Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6697827, 10 pages
https://doi.org/10.1155/2021/6697827

Research Article

Exploring the Citywide Human Mobility Patterns of Taxi Trips through a Topic-Modeling Analysis

Hui Xiong,[1] Kaiqiang Xie,[1] Lu Ma,[2] Feng Yuan,[2] and Rui Shen[2]

[1]*School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China*
[2]*Ministry of Transport Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Lu Ma; lma@bjtu.edu.cn

Received 2 December 2020; Revised 22 May 2021; Accepted 12 July 2021; Published 20 July 2021

Academic Editor: Lelitha Vanajakshi

Understanding human mobility patterns is of great importance for a wide range of applications from social networks to transportation planning. Toward this end, the spatial-temporal information of a large-scale dataset of taxi trips was collected via GPS, from March 10 to 23, 2014, in Beijing. The data contain trips generated by a great portion of taxi vehicles citywide. We revealed that the geographic displacement of those trips follows the power law distribution and the corresponding travel time follows a mixture of the exponential and power law distribution. To identify human mobility patterns, a topic model with the latent Dirichlet allocation (LDA) algorithm was proposed to infer the sixty-five key topics. By measuring the variation of trip displacement over time, we find that the travel distance in the morning rush hour is much shorter than that in the other time. As for daily patterns, it shows that taxi mobility presents weekly regularity both on weekdays and on weekends. Among different days in the same week, mobility patterns on Tuesday and Wednesday are quite similar. By quantifying the trip distance along time, we find that Topic 44 exhibits dominant patterns, which means distance less than 10 km is predominant no matter what time in a day. The findings could be references for travelers to arrange trips and policymakers to formulate sound traffic management policies.

1. Introduction

Gaining a deeper understanding of human mobility is a prerequisite for a broad range of possible studies in many research fields, such as city structures [1], urban planning [2, 3], and traffic forecasting [4, 5]. While human mobility has traditionally been studied using surveys that deliver snapshots of population displacement patterns, the growing availability of massive geo-related data sets, such as cell phone calls and vehicle GPS tracking, has made it possible to explore human mobility at high spatiotemporal resolution in recent years [6]. Exploration of these empirical data sets has revealed some interesting facts. For example, through an analysis of banknotes, the distribution of displacement is a power law and it was concluded that human travel behavior can be described in terms of Lévy walks with a heavy-tailed pause time [7, 8]. However, while the aggregated displacement distribution follows a power law with an exponential cut-off, the absence of scaling properties in the displacement distribution at an individual level has also been demonstrated [8, 9]. Additionally, it has been argued that the scaling law in human mobility is exponential rather than a power law [9, 10].

Previous studies on human mobility have explored, for the most part, the macroscopic properties of mobility patterns using various data sets. For example, using banknotes in the United States as a proxy for human mobility, it was concluded that the distribution of traveling distances decayed as a power law and that human travel is an ambivalent and effectively superdiffusive process [7]. Based on GPS mobility data from private vehicles in Florence, Italy, it was also shown that long trip length distribution differed from exponential behavior and, instead, seemed to follow a power law [10]. In contrast, the distribution of individuals' intraurban travel was found to be exponential, using mobile phone data [11]. With the trajectories of 100,000 mobile

phone users, it was found that, contrary to the Lévy flight and random walk models, humans tended to return to a few highly frequented locations [12]. Using maximum entropy theory, human mobility was found to be highly predictable, independently of the distance that individuals covered [13]. Similarly, based on a mobile phone data set in Portugal, it was reported that most people spent most of their time at only a few locations [14]. To explore the underlying mechanism of empirical scaling laws, based on mobile phone traces, a model that not only accounted for the observed scaling laws but also allowed analytical predictions of the scaling exponents was used [15]. Also, algorithms have been proposed for land use identification and clustering [14, 16] and making travel mode inferences from the original GPS data [17].

Apart from banknotes and mobile phone data sets involving multiple travel modes, taxi GPS data sets reflect directly how people move within an urban area, serving as a reliable proxy for human mobility. A major advantage of taxi GPS data sets is that they provide accurate spatial-temporal information on the start and endpoints of every single trip. Consequently, there have been several studies on human mobility using taxi GPS data. Based on a weighted AIC criterion, the scaling law of taxi trip displacement, travel time, and the average running speed in occupied status was concluded to be exponential, rather than a power law, in Beijing, China [9]. In contrast, the trip displacement distribution was shown to be statistically greater than an exponential distribution but smaller than a truncated power law distribution. Specifically, the distribution of short trips (<30 miles) was reported to be best fitted with a power law while long trips followed exponential decay [18]. It has also been shown that two regimes (exponential power law and truncated power law) exist in the distribution of travel time, divided by a breakpoint [19]. Similarly, two regimes, power law and truncated power law patterns, were also found in the distribution of occupied taxi trips [16]. Based on large-scale taxi GPS trace data, urban taxi drivers' temporal and spatial distributions were analyzed. Compared with workdays, where three peaks were identified, there were only two peaks on weekdays concerning the time evolution of a taxi's daily operation status. In particular, it has been shown that the distance between each taxi driver's pick-up locations' mean center and drop-off locations' mean center was around 1000 m [20]. However, using the nonnegative matrix factorization method, three basic mobility patterns on workdays were noted, corresponding to three different travel purposes: commuting between home and workplace, traveling from workplace to workplace, and others, such as leisure activities [21].

To investigate the microscopic characteristics of human mobility patterns of taxi trips [22–24], a topic model analysis was proposed. Topic models [25–29] are used mostly in machine learning and natural language processing to find the latent topics in a large corpus of documents. In recent years, topic models [30–32] have been widely used in a variety of applications, e.g., ad hoc information retrieval [33] and geographical information retrieval [34]. In the field of human mobility research, there are a few existing studies using topic models in terms of different aspects. For example, Farrahi and Gatica–Perez (2011) [35] applied topic models on a reality-mining dataset to discover daily routines. In Montoliu (2011) [36], the spatiotemporal behavior of bicycle-sharing systems was extracted with topic models based on station-occupancy statistics. Topic models have also been investigated to analyze OD matrices generated by a bicycle-sharing system in Paris, France [37].

Generally, previous studies have explored, primarily, the statistical characteristics of mobility patterns at a macrolevel, and the high spatiotemporal resolution of these massive data sets has not been well leveraged to characterize the mobility patterns in a more detailed way. Thus, in this paper, we focused on characterizing human mobility in more detail. In this study, human mobility was quantified using the travel displacement and time for each trip. Based on the two measurements, two research questions were brought up. First, it is interesting to explore the distributional characteristics of travel displacement and time, which provide the baseline understanding of human mobility patterns and the reference for the classification of the two measurements in the topic model analysis. The second research question of this study focuses on exploring deeper mobility patterns with the aid of topic models.

This paper is organized in the following way. Section 2 provides an introduction of the GPS data of taxi trips with an emphasis on fitting the trip displacement and time with certain parametric distributions. Section 3 introduces the methodology of topic modeling and the proposed method of the transformation between the GPS information and the text data. Section 4 presents the results and analyses using topic models and illustrates several important human mobility patterns, and Section 5 ends this study with conclusions.

2. About the GPS Data of Taxi Trips

2.1. Description of the Data. The GPS data used in this study was collected by taxi vehicles in Beijing, China, from March 10 to 23, 2014. There were nearly 18, 000 taxis, equipped with GPS devices, running on the streets in Beijing daily, sending their real-time geographical position information l (latitude and longitude) back to the dispatching center every ten seconds. Besides, the taxi driver ID i, timestamp t, service status s (occupied or vacant), and instantaneous running speed v were also collected. In specific, s is a binary variable, as zero and one, representing the vacant and occupied status, respectively. The service status of a taxi is useful for identifying whether it is on service. Note that only the GPS data during service trips were adopted in this study. The time the variable s switches to one from zero means the start of a trip with taxi customers. On the other hand, the time the variable s switches to zero from one means the end of a trip with taxi customers. The collected information from taxi trip indexed by i was organized as tuple sequences $(i, t_{o1}, l_{o1}, t_{d1}, l_{d1})$ and $(i, t_{o2}, l_{o2}, t_{d2}, l_{d2})$. The notations t and l represent the time stamp (in YYYY-MM-DD HH : MM:SS format) and geographical location for the start and endpoints of each trip. In

this study, 2,572,193 taxi trips were extracted from the original dataset during the study period.

Two commonly used indicators were employed to characterize human mobility patterns by taxi, namely, trip displacement Δr and travel time Δt. In this study, trip displacement refers to the great-circle distance between the origin and destination points of a trip. Equation (1) presents the calculation method of the trip displacement, where R is Earth's radius, and lat and long are the latitude and longitude, respectively. The great-circle distance is the shortest distance between two points on the surface of the Earth. Travel time refers to the total time used for finishing the trip. Considering the geographical scope and measurement accuracy of GPS devices, trips with displacements shorter than 1 km or longer than 100 km were excluded. Similarly, trips with travel times shorter than 5 mins or longer than 3 hours were also removed from the dataset. In total, 2,240,932 trips were considered in the following analyses:

$$\begin{aligned}\Delta r = R \times\ & \arccos\left[\sin\left(\text{lat}_o\right)\sin\left(\text{lat}_d\right)\right. \\ & \left. + \cos\left(\text{lat}_o\right)\cos\left(\text{lat}_d\right)\cos\left(\text{long}_d - \text{long}_o\right)\right].\end{aligned} \tag{1}$$

Power law, exponential, and log-normal distributions are the heavy-tailed distributions that have been commonly used in characterizing human mobility patterns in terms of trip displacement and travel time. The probability density function of the three distributions is given in equations (2)–(4), respectively, where α, λ, σ, and μ are the parameters.

$$p(x) \sim x^{-\alpha}, \tag{2}$$

$$p(x) \sim \exp(-\lambda x), \tag{3}$$

$$p(x) \sim \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \tag{4}$$

In this study, we first fitted these distributions to the data, and then model selections were made concerning the AIC criterion [38]. The model with the smallest AIC value was then selected as the best-fitted model. For the power law model, the methodology described in Clauset et al. (2009) [39] was adopted via the package *poweRlaw* in the *R* statistical software. The parameters of the exponential and log-normal distributions were estimated with the maximum-likelihood method, using *fitdistrplus* package in *R*.

To provide a preliminary understanding of the data, Figure 1 illustrates the spatial distributions of the origin and destination points of the taxi trips within the urban area of Beijing. It is noticeable that most of these points are scattered within the city boundary while only a few points are located outside of the city boundary. It clearly shows the monocentric city structure. Those points also reflect the shape of the traffic road network in Beijing, to some extent.

Figure 2 illustrates the hourly dynamics of the number of taxi trips on different days of a week. On weekdays, a bimodal pattern is observed which is significantly different than the pattern on weekends. In general, the number of taxi trips on weekdays is greater than that on weekends. On weekdays, the number of taxi trips increases steadily to a peak at 9 a.m. since 4 a.m. It then declines slightly until noon and it rises to the section peak at 1 p.m. The number of taxi trips declines steadily until 24 p.m. from 1 p.m. Such a bimodal pattern is in line with the previous studies [1, 40, 41]. In contrast, the hourly dynamics of the number of taxi trips on weekdays exhibit different patterns. On Sunday, the number rises steadily to a peak at 1 p.m. and then fluctuates slightly until 5 p.m. On Saturday, after the peak at 10 a.m., a sudden drop at noon was observed.

2.2. Distributional Characteristics of Trip Displacement and Travel Time. Figure 3 illustrates the observed probability density of trip displacement and travel time. Trip displacement provides an important and widely used indicator to characterize the travel behaviors of taxi trips. By visual observation, the probability density of trip displacement $\mathbf{P}(\Delta\mathbf{r})$ increases until reaching a peak where the trip displacement $\Delta\mathbf{r}$ is about 5 km and then it declines dramatically. For shorter trips, travelers would be less likely to choose taxis while for longer trips travelers have less tendency to choose taxis due to the cost of the trip. Besides, the overall proportion of long trips is low. Therefore, we could observe the dramatic drops of probabilities for long trip displacement and travel time, on the tail part of the probability density function.

The phenomena indicate that some threshold values of trip displacement and travel time exist. Certain physical or social boundaries would limit the usage of taxis for long-distance trips. One reason is that taxi trips are usually for traveling within the city which inherently restricts the trip distance as well as travel time of taxi trips. Another reason would be the monetary cost of taxi trips. There might exist certain psychologically tolerable values of the price of taxi trips for most travelers. The drop for travel time is steeper than that of trip displacement, which means that travel time is a more important factor affecting the choice of taxis than trip displacement. It is realistic because travelers concern travel time of the trip. If the travel time is beyond certain tolerable values, they might switch to other modes.

Specifically, 96.71% of the trips had a displacement of fewer than 20 km. To enhance the flexibility of the model, the probability density function of trip displacement was cut into two regimes at the point of 20 km. The two parts of the function are allowed to be fitted by different models. Table 1 reports the estimated parameters for fitting the distributions of trip displacement and travel time. According to the AIC values, trip displacement is fitted the best by the power law distribution for both the first and tail parts.

Travel time is widely used as another basic indicator for reflecting human mobility patterns. It is highly related to trip displacement but with a high level of uncertainty due to the interference of traffic conditions, driving behaviors of taxi drivers, and the choice of path. Figure 3 illustrates the probability density function of travel time $\mathbf{P}(\Delta\mathbf{t})$. Similar to that of the trip displacement, the two-regime pattern was also observed. The function increases steadily until the travel time reaches 20 minutes and it drops dramatically for travel time greater than 20 minutes. For the region with a very

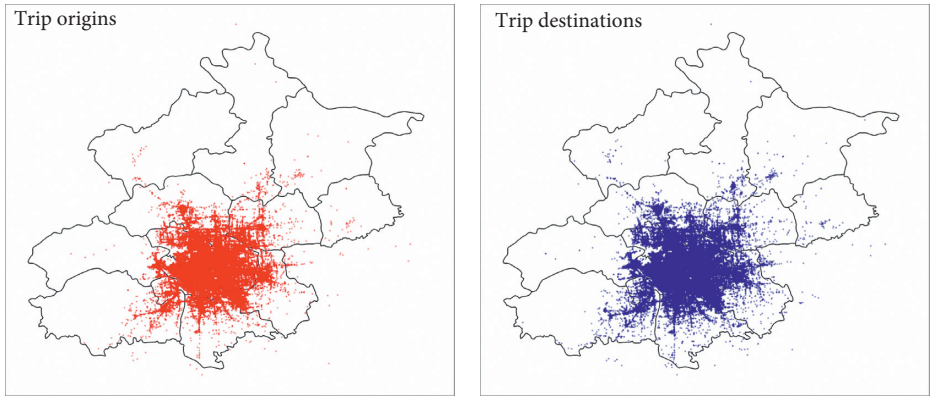

Figure 1: Spatial distributions of the origin and destination points of the taxi trips.

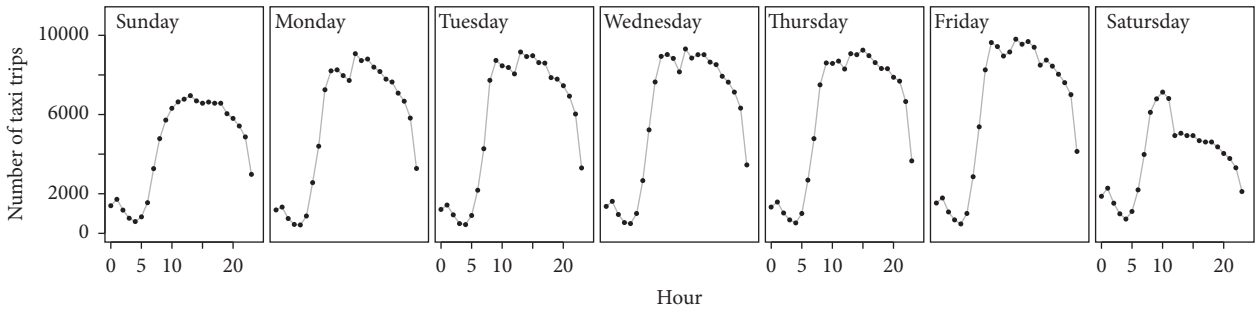

Figure 2: Hourly dynamics of the number of taxi trips on different days of a week.

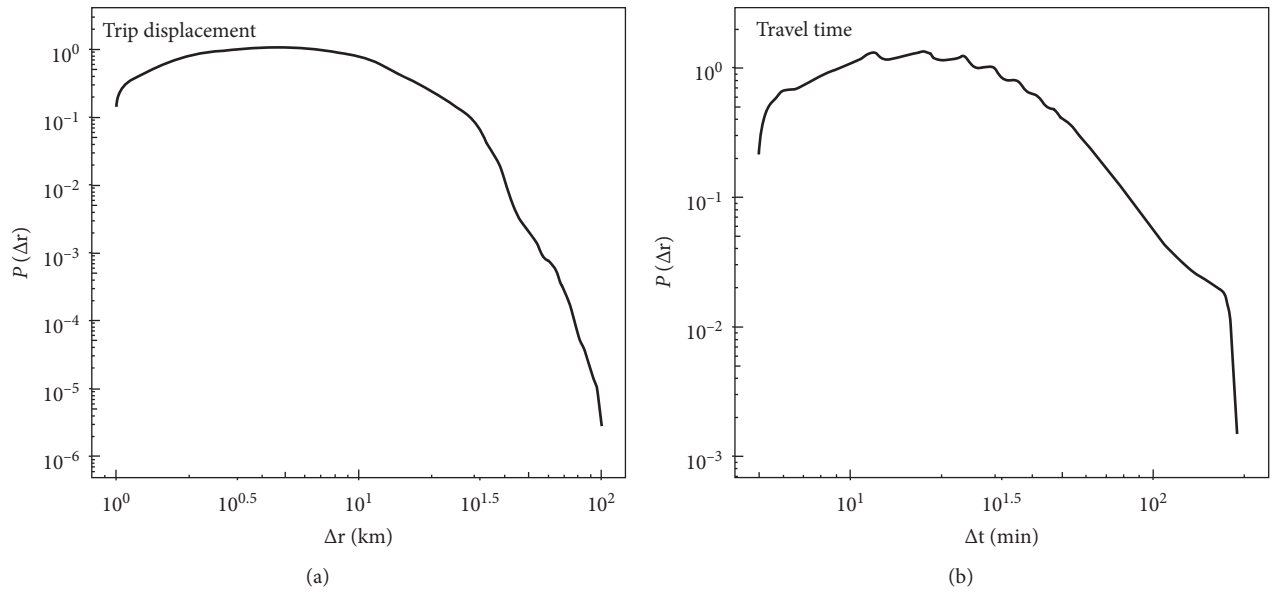

Figure 3: Observed probability density of trip displacement and travel time.

short travel time, a sharp increase in the function is observed. People tend to travel by other modes, e.g., bicycle and pedestrians for trips with very short travel time. There are 96.27% of trips with travel time less than 60 minutes. Likewise, the probability density function of travel time was cut into two regimes at the point of 60 minutes. According to the AIC values, travel time is fitted the best by the power law distribution for the tail part and the exponential distribution

Table 1: Estimated parameters for fitting the distributions of trip displacement and travel time.

Distribution	Power law (α)		Exponential (λ)		Log-normal (μ, σ)	
	First part	Tail part	First part	Tail part	First part	Tail part
Trip displacement	4.017	8.294	0.176	0.039	1.490 0.716	3.229 0.182
Travel time	3.883	4.262	0.051	0.011	2.794 0.601	4.401 0.269

for the first part, meaning that the probability density decayed more quickly than that of a power law. This result is consistent with Liang et al. (2012) [9].

3. Methods of Topic Modeling

3.1. Latent Dirichlet Allocation Modeling. As one of the most commonly used topic models, the latent Dirichlet allocation (LDA) was proposed originally in Blei et al. (2003) [42] as a Bayesian generative probabilistic framework. In LDA terminology, the entity "word" represents the basic unit of discrete data, a "document" consists of a sequence of N words, and a "corpus" is a collection of M documents. In topic models, each document is modeled as a mixture of topics and each topic is modeled as a distribution of words. The main objectives of LDA inference are to find the probability of a word given each topic k, $p(w = t|z = k)$, and find the probability of a topic given each document m, $p(z = k|d = m)$, where w, d, and z represent word, document, and topic, respectively. As shown in Figure 4, for the LDA model, the generative steps for each document can be summarized as follows:

Step 1: the term distribution for each topic is determined by $\beta \sim \text{Dirichlet}(\delta)$

Step 2: the proportions θ of the topic distribution for the document are determined by $\theta \sim \text{Dirichlet}(\alpha)$

Step 3: for the i^{th} word w_i in a document, choose a topic, $z_i \sim \text{Multinomial}(\theta)$

Step 4: choose a word w_i from a multinomial probability distribution conditioned on the topic z_i: $p(w_i|z_i, \beta)$, where β is the term distribution of topics and contains the probability of a word occurring in a given topic

Generally, there are two commonly used methods for LDA model estimation, i.e., the variational expectation-maximization method [42] and the Gibbs sampling method. In this study, we adopted the Gibbs sampling approach due to its ease of understanding and simple implementation.

3.2. Transformation between the GPS Information and the Text Data. To apply the LDA model to the taxi GPS dataset, an analogic transformation between the GPS information and the text data is required. The latter is the subject in the LDA model. As shown in Table 2, multiday data are analogically treated as a corpus of documents, single-day data as a single document, and different mobility patterns as latent topics. In this way, we may build a "word" containing the key attributes of a trip, such as the day of the week, the hour of the day, trip displacement, and travel time. In this study, we use the combination of the hour of the day, trip displacement, and travel time as a "word" defining a trip.

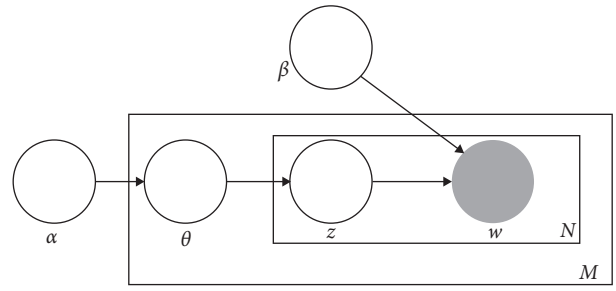

Figure 4: A graphical representation of the latent Dirichlet allocation.

Finding the topic patterns from the word will lead us to the human mobility patterns. Considering the categorical nature of real texts, we transformed the continuous variables (trip displacement and travel time) into categorical variables by partitioning the domain into several successive bins, as shown in Table 3. Each trip is represented as a word in the format of "hour of day + trip displacement + travel time". For example, the word "13 + (15, 20] + (30, 45]" represents a taxi trip that started during the period 13 : 00–14 : 00 with a displacement of $15 < \Delta r \leq 20$ (km) and travel time of $30 < \Delta t \leq 45$ (mins). Prior to the LDA modeling, we removed the words of which the count was less than five times in the vocabulary. Eventually, there are 1316 different words left in the vocabulary. For the analysis, we seek to understand two questions: what types of mobility patterns does the LDA discover? How do the discovered mobility patterns characterize the dataset of a single day? In this study, the LDA modeling was accomplished with the aid of the *topicmodels* package in *R*.

Prior to the estimation of the LDA model, the number of latent topics K has to be predecided. Perplexity has been used as an effective measure to determine the optimal number of latent topics K. Generally, the smaller the perplexity, the better the performance of the LDA model on an unseen dataset. For that purpose, we randomly chose 70% of the data as a training set and the remainder as the test set. Then, we computed the model perplexity on the test documents after training the LDA model, with K ranging from 5 to 100 in increments of 5 ($K = 5, 10, \ldots, 100$). For all the values of K, 1000 iterations of the Gibbs sampling algorithm were performed. LDA hyperparameters (α, β) were set to $\alpha = 50/K$ and $\beta = 0.1$, according to the suggestion of Griffiths and Steyvers (2004) [43]. The perplexity for a particular number of latent topics K can be estimated as follows:

Table 2: The analogy between the GPS information and the text data.

Text data	Corpus	Document	Topic	Word
GPS information	Multiday data	Single-day data	Mobility pattern	Single trip

Table 3: Categorization of continuous variables.

Variables	Min	Max	Interval	Examples of the categories
Trip displacement Δr (km)	1	95	5	(15, 20], (20, 25]
Travel time Δt (min)	4	180	15	(15, 30], (30, 45]

$$\text{perplexity}(K) = \exp[\text{entropy}(K)], \tag{5}$$

$$\text{entropy}(K) = -\sum_{k=1}^{K} p(z_k)\left(\sum_{l} p(w_l|z_k)\log(p(w_l|z_k))\right), \tag{6}$$

$$p(z_k) = \sum_{m=1}^{M} p(z_k|d_m)p(d_m). \tag{7}$$

Here, $p(z_k)$ is the probability of the k^{th} topic and $p(d_m)$ is the probability of the m^{th} document, and l is the index for all words in the vocabulary. The model perplexity over the number of topics is plotted in Figure 5. The perplexity drops dramatically from the beginning until $K = 30$ and then the perplexity stabilizes after $K = 65$. Thus, we chose $K = 65$ as the optimal number of latent topics for the following experiments.

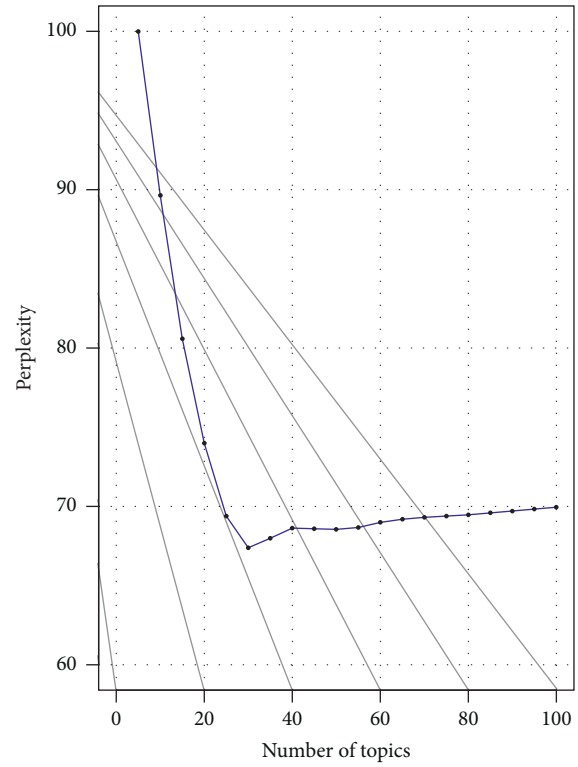

Figure 5: Perplexity versus the number of latent topics.

4. Results and Analyses

The LDA model successfully found meaningful latent topics over different days. The unsupervised discovery of these latent topics revealed different types of mobility patterns, assigning days to different topics and topics to different words with a probability measure. To illustrate the different topics discovered, we ranked the five most probable words for each topic and the five most probable topics for each day, ranked by $P(w|z)$ and $P(z|d)$, respectively.

4.1. Latent Topics. In Table 4, we illustrate some of the discovered mobility patterns and list the top words with corresponding probability $P(w|z)$. Topic 9 captures the mobility patterns for trips during the morning peak hours. The top word for this topic is "8 + (0, 10] + (30, 45]", followed by "7 + (0, 10] + (30, 45]" and "8 + (0, 10] + (15, 30]". This pattern matches commute trips traveling from home to workplaces in morning peak hours. The displacement interval for these top words is $0 < \Delta r \leq 10$ km, indicating that the displacement of taxi trips is relatively small. However, the varying travel time intervals for these top words may be since traffic conditions vary for different hours and different locations within the city boundaries.

Topic 18 captures the mobility patterns at noon. The top word for this topic is "14 + (0, 10] + (0, 15]", followed by "12 + (0, 10] + (15, 30]", which matches the trips from workplaces to nearby restaurants for lunch by taxi. It is observed that the trips started at 14 and 12 have similar trip displacement while different travel times. The trips started at 12 require additional travel time because the traffic condition in the city is worse at noon. There are some lunch trips and business trips around noon.

Topic 28 mainly captured the mobility pattern before noon. The top word for this topic is "11 + (0, 10] + (0, 15]", indicating trips that started before noon with a displacement smaller than 10 km and with travel times shorter than 15 mins. However, the second most probable word for this topic is "11 + (0, 10] + (15, 30]", which can be regarded as indicating the evolution of traffic conditions as travel time gets longer.

Topic 43 mainly captures the mobility pattern during evening peak hours. The top word for this topic is "18 + (10, 20] + (135, 150]", representing taxi trips that were no further than 20 km but the travel time is extremely long during rush hours. The second most probable word is "19 + (10, 20] + (15, 30]" reflecting better traffic conditions.

Table 4: Examples of the discovered interesting mobility patterns.

Words	$P(w\|z)$	Words	$P(w\|z)$
Topic 9		Topic 18	
8 + (0, 10] + (30, 45]	0.10764	14 + (0, 10] + (0, 15]	0.08191
7 + (0, 10] + (30, 45]	0.07833	12 + (0, 10] + (15, 30]	0.07604
8 + (0, 10] + (15, 30]	0.07017	13 + (0, 10] + (15, 30]	0.04734
17 + (0, 10] + (15, 30]	0.04827	17 + (0, 10] + (0, 15]	0.04361
9 + (0, 10] + (30, 45]	0.04406	14 + (0, 10] + (15, 30]	0.04282
Topic 28		Topic 43	
11 + (0, 10] + (0, 15]	0.26593	18 + (10, 20] + (135, 150]	0.01717
11 + (0, 10] + (15, 30]	0.16581	19 + (20, 30] + (15, 30]	0.01163
10 + (0, 10] + (15, 30]	0.06649	15 + (10, 20] + (90, 105]	0.01163
16 + (0, 10] + (0, 15]	0.04220	1 + (10, 20] + (90, 105]	0.00609
10 + (0, 10] + (0, 15]	0.03817	10 + (0, 10] + (45, 60]	0.00609
Topic 27		Topic 30	
6 + (0, 10] + (75, 90]	0.01217	4 + (20, 30] + (75, 90]	0.02014
6 + (10, 20] + (135, 150]	0.01217	3 + (10, 20] + (120, 135]	0.01523
0 + (0, 10] + (120, 135]	0.00637	18 + (30, 40] + (30, 45]	0.01523
0 + (0, 10] + (135, 150]	0.00637	7 + (0, 10] + (120, 135]	0.01523
0 + (10, 20] + (105, 120]	0.00637	8 + (0, 10] + (150, 165]	0.01523

Topics 27 and 30 capture the mobility pattern during the early morning. The top word for Topic 30 is "4 + (20, 30] + (75, 90]", probably representing the mobility of the so-called "early birds". However, the top words for Topic 27 are "6 + (0, 10] + (75, 90]" and "6 + (10, 20) + (135, 150]", which can be interpreted as the traffic conditions evolving more slowly than with the top word of Topic 30. Besides, these two topics can also represent trips that cover medium displacement but with extremely long travel times.

Topic 59 captures trips over the period 21 : 00–22 : 00, which might correspond to trips back home at night. Topic 54 captures trips taken in bad traffic conditions, as the top words represented trips with medium displacement but extremely long travel times. Topic 26 captures trips that started at 9 a.m., with short displacements and the travel time around 60 minutes.

These selected topics are representative of the meaningful topics discovered. Some of the other interesting topics are also listed below. However, we should also note that the probability of the top words in the topics shown above is not very large, indicating that the topics are not discriminant of featured patterns and this is possibly due to the relatively large vocabulary in this study.

4.2. Daily Patterns. The LDA method also allows the extraction of daily patterns according to the probability assigned to different topics and it is meaningful to explore the relationships among these daily patterns. For that purpose, the dot-product method was adopted as a measure of similarity. First, we extracted the probabilities for each day as a vector, $x_i = (x_{i1}, x_{i2}, \ldots, x_{i65})$, where the k^{th} ($k = 1, 2, 3 \ldots, 65$) element in this vector x_{ik} represents the probability of topic k assigned to i^{th} day and $\sum_{k=1}^{65} x_{ik} = 1$, $i = 1, 2, 3, \ldots, 14$. In this way, the similarities s_{i-j} between two different daily patterns (x_i and x_j) can be calculated as shown in equation (8) where "·" represents the dot-product. The higher the value of s_{i-j}, the more similar the two daily patterns, x_i and x_j.

$$s_{i-j} = \frac{x_i \cdot x_j}{\|x_i\|_2 \times \|x_j\|_2},$$
$$\|x_i\|_2 = \sqrt{\sum_{k=1}^{65} x_{ik}^2}. \tag{8}$$

We first examined the similarities between different daily patterns, according to the day of the week. With the dot-product method, the weekly similarities between the same days in different weeks were found to be significant, with $s_{i-j} \approx 1$. There is strong evidence for weekly regularities in daily taxi mobility. It was also noticed that the similarities for the weekdays were larger than those for the weekends, reflecting the fact that the weekend daily mobility pattern distributions were more diverse than those on weekdays.

We also explored similarities among different days in the same week. As shown in Figure 6, there were stronger similarities among weekdays, and the similarities between weekdays and weekends were weaker. That is, the topic distributions of the weekdays differ significantly from those of the weekends, which can be interpreted as the result of different daily living styles on weekdays and weekends. This is consistent with our common understanding and also shows the capability of the LDA model to characterize daily mobility patterns.

According to the generative framework of the topic model, a document can be viewed as a mixture of topics. In Table 5, we list the top five topics for each day, as well as the corresponding probabilities. As can be seen, these top topics account for nearly 70% of the probability mass each day. That is, the daily mobility patterns can be represented by only a few topics. Topic 44 was the top topic for all these days with differing probability measures. Also, Topic 44 is not very discriminating (Table 6), but it can be interpreted as a

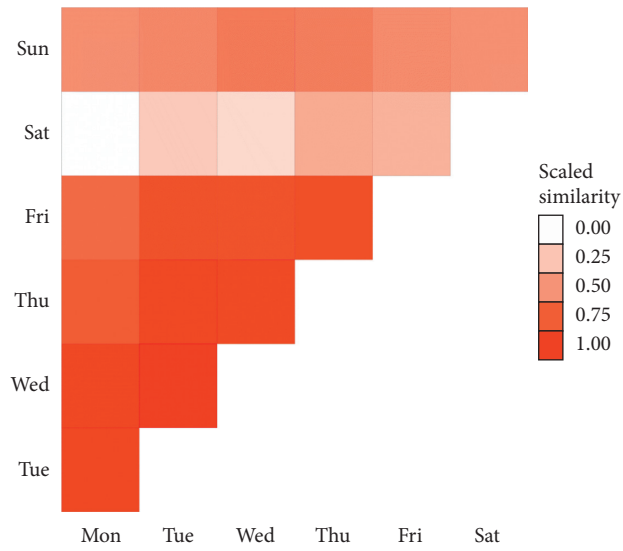

Figure 6: Similarities among different days in the same week.

Table 5: Top topics for different days.

Topic	$P(z\|d)$	Topic	$P(z\|d)$	Topic	$P(z\|d)$
Day 1		Day 2		Day 3	
44	0.21	44	0.24	44	0.22
33	0.19	33	0.18	33	0.17
63	0.12	41	0.10	63	0.11
41	0.11	63	0.09	41	0.11
18	0.09	18	0.09	18	0.09
Day 4		Day 5		Day 6	
44	0.24	44	0.22	44	0.22
33	0.17	41	0.16	18	0.15
63	0.12	33	0.15	63	0.14
18	0.10	18	0.10	33	0.11
41	0.08	63	0.10	5	0.09
Day 7		Day 8		Day 9	
44	0.19	44	0.22	44	0.22
63	0.19	33	0.19	33	0.17
18	0.15	63	0.13	63	0.12
33	0.12	18	0.10	41	0.09
41	0.06	41	0.09	18	0.09
Day 10		Day 11		Day 12	
44	0.22	44	0.26	44	0.26
33	0.18	63	0.15	33	0.16
63	0.10	33	0.14	63	0.15
41	0.09	18	0.10	41	0.13
18	0.09	41	0.09	18	0.09

Table 6: Top words for Topic 44.

Word	$P(w\|z_{44})$
20 + (0, 10] + (0, 15]	0.071107
16 + (0, 10] + (0, 15]	0.063276
8 + (0, 10] + (0, 15]	0.060526
17 + (0, 10] + (15, 30]	0.052038
14 + (0, 10] + (15, 30]	0.051296
9 + (0, 10] + (15, 30]	0.042155
22 + (0, 10] + (0, 15]	0.040981
23 + (0, 10] + (0, 15]	0.038269
15 + (0, 10] + (0, 15]	0.036135
10 + (0, 10] + (15, 30]	0.033785

basis for daily patterns. Thus, the daily patterns are similar to each other, due to the presence of the basic topics.

5. Conclusions

Understanding human mobility is of great importance for various applications, such as urban planning and traffic forecasting. Advances in data collection and analytics have shed light on new methods to explore human mobility with high spatiotemporal resolution. In this study, using a massive GPS data set collected by taxis in Beijing over two weeks, we explored both the macroscopic and microscopic characteristics of human mobility by taxi. A two-peak pattern in the number of taxi trips was found to be universal for both weekdays and weekends. In particular, we used the LDA model to find latent mobility patterns in an unsupervised manner. The findings showed that the mobility pattern followed a power law for macroscopic characteristics. Regarding the microscopic characteristics, the daily mobility patterns were similar to each other and could be represented well by only a few top mobility patterns.

The contribution of this paper can be summarized by the following two explanations. First, we introduced a framework for characterizing mobility patterns using the LDA method in a detailed way, which is beneficial for understanding latent mobility patterns. Second, we provide evidence to support the idea that the scaling law for human mobility is a power law.

In future work, we have several ideas for extending this study. We would like to mix weekday attributes into the current word combination to investigate weekly mobility patterns over a longer period. Other topic models based on LDA could also be applied. One possible choice may be the Author Topic Model (ATM) [44], where we would treat the weekday as the author of a text document and explore the differences in the authors' writing styles. It is also potential to use the GPS data from ride-sourcing trips [45, 46] as they might have larger sample sizes. Finally, we would also like to implement topic models for other kinds of mobility data sets, such as mobile phones and smart card data.

Data Availability

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Key R&D Program of China (no. 2019YFF0301400) and National Natural Science Foundation of China (no. 71971023).

References

[1] T. Louail, M. Lenormand, O. G. Cantu Ros et al., "From mobile phone data to the spatial structure of cities," *Scientific Reports*, vol. 4, no. 5276, p. 5276, 2014.

[2] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th ACM International Conference on Ubiquitous Computing*, pp. 89–98, Beijing, China, September 2011.

[3] J. Yuan, Y. Zheng, and X. Xie, "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*, pp. 186–194, Beijing, China, August 2012.

[4] W.-S. Jung, F. Wang, and H. E. Stanley, "Gravity model in the Korean highway," *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48005, 2008.

[5] S. Goh, K. Lee, J. S. Park, and M. Y. Choi, "Modification of the gravity model and application to the metropolitan Seoul subway system," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 86, no. 2, Article ID 26102, 2012.

[6] Y. Liu, C. Lyu, A. Khadka, W. Zhang, and Z. Liu, "Spatiotemporal ensemble method for car-hailing demand prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 12, pp. 5328–5333, 2020.

[7] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[8] X. Y. Yan, X. P. Han, B. H. Wang, and T. Zhou, "Diversity of individual mobility patterns and emergence of aggregated scaling laws," *Scientific Reports*, vol. 3, pp. 2678–2685, 2013.

[9] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu, "The scaling of human mobility by taxis is exponential," *Physica A: Statistical Mechanics and Its Applications*, vol. 391, no. 5, pp. 2135–2144, 2012.

[10] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini, "Statistical laws in urban mobility from microscopic GPS data in the area of Florence," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, no. 5, p. P05001, 2010.

[11] C. Kang, X. Ma, D. Tong, and Y. Liu, "Intra-urban human mobility patterns: an urban morphology perspective," *Physica A: Statistical Mechanics and Its Applications*, vol. 391, no. 4, pp. 1702–1717, 2012.

[12] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[13] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010a.

[14] B. C. Csáji, A. Browet, V. A. Traag et al., "Exploring the mobility of mobile phone users," *Physica A: Statistical*

Mechanics and its Applications, vol. 392, no. 6, pp. 1459–1473, 2013.

[15] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[16] J. Tang, F. Liu, Y. Wang, and H. Wang, "Uncovering urban human mobility from large scale taxi GPS data," *Physica A: Statistical Mechanics and Its Applications*, vol. 438, pp. 140–153, 2015.

[17] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Y. Ma, "Understanding mobility based on GPS data," in *Proceedings of the 10th ACM international conference on Ubiquitous computing*, pp. 312–321, Seoul, Korea, September 2008.

[18] H. Cai, X. Zhan, J. Zhu, X. Jia, A. S. F. Chiu, and M. Xu, "Understanding taxi travel patterns," *Physica A: Statistical Mechanics and its Applications*, vol. 457, no. 1, pp. 590–597, 2016.

[19] Z. Zheng, S. Rasouli, and H. Timmermans, "Two-regime pattern in human mobility: evidence from GPS taxi trajectory data," *Geographical Analysis*, vol. 48, no. 2, pp. 157–175, 2016.

[20] X. Hu, S. An, and J. Wang, ""Exploring urban taxi drivers' activity distribution based on GPS data," *Mathematical Problems in Engineering*, vol. 2014, Article ID 708482, 13 pages, 2014.

[21] C. Peng, X. Jin, K. C. Wong, M. Shi, and P. Liò, "Collective human mobility pattern from taxi trips in urban area," *PLoS One*, vol. 7, no. 4, Article ID e34487, 2012.

[22] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 363–381, 2014.

[23] N. Davis, G. Raina, and K. Jagannathan, "Taxi demand forecasting: a hedge-based tessellation strategy for improved accuracy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 11, pp. 3686–3697, 2018.

[24] N. Davis, G. Raina, and K. Jagannathan, "A framework for end-to-end deep learning-based anomaly detection in transportation networks," *Transportation Research Interdisciplinary Perspectives*, vol. 5, Article ID 100112, 2020.

[25] Z. Liu, Y. Liu, C. Lyu, and J. Ye, "Building personalized transportation model for online taxi-hailing demand prediction," *IEEE Transactions on Cybernetics*, pp. 1–9, 2020.

[26] F. Ali, D. Kwak, P. Khan et al., "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowledge-Based Systems*, vol. 174, pp. 27–42, 2019.

[27] Z. Cheng, M. Trépanier, and L. Sun, "Probabilistic model for destination inference and travel pattern mining from smart card data," *Transportation*, pp. 1–19, 2020.

[28] S. Das, A. Dutta, and M. A. Brewer, "Case study of trend mining in transportation research record articles," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2674, no. 10, pp. 1–14, 2020.

[29] Y. Sun and S. Kirtonia, "Identifying regional characteristics of transportation research with transport research international documentation (TRID) data," *Transportation Research Part A: Policy and Practice*, vol. 137, pp. 111–130, 2020.

[30] B. Qi, A. Costin, and M. Jia, "A framework with efficient extraction and analysis of twitter data for evaluating public opinions on transportation services," *Travel Behaviour and Society*, vol. 21, pp. 10–23, 2020.

[31] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 49–66, 2017.

[32] K. D. Kuhn, "Using structural topic modeling to identify latent topics and trends in aviation incident reports," *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 105–122, 2018.

[33] X. Wei and W. B. Croft, ""LDA–based document models for ad–hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178–185, Berkeley, CA, USA, August 2006.

[34] Z. Li, C. Wang, X. Xie, X. Wang, and W. Y. Ma, ""Exploring LDA–based document model for geographic information retrieval," in *Advances in Multilingual and Multimodal Information Retrieval, Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 842–849, Budapest, Hungary, September 2007.

[35] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, pp. 1–27, 2011.

[36] R. Montoliu, "Discovering mobility patterns on bicycle–based public transportation system by using probabilistic topic models," *Ambient Intelligence-Software and Applications*, vol. 153, pp. 145–153, 2012.

[37] E. Côme, N. A. Randriamanamihaga, L. Oukhellou, and P. Aknin, "Spatio–temporal analysis of dynamic origin–destination data using latent Dirichlet allocation: application to Vélib' bike sharing system of Paris," in *Proceedings of 93rd Annual Meeting of Transportation Research Board*, Washington, NJ, USA, January 2014.

[38] K. P. Burnham and D. R. Anderson, "Multimodel inference," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, Nov. 2004.

[39] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[40] N. Breyer, C. Rydergren, and D. Gundlegård, "Comparative analysis of travel patterns from cellular network data and an urban travel demand model," *Journal of Advanced Transportation*, vol. 2020, Article ID 3267474, 17 pages, 2020.

[41] X. Liu, L. Sun, Q. Sun, and G. Gao, "Spatial variation of taxi demand using GPS trajectories and POI data," *Journal of Advanced Transportation*, vol. 2020, Article ID 7621576, 20 pages, 2020.

[42] D. Blei, M. A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[43] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, 2004.

[44] M. Rosen–Zvi, T. Griffiths, M. Steyvers, and P. Smyth, ""The author–topic model for authors and documents, "" in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 487–494, Banff, Canada, July 2004.

[45] D. Sun and X. Ding, "Spatiotemporal evolution of ride-sourcing markets under the new restriction policy: a case study in Shanghai," *Transportation Research Part A: Policy and Practice*, vol. 130, pp. 227–239, 2019.

[46] F. Chen, Z. Yin, Y. Ye, and D. Sun, "Taxi hailing choice behavior and economic benefit analysis of emission reduction based on multi-mode travel big data," *Transport Policy*, vol. 97, pp. 73–84, 2020.