th Ring freeway, the East 2th Ring freeway, etc.
- Pattern 8 mainly includes some main roads surrounding the east-south of the 2th Ring freeway.
- Pattern 9 mainly includes the East 3th Ring freeway and some main roads surrounding or crossing it.
- Pattern 10 mainly includes the East 4th Ring freeway, the East 5th Ring freeway, and some main roads crossing it.
- Pattern 11 mainly includes the South 3th Ring freeway, the South 4th Ring freeway and some main roads surrounding and crossing them.

The identified topic patterns are almost the same in different time periods, indicating that the spatial interactions do not change a lot dynamically even though the movement flows and traffic states usually vary dramatically over time. This may be because stroke is chosen as the representation granularity of roads, which is a relatively larger geographical unit compared to road segment. As is explored in our previous work (Liu, Gao, et al., 2017), the traffic interactions among neighboring road segments show prominent dynamic characteristics that the interactions on workday morning rush hours, workday evening rush hours and holidays are stronger than that on workday non-rush hours and weekends.

3.5. Result verification

To verify the effectiveness of our proposed approach, we use the community detection-based method as a comparison, which is widely used in recent years to discover the hidden spatial interaction patterns (Gao et al., 2013; Hawelka et al., 2014; Liu et al., 2014; Liu et al., 2015). With the areal units being represented as nodes, and the interacting flows being represented as weighted edges, the community detection algorithms can divide the spatially embedded network into sub-regions, within which the nodes (i.e., areal units) are interacted with each other more strongly.

In our case, firstly, the strokes are mapped to nodes and the connections between strokes are mapped to edges; while the weight of each edge is determined by the frequency of vehicles moving between the two ending nodes (i.e., strokes). Secondly, we apply a classical community detection algorithm proposed by Clauset, Newman, and Moore (2004) to the weighted stroke-stroke network. This algorithm merges individual nodes into communities one by one in a way that greedily maximizes the modularity score of the network. The algorithm can be stopped until no merge can increase the current modularity score. The modularity Q is defined as:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v, c_w) \tag{5}$$

where m is the number of edges in the network; A_{vw} is 1 if node v and w are connected and 0 otherwise; k_v is the degree of node v which is defined to be the number of edges incident upon it; δ-function $\delta(i,j)$ is 1 if $i = j$ and 0 otherwise.

Fig. 9 shows the results of the community detection-based method. Fig. 9(1) depicts how Q changes with the number of communities, based on which, we choose eight as the optimal number of communities. While Fig. 9(2) shows the eight communities, i.e., the derived spatial interaction patterns. Strokes within the same communities are deemed to be interacted with each other more strongly.

In order to compare our proposed approach and the community detection-based method in a quantitative way, we calculate the average frequency of vehicle movements between strokes within the same topic patterns or communities. As shown in Table 5, the average transfer frequencies between strokes within the same topic patterns are significantly higher than that within the same communities, indicating that our proposed approach performs much better in identifying the spatial interaction patterns. This may be because in our approach, the spatial interaction patterns are discovered from the moving paths directly based on the co-occurrences of the strokes. While in the community detection-based method, the intrinsic structure of the road network would affect the results greatly, which may make some weakly interacted but well-connected strokes being identified into the same communities. Besides, compared to the community detection-based method, our approach can discover the spatial interaction patterns in a hierarchical way, as the topic model can draw topic-stroke distributions instead of the fixed memberships of strokes to topics.

In addition, to help understand the derived spatial interaction patterns, we utilize the chord diagram to visualize the actual vehicle moving flows among strokes in the 11 topic patterns. A chord diagram is a graphical method of displaying the inter-relationships and flows between several entities (e.g., topic patterns in our case). As shown in Fig. 10, when we only consider the strokes at highest probability level in each topic pattern (i.e., the red strokes in Fig. 8), the majority of vehicle movements among strokes happen inside the same topic patterns with average percentage of intra-flows of 60.02%, proving again that our proposed approach can well identify the spatial interaction patterns.

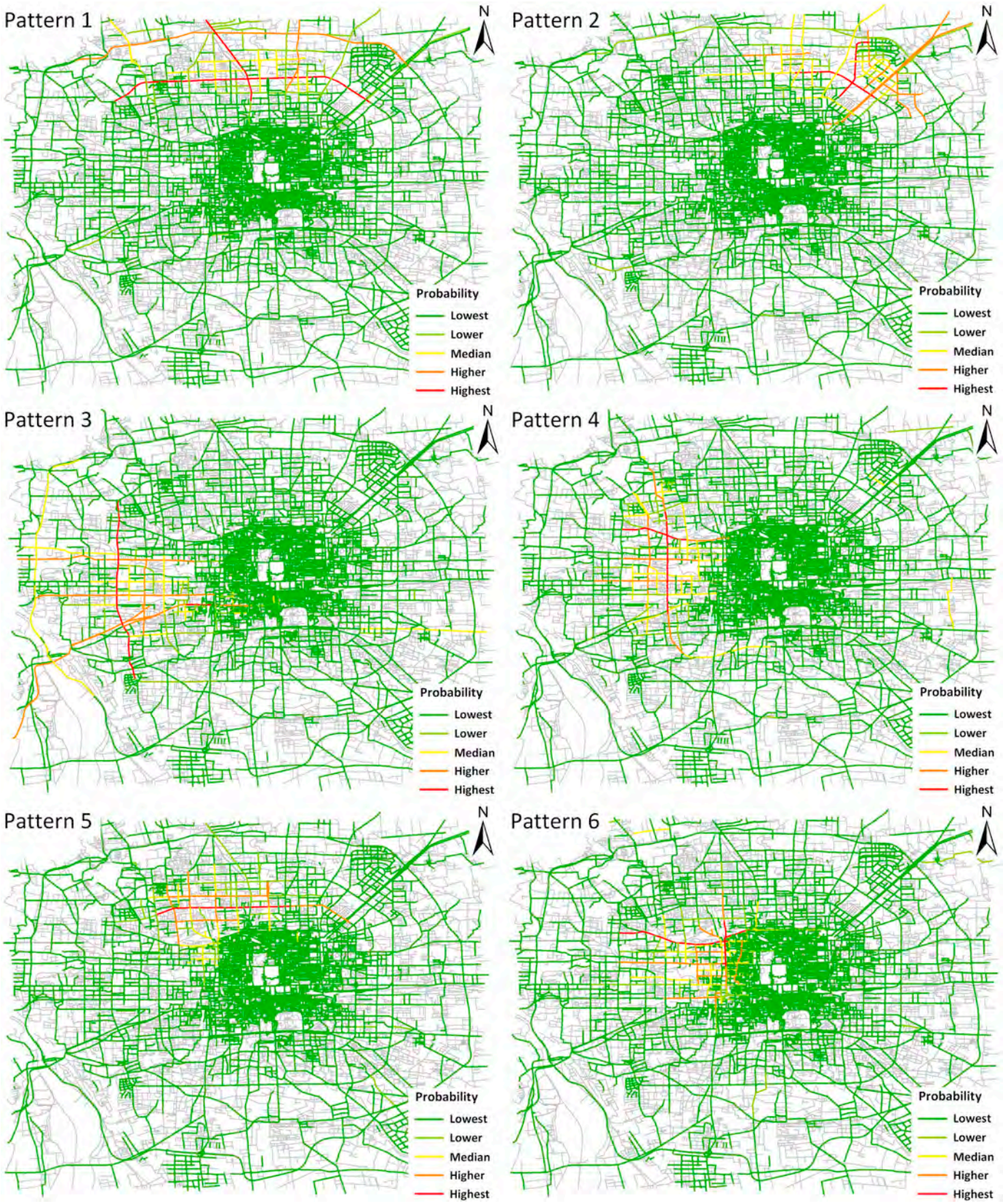

Fig. 8. Common spatial interaction patterns in different time periods.

4. Discussion

This paper investigated the spatial interactions of vehicle movements on urban road network. The innovation points of this paper can be summarized as follows.

Firstly, as the existing literatures mainly focused on the spatial interactions among areal units, neglecting that most of human movements are constrained by the underlying road network, in this paper,

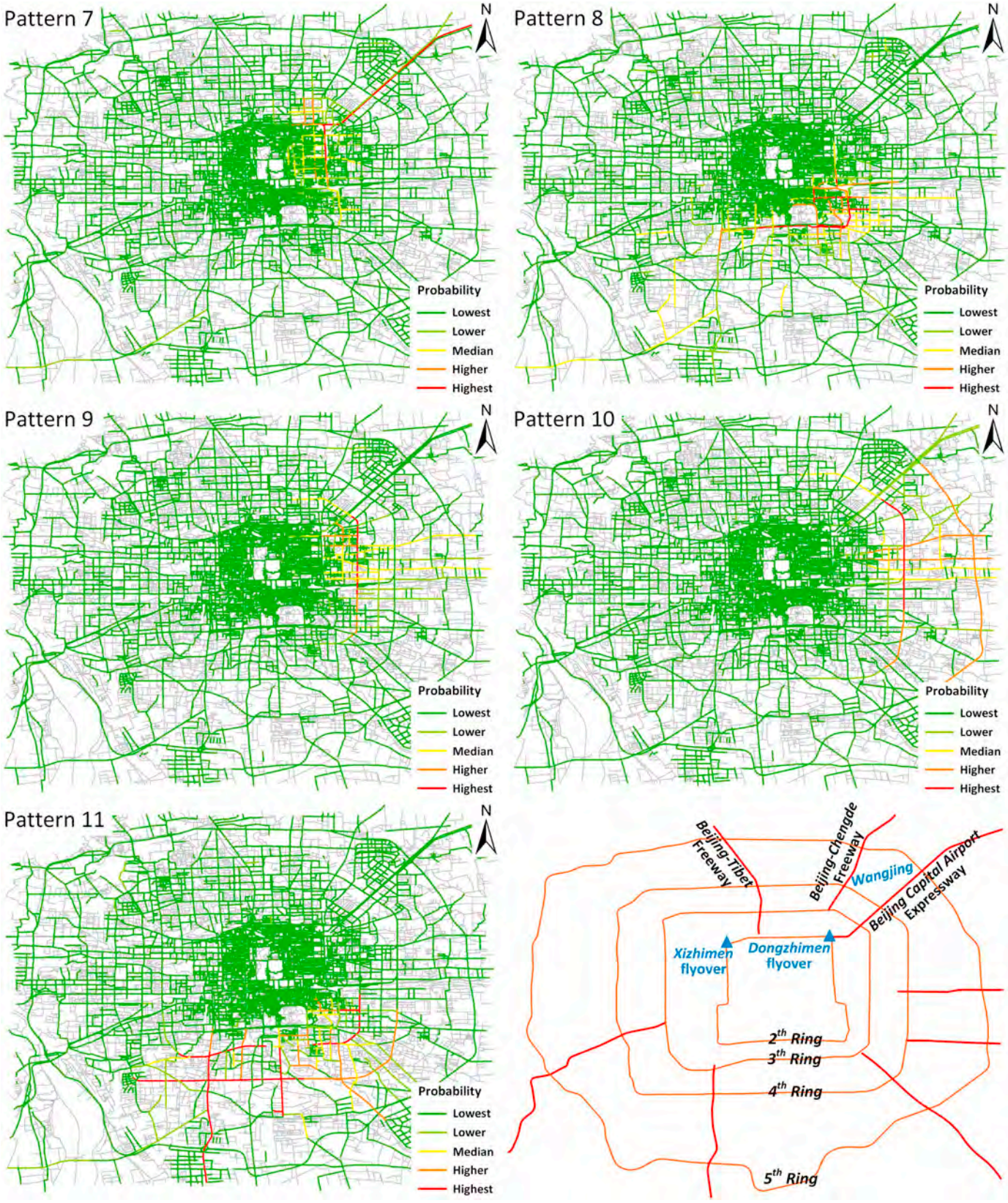

Fig. 8. (*continued*)

we investigated the detailed vehicle moving process on road network and identified the spatial interaction patterns at linear instead of areal geographical units, which filled the gap of current studies.

Secondly, in order to better identify the spatial interaction patterns, we revisited human spatial cognition on road network space first. Based on which, we chose "stroke" as geographical unit to express road, and proposed an innovative method to represent vehicle moving paths, which can well reflect human driving behavior.

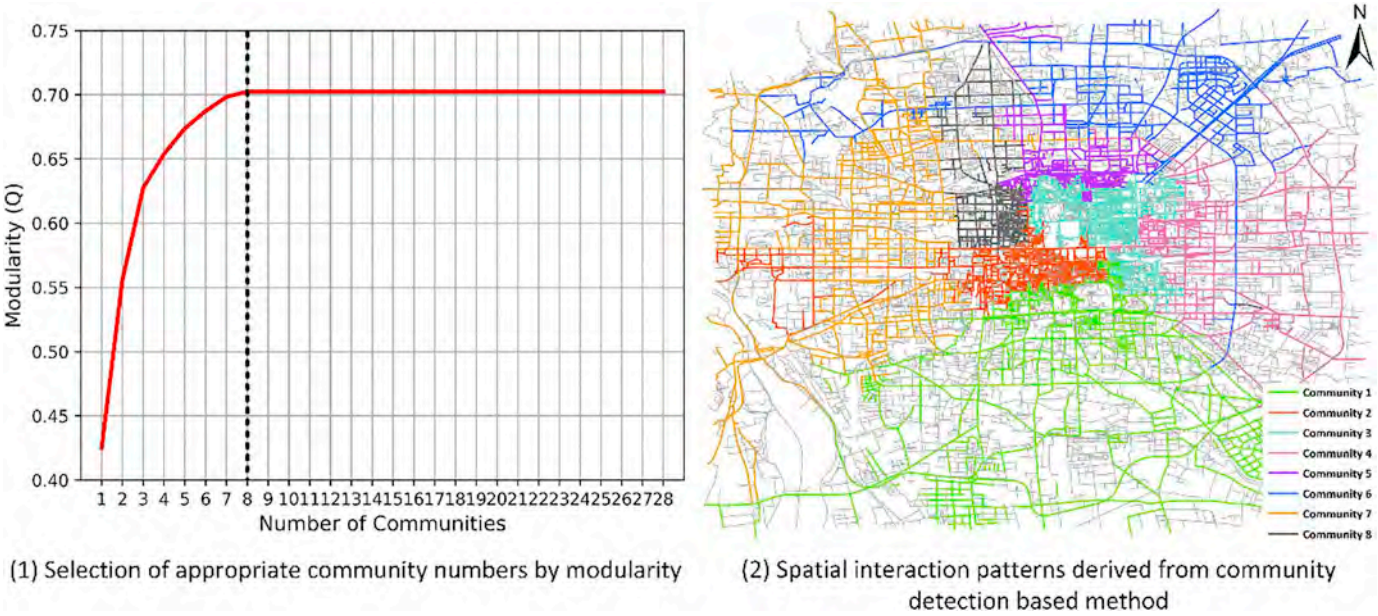

Fig. 9. Results of the community detection based method as a comparison.

Table 5
Average frequency of vehicle movements between strokes within the same topic patterns or communities.

	Topic patterns				Communities
	Highest	≥Higher	≥Median	≥Lower	
Frequency	29,681.7	11,017.1	4484.5	1783.5	72.5

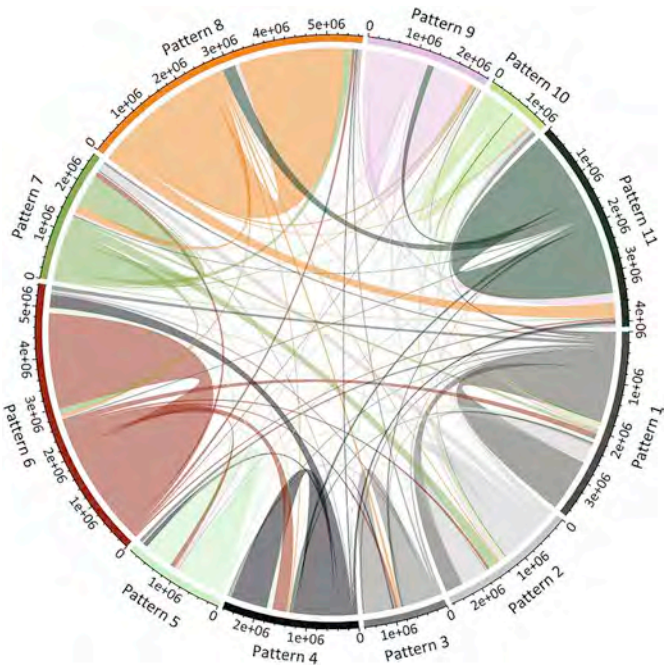

Fig. 10. Vehicle moving flows among strokes with highest probability level in the 11 spatial interaction patterns.

Thirdly, we grasped the common nature of "moving path-stroke" and "document-word", and took the advantage of topic model in extracting semantic relations of words to extract the interaction relations of strokes by applying a topic model to a mass of moving paths.

Our identified spatial interaction patterns are proved effective that vehicles do move much more frequently among strokes within the same topic patterns than that within the same communities derived by the comparison method. Actually, as the topics are extracted according to the co-occurrences of strokes in moving paths, the derived topic patterns can also be interpreted from other perspectives as follows.

(1) They can reveal the hotspots of vehicle movements and human driving behavior in the city.
(2) They can reflect the traffic patterns of the city, as spatially interacted strokes would also influence each other more easily in traffic states, such as volumes and velocities.
(3) They can reflect the underlying urban structures of the city in the road network space, providing a new viewpoint for urban planning.

5. Conclusion

Most spatial interaction studies have been focused on areal spatial units such as cities and administrative districts, neglecting that human movements are usually constricted by the underlying road network, causing spatial interactions among the linear geographical units, i.e., roads. In order to fill the gap of current studies, this paper proposed an innovative approach to identify the spatial interaction patterns of vehicle movements on road network. Firstly, "strokes" are chosen as geographical units to represent the vehicle moving paths. Then, an analogy between stroke-moving path and word-document is established, and a topic model is applied to the moving paths to identify the spatial interaction patterns on road network. It indicates that our extracted topics can better reflect the spatial interaction patterns compared to the community detection based method. Our proposed approach provides an innovative viewpoint for spatial interaction modelling on road network space, and enriches the current spatial interaction studies, which can help the traffic managers and urban planners better understand the traffic patterns and urban structures of the city, and plan an efficient traffic control or spatial configuration for the city. In future work, we would like to extend our proposed method in other cities with different underlying road network structures and urban environment.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant Nos. 41631177, 41601421, 41701167), the National Key Research and Development Program (Grant No. 2016YFB0502104). Their supports are gratefully acknowledged. We also thank the anonymous referees for their helpful comments and suggestions.

References

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences, 106*(51), 21484–21489.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., ... Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports, 734*, 1–74.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7–9), 1775–1781.

Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., & Chen, G. (2014). Visualizing hidden themes of taxi movement with semantic transformation. *Visualization Symposium (PacificVis), 2014 IEEE Pacific* (pp. 137–144). .

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111.

Conroy, R. A. (2001). *Spatial navigation in immersive virtual environments.* University of London.

Couclelis, H., Golledge, R. G., Gale, N., & Tobler, W. (1987). Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology, 7*(2), 99–122.

Dalton, R. C. (2003). The secret is to follow your nose: Route path selection and angularity. *Environment and Behavior, 35*(1), 107–131.

Fotheringham, A. S., & O'Kelly, M. E. (1989). *Spatial interaction models: Formulations and applications.* Dordrecht: Kluwer Academic Publishers.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS, 21*(3), 446–467.

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

Golledge, R. G. (1995). Path selection and route preference in human navigation: A progress report. *International Conference on Spatial Information Theory* (pp. 207–222). Berlin, Heidelberg: Springer.

Golledge, R. G., & Spector, A. N. (1978). Comprehending the urban environment: Theory and practice. *Geographical Analysis, 10*(4), 403–426.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228–5235.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271.

Hillier, B., & Iida, S. (2005). Network and psychological effects in urban movement. *International Conference on Spatial Information Theory* (pp. 475–490). Berlin, Heidelberg: Springer.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems. Vol. 23. Advances in Neural Information Processing Systems* (pp. 856–864). (NIPS 2010).

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography, 7*, 186–190.

Jiang, B. (2007). A topological pattern of urban street networks: Universality and peculiarity. *Physica A: Statistical Mechanics and its Applications, 384*(2), 647–655.

Jiang, B. (2009). Street hierarchies: A minority of streets account for a majority of traffic flow. *International Journal of Geographical Information Science, 23*(8), 1033–1048.

Jiang, B., & Jia, T. (2011). Agent-based simulation of human movement shaped by the underlying street structure. *International Journal of Geographical Information Science, 25*(1), 51–64.

Jiang, B., & Liu, C. (2009). Street-based topological representations and analyses for predicting traffic flow in GIS. *International Journal of Geographical Information Science, 23*(9), 1119–1137.

Jiang, B., Zhao, S., & Yin, J. (2008). Self-organized natural roads for predicting traffic flow: A sensitivity study. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(07), P07008.

Lai, J., Cheng, T., & Lansley, G. (2017). Improved targeted outdoor advertising based on geotagged social media data. *Annals of GIS, 23*(4), 237–250.

Lenormand, M., Huet, S., Gargiulo, F., & Deffuant, G. (2012). A universal model of commuting networks. *PLoS One, 7*(10), e45985.

Liu, K., Gao, S., Qiu, P., Liu, X., Yan, B., & Lu, F. (2017). Road2Vec: Measuring Traffic Interactions in Urban Road System from Massive Travel Routes. *ISPRS International Journal of Geo-Information, 6*(11), 321.

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography, 43*, 78–90.

Liu, X., Liu, K., Li, M., & Lu, F. (2017). A ST-CRF map-matching method for low-frequency floating car data. *IEEE Transactions on Intelligent Transportation Systems, 18*(5), 1241–1254.

Liu, Y., & Cheng, T. (2018). Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science,* 1–28.

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS One, 9*(1), e86026.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). .

Mohamed, K., Côme, E., Oukhellou, L., & Verleysen, M. (2017). Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems, 18*(3), 712–728.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS One, 7*(5), e37027.

Passini, R. (1984). Spatial representations, a wayfinding perspective. *Journal of Environmental Psychology, 4*(2), 153–164.

Peng, P., Cheng, S., Chen, J., Liao, M., Wu, L., Liu, X., & Lu, F. (2018). A fine-grained perspective on the robustness of global cargo ship transportation networks. *Journal of Geographical Sciences, 28*(7), 881–889.

Simini, F., González, M. C., Maritan, A., & Barabási, A. L. (2012). A universal model for mobility and migration patterns. *Nature, 484*(7392), 96.

Stouffer, S. A. (1940). Intervening opportunities: A theory relating mobility and distance. *American Sociological Review, 5*(6), 845–867.

Thomson, R. C. (2004). Bending the axial line: Smoothly continuous road centre-line segments as a basis for road network analysis. *Proceedings 4th International Space Syntax Symposium, London, UK*.

Thomson, R. C., & Richardson, D. E. (1999). The 'Good Continuation' principle of perceptual organization applied to the generalization of road networks. *Proceedings of 19th International Cartographic Conference* (pp. 1215–1223). .

Tizzoni, M., Bajardi, P., Decuyper, A., King, G. K. K., Schneider, C. M., Blondel, V., ... Colizza, V. (2014). On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology, 10*(7), e1003716.

Wan, L., Gao, S., Wu, C., Jin, Y., Mao, M., & Yang, L. (2018). Big data and urban system model-Substitutes or complements? A case study of modelling commuting patterns in Beijing. *Computers, Environment and Urban Systems, 68*, 64–77.

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

Yan, X. Y., Zhao, C., Fan, Y., Di, Z., & Wang, W. X. (2014). Universal predictability of mobility patterns in cities. *Journal of the Royal Society Interface, 11*(100), 20140834.

Yang, B., Luan, X., & Li, Q. (2011). Generating hierarchical strokes from urban street networks based on spatial pattern recognition. *International Journal of Geographical Information Science, 25*(12), 2025–2050.

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 186–194). .

Zhang, F., Zhu, X., Guo, W., Ye, X., Hu, T., & Huang, L. (2016). Analyzing urban human mobility patterns through a thematic model at a finer scale. *ISPRS International Journal of Geo-Information, 5*(6), 78.

Zhu, D., Huang, Z., Shi, L., Wu, L., & Liu, Y. (2018). Inferring spatial interaction patterns from sequential snapshots of spatial distributions. *International Journal of Geographical Information Science, 32*(4), 783–805.

Zhu, D., Wang, N., Wu, L., & Liu, Y. (2017). Street as a big geo-data assembly and analysis unit in urban studies: A case study using Beijing taxi data. *Applied Geography, 86*, 152–164.

Zipf, G. K. (1946). The P1 P2/D hypothesis: On the intercity movement of persons. *American Sociological Review, 11*(6), 677–686.