International Journal of
Geo-Information

MDPI

Article

Analyzing Urban Human Mobility Patterns through a Thematic Model at a Finer Scale

Faming Zhang [1], Xinyan Zhu [1,2], Wei Guo [1,2,*], Xinyue Ye [3,*], Tao Hu [1,2] and Liang Huang [4]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; zhang.fa.ming@163.com (F.Z.); geozxy@263.net (X.Z.); taohu07@hotmail.com (T.H.)
[2] Collaborative Innovation Center of Geospatial Technology, 129 Luoyu Road, Wuhan 430079, China
[3] Department of Geography, Kent State University, Kent, OH 44240, USA
[4] School of Navigation, Wuhan University of Technology, Wuhan 430063, China; plaquemine@whu.edu.cn
* Correspondence: guowei-lmars@whu.edu.cn (W.G.); xye5@kent.edu (X.Y.);
Tel.: +86-18607119346 (W.G.); +1-419-494-7825 (X.Y.)

Academic Editors: Constantinos Antoniou and Wolfgang Kainz
Received: 7 April 2016; Accepted: 23 May 2016; Published: 1 June 2016

Abstract: Taxi trajectories reflect human mobility over a road network. Pick-up and drop-off locations in different time periods represent origins and destinations of trips, respectively, demonstrating the spatiotemporal characteristics of human behavior. Each trip can be viewed as a displacement in the random walk model, and the distribution of extracted trips shows a distance decay effect. To identify the spatial similarity of trips at a finer scale, this paper investigates the distribution of trips through topic modeling techniques. Firstly, trip origins and trip destinations were identified from raw GPS data. Then, different trips were given semantic information, *i.e.*, link identification numbers with a semantic enrichment process. Each taxi trajectory was composed of a series of trip destinations corresponding to the same taxi. Subsequently, each taxi trajectory was analogous to a document consisting of different words, and all taxi's trajectories could be regarded as document corpora, enabling a semantic analysis of massive trip destinations. Finally, we obtained different trip destination topics reflecting the spatial similarity and regional property of human mobility through LDA topic model training. The effectiveness of this approach was illustrated by a case study using a large dataset of taxi trajectories collected from 2 to 8 June 2014 in Wuhan, China.

Keywords: mobility; semantic enrichment; LDA topic model; finer scale; China

1. Introduction

Taxi trajectories are constituted by a series of locations that might reflect what and where activities occur and, therefore, are used to analyze human dynamics. Thus, an increased understanding of human dynamics has motivated many studies on transportation management and urban planning [1–5]. Taxi GPS location records, social media check-ins, public transportation smart card data and mobile phone data have offered great research opportunities because of the increasing use of LBS (Location-Based Service) [6–10]. Compared to traditional data from questionnaires or statistical yearbooks, the new data are much richer at the finer spatiotemporal scale [8,10–12].

Taxi trajectories have been widely studied in many fields, such as urban planning [13,14], land use modeling [15–17] and traffic flow prediction [15,18]. Such data are collected by taxis equipped with GPS (Global Positioning System) devices. Methods, such as random walk, random direction, random way point and an obstacle model [19,20], have been proposed. Most studies indicate that human mobility can be expressed using a Lévy flight or truncated Lévy flight model [2,21]. Liang *et al.* [22] studied the displacement distribution of human mobility according to the trajectory of taxicabs, while

Liu *et al.* [23,24] investigated taxi trajectory data by taking into account the influence of geographic heterogeneity and distance decay. Ballatore *et al.* [25] described a knowledge-based approach to quantify the semantic similarity of lexical definitions. Chu *et al.* [26] analyzed the hidden knowledge of massive taxi movement using street names, but ignored the spatial difference when a same-name road was too lengthy. Gong *et al.* [27] inferred the visit probability of POI (Point Of Interest) and uncovered travel patterns from raw taxi data lacking activity information in the context of temporal and spatial constraints.

Roads intersect with each other at crossroads. To express the spatial distribution of mobility patterns, we split road network at crosses, generating small road segments, called links. Each link was given a unique identification number. Most previous studies investigated the spatial distribution of trajectories based on regions and road names. However, fewer studies have focused on the finer scale of the links in a road network.

This paper investigates the spatial distribution of different trip destinations through topic modeling techniques [26,28], such as Latent Dirichlet Allocation (LDA) at the link level. Trips with different trajectories were extracted based on a trip extraction algorithm from raw GPS data. However, the location of taxi trajectories alone cannot explicitly express human mobility patterns. We preprocess raw trajectory data from taxicabs by adding semantic information before using LDA to give locations unique identifiers that we can understand. Thus, a link identification number was attached to corresponding geographic locations by a map-matching process [11,29–32]. Taxi trajectories were transformed into a series of semantic trajectories consisting of trip destinations with link identification numbers. Each taxi semantic trajectory is analogous to a document with series of link IDs, while the massive taxi semantic trajectories are grouped together as a corpus by link IDs. Links with similar link IDs were aggregated into topics expressing a common hidden meaning, using the LDA method. Subsequently, these topics were used to analyze the spatial similarity of mobility patterns and their relationship to daily activities. The article is outlined as follows. In Section 2, the detailed analysis about trips is provided. In Section 3, the semantic analysis about trips is given. In Section 4, we analyze the model of latent Dirichlet allocation and give the definition of significance. In Section 5, we describe our experiment, including trip topic extraction and visualization, trip topic and urban dynamics analysis and trip topic evolution analysis. A discussion of the results and some conclusions are outlined at the end.

2. Trip Extraction and Analysis

2.1. Trip Extraction

Wuhan is the largest city in Central China. Many taxi companies have their vehicles equipped with GPS receivers in order to monitor the operation of each taxi. Taxis equipped with GPS are called "floating" cars, which can monitor the running status of real-time traffic. With the help of GPS devices, historical trajectories of taxis can be recorded as a series of locations sampled at small periodic intervals. In this research, the dataset includes more than 2050 floating cars records from Wuhan. The data cover seven consecutive days of a week, from 2 to 8 June 2014, from Monday to Sunday. For each taxi, its longitude, latitude, time stamp, instantaneous velocity, azimuth angle and occupancy are automatically collected approximately every 40 s. Each taxi reports nearly 2160 GPS sample points every day. In fact, the amount of GPS records is slightly less, as GPS receivers are shut down by drivers or become disconnected. The accumulated observations create a very large dataset for research, averaging 2,485,000 records per day. Subsequently, a record of GPS information can be denoted by $p\,(t, id,\ lat, lon,\ v, h, s)$, among them t denoting the time instant corresponding to the current position of the taxi, id denoting the taxi identification number, lat and lon representing the position of the taxi, namely latitude and longitude, v representing the instant velocity of the taxi, h representing the driving direction of the taxi and s indicating service status of a taxi, vacant or occupied. Table 1 shows some continuous sampling points of a taxi trajectory used in this research. As depicted in the table, a taxi is occupied when the status equals one; otherwise, it is vacant.

Table 1. The taxi trajectory data.

Date	ID	Longitude	Latitude	Velocity	Heading	Status
2 June 2014 01:04:44	40416	114.27183	30.59821	6.284861	316.78	0
2 June 2014 01:04:49	40416	114.271583	30.598418	8.756667	329.19	0
2 June 2014 01:05:09	40416	114.27093	30.59895	5.190278	310.57	1
2 June 2014 01:05:41	40416	114.27055	30.600498	8.756667	51.46	1
2 June 2014 01:06:24	40416	114.273651	30.601985	12.759861	62.53	1

GPS information of the taxi not only reflects the running state of traffic, but also can be used to investigate human mobility patterns based on the taxi occupancy [10–12]. Therefore, each extracted trip could be simplified to be a vector, $< t_i, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}) >$, the term t_i denoting the time instant corresponding to the trip origin of the taxi, (x_{i1}, y_{i1}) representing the trip origin geographic coordinates of the taxi and (x_{i2}, y_{i2}) indicating the geographic coordinates of the trip destination. Therefore, trips taken during different periods could be extracted to study human mobility patterns.

Definition 1: *In the research, a trip is simplified to be a vector,* $< t_i, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}) >$*, the term* t_i *denoting the time instant corresponding to the trip origin of the taxi,* (x_{i1}, y_{i1}) *representing the trip origin geographic coordinates of the taxi and* (x_{i2}, y_{i2}) *indicating the geographic coordinates of the trip destination.*

Definition 2: *A taxi trajectory is constituted by a series of trip tuples,* $< t_i, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}) >$, $< t_{i+1}, (x_{i+1,1}, y_{i+1,1}), (x_{i+1,2}, y_{i+1,2}) >, \ldots, < t_m, (x_{m,1}, y_{m,1}), (x_{m,2}, y_{m,2}) >, \ldots, < t_n, (x_{n,1}, y_{n,1}), (x_{n,2}, y_{n,2}) >$.

Table 2 summarizes the statistics of the dataset that we selected. Figure 1 shows a map of Chinese cities, indicating where Wuhan is located. As depicted in Figure 2a, a taxicab trajectory within a loop highway was plotted in the map on 2 June 2014. According to the characteristic of taxicab trajectories, especially the recorded GPS points where anonymous passengers were picked up and dropped off during different time periods, we extracted quantities of trips from more than 2050 taxis on different days. Subsequently, each trip was simplified to be a point pair and trip distance, which are represented by a Pick-Up Point (PUP), a Drop-Off Point (DOP) and the Euclidean distance between the two points. At the same time, the two points, PUP and DOP, can be viewed as the origin and destination of a trip, respectively. Therefore, a trip destination represents the purpose of a trip and reflects human mobility. However, it is worth noting that trips less than a certain distance should be removed, as they are often caused by false driver operations or data errors. In this research, the distance threshold was set to be 0.5 km. Figure 2b demonstrates the spatial distribution of all taxicab trajectories on 2 June 2014; yellow points and red points denote the positions of trip origins and destinations on 2 June 2014, respectively, which correspond to pick-up points and drop-off points. Figure 2 illustrates the spatial distribution of pickups and drop offs from the perspective of the number of taxicabs. As shown in Figure 2, the numbers of red points and yellow points present certain differences because of the coverage. In fact, the number of red points is equal to the yellow points because one pickup point corresponds to one drop off point.

Table 2. Statistics of seven-day's taxi records from 2 to 8 June 2014.

Date	Number of Records	Number of Taxies	Number of Trips	Number of Valid Trips
2 June	2,446,961	2057	56,770	41,134
3 June	2,468,565	2059	55,487	40,992
4 June	2,449,164	2069	54,472	41,048
5 June	2,483,043	2073	55,598	41,698
6 June	2,510,001	2064	56,936	41,098
7 June	2,539,803	2049	59,220	42,989
8 June	2,498,303	2063	58,353	43,233
Total	17,395,840		396,836	292,192

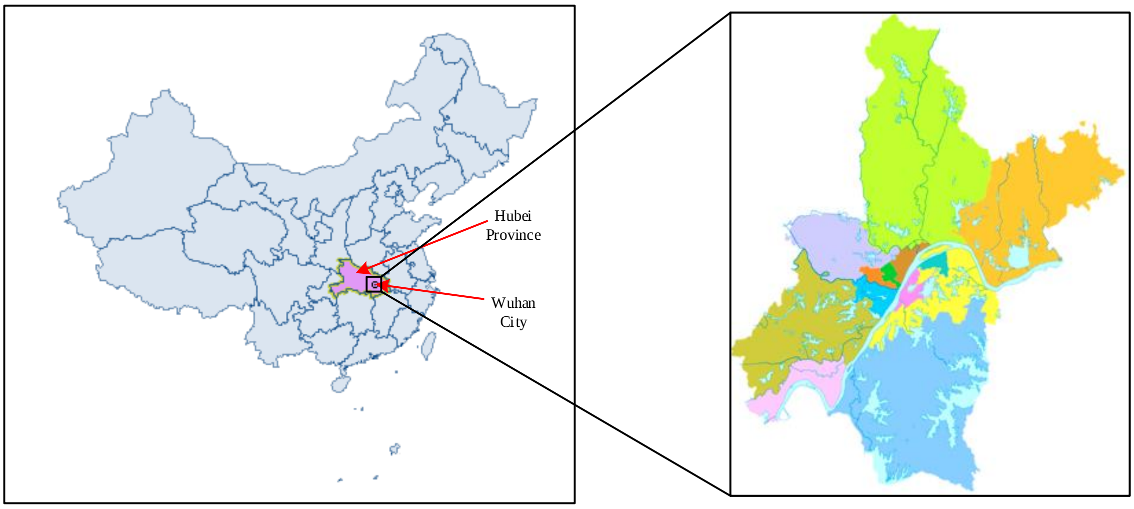
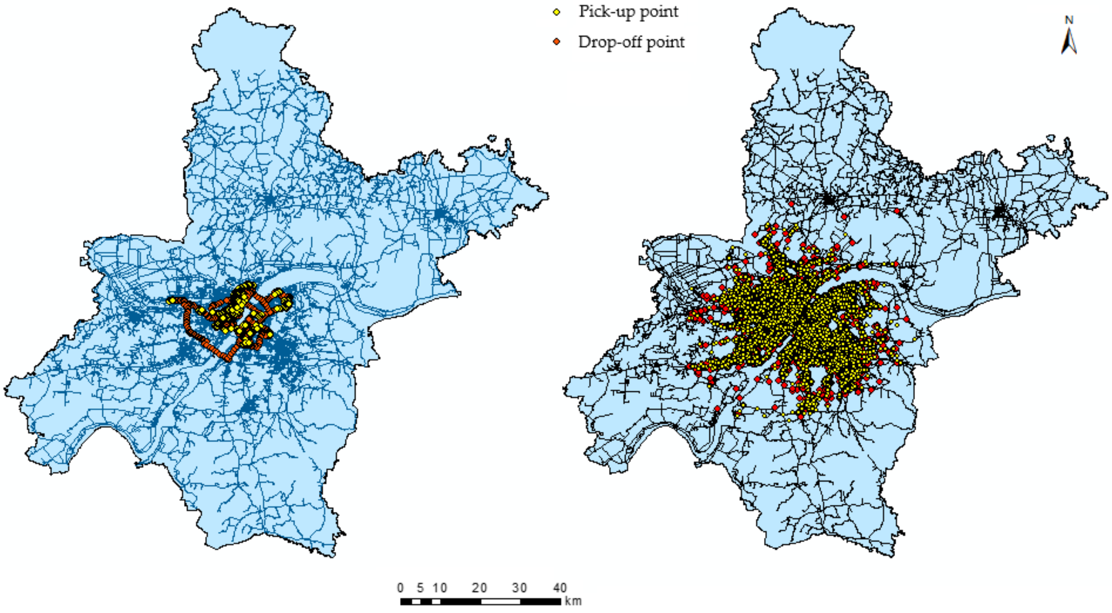


Figure 1. The location of Wuhan in China.



(**a**) (**b**)

Figure 2. (**a**) Spatial distribution of pickups and drop offs of a taxicab on 2 June 2014; (**b**) spatial distribution of pickups and drop offs of all taxicab on 2 June 2014. Yellow points and red points denote pick-up points and drop-off points, respectively.

2.2. Distribution of Trips

In this paper, we explore similarities in human behavior in terms of the temporal distribution of trip origins and trip distance. The occurrences of trip origins during each hour every day can be easily obtained and represent the characteristics of human activities over time. Consistent with many previous studies, as depicted in Figure 3, there are strong daily rhythms and day-to-day trip similarities [33–37]. People take more trips during the day than at night, and the temporal patterns on weekends are significantly different from those on workdays. Therefore, there were different mobility patterns on workdays and non-workdays. On weekends, more entertainment, parties, shopping and other recreational activities contribute a large proportion to trip purposes. There were more trips from zero o'clock to five o'clock on weekends than any other time.

Each trip can be viewed as a displacement of an individual trajectory, and the distance distribution reflects the mobility patterns of people. The observed distribution of extracted distance on 2 June was drawn in Figure 4, as a previous study used an exponentially-truncated power law distribution to fit the distance distribution of taxi trips [23]. It shows that there are more trips corresponding to short distant travel, while it is the opposite for a long trip. In order to better reflect mobility patterns, in Figure 4, we choose one hundred meters as the trip unit instead of one thousand meters [23] because it will appear as a negative value. People travel different distances for different purposes; we divided distance into four categories according to elbows where there exist sharp changes, 1 km, 7 km and 20 km, respectively, as depicted in Figure 4. Taxi trips with different distances could also prove reasonable by the differences in their temporal variations. As depicted in Figure 5, although the temporal variations of PUP with different distances from four groups were different, the distribution of trip distance also presents day-to-day similarity in a week, implying the similarity of human daily activity.

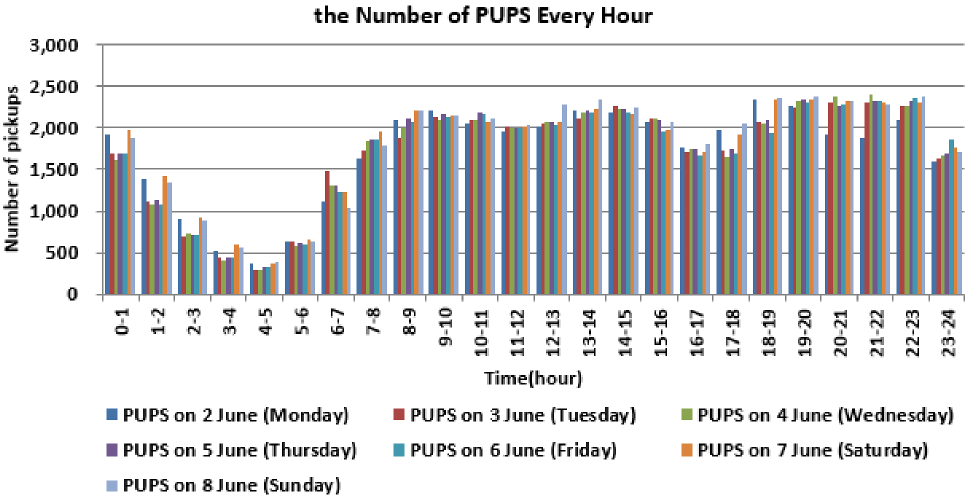


Figure 3. Number of PUPS (Pick-Up Points) every hour and the temporal variation during a week.

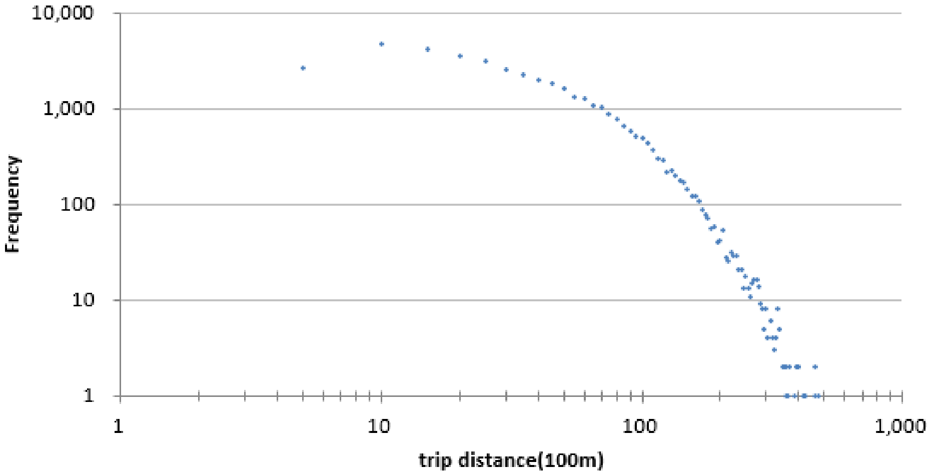


Figure 4. A log-log plot of the distance distribution of trips on 2 June.

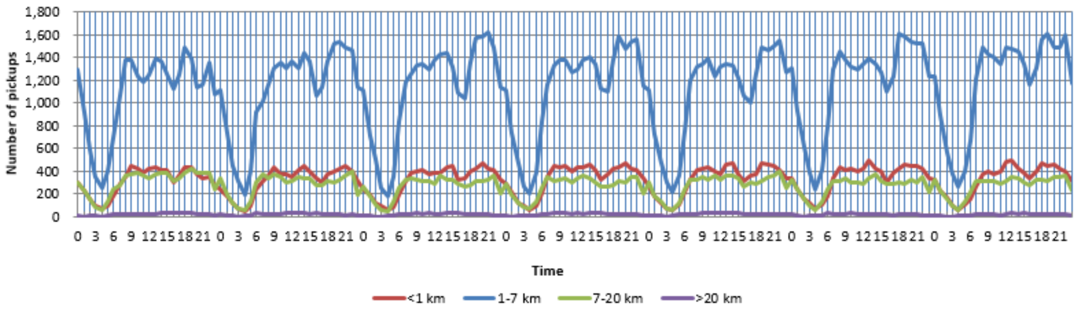


Figure 5. Temporal variations of PUP with different distances.

3. Semantic Analysis

It is difficult to discover trip knowledge directly from the geometric expression of trips by their coordinates alone. In contrast, finding knowledge in semantic information is not difficult. Therefore, in order to search hidden knowledge using semantic information, geometric coordinates must be processed by a semantic enrichment process. Semantic enrichment is based on additional information, such as the annotation or road segment's name corresponding to the GPS recorded locations, which can add to our understanding of trips [38]. However, some road segments are so long that many roads may have the same name when a road splits into many smaller roads. Similarly, several origins and trip destinations may correspond to the same road name. Such road name information cannot reflect the spatial distribution of trip origins and destinations in a certain sense.

In this research, a road network constituted by link IDs was used as the supplemental information to analyze human mobility patterns. Each geographic coordinate recorded by the GPS device attached to a taxi should correspond to the link ID to which it belongs. To solve this problem, taxi trajectories were matched by a map-matching algorithm [11,39,40] to find the best fit link attached to each GPS location. A unique link identification number was applied as supplemental data of the semantic enrichment for each GPS point. Generally, any trip is constituted by the trip origin and the trip destination, corresponding to different GPS points, respectively. In other words, if a link identification number were regarded as the origin of a trip, another, different link identification number should correspond to the destination of the trip. A link identification number also has a street name. Table 3 shows a few examples for extracted trip origins, trip destinations, matched link identification number and streets name to these trips as the trip semantic information. Subsequently, each semantic trip can be represented as a tuple, $trip\langle id, t, \ linkID, streetName, ODState\rangle$, in which id is the taxi identification number, t denotes the time instant, $linkID$ represents the matched link identification number, $streetName$ indicates the matched street name and $ODState$ denotes an origin or destination attached to the same extracted trip. Through semantic enrichment, a taxi trajectory is constituted by a series of trip tuples.

Table 3. Examples of extracted trip origins and trip destinations.

Taxi ID	Date	Latitude	Longitude	Link ID	Street Name	*ODState*
10319	2 June 2014 07:35:30	30.627251	114.381743	16373	Gongye Road	origin
10319	2 June 2014 07:40:02	30.63174	114.3775	16748	Heping Road	destination
16657	2 June 2014 18:24:44	30.515136	114.313965	9208	Ping'an Road	origin
16657	2 June 2014 18:44:42	30.548246	114.296945	10838	Minzhu Road	destination

4. Trip Topic Modeling with LDA

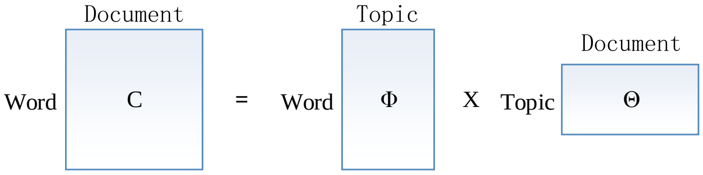
4.1. Latent Dirichlet Allocation

The topic model can be applied to analyze the quantities of documents to find hidden information in the corpus [41]. It overcomes the shortcomings of document similarity calculation in the field of traditional information retrieval and automatically searches for the textual semantic topic among massive words. In the topic model, a topic indicating a concept or an aspect denotes a series of related words and expresses the conditional probability of those words under the topic. From the mathematical perspective, the topic is a conditional probability distribution on the vocabulary; the more correlated with the topic the word is, the greater the conditional probability of the word. Otherwise, it is smaller. We can imagine that the topic is a bucket containing those words having higher probability, and these words have a strong correlation with this topic. If all of the words of each document in a corpus could be thought of as observations, a topic model can be created by inference from the observations to derive the hidden thematic structure [8,17]. A document may include several topics, and words in a document have different occurrence probabilities in each topic.

Latent Dirichlet Allocation (LDA) is the simplest topic model [28] and helps to dig out hidden topics. LDA statistically groups words into potential topics by studying their occurrences among a large collection of documents [26]. Consequently, each document represents a probability distribution of certain topics; each topic also denotes a probability distribution of many words. We can explain the relationship between document and topic using a generative model. In a generative model, each word in a document is obtained by a special process, for which each word chooses a topic at a certain probability and a topic also selects a certain word at a certain probability. Consequently, as for a document, the probability of each word appearing could be denoted as follows:

$$p\left(word|document\right) = \sum_{topic} p\left(word|topic\right) * p\left(topic|document\right) \tag{1}$$

Subsequently, this probability formula can also be expressed in a matrix as follows:



namely,

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} = \begin{bmatrix} \varnothing_{11} & \varnothing_{12} & \cdots & \varnothing_{1t} \\ \varnothing_{21} & \varnothing_{22} & \cdots & \varnothing_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ \varnothing_{m1} & \varnothing_{m2} & \cdots & \varnothing_{mt} \end{bmatrix} \times \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \theta_{t1} & \theta_{t2} & \cdots & \theta_{tn} \end{bmatrix} \tag{2}$$

As described in Equation (2), c_{mn} denotes the probability of the m-th word in the n-th document; $\varnothing_{mt}$ denotes the probability of the m-th word in the t-th topic; and θ_{tn} denotes the probability of the t-th topic in the n-th document. Consequently, the document-word matrix represents the probability of each word in each document; the topic-word matrix denotes the probability of each word in each topic; and the document-topic matrix indicates the probability of each topic in every document. Given a series of documents, we can obtain a document-word matrix through a tokenizer splitting document. The function of the topic model is to gain those two matrixes through training this document-word matrix.

In this research, the taxi's trajectory was generally constituted by more than one trip. That means a trajectory may contain a series of trips corresponding to different periods of time. Therefore, a trajectory including trips of different times was modeled as a document, and different semantic trip destinations attached to the link identification number in different periods of time were viewed as words. Thus, the LDA model can be used for topic modeling for trip destination to analyze the spatial distribution. At the same, different trajectories in different periods of time could be used to study the temporal distribution of trip destinations. We implemented the LDA inference using the Stanford Topic Modeling Toolbox [42]. Subsequently, we visualized different topics using different colors in an obvious way for exploring geographical information [43,44]. In section 5, it illustrates the results applying LDA to ten trip topics generated from a trip destination in Wuhan in morning rush hour (6:00 to 10:00) and evening rush hour (17:00 to 21:00), using trip destinations from 2 to 8 June 2014.

4.2. Significance Definition

Topics and trip destinations may belong to different level of significance, depending on their hidden thematic knowledge. Specifically, trip topic significance and trip destinations significance are discussed as follows.

Trip topic significance: The topic is a concept or an aspect of a document, and it is characterized by a series of related words. The more correlated with the theme the words are, the more possible it is that they gather into a theme. The significance of a trip topic, F_t, is defined as the total frequency of the trip destination (F) that supports the topic (t) in the hidden thematic knowledge, representing the topic importance. The greater the F_t is, the more trip destinations the topic attracts. In Table 4, 10 topics are presented by the order of their topic significance (*i.e.*, total frequency) from Topic00 to Topic09. The frequency of Topic08, which has the highest trip topic significance, F_t, is 2460, while Topic05 saw the lowest F_t, only having less than half of that at 1003. Therefore, among the 10 topics, Topic08 attracted more trip destinations attached to people's trip behavior than any other topic.

Trip destination significance: Trip destination significance, denoted by $F_w\,(t)$, is the significance level of a trip destination (*i.e.*, words) w in a given trip topic. A bigger value, $F_w\,(t)$, means that the word of a trip destination (w) offers more contributions to the production of trip topic t than other topics. The same trip destination may contribute to the production of several trip topics with different trip destination significance. For example, in Table 4, Link 22921, as a trip destination, contributes greatly to Topic01 and Topic06 with 25 percent and 44 percent support, respectively. At the same time, it plays a less important role in Topic00, Topic02, Topic04 and Topic09. However, it has no significance for Topic03, Topic05 and Topic07. In contrast, Link 10229 mainly contributes to two topics, amounting to 99% in total. One topic is Topic08 with 60% support, and another is Topic01 with 39%. Only one percent contributes to Topic00. All in all, this shows that the same link is more associated with some topics than other topics.

Table 4. Latent Dirichlet Allocation (LDA) results: trip destination significance in topics. *Freq* and *Prob* stand for frequency and probability, respectively.

Topic	Total Frequency of Topic	Link 22921		Link 10229		Link 14346		Link 10139	
		Freq	Prob	Freq	Prob	Freq	Prob	Freq	Prob
Topic08	2460	11	0.08	39	0.60	1	0.02	109	0.53
Topic06	2173	60	0.44	0	0.00	18	0.28	0	0.00
Topic09	2066	9	0.06	0	0.00	0	0.00	30	0.15
Topic04	1728	5	0.04	0	0.00	14	0.22	28	0.14
Topic03	1641	0	0.00	0	0.00	10	0.16	0	0.00
Topic00	1543	4	0.03	0	0.01	0	0.00	2	0.01
Topic01	1449	33	0.25	26	0.39	21	0.32	21	0.10
Topic02	1381	13	0.10	0	0.00	0	0.00	0	0.00
Topic07	1244	0	0.00	0	0.00	0	0.00	16	0.08
Topic05	1003	0	0.00	0	0.00	0	0.00	0	0.00

Trip topic probability distribution: Each trip topic is in fact a probability distribution of different trip destinations attached to various mobility patterns under the condition of this topic, namely

$\varphi_t < p_{w_1}, \ldots, p_{w_m} >$, in which p_{w_i} represents the probability of a trip destination, w_i, generating topic t. This is the ratio of trip destination significance over trip topic significance, $p\left(w_i|t\right) = F_{w_i}\left(t\right)/F_t$; a conditional probability of trip destination under the condition of topic t. Here, the conditional probability of all trip destinations over the topic sum to one, $\sum_i p_{w_i} = 1$. The bigger the value $p\left(w_i|t\right)$, the more representative the trip destination is expressed by topic t. Consequently, if a threshold δ is defined, then all trip destinations with a conditional probability higher than the threshold are considered as representative trip destinations for topic t, which will be used for the visual analysis of topics.

Considering trips with different time periods, such as working, shopping and recreation, we extracted trips from taxi trajectories during morning rush hour (6:00 to 10:00) and evening rush hour (17:00 to 21:00), respectively. We utilized semantic trip destinations that indicate where an activity occurs, like link IDs (link identification numbers) to analyze human behaviors. The analysis results of the topic model can help drivers to know where to efficiently pick up passengers, while passengers also can know where to take a taxi. Figure 6 shows the process of the sematic analysis framework for trip destination, and the framework includes trips manager, semantic enrichment, topic modeling and topic analysis.

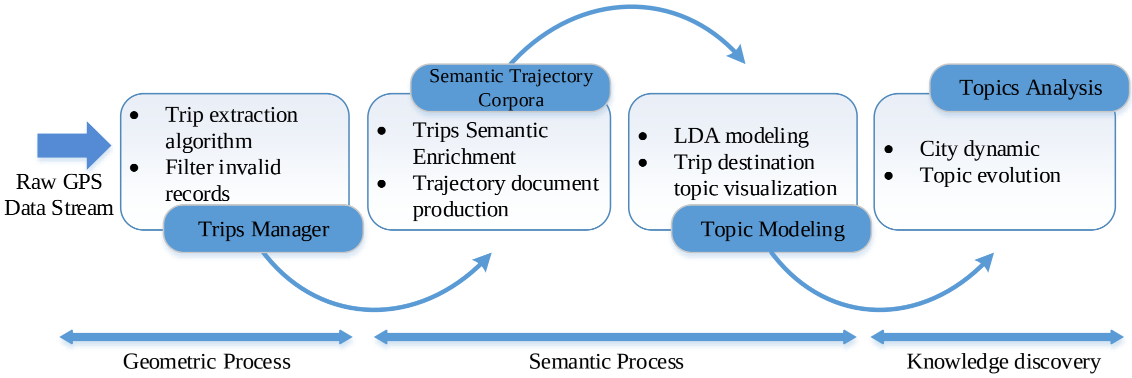


Figure 6. The process of the sematic analysis framework for trip destination.

5. Trip Topic Analysis of Mobility Patterns in Wuhan

5.1. Trip Topic Extraction and Visualization

The Yangtze River and Han Jiang divide Wuhan into three territories, Wuchang, Hankou and Hanyang, respectively. Administratively, Wuhan is composed of 13 districts: seven urban districts (Jiang'an, Jianghan, Qiaokou, Hanyang, Wuchang, Hongshan and Qingshan) and six suburban districts (Xinzhou, Huangpi, Dongxihu, Hannan, Jiangxia and Caidian). In addition, there are three state-level economic development zones, including the Wuhan economic and technological development zone, the east lake new technology development zone and the Wuhan Wujiashan Taiwanese investment zone. GPS records from 2 to 8 June 2014 were analyzed and preprocessed. The amount of records was 9,847,733 in total. Using the trip extraction algorithm, 164,872 valid trips were identified considering the trip availability, because some trips may be caused by false operations or data transfer errors. Intersections in the road network were used as semantic supplementary information for identifying the origins and destinations of trips. Trips were extracted according to the service state, whether a taxi were occupied or vacant. Each taxi trajectory may contain several trips, more than one trip destination in the morning rush hour (6:00 to 10:00) and evening rush hour (17:00 to 21:00) included in a taxi trajectory. These were collected to formulate two trip destination documents, respectively. Finally, the Stanford Topic Modeling Toolbox (TMT) was used to build an LDA model for these two trip destination documents. In the process of topic modeling, CVB0 (in the toolbox) was used to train the LDA model with ten topics for 2000 iterations. The topic smoothing parameter was set to 0.01, and all of the trained parameters and results were saved every 50 iterations. For each topic, the probability threshold (δ) for trip destination was set at 0.5% to identify a representative trip destination.

The spatial distribution of ten trip topics during morning rush hour were extracted and visualized (Figure 7). The base map is seven urban districts in Wuhan whose boundaries were drawn with bold black solid lines and roads drawn as gray lines. As we can see, each topic was drawn with a different color. For each trip topic, only representative trip destinations with $p(w_i|t) \geqslant \delta$ were rendered. The width of each link in the same topic containing a series of link IDs was proportional to $p(w_i|t)$. For a link that does not belong to any topic or $p(w_i|t) < \delta$, the color of the base map was attached to it. As shown in Table 4, some road segments were included in multiple trip topics. For example, Link 10229 was significant in both Topic01 and Topic08. Any link contributing to several topics was drawn only with the color corresponding to the topic that had the highest $p(w_i|t)$ value. As depicted in Figure 8, all of the links belonging to Topic08 mean that they were included for a common reason, such as similar travel districts or trip destinations.

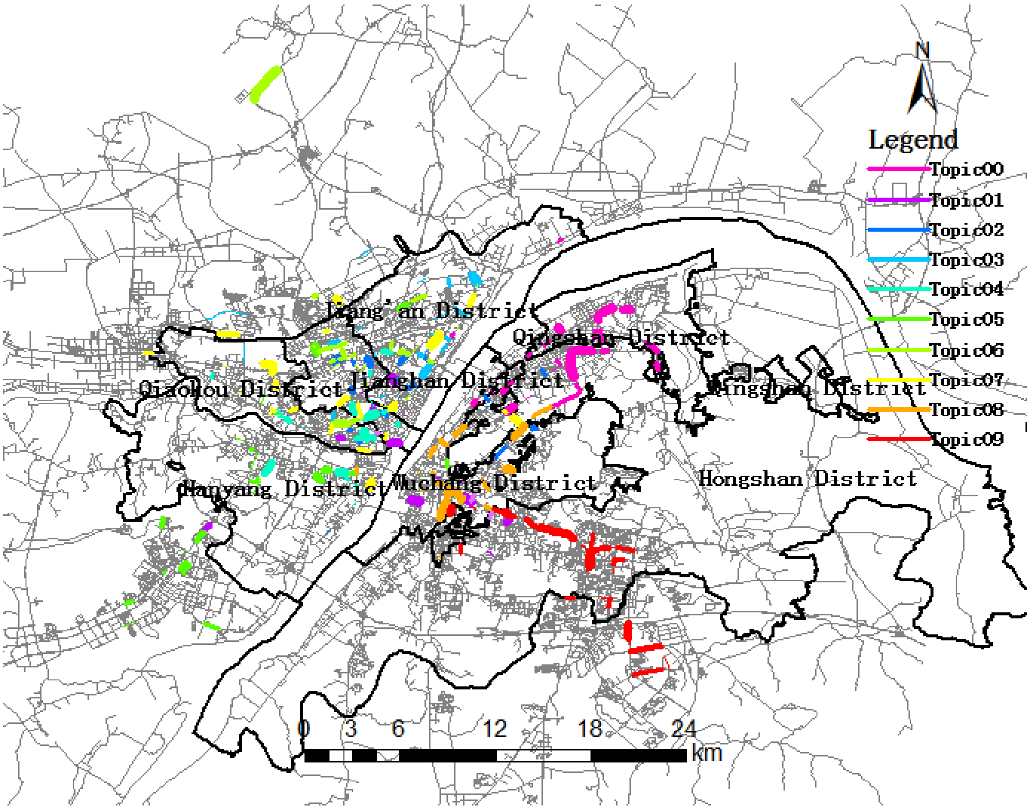


Figure 7. Ten topics during the morning rush hour period.

5.2. *Trip Topic and Urban Dynamics*

Trip destinations are aggregated into a topic having some common characteristics, revealing the spatial similarity of trip destinations. The distribution of topics showed trip patterns of passengers and the dynamics of human behavior. As depicted in Figure 7, seven urban districts are denoted by a thick black solid line, and the background of the road network is expressed in light grey lines. The same color lines form into a topic; the wider the line is, the more trip destinations. This reveals hot spot regions for trip destinations from the perspective of probability. Figure 7 shows that all ten topics almost covered all of the urban districts of Wuhan. For example, Topic07 (in yellow) mainly covers Jianghan, Jiang'an and Qiaokou, implying that a great number of taxi rides crossed district boundaries. However, there also exists Topic05 and Topic09 outside the urban districts, implying that some taxis move between urban districts and suburban districts.

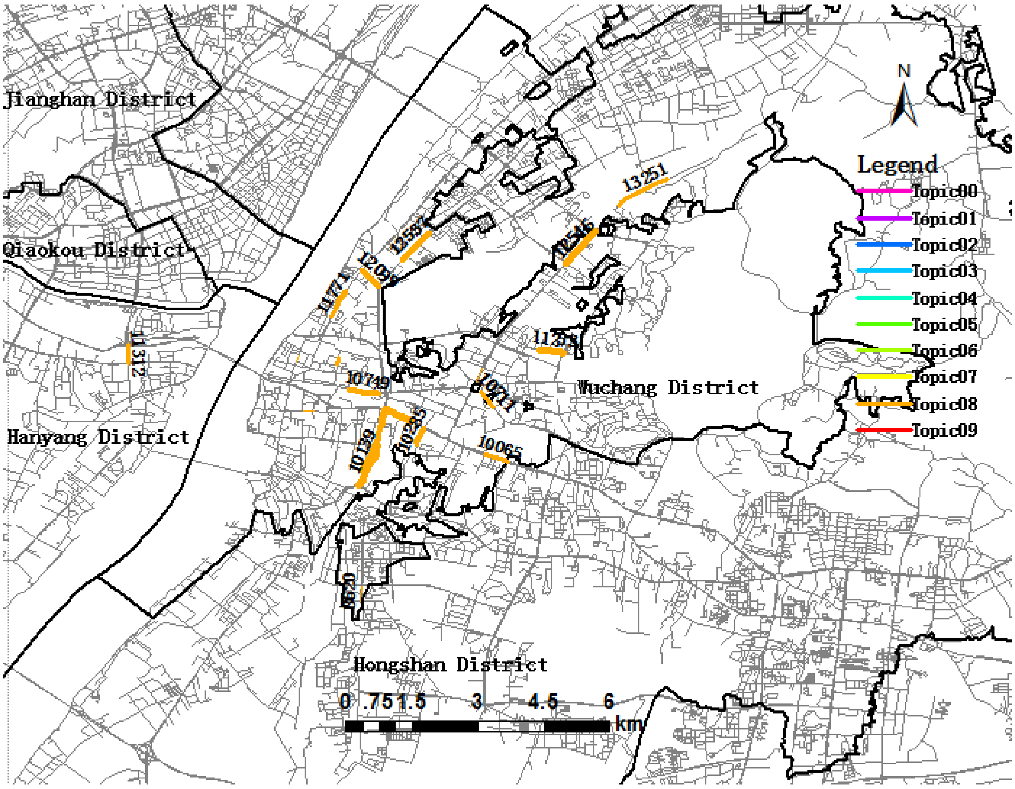


Figure 8. Visualization of trip Topic08 during the morning rush hour period.

Table 4 shows that Topic08 (in gold) has the highest topic significance, suggesting that it was the focus of the trip destination in the morning rush hour period. As depicted in Figure 8, the links belonging to Topic08 are mainly distributed in Wuchang district, including Link 10139, Link 10514, Link 11253, Link 12546 and other links. Wuchang is the political, cultural and information center of Hubei province and has the highest topic significance. In the morning, great quantities of people will go to work in this region. Among them, Link 10139 (Figure 8) has the highest trip destination probability in Topic08, and the most important reason is that Wuchang railway station is located there, attracting a large number of trip destinations. Figure 7 shows the hot spots of trip destinations from 6 a.m. to 10 a.m., which have a higher probability among topics. At the top-left of Figure 7, we could note that there is a link isolated from the urban districts, namely Tianhe international airport, meaning that many people take planes during the morning rush hour.

Topic09 (in mars red) and Topic00 (in ginger pink) will be explained in detail. As depicted in Figure 9, all links related to Topic09 are aggregated in the same area, showing the spatial distribution of trip destination in the morning peak period. The aggregated roads include Link 9933, Link 9734, Link 9502, Link 8669, Link 8750, Link 8803, and so on, which is a main road in the southeastern intra-urban area with many high tech enterprises and containing the Guanggu commercial area. Therefore, it is a hotspot for activities in the Hongshan district, and many trip destinations contribute to Topic09. Among them, Link 8669, Minzu Avenue, has the highest probability of trip destinations, revealing that most passengers choose this link to be the trip destinations of their trips in Topic09.

Figure 10 shows that Topic00 was distributed in Qingshan district and Hongshan district. Topic00 mainly contains Link 17786, Link 17724, Link 16992, Link 15500, Link 15830, and so on, including Heping Avenue, Jianshesi Road, Youyi Avenue and Gongye Road. As depicted in Figure 10, some links attract more trip destinations than other links. Link 15500, Link 17786 and Link 16992 are wider, indicating that they attract more trip destinations.

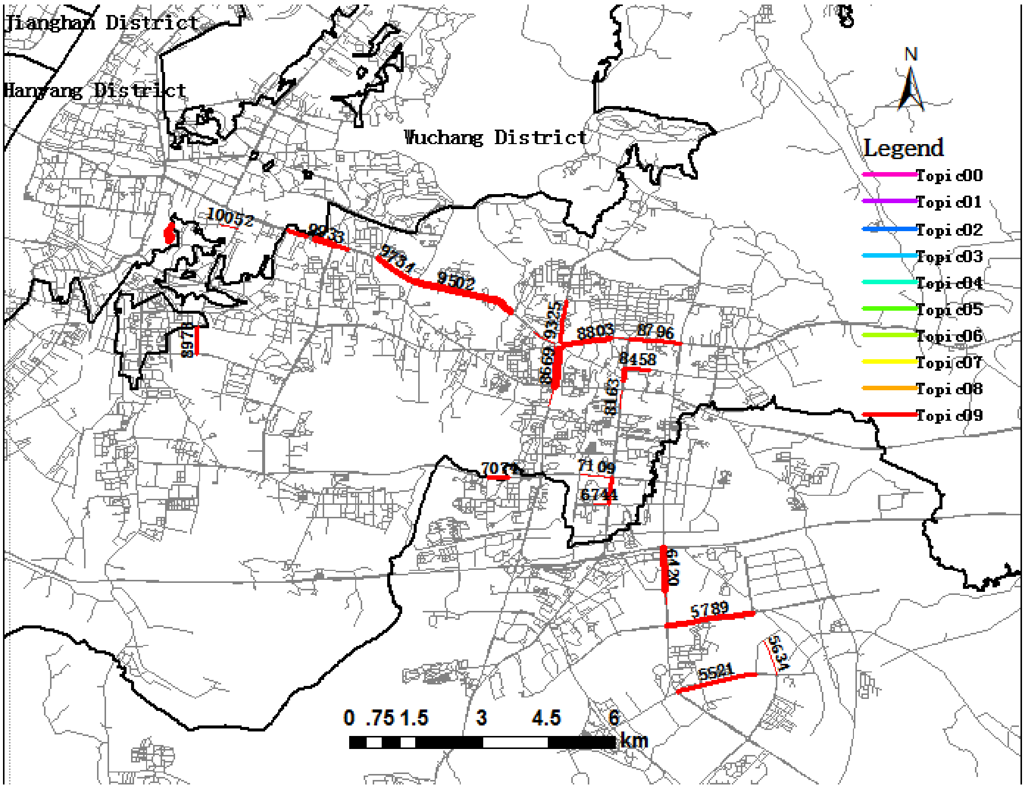


Figure 9. Visualization of trip Topic09 during the morning rush hour period.

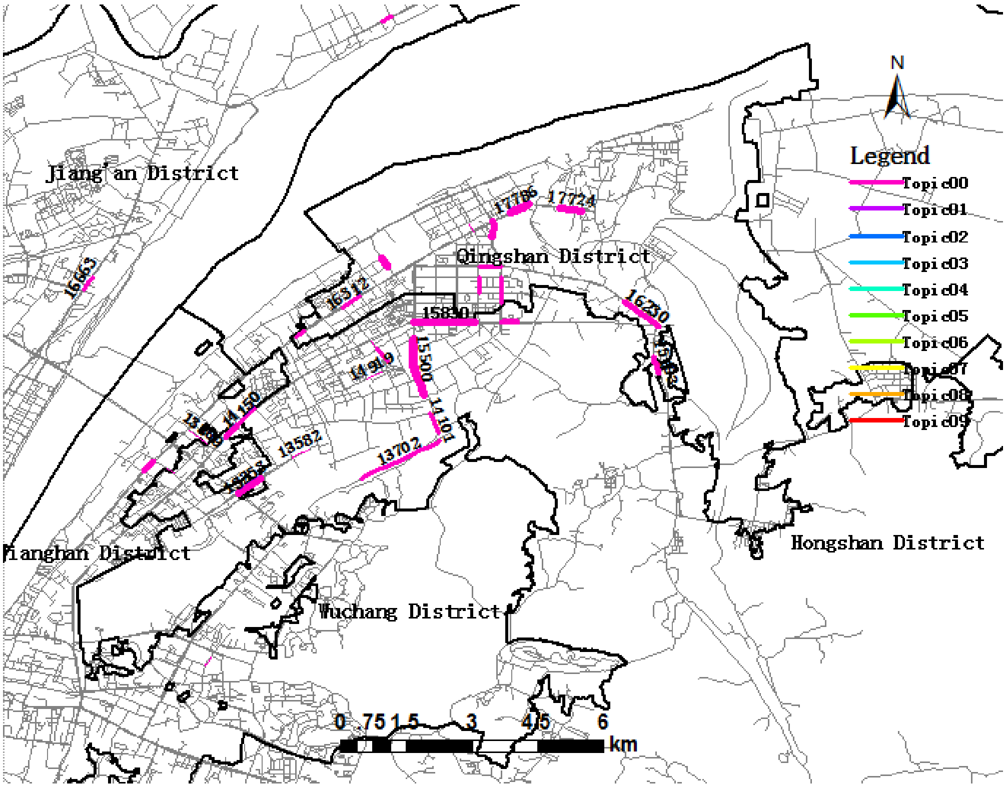


Figure 10. Visualization of trip Topic00 during the morning rush hour period.

5.3. Trip Topic Evolution

In this research, trip topics were created in the periods of morning rush hour and evening rush hour, respectively. As depicted in Figure 11, 10 trip topics were generated during the evening rush hour distinct from the 10 topics generated during the morning rush hour in Figure 7. It is easy to know that different topics reflect different mobility patterns. For example, people usually travel from home to work in the morning rush hour and reversely in the evening rush hour. Therefore, it is important to find the evolutionary trends of trip topics in different periods implying the difference of mobility patterns. Therefore, we define a topic similarity to find two similar topics between two trip topics of different periods. Given two topics i and j, the topic similarity is defined as follows:

$$S_{i,j} = \frac{Size\left(T_i \cap T_j\right)}{Min\left(Size\left(T_i\right), Size\left(T_j\right)\right)} \tag{3}$$

where T_i and T_j are the sets of the link identification number with a high probability $p\left(w|z\right) > c$ in topics i and j, respectively. Based on this similarity, closely related topics are identified in the morning rush hour and the evening rush hour, which reflects the temporal evolution of topics. Such similarity was calculated according to Equation (3) to find the variation of the topic content, the emergence of links in a topic and their fading out. Given two similar topics, the topic significance change is defined as follows:

$$SC_{i,j} = \frac{Size\left(T_i\right) - Size\left(T_j\right)}{Size\left(T_j\right)} \times 100\% \tag{4}$$

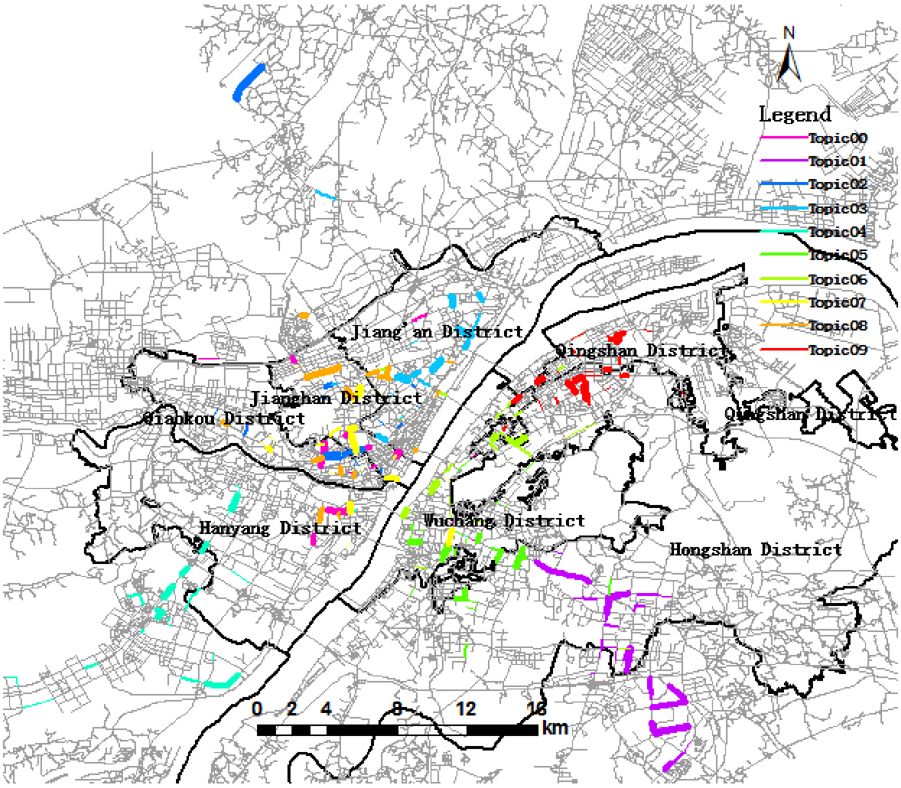


Figure 11. Ten topics during the evening rush hour period.

As depicted in Figure 7 and 11, they reflect the spatial distribution of the trip topic in the morning rush hour (6:00 to 10:00) and the evening rush hour (17:00 to 21:00), respectively, and Table 5 shows the temporal evolution of the trip topic during different time periods. As we can know from Figure 6 and Figure 10, there is a similar spatial distribution of topics between "morning rush hour topics" and "evening rush hour topics". For convenience, we defined the topics of morning rush hour (6:00 to 10:00) and evening rush hour (17:00 to 21:00) as "morning rush hour topics" and "evening rush hour topics", respectively. In Table 5, the first column denotes ten topics in the morning rush hour period, and the column "evening rush hour topics" contains the most similar topics corresponding to the morning rush hour topics. Much knowledge can be inferred from Table 5. Table 5 illustrates that the taxi activities are higher during the evening rush hour than during the morning rush hour, because the total topic significance of the evening rush hour is bigger than that during the morning rush hour. People have more activities in the evening, such as recreation, dining and shopping.

Table 5. Temporal evolution of trip topics.

Morning Rush Hour Topics	Topic Significance	Evening Rush Hour Topics	Topic Significance	Topic Similarity	Significance Change
Topic08	2460	Topic05	2110	0.43	−14.23%
Topic06	2173	Topic08	1805	0.40	−16.94%
Topic09	2066	Topic01	2404	0.53	+16.41%
Topic04	1728	Topic00	1426	0.29	−17.49%
Topic03	1641	Topic03	1976	0.42	+20.43%
Topic00	1543	Topic09	1489	0.47	−3.45%
Topic01	1449	Topic00	1426	0.25	−1.64%
Topic02	1381	Topic02	1672	0.13	+21.10%
Topic07	1244	Topic00	1426	0.17	+14.59%
Topic05	1003	Topic04	1073	0.22	+7.02%
Total topic significance	16,687	–	17,681	–	–

From Table 5, many useful insights can be derived from topic evolution. Most topic similarity is less than 0.5, a smaller value that indicates the difference between "morning rush hour topics" and "evening rush hour topics". Among them, the biggest topic similarity is 0.53, and the topics are distributed at Luoyu Road and Guanggu Square, which is close to the high technological development zone with many scientific research institutions, enterprises and restaurants. Therefore, it attracts more workers in the morning rush hour and contributes to trip destinations for dinner during the evening rush hour. The topic significance increases by 16.41%, which means more activities happen in the evening. Figure 12 demonstrates the spatial distribution of similar topics during the morning rush hour and the evening rush hour showing the evolution of these two topics. As compared to Figure 12a, new links, like Link 5521, Link 5550, Link 4713 and Link 5079, contribute to trip Topic01 during the evening rush hour. That means there are more trip destinations in this region during evening rush hour. At the same time, there are also links from the evening rush hour fading out compared to those of the morning rush hour, such as Link 10045, Link 8973 and Link 9325. The topic significance of Topic05 during the evening rush hour decreased by 14.23% corresponding to and similar to Topic08 during the morning rush hour, illustrating that the region containing Topic08 has fewer activities in the evening. In addition, the significance of Topic03 in the evening rush hour increased by 20.43%, corresponding to and similar to Topic03 during the morning rush hour, which indicates an activity region in the Jian'an district in the evening.

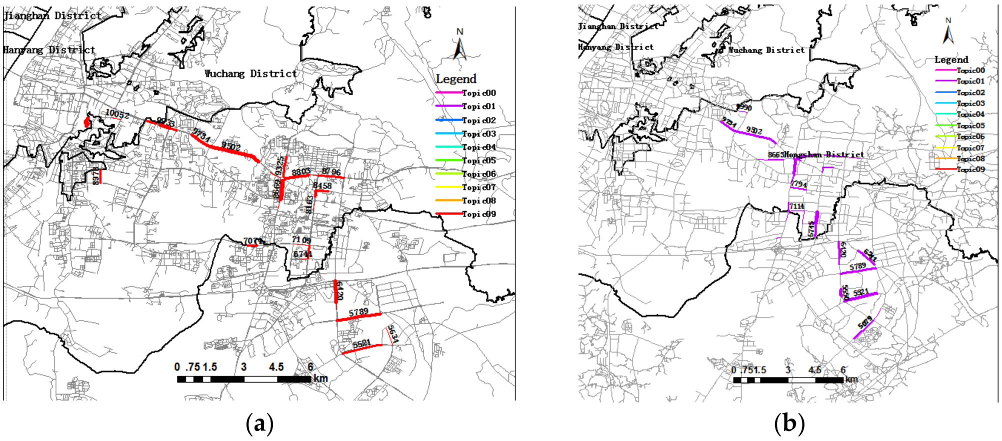


Figure 12. (**a**) Visualization of trip Topic09 during the morning rush hour; (**b**) visualization of trip Topic01 during the evening rush hour.

6. Conclusions

Previous studies have focused on analyzing the raw data from GPS sensors to promote the accuracy of GPS data. Recent research has shifted towards developing methods to study human mobility incorporating semantic information, like road segment name and annotation. Generally, a taxi trajectory is constituted by a series of trip destinations, which might reflect human mobility patterns. Such mobility patterns can be further used to help taxi drivers to take more passengers and to tell passengers where they can call a taxi quickly.

This paper analyzed semantic trip destinations extracted from raw taxi trajectories at a finer scale using an LDA method in order to convey the spatial similarities of mobility patterns and the temporal evolution of topics. We first introduced the road segment identification number of pickups and drop offs as semantic information on massive GPS trip data, instead of using street names [26]. Then, the LDA model was applied to construct trip topics in order to find human mobility patterns. Each trip topic contains a series of links with different trip destination significance in expressing a common aspect, such as belonging to the same district or belonging to the same functional region. Different links in a topic were rendered using the same color, and the width of a link is proportional to the probability in this topic. The wider the link is, the more trip destinations the link will attract. Then, we tell what links are hot spots links and find the relationship between links and functional regions. The temporal evolution of topics can be detected by analyzing topic relationships between the morning rush hour and the evening rush hour. Consequently, a potential application of this research could help taxi drivers know when and where they can pick up more passengers according to the result of topic analysis if time intervals are divided properly. Additionally, it could tell passengers how to take a taxi more quickly.

Our results are influenced by techniques. For example, the accuracy of location is influenced by GPS devices, and the map-matching algorithm affects the precision of GPS trajectories. This paper only focused on the group topics of passenger mobility. Such group topics imply patterns of human activities and travel. Moreover, the representation as trip topic distribution across different periods expresses the temporal evolution of topics, and the topic similarity reflects human mobility patterns over time. In the future, we will interpret human mobility through integrating features of the city (e.g., points of interest, mobile phone data positioning data, land use plan, urban planning) to measure the relationships between trip destinations and POIs or population density.

Acknowledgments: This work was supported by National Natural Science Foundation of China No. 41271401 and the National 863 project, Multi-source Information Real-time Access and Heterogeneous Information Autonomous Loading Technology under the Unified Spatiotemporal System, No. 2013AA122301, Pan-information map fusion technology Based on the Internet superposition protocol, No. 2013AA12A203, National Science and Technology Support Plan, the Key Technology and Applications of Location-based Sensor Network and Pan-information Map, No. 2012BAH35B03, National Science Foundation No. 1416509, ACI-1535031,1535081..

Author Contributions: Faming Zhang performed the research, analyzed the data and wrote the paper. Xinyan Zhu co-designed the research. Xinyue Ye co-designed the research and extensively updated the paper. Wei Guo developed some earlier prototypes. Tao Hu and Liang Huang edited the paper. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Candia, J.; Gonzalez, M.C.; Wang, P.; Schoenharl, T.; Madey, G.; Barabasi, A.-L. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A Math. Theor.* **2008**. [CrossRef]
2. Jiang, B.; Yin, J.; Zhao, S. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E* **2009**, *80*. [CrossRef] [PubMed]
3. Bai, F.; Helmy, A. A Survey of Mobility Models. In *Wireless Adhoc Networks*; University of Southern California: Los Angeles, CA, USA, 2004; pp. 1–30.
4. Chaix, B.; Kestens, Y.; Perchoux, C.; Karusisi, N.; Merlo, J.; Labadi, K. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. *Am. J. Prev. Med.* **2012**, *43*, 440–450. [CrossRef] [PubMed]
5. Miller, H.J.; Shaw, S.L. Geographic information systems for transportation in the 21st Century. *Geogr. Compass* **2015**, *9*, 180–189. [CrossRef]
6. Song, C.; Koren, T.; Wang, P.; Barabasi, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **2010**, *6*, 818–823. [CrossRef]
7. Song, C.; Qu, Z.; Blumm, N.; Barabasi, A.-L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021. [CrossRef] [PubMed]
8. Yuan, Y.; Raubal, M.; Liu, Y. Correlating mobile phone usage and travel behavior: A case study of Harbin, China. *Comput. Environ. Urban Syst.* **2012**, *36*, 118–130. [CrossRef]
9. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-Located Twitter as Proxy for Global Mobility Patterns. *Cartogr. Geogr. Inf. Sci.* **2014**. [CrossRef] [PubMed]
10. Rhee, I.; Shin, M.; Hong, S.; Lee, K.; Kim, S.J.; Chong, S. On the Levy-Walk Nature of Human Mobility. *IEEE/ACM Trans. Netw.* **2011**. [CrossRef]
11. Chen, C.; Bian, B.; Ma, J. From Traces to Trajectories: How Well Can We Guess Activity Locations from Mobile Phone Traces? *Transp. Res. Part C* **2014**. [CrossRef]
12. Lu, Y.; Liu, Y. Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Comput. Environ. Urban Syst.* **2012**, *36*, 105–108. [CrossRef]
13. Veloso, M.; Phithakkitnukoon, S.; Bento, C. Sensing urban mobility with taxi flow. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, IL, USA, 1–4 November 2011.
14. Zheng, Y.; Liu, Y.; Yuan, J.; Xie, X. Urban computing with taxicabs. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011.
15. Gao, S.; Wang, Y.; Gao, Y.; Liu, Y. Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environ. Plan. B Plan. Des.* **2013**, *40*, 135–153.
16. Qi, G.; Li, X.; Li, S.; Pan, G.; Wang, Z.; Zhang, D. Measuring social functions of city regions from large-scale taxi behaviors. In Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), Seattle, WA, USA, 21–25 March 2011; pp. 384–388.
17. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.
18. Zheng, F.; Van Zuylen, H. Urban link travel time estimation based on sparse probe vehicle data. *Transp. Res. Part C Emerg. Technol.* **2013**, *31*, 145–157. [CrossRef]
19. Cheng, Z.Y.; Caverlee, J.; Lee, K.; Sui, D.Z. Exploring millions of footprints in location sharing services. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011; pp. 81–88.
20. Lee, K.; Hong, S.; Kim, S.J.; Rhee, I.; Chong, S. SLAW: A mobility model for human walks. In Proceedings of the IEEE International Conference on Computer Communications, Rio de Janeiro, Brazil, 19–25 April 2009; pp. 855–863.

21. Rhee, I.; Shin, M.; Hong, S.; Lee, K.; Chong, S. On the levy-walk nature of human mobility. In Proceedings of the IEEE INFOCOM, Phoenix, AZ, USA, June 2011; pp. 630–643.
22. Liang, X.; Zheng, X.; Lv, W.; Zhu, T.; Xu, K. The scaling of human mobility by taxis is exponential. *Phys. A* **2012**, *391*, 2135–2144. [CrossRef]
23. Liu, Y.; Kang, C.; Gao, S.; Xiao, Y.; Tian, Y. Understanding intra-urban trip patterns from taxi trajectory data. *J. Geogr. Syst.* **2012**, *14*, 463–483. [CrossRef]
24. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic "source-sink areas": Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]
25. Ballatore, A.; Wilson, D.C.; Bertolotto, M. Computing the semantic similarity of geographic terms using volunteered lexical definitions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2099–2118. [CrossRef]
26. Chu, D.; Sheets, D.A.; Zhao, Y.; Wu, Y.Y.; Yang, J.; Zheng, M.; Chen, G. Visualizing Hidden Themes of Taxi Movement with Semantic Transformation. In Proceedings of the IEEE Pacific Visualization Symposium (PacificVis), Yokohama, Japan, 4–7 March 2014.
27. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2015**, *43*, 1–12.
28. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
29. Hashemi, M.; Karimi, H.A. A critical review of real-time map-matching algorithms: Current issues and future directions. *Comput. Environ. Urban Syst.* **2014**, *48*, 153–165. [CrossRef]
30. Quddus, M.A.; Ochieng, W.Y.; Noland, R.B. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transp. Res. Part C Emerg. Technol.* **2007**, *15*, 312–328. [CrossRef]
31. Zheng, Y.; Lou, Y.; Zhang, C.; Xie, X.; Wang, W.; Huang, Y. Map-Matching for Low-Sampling-Rate GPS Trajectories. U.S. Patent Application 12/712,857, 25 February 2010.
32. Chen, B.Y.; Yuan, H.; Li, Q.; Lam, W.H.K.; Shaw, S.-L.; Yan, K. Map-matching algorithm for large-scale low-frequency floating car data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 22–38. [CrossRef]
33. Ahas, R.; Aasa, A.; Silm, S.; Tiru, M. Daily rhythms of suburban commuter's movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transp. Res. C Emerg.* **2010**, *18*, 45–54. [CrossRef]
34. Schonfelder, S.; Axhausen, K.W. *Urban Rhythms and Travel Behaviour: Spatial and Temporal Phenomena of Daily Travel*; Ashgate Publishing: London, UK, 2010.
35. Härri, J.; Filali, F.; Bonnet, C. Mobility models for vehicular ad hoc networks: A survey and taxonomy. *Commun. Surv. Tutor.* **2009**, *11*, 19–41. [CrossRef]
36. Zonoozi, M.M.; Dassanayake, P. User mobility modeling and characterization of mobility patterns. *IEEE J. Sel. Areas Commun.* **1997**, *15*, 1239–1252. [CrossRef]
37. Fang, Z.; Li, Q.; Shaw, S.L. What about people in pedestrian navigation? *Geo-Spat. Inf. Sci.* **2015**, *18*, 135–150. [CrossRef]
38. Parent, C.; Spaccapietra, S.; Renso, C.; Andrienko, G.; Andrienko, N.; Bogorny, V.; Damiani, M.L.; Gkoulalas-Divanis, A.; Macedo, J.; Pelekis, N.; Theodoridis, Y.; *et al.* Semantic trajectories modeling and analysis. *ACM Comput. Surv.* **2013**, *45*, 1–32. [CrossRef]
39. Li, Q.; Bo, H.; Yang, Y. Flowing car data map-matching based on constrained shortest path algorithm. *Geomat. Inf. Sci. Wuhan Univ.* **2013**, *38*, 805–808.
40. Zhang, Y.; Yang, B.; Luan, X. Automated matching urban road networks using probabilistic relaxation. *Acta Geod. Catogr. Sin.* **2012**, *41*, 933–939.
41. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
42. Ramage, D.; Rosen, E. Standford TMT (online). Palo Alto: The Stanford Natural Language Processing Group, 2015. Available online: http://nlp.stanford.edu/software/tmt/tmt-0.4/ (accessed on 31 January 2015).
43. Wang, H.; Zou, H.; Yue, Y.; Li, Q. Visualizing hot spot analysis result based on mashup. In Proceedings of the 2009 International Workshop on Location Based Social Networks, LBSN '09, Seattle, WA, USA, 3 November 2009; pp. 45–48.
44. Speckmann, B.; Verbeek, K. Necklace maps. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 881–889. [CrossRef] [PubMed]