Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 7375327, 1 page
https://doi.org/10.1155/2020/7375327

Retraction

Retracted: Understanding Regional Mobility Patterns Using Car-Hailing Order Data and Points of Interest Data

Journal of Advanced Transportation

Received 28 May 2020; Accepted 28 May 2020; Published 16 July 2020

Journal of Advanced Transportation has retracted the article titled "Understanding Regional Mobility Patterns Using Car-Hailing Order Data and Points of Interest Data" [1]. Following publication of the article, the authors identified that the method was verified only by comparison of the model and statistical results. A more robust analysis was required to substantiate the conclusions made in the article, and it is therefore being retracted at the request of the authors and with the agreement of the editorial board.

References

[1] Z. Zhang, Y. Chen, J. Xiong, and T. Liang, "Understanding Regional Mobility Patterns Using Car-Hailing Order Data and Points of Interest Data," *Journal of Advanced Transportation*, vol. 2020, Article ID 1410808, 13 pages, 2020.

Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 1410808, 13 pages
https://doi.org/10.1155/2020/1410808

Research Article

Understanding Regional Mobility Patterns Using Car-Hailing Order Data and Points of Interest Data

Zheng Zhang,[1] **Yanyan Chen**,[1] **Jie Xiong**,[1] **and Tianwen Liang**[2]

[1]*College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China*
[2]*Research Institute of Highway Ministry of Transport, No. 8, Road Xitucheng, Haidian District, Beijing 100088, China*

Correspondence should be addressed to Yanyan Chen; cdyan@bjut.edu.cn and Tianwen Liang; tw.liang@rioh.cn

Received 28 August 2019; Accepted 18 January 2020; Published 18 February 2020

Academic Editor: Rocío de Oña

Car hailing is undergoing rapid global development, thereby providing new opportunities and challenges to operators and transport engineers due to uneven or irregular demand in certain areas. To date, only a limited number of studies have analyzed regional mobility patterns or anomaly detection. This study therefore proposes a methodology for recognizing regional mobility patterns using car-hailing order datasets and point of interest datasets. More specifically, we detect regional mobility patterns by incorporating regional intrinsic properties to a hierarchical mixture model termed latent Dirichlet allocation (LDA). This model can simulate the process of generating car-hailing order data and yield regional mobility patterns from spatial, temporal, and spatiotemporal perspectives. Moreover, by combining the trained results with future mobility records, we can measure similarities between areas and detect anomalous areas by calculating the perplexity. We also implement our workflow on a real-word car-hailing order dataset and reveal that it is possible to identify areas with similar or anomaly mobility patterns. This research will contribute to the design of regional transportation policies and customized bus services.

1. Introduction

With the rapid development of information and communication technologies, many online car-hailing service platforms, such as Didi, Uber, and Lyft, have experienced rapid global growth and led to significant changes in people's lifestyles and travel behavior [1–3]. For example, in Beijing, the number of registered drivers of a car-hailing company reached 27,187 in 2018 [4]. Compared to traditional taxi services on the street, online car hailing provides a complete door-to-door service with the advantages of easy payment, comfort, and minimal waiting times. Moreover, large spatiotemporal datasets such as GPS trajectory and operation order data are generated from people's transport behavior, which provides an opportunity to investigate car-hailing mobility patterns.

Despite being a popular and convenient service, car hailing inevitably has some limitations; for example, cars will not always be available for passengers, especially during rush hours or in bad weather, whereas some locations or times will see many drivers looking for passengers and few people requiring rides. Therefore, a regionally oriented management policy or scheduling plan is essential to alleviate this issue. Although detection of the regional patterns and anomalies of car-hailing trips is a challenging task, it is essential to allow service providers and transport planners to predict long-term land use characteristics.

Many previous studies of mobility patterns have relied primarily on large-scale spatiotemporal datasets. Such datasets include detailed call records [5, 6], in-vehicle GPS data [7, 8], transit smart card transactions [9, 10], and Wi-Fi data [11]. As these datasets exhibit heterogeneity and high dimensionality, statistical learning methods such as cluster analysis and matrix/tensor decomposition are adopted to investigate mobility patterns. For example, Kang and Qin [12] analyzed taxicab operation patterns using a matrix factorization method and classified typical taxi demand and supply regions. In addition, Demissie et al. [13] applied a fuzzy c-means clustering algorithm to categorize locations with the same features using cell phone data instead of car

trips. They then identified the patterns and intensities of urban activities with similar features. They used a cell-based method to extract dynamic traffic information and identify the bottleneck in a netwide scale. Furthermore, Yong et al. [14] employed matrix factorization and correlation analysis to extract some of the stable/occasional components of human movement patterns in the Beijing subway. However, large-scale datasets and issues of sparsity and high-dimensionality may distort the results [15]. Moreover, car-hailing order data exhibit spatiotemporal dependence, and the temporal mobility profile is the result of all regional data properties combined [16]. Specifically, the majority of trips depart from residential areas during morning peak hours, whereas the central business district (CBD) is the main source of passengers during afternoon peak hours [17]. These two problems must be considered when investigating mobility patterns in a real-world spatiotemporal dataset.

To tackle the above issues, hierarchical mixture models (such as topic models) have been designed to capture the structure of spatiotemporal mobility patterns. More specifically, hierarchical mixture models define the underlying pattern from a collection of data points with respect to its probability distribution over a set of predefined latent variables. Sun and Axhausen [18] proposed an approach to model large-scale human mobility spatiotemporal data in a probabilistic setting and investigate multidimensional mobility interactions using several latent variables. In addition, Hasan and Ukkusuri [19] proposed a generative method based on a topic model to classify individual activity patterns. The algorithm defined each entry as a combination of several attributes, which resulted in a large vocabulary size and ignored the interactions among attributes. By analyzing a real-world driving behavior dataset, Qi et al. [16] revealed the underlying driving styles in a probabilistic framework based on a topic model. Furthermore, Fan et al. [20] detected individual mobility patterns using separate topic models for the day, time of day, and location dimensions. Probabilistic models can overcome the sparsity problems of spatiotemporal datasets, and in this context, Matsubara et al. [21] detected web-click log patterns with a tensor topic model framework. However, probabilistic models in a matrix/tensor factorization framework are widely used data imputation.

Overall, although there is increasing interest in capturing the underlying structure and patterns within a human mobility dataset, the following limitations remain: (1) the regional (i.e., traffic analysis zone scale) temporal mobility patterns of car-hailing riders, which can provide more macroscopic insights into car dispatching, have not been fully evaluated and (2) the detection of regional mobility patterns requires improvement, especially by the incorporation of regional intrinsic properties to enhance pattern interpretation.

To address these challenges, this study proposes a probabilistic methodology that can extract hidden patterns and explore the combined spatial and temporal patterns in car-hailing trips and then detect anomalies based on the obtained patterns. More specifically, we incorporate point of interest (POI) data as the intrinsic properties of traffic analysis zones (TAZs) into a two-dimensional latent Dirichlet allocation (LDA) model. In addition, an efficient collapsed Gibbs sampling method is developed for statistical inference of the two-dimensional probabilistic model. Furthermore, the effectiveness of the algorithm is finally illustrated using a real-world car-hailing order dataset. The combined spatial and temporal patterns of a TAZ can be depicted using this hierarchical mixture probabilistic model, and the trained result reveals hidden mobility patterns and detects anomalies at the TAZ scale. This study therefore constitutes an important contribution to the literature since a method is developed that combines temporal, local intrinsic attributes to unravel regional mobility patterns, and subsequently, the mined results are validated by studying the mobility patterns of routine users.

The remainder of this paper is organized as follows. Initially, we present the proposed method for detecting regional car-hailing mobility patterns and anomalies. The results of our empirical analysis are then discussed, and finally, we present the conclusions and implications for transport planners.

2. Probabilistic Model for Detecting Hidden Mobility Patterns

2.1. Background and Notation. Departure and arrival trips derived from car-hailing order datasets are defined as mobility patterns in a TAZ. The ability to reproduce future mobility records using uncovered hidden variables is defined as anomaly detection. In other words, anomaly TAZs are characterized by irregular mobility patterns or hard to predict departure/arrival trips [22]. Based on this, we develop a hierarchical mixture model, which incorporates POI data into a two-dimensional LDA model, for uncovering hidden mobility patterns and detecting anomalous zones. The LDA model was originally proposed by Blei et al. [23] and has since been widely used in the fields of text mining [24], imagine classification [25], activity inference [19], and behavior recognition [26], among others. LDA models are generative models that can specify a probabilistic process for generating discrete datasets (e.g., documents, spatial-temporal datasets, and behavior datasets). LDA models have powerful skill in mining latent topics from a discrete dataset. Based on this, a spatial-temporal dataset needs to be discretized to the "corpus-document-word" form for mining latent topics, which means that departure and arrival trips in a TAZ are different words, and all these trips constitute a document. Trips across all the TAZs constitute a corpus.

By analogizing the car-hailing order data to the "corpus-document-word" form, we first map all variables in the car-hailing order records into different categories. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$ denote a car-hailing order, where $m = 1, \ldots, M$ indicates the field index of the order record (i.e., M is the number of fields; each index indicates a specific field such as pick-up location) and $d = 1, \ldots, D$ indicates the

different attribute indices of field m. Thus, we define $x_{md} \in \{1, \ldots, D\}$ as discrete values for field m, beginning from a value of 1. We use $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to represent the entire spatial-temporal car-hailing order dataset, with $i = 1, \ldots, n$ denoting the index of each record. With these notations, the entire car-hailing order data can be reorganized into a "corpus-document-word" format where each record in the dataset is regarded as a document and all attributes are categorized into spatial and temporal words. All records together comprise the entire corpus.

A flowchart of the proposed methodology is shown in Figure 1. Firstly, the POI dataset and discretized car-hailing order dataset are aggregated at the TAZ level. Subsequently, a two-dimensional LDA model is trained using the historical travel information of the selected study area. The concept of perplexity is adopted to measure the performance of the trained model. By combining the result of the trained LDA model and the future dataset, TAZ anomaly detection is implemented using predictive perplexity. The TAZs with similar mobility patterns can be effectively identified by measuring the similarity of the distributions of both spatial and temporal words between each pair of TAZs.

2.2. Model Specification. The probabilistic generation process of a spatial or temporal word in a TAZ begins by assuming that TAZs are represented as random mixtures over latent spatial and temporal topics, where each spatial/temporal topic is characterized by a distribution over the spatial/temporal word. Considering that TAZ topics are the products of both intrinsic properties and mobility patterns, we incorporate local POI information into a two-dimensional LDA model.

For each TAZ, let α be the prior parameter for the Dirichlet document-topic distribution. β and γ are the prior parameters for the Dirichlet temporal topic-word and spatial topic-word distribution, respectively. We assume that there are J temporal topics and K spatial topics. ψ is a $J \times V^t$ matrix where V^t represents the number of different temporal words. Similarly, φ is a $K \times V^s$ matrix where V^s represents the number of different spatial words. Each $\psi_{tj}(\varphi_{sk})$ is a distribution over the temporal/spatial vocabulary. The topic proportions for the h-th taz are θ_h, where θ_{hjk} is the topic proportion for topic (j,k) in the h-th taz. The topic assignments for the h-th taz are z_h, where $z^t_{h,g}(z^s_{h,g})$ represents the topic assignment for the g-th temporal/spatial word in the h-th taz. Finally, the observed words for taz are w^t_h and w^s_h. The number of arrival and departure trips in a taz can be labeled as N_{taz}. The graphical model is shown in Figure 2, and the probabilistic process for generating the spatio-temporal topic model is as follows:

(1) For each topic J, K,

Draw $\lambda \sim N(0, \sigma^2)$
Draw the spatial word distribution for each spatial topic $\varphi_k \sim \text{Dirichlet}_S(\gamma)$
Draw the temporal word distribution for each temporal topic $\psi_j \sim \text{Dirichlet}_T(\beta)$

(2) For the h-th taz,

Let $\alpha_h = \exp(p_h^T \lambda)$
Draw the topic distribution for taz $\theta_h \sim \text{Dirichlet}(\alpha_h)$
For the i-th mobility pattern in the h-th taz:
Draw a topic $z_h \sim \text{multinomial}_{J\times K}(\theta_h)$
Let $z^s_{h,i} = \text{mod}(z_h, K)$
Let $z^t_{h,i} = (z_h - z^s_{h,i})/J$
Draw a word $w^t_i \sim \text{multinomial}_J(\psi^t_j)$
Draw a word $w^s_i \sim \text{multinomial}_K(\varphi^s_k)$

where N is the Gaussian distribution with σ as a hyperparameter and λ is a vector with the same length as the POI vector.

In contrast to the standard LDA model, the hyperparameter α is assigned to a specific TAZ based on the observed POI features of each region. Thus, the values of α vary for different combinations of POI category distributions. Therefore, hidden car-hailing mobility patterns can be determined using both the mobility patterns and POI features.

3. Statistical Inference via Gibbs Sampling

Exact inference for an LDA-like model is difficult; therefore, approximate inference algorithms can be used, such as variational expectation maximization, expectation propagation, and Gibbs sampling [23, 25, 27]. Gibbs sampling is a unique example of a Markov-chain Monte Carlo (MCMC) simulation [28] that often yields a simple algorithm for approximate inference in high-dimensional models such as LDA. Therefore, Gibbs sampling is used in this study for model inference.

Gibbs sampling inherits the stationary behavior of the Markov chain; therefore, one sample x_i is sampled for each transition in the chain after a stationary state has been reached, according to the values of all other dimensions of $\mathbf{x}_i$. To build a Gibbs sampler, the full conditionals must be found (refer to [23] for the detailed Gibbs sampling procedure). In our model, the full conditional distribution is identified using

$$P\left(z^s_i = k,\, z^t_i = j \mid w^s_i = s,\, w^t_i = t,\, \mathbf{z}^s_{\neg i},\, \mathbf{z}^t_{\neg i}\mathbf{w}^s_{\neg i}\mathbf{w}^t_{\neg i}\right)$$

$$\propto \frac{C^{SK}_{sk} + \gamma}{\sum_{s^-} C^{SK}_{s^- k} + K\gamma} \times \frac{C^{TJ}_{tj} + \beta}{\sum_{t^-} C^{TJ}_{t^- j} + J\beta} \times \frac{C^{JK}_{hjk} + \alpha_{jk}}{\sum_{j^-}\sum_{k^-} C^{JK}_{hj^-k^-} + JK\alpha_{jk}}, \tag{1}$$

where C^{TJ}_{tj} represents the number of tokens of temporal word t assigned to topic j, C^{SK}_{sk} represents the number of tokens of spatial word s assigned to topic k, and C^{JK}_{hjk} represents the number of words in the TAZ assigned to topic (j,k). Note that the current instance i is excluded when computing C^{TJ}_{tj}, C^{SK}_{sk}, and C^{JK}_{hjk}. Using this Gibbs sampler, each (z^s_i, z^t_i) in the dataset can be updated sequentially in each iteration. The sampler can reach stationary behavior after a number of iterations. Finally, we obtain the multinomial parameter sets Θ, Φ, and Ψ as follows:

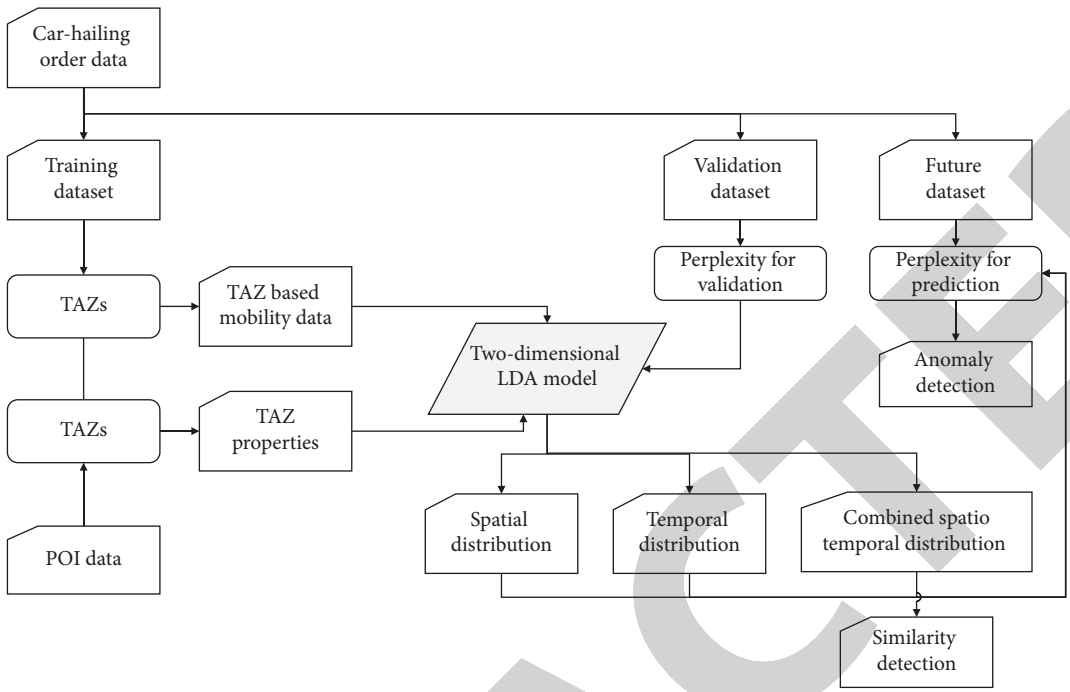


Figure 1: Flowchart of the proposed methodology for analyzing TAZ-based mobility patterns.

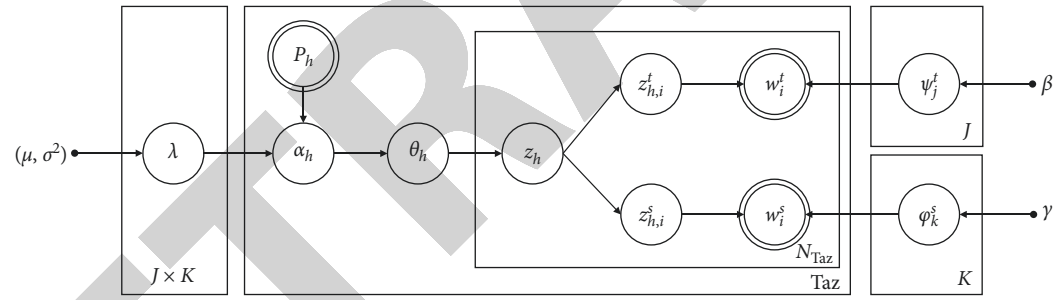


Figure 2: Graphical representation of the model. Single circles indicate random variables, double circles indicate observed variables, boxes indicate replicates of the data, and arrows represent the dependency between entities.

$$\theta_{hjk} = \frac{C^{JK}_{hjk} + \alpha_h}{\sum_{j'}\sum_{k'} C^{JK}_{hj'k'} + JK\alpha_h},$$
$$\psi_{tj} = \frac{C^{TJ}_{tj} + \beta}{\sum_{t'} C^{TJ}_{t'j} + T\beta}, \tag{2}$$
$$\varphi_{sk} = \frac{C^{SK}_{sk} + \gamma}{\sum_{s'} C^{SK}_{s'k} + S\gamma}.$$

$$P\left(\overline{\mathbf{z}}^s_i = k, \overline{\mathbf{z}}^t_i = j \,\middle|\, \overline{\mathbf{w}}^s_i = s, \overline{\mathbf{w}}^t_i = t, \overline{\mathbf{z}}^s_{\neg i}, \overline{\mathbf{z}}^t_{\neg i}\overline{\mathbf{w}}^s_{\neg i}\overline{\mathbf{w}}^t_{\neg i}; \mathbf{z}_{\text{train}}, \mathbf{w}_{\text{train}}\right) \propto$$
$$\cdot \frac{C^{SK}_{sk} + \overline{C}^{SK}_{sk} + \gamma}{\sum_{s^-}\left(C^{SK}_{s^-k} + \overline{C}^{SK}_{s^-k}\right) +)K\gamma} \times \frac{C^{TJ}_{tj} + \overline{C}^{TJ}_{tj} + \beta}{\sum_{t^-}\left(C^{TJ}_{t^-j} + \overline{C}^{TJ}_{t^-j}\right) + J\beta}$$
$$\times \frac{\overline{C}^{JK}_{hjk} + \alpha_{jk}}{\sum_{j^-}\sum_{k^-}\overline{C}^{JK}_{hj^-k^-} + JK\alpha_{jk}}, \tag{3}$$

For an unseen TAZ that does not occur in the training dataset, we can also apply Gibbs sampling to infer its topic composition, $\overline{\theta}_h$. Given a set of training data and the corresponding topic assignment for each car-hailing record from Gibbs sampling $(\mathbf{z}_{\text{train}}, \mathbf{w}_{\text{train}})$, we sample the topic assignment $(\overline{\mathbf{z}}^t_i, \overline{\mathbf{z}}^s_i)$ for each car-hailing record of the TAZ that does not occur in the training dataset as follows:

where C^{TJ}_{tj} represents the number of tokens of temporal word t assigned to topic j in the training dataset and $\overline{C}^{TJ}_{tj}$ is the number of tokens of temporal word t assigned to topic j in the unseen dataset excluding the current instance i. C^{SK}_{sk} and $\overline{C}^{SK}_{sk}$ can be defined in a similar way. This can be used for the calculation of perplexity, which is a measurement of the quality of the model.

3.1. Model Selection. For model selection, we run our algorithm with varying (*J, K*) compositions and compute the perplexity, which identifies the performance of a probabilistic model. This function calculates the average likelihood of observing a test dataset given a set of model parameters. The validation dataset including a randomly selected TAZ is therefore used to calculate the perplexity. More specifically, the perplexity is defined as the exponential of the negative of the average predictive likelihood of a test data [25]:

$$\text{perplexity}\left(\mathbf{w}_h^t, \mathbf{w}_h^s \mid \mathbf{w}_{\text{train}}\right) = \exp\left[-\frac{\ln p\left(\mathbf{w}_h^t, \mathbf{w}_h^s \mid \mathbf{w}_{\text{train}}\right)}{N_{\text{taz}}}\right]. \tag{4}$$

Computing $p(\mathbf{w}_h^t, \mathbf{w}_h^s \mid \mathbf{w}_{\text{train}})$ is possible using

$$p\left(\mathbf{w}_h^t, \mathbf{w}_h^s \mid \mathbf{w}_{\text{train}}\right) = \int \prod_{n=1}^{N_{\text{taz}}}\left[\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{hjk}\psi_{w_{un}^t j}\varphi_{w_{un}^s k}\right] p\left(\theta \mid \mathbf{w}_{\text{train}}\right) \cdot p\left(\psi \mid \mathbf{w}_{\text{train}}\right) p\left(\varphi \mid \mathbf{w}_{\text{train}}\right) \mathrm{d}\theta\, \mathrm{d}\psi\, \mathrm{d}\varphi. \tag{5}$$

This integration is hard to compute; however, an efficient solution is the Monte Carlo simulation. We therefore use M point estimates from the Markov chain and compute the average over M samples:

$$p\left(\mathbf{w}_h^t, \mathbf{w}_h^s \mid \mathbf{w}_{\text{train}}^t, \mathbf{w}_{\text{train}}^s\right) = \frac{1}{M}\sum_{m=1}^{M}\prod_{i=1}^{N_{taz}}\left[\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{hjk}\psi_{w_i^t j}^m\varphi_{w_i^s k}^m\right]. \tag{6}$$

Finally, we apply the average perplexity of the validation dataset to evaluate the performance of the model; thus, the optimal J and K values can be obtained according to the perplexity score.

3.2. Anomaly Detection. Following model inference and selection, the model yields two sets of spatial and temporal patterns that characterize the mobility patterns of each TAZ in the training set. As LDA-like models are mixture models, they use a convex combination of a set of component distributions to model observations. Therefore, future mobility patterns of a TAZ can be reconstructed using the trained spatial and temporal patterns. When future TAZ-based mobility patterns cannot be inferred by the trained latent patterns, we consider that the TAZ is hard to predict; i.e., anomalous mobility patterns are more frequent than in other TAZs. More specifically, the perplexity of a TAZ's future mobility records with respect to its previous mobility records indicates the degree of anomalous mobility patterns:

$$\text{perplexity}\left(\overline{\mathbf{w}}_h^t, \overline{\mathbf{w}}_h^s \mid \mathbf{w}_h^t, \mathbf{w}_h^s\right) = \exp\left[-\frac{\ln\, p\left(\overline{\mathbf{w}}_h^t, \overline{\mathbf{w}}_h^s \mid \mathbf{w}_h^t, \mathbf{w}_h^s\right)}{\overline{N}_{\text{taz}}}\right], \tag{7}$$

where $(\overline{\mathbf{w}}_h^t, \overline{\mathbf{w}}_h^s)$ denotes a set of future car-hailing order records in a TAZ, $(\mathbf{w}_h^t, \mathbf{w}_h^s)$ indicates the observed records in the TAZ, and $\overline{N}_{\text{taz}}$ represents the total number of future records in the TAZ. The conditional probability can be obtained as follows:

$$p\left(\mathbf{w}_h^t, \mathbf{w}_h^s \mid \mathbf{w}_h^t, \mathbf{w}_h^s\right) = \frac{1}{M}\sum_{m=1}^{M}\prod_{i=1}^{N_{\text{taz}}}\left[\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{hjk}\psi_{w_i^t j}^m\varphi_{w_i^s k}^m\right]. \tag{8}$$

The obtained value can be used to measure the intrinsic regularity of mobility patterns in the TAZ. The higher the perplexity is, the more difficult it is to predict future mobility patterns based on historical mobility patterns. In this way, we can determine the reliability of the mobility patterns in the TAZ.

4. Results and Analysis

4.1. Data Description. The data sources used in this study are predominantly car-hailing order data and POI data. A TAZ, which is commonly used for comprehensive urban transportation planning, is regarded as the basic unit of regional mobility pattern analysis. TAZ-based spatiotemporal mobility patterns can be extracted from car-hailing order data, and POI data are regarded as auxiliary information representing the land use characteristics within each TAZ.

4.2. Car-Hailing Order Data. The selected car-hailing order data cover trips from July 6$^{\text{th}}$ to August 20$^{\text{th}}$, 2018, in the metropolitan areas of Beijing, China, and were provided by a large Chinese transportation network company, namely, (TNC)-DiDi Inc. The multiday order datasets have similar trip volumes with an average number of daily trips of 812,371. The order dataset includes the *order ID, passenger ID, pick-up location (longitude, latitude), pick-up time, drop-off location (longitude, latitude), drop-off time, and passenger miles*. The data sample is presented in Table 1. The spatial connections between pick-up or drop-off locations and TAZ are revealed by mapping each record to the TAZ layers on the ArcGIS platform. For the purpose of this study, the pick-up and drop-off locations are labeled with a TAZ code.

As can be seen from Table 1, flawed records (the first record from Table 1) may occur due to pseudotrips, which are trips registered by the TNC test driver that consist of an abnormal distance, time, or missing data. Consequently, order data with a speed (passenger_mile/(off_time-on_time)) of greater than 120 km/h are eliminated. As described previously, we set spatial and temporal identifiers for each record when constructing the spatiotemporal corpus. To detect hidden mobility patterns and anomalies, we discretize the car-hailing dataset to construct the spatiotemporal corpus using these flawed records. An original order record can include direction, time, and distance of a trip. Using trip direction as an example, we employ "1" to indicate departure trips and "2" to indicate arrival trips. The discretized process of these information is described as follows.

Spatial words merge the discretized variables as follows:

(i) *Direction*: "1" indicates that the order departs from this TAZ, whereas "2" indicates that the order arrives at this TAZ.

Table 1: Sample of car-hailing data.

Order ID	Passenger ID	Drop-off lon	Drop-off lat	Pick-up time	Drop-off time	Passenger mile	Pick-up lon	Pick-up lat
6c7c988f71db4	a339f64c5	116.514	39.904	2018/7/14 23:33	2018/7/14 23:33	0.0	116.459	N/A
6e87b62ebbff4	7bc83c7bf	116.674	39.902	2018/7/14 23:51	2018/7/15 0:00	1.6	116.656	39.903
e0fcde3527ab4	01ebc32b7	116.207	40.234	2018/7/14 23:39	2018/7/15 0:00	1.4	116.203	40.242
82cc993d5eae4	185a6f844	116.482	39.921	2018/7/14 23:42	2018/7/15 0:00	9.6	116.584	39.911
9ce350e550bc	43f49c88f3	116.089	39.954	2018/7/14 23:56	2018/7/15 0:00	1.5	116.101	39.956
f60283ca84654	b60b757f9	116.285	39.906	2018/7/14 23:36	2018/7/15 0:00	10.3	116.318	39.880

(ii) *Distance*: "1" indicates a short travel distance (within 3 km), "2" indicates a medium travel distance (3–30 km), and "3" indicates a long travel distance (beyond 30 km).

Temporal words merge the discretized continuous timestamp variables as follows:

(i) *Week*: "1" indicates a weekday and "2" indicates the weekend.

(ii) *Day*: each day is divided into 24-hour-long windows, and the corresponding time window of the pick-up or drop-off timestamp is used as the day identifier.

For the sake of unraveling the combined spatial and temporal hidden mobility patterns at a TAZ scale, we categorized direction and distance information into spatial variables while time information belongs to temporal variables. Here, we regarded spatial variables and temporal variables as spatial words and temporal words, respectively. Table 2 therefore provides an example of the reconstructed spatial-temporal corpus; e.g., the second entry describes a trip that arrived at TAZ_{10101} after a short distance between 08:00 and 09:00 on the weekend. With these data processing procedures, the TAZ-based corpus can be constructed. We consider the data for each order to have three elements, where taz, t, and s indicate the TAZ ID, temporal identifier, and spatial identifier, respectively.

Interpreting TAZ-based car-hailing patterns involves determining multiday mobility patterns from the daily mobility patterns in a TAZ. Car trips can be represented as a distribution of spatial and temporal topics. The defining characteristics and assumptions of the car-hailing mobility pattern inference problem in a TAZ are as follows:

(i) The anomaly detection problem is applicable for a TAZ

(ii) Passenger arrivals or departures within a TAZ are assumed to be independent

(iii) Passenger arrivals at or departures from different TAZs are assumed to be independent

(iv) The travel intensity of a TAZ indicates the total number of departure and arrival trips during a time slot (e.g., 1 hour in this study)

4.3. POI Data. The POI dataset, which was collected by Google Place API, has typically been used to represent land use characteristics [29]. In addition to recording the name and location (longitude and latitude) of each point of interest, the dataset also categorizes POIs into 20 groups, such as administrative agencies, train and metro stations, shopping areas, and residential areas. We employ these POI categories to represent the land use characteristics in a TAZ by computing frequency-inverse document frequency (TF-IDF) [30]. Below is the procedure used to compute the POI composition in a TAZ.

Table 2: Spatiotemporal mobility pattern corpus.

ZoneID	Temporal_Code	Spatial_Code
10101	222	22
10101	208	21

For a given TAZ T_i, a POI vector, $f_i = (v_{i1}, \ldots, v_{ij}, \ldots, v_{iC})$, can be organized, where v_{ij} $(j = 1,\ldots,C)$ is the TF-IDF value of the j-th POI category and C is the number of POI categories. The TF-IDF value v_{ij} is given by

$$v_{ij} = \frac{n_j}{N_i} \times \log \frac{m}{\left\| T_{ij} \right\|}. \tag{9}$$

The TF term is the left part in equation (9), where n_j represents the number of POIs belonging to the j-th category, whereas N_i represents the number of POIs pinpointed in TAZ T_i. The IDF term is the logarithm of the total number of TAZ m divided by the number of TAZs in the j-th POI category. T_{ij} indicates the number of TAZs in the j-th POI category.

In this study, the 3rd Ring Road in Beijing represents the boundary of the study area (Figure 3(a)); car-hailing order data generated within two weeks in the 3rd Ring Road in Beijing plus POI data are used as the input of the model. The model is run during the first week for pattern identification, and the data from the last week are used to compute the predictive perplexity. For the data from the first week, we split the entire car-hailing order dataset into two parts, where 75% of the TAZs are used as a training dataset and the remaining 25% are used as a test dataset. Each TAZ has 3000–5000 car-hailing records, with an average per TAZ of 3,770. We select the optimal number of patterns based on the perplexity values. For model selection, we run our algorithm on a grid with $J = 3, 4, 5, 6$ and $K = 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ and compute the average perplexity for the test dataset. Finally, we obtain the best performance of our model with the lowest average perplexity when $J = 3$ and $K = 9$ (Table 3), and so the analysis is performed using these parameters.

4.4. Mobility Pattern Annotation. Each TAZ typically covers one or more spatial and temporal topics, and each trip is

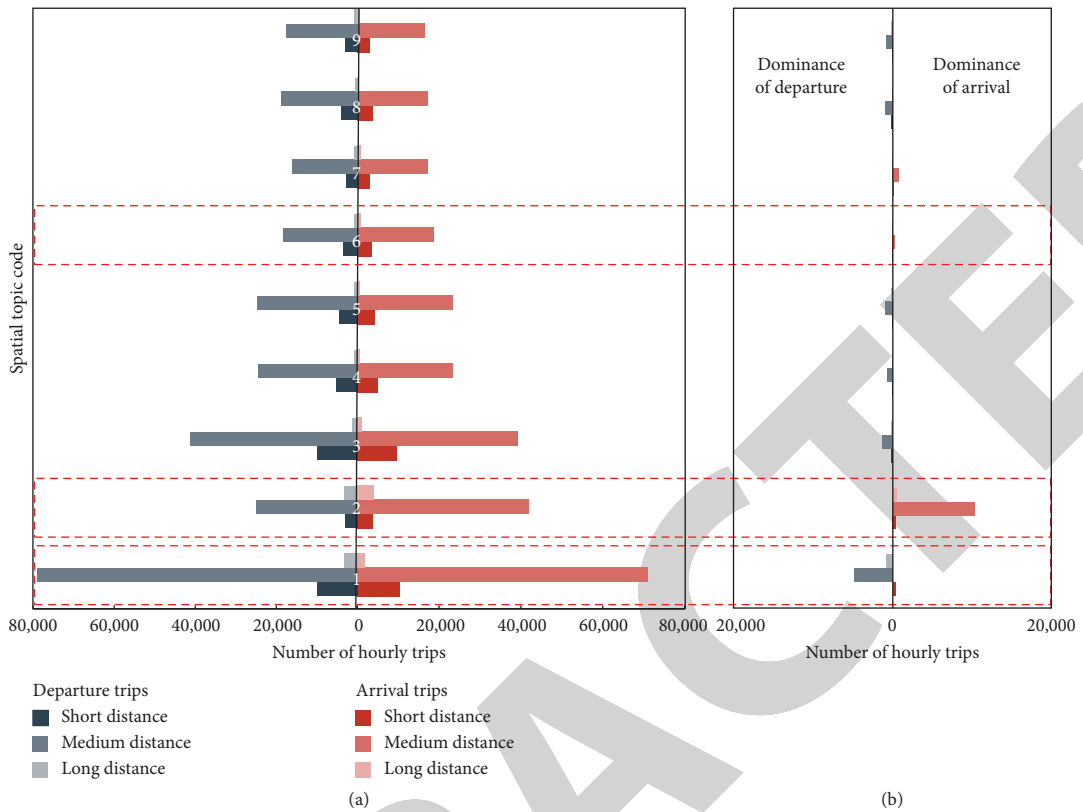


Figure 3: Distribution of spatial topics of car-hailing mobility patterns. (a) Spatial word distribution. (b) Difference between hourly departure and arrival trips.

associated with a spatial topic and a temporal topic. Thus, Figures 4 and 5 show the results of the algorithm applied to the real car-hailing order dataset. Figure 4 shows the word distribution corresponding to each hidden topic of car-hailing mobility patterns from a spatial perspective. Departure and arrival words are depicted in different colors. For each hidden topic, spatial words with different travel distances for departure and arrival trips are portrayed in different shades of the same color (Figure 4(a)). Figure 4(b) shows the result of hourly departure trips without arrival trips with the corresponding spatial word within each hidden topic.

From these results, we annotate each spatial topic using semantic terms that can contribute toward a better understanding of actual hidden patterns. More specifically, we use the most frequent words in a discovered topic to annotate each topic. According to Figure 4(a), the most common spatial words for all topics are medium-distance trips, followed by short-distance trips, while long-distance travel by car hailing represents a small proportion of all topics. In general, the mobility patterns in each topic exhibit similar trends, and the travel intensity decreases from topic 1 to topic 9. However, the red dashed box around topic 1 depicts a greater demand for departure trips, whereas the box around topic 2 reveals a high demand for arrival trips, regardless of the travel distance (Figure 4(b)). A hidden pattern also exists depicting a balance between departure trips and arrival trips, as shown by topic 6. In addition, topic 7 shows similar trends to topic 2 but with a lower travel intensity, and the remaining topics show typical representations of dominant departures with low travel intensities. Overall, the spatial topics can be categorized as follows:

Departure dominance topics: topics 1, 3, 4, 5, 8, and 9, where topic 1 depicts a significant departure trend among all departure dominance topics. The departure intensities of the remaining topics decrease in the order of topic 3 > topic 8 > topic 5 > topic 9 > topic 4.

Arrival dominance topics: topics 2 and 7.

Balance between departure and arrival topic: topic 6.

Figure 5 shows the temporal evolution of three temporal topics, which indicates the average number of hourly trips for both weekdays and weekends. The three temporal topics exhibit similar fluctuation trends but different travel intensities; the travel intensity decreases from topic 1 to 3, which are defined as high, normal, and low travel intensities for the purpose of this study. Clear differences are observed for all topics between weekdays and weekends. More specifically, travel intensities remain relatively constant from 07:00 to 23:00 on weekdays, while four obvious peaks can be observed at 9:00–10:00, 14:00–15:00, 21:00–22:00, and 18:00–19:00. In contrast, only one significant peak is observed on weekends, namely, between 18:00 and 19:00. However, topics are hard to distinguish from one another between 7:00 and 8:00 both on weekdays and on weekends. The temporal

Table 3: Average perplexity distribution on different (*J*,*K*).

J	K									
	3	4	5	6	7	8	9	10	11	12
3	249.854	200.1	251.838	241.72	279.497	244.093	**162.178**	205.022	248.705	230.041
4	183.725	203.918	212.886	247.606	255.804	262.605	253.499	222.798	246.647	195.523
5	257.275	221.73	246.606	219.423	185.079	223.306	215.182	278.501	218.187	242.082
6	246.577	227.542	254.104	261.555	175.647	192.918	207.824	166.597	171.033	192.878

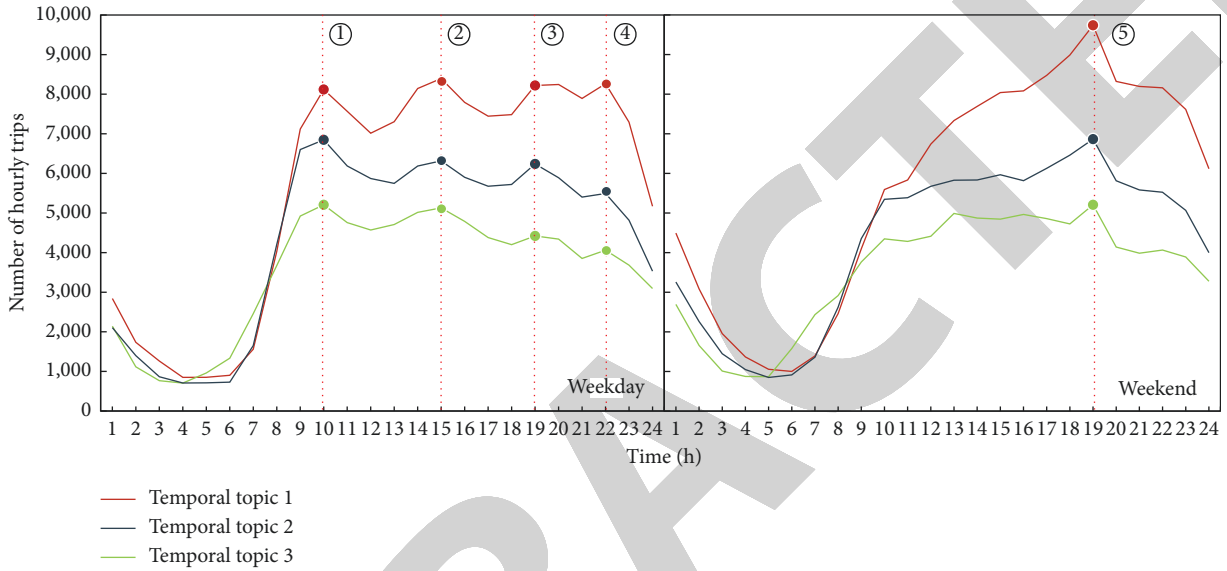
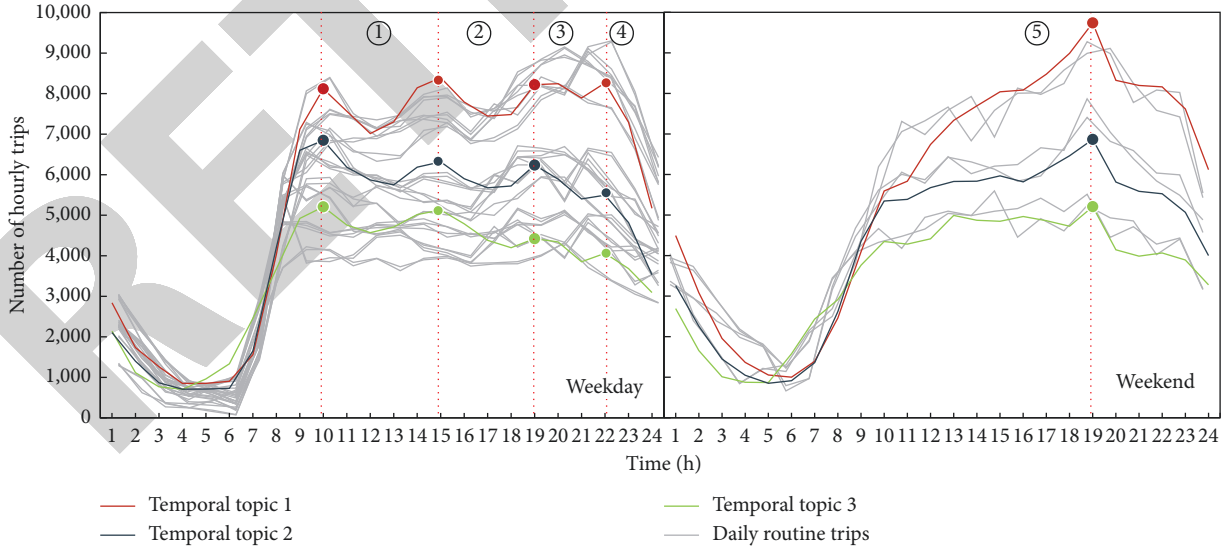
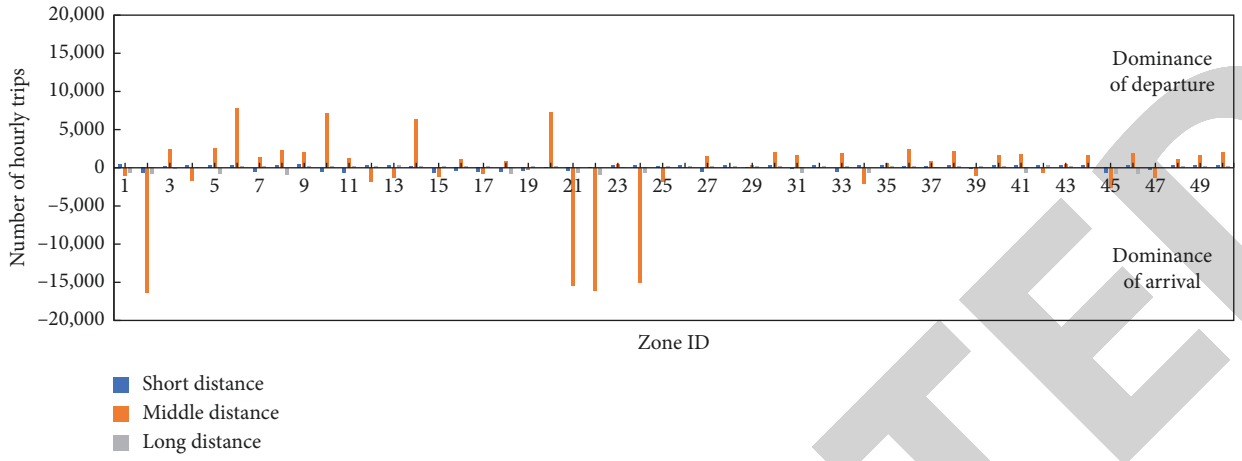


Figure 4: Distribution of temporal topics of car-hailing mobility patterns.



(a)

Figure 5: Continued.



(b)

Figure 5: Validation of the results. (a) Comparison between temporal topics and user routines. (b) Comparison between spatial topics and user routines.

evolution of the car-hailing travel intensity differs from that of general urban transport mobility routines, such as bus travelers [16], where two travel intensity peaks emerge in the morning and evening.

In view of the above-described four peaks observed on weekdays (peaks 1–4) and these single peak observed on weekends (peak 5), peaks 1–4 were considered to correspond to commuting trips, business trips, leisure trips, and home trips, while peak 5 was attributed to a burst in travel intensity for leisure purposes. Since car-hailing services represent a "door-to-door" travel mode with many advantages despite their higher cost compared to public transport, more specific travel intensity peaks occur because it only appeals to specific travelers [31]. In addition, the travel intensity of car hailing between 00:00 and 03:00 (Figure 5) remains high because of the unavailability of other modes of public transport.

As described previously, the distribution of both spatiotemporal topics of a TAZ can be used as the soft clustering among the mobility patterns within a zone, which is a two-dimensional matrix. Thus, Table 4 lists an example of combining spatiotemporal topic distributions, where each entry indicates the possibility assigned to a specific combined topic (*spatial topic code, temporal topic code*). For example, the grey entry denotes the largest combined spatiotemporal proportion given in Table 4, which exhibits a salient feature of spatial topic 2 and temporal topic 2. Based on this information, we can label this TAZ as an arrival dominance TAZ with a normal travel intensity (the hourly trips can reach 7000 as indicated in Figure 5). The combined spatiotemporal distribution can therefore be derived for each TAZ after model selection and inference.

5. Validation

To validate the result obtained using the described method, statistical analysis was conducted based on the entire dataset. As a temporal/spatial topic indicates the regularity of regional travel activities, it reflects the travelling routines of car-hailing services. Based on this, we select routine travelers from the entire car-hailing dataset, where a routine car-hailing traveller is defined as hailing a car at least three days on weekdays and one day on weekends. Figure 3 shows a comparison of the results obtained for a routine traveller in addition to the extracted temporal and spatial topics. More specifically, Figure 3(a) shows the temporal mobility profiles of routine travelers, while Figure 3(b) shows the mobility patterns of routine travelers among 50 random selected TAZs. As indicated, temporal topics correspond to actual temporal mobility patterns, while spatial topics capture the actual spatial mobility patterns that are categorized into the arrival dominance, departure dominance, and balance, respectively. Thus, according to the data presented in Figure 3, the extracted temporal/spatial topics can be regarded as reasonable representations of regional mobility patterns.

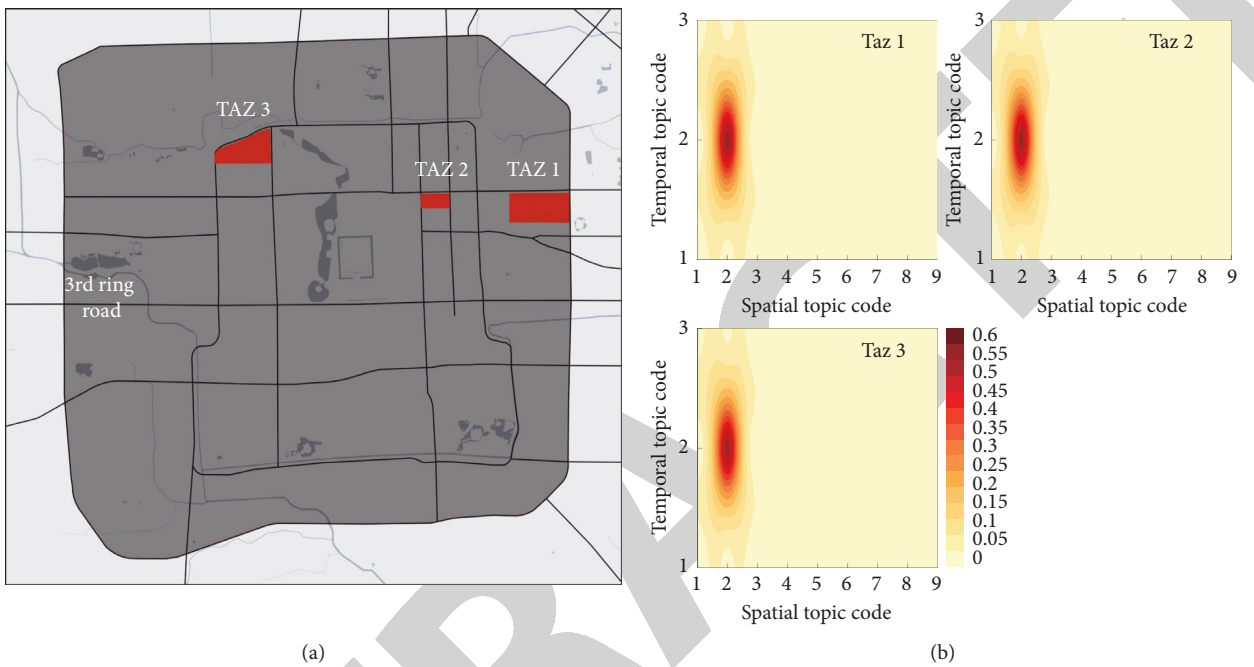
6. Application

The obtained results will likely contribute to long-term transportation planning, such as regionally oriented transportation policy design, region-based customized buses, and the development of car-hailing service monitoring systems. The primary objectives of these applications include detecting TAZs with similar spatiotemporal topics and detecting TAZs with anomaly mobility patterns through use of the derived results of the probabilistic model. Here, a brief introduction to the method employed for detecting similarities and anomalies is given below.

6.1. Detection of Similar TAZs. The similarity between two TAZs can be computed using the Kullback–Leibler divergence, Jensen–Shannon divergence (JSD), and Wasserstein distance [23]. These measures can be applied to determine the similarity of the distribution of combined spatiotemporal topics among all TAZs, in which a combined spatiotemporal topic distribution reveals the internal hidden

Table 4: Combined spatiotemporal topics of an example TAZ.

Temporal topics	Spatial topics								
	1	2	3	4	5	6	7	8	9
1	1.85%	1.94%	1.38%	1.94%	1.56%	1.66%	0.94%	2.88%	0.56%
2	2.91%	52.66%	3.38%	2.25%	0.38%	1.03%	0.39%	3.45%	1.88%
3	1.31%	3.12%	2.13%	1.13%	2.66%	0.66%	0.75%	2.82%	2.38%



(a) (b)

Figure 6: TAZs with similar spatial and temporal topics: (a) spatial distribution of TAZs and (b) probability distributions for the possible topics.

patterns in a TAZ. For the purpose of this study, we adopt the JSD to compute pairwise similarity among TAZs from the training dataset:

$$\mathrm{JSD}(\mathrm{taz1}, \mathrm{taz2}) = \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[\theta_{\mathrm{taz1}\,jk}\log\frac{\theta_{\mathrm{taz1}\,jk}}{\overline{\theta}_{jk}} + \theta_{\mathrm{taz2}\,jk}\log\frac{\theta_{\mathrm{taz2}\,jk}}{\overline{\theta}_{jk}}\right], \tag{10}$$

where $\overline{\theta}_{jk} = (1/2)(\theta_{\mathrm{taz1}\,jk} + \theta_{\mathrm{taz2}\,jk})$. We illustrate the obtained results by randomly selecting a TAZ, namely, TAZ1, from the training dataset, and the TAZs with similar topic distributions to TAZ1 are depicted in Figure 6. According to Figure 6(b), these TAZs are characteristic of a normal travel intensity and arrival dominance, and the majority of arrival trips constitute a medium travel distance (i.e., 3–30 km). Moreover, the average hourly arrival trips, without including departure trips, could approximate 10,000 (Figure 4(b)). With this knowledge, we can design region-oriented demand management strategies specific to this kind of TAZ. To alleviate traffic congestion and energy consumption, these TAZs could include the following:

(i) Implementing congestion fees

(ii) Meeting departure demands without dispatching abundant number of vehicles to these areas

(iii) Guiding travelers to take customized buses rather than hailing a car

(iv) Optimization of the transit network

6.2. Anomaly Detection. A further application of this method is the detection of anomalous TAZs using predictive perplexity according to equation (7). We infer the predictive perplexity of the future dataset using the trained hidden patterns. As described previously, predictive perplexity can serve as a reliable proxy for temporal changes in mobility patterns, whether routine or anomalous. The average perplexity of all TAZs in the study areas is 266.69, which can be explained by the regular daily patterns of people living in the central area of Beijing. A higher perplexity indicates that the TAZ is prone to abnormal mobility patterns, whereas a lower perplexity indicates routine mobility patterns.

The spatial distribution of predictive perplexity for each TAZ is shown in Figure 7. Light colors indicate regular travel patterns across all future datasets. Mobility patterns in these TAZs are typically high; i.e., trips are stable within an acceptable range. Dark colors indicate uncertain and random mobility patterns. Figure 7(a) shows six areas with high and low perplexities. i, ii, and iv are typical areas of high perplexity; thus, future mobility patterns are difficult to capture

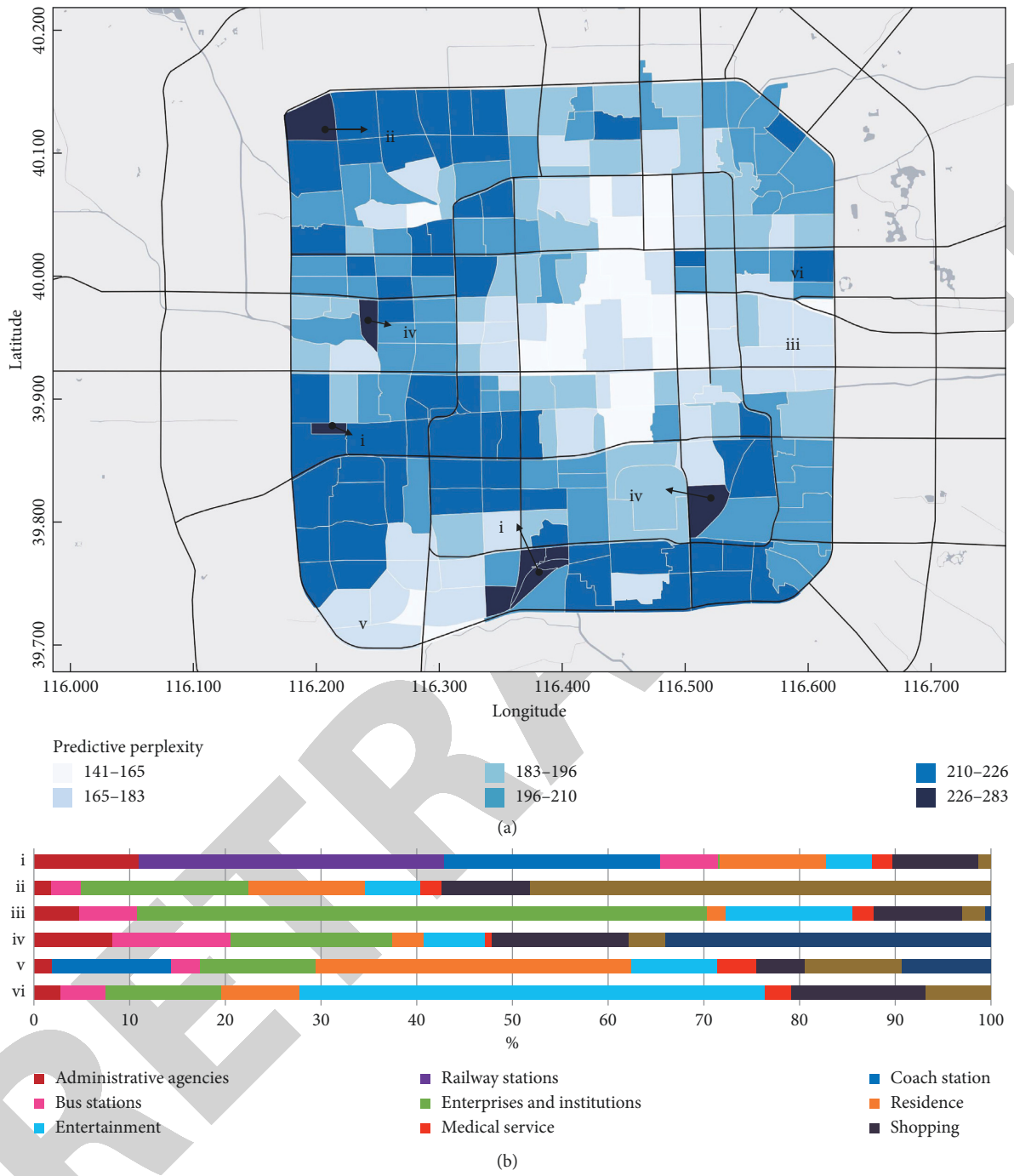


Figure 7: (a) Results of anomaly detection for the study areas and (b) POI statistics for areas i–vi showing the composition of each TAZ.

using the trained hidden patterns. In contrast, iii, v, and vi are areas of low perplexity; thus, mobility patterns in these areas are more regular.

As mobility patterns are coupled with land use characteristics [6, 17], the degree of predictive perplexity in a TAZ may indicate the land use composition, which is compiled using the POI dataset. As shown in Figure 7(b), railway stations comprise a dominant proportion of area i, and they tend to attract and generate large crowds. Moreover, the random mobility patterns in this type of area could lead to a high predictive perplexity. Another cause of a high predictive perplexity is a peak in the passenger volume. Since area ii is a characteristic area for education, working times are more flexible than elsewhere. Area iv is a scenic area, which is also hard to predict because of the nonroutine behavior of tourists. Areas iii and v represent TAZs with regular mobility patterns and are dominated by businesses and institutions and residences, respectively. The perplexity of area vi is higher than those of iii and v, but lower than those of i, ii, and iv; the predominance of entertainment land

use may explain this medium perplexity. Overall, transportation hubs, scenic spots, and entertainment areas may lead to irregular mobility patterns and difficulty in predicting future mobility patterns because of the stochastic behavior of travelers. In contrast, mobility patterns in business and residential areas are usually highly predictive.

The anomaly detection of TAZs could therefore provide operators with prior knowledge of travel demand in some areas, thereby allowing them to respond timely to unexpected passenger flows. From the perspective of long-time operations, these results can aid in the design of dynamic scheduling strategies or reference pick-up/drop-off locations to alleviate waiting times. For example, if someone was unable to hail a car at the train station, an online car-hailing system could refer a pick-up location nearby with a low predictive perplexity. Compared with previous studies on the similarity and anomaly detection of regions [22, 32], the method employed herein could overcome the limitation of dataset sparsity and combine spatiotemporal features simultaneously.

7. Conclusion

Car-hailing services are undergoing rapid development worldwide, and this will have a significant influence on travel activities while posing substantial challenges to transportation operators, thereby affecting how transport policies and schedules are designed. Currently, only limited research has been carried out to evaluate the regional spatiotemporal mobility patterns of car hailing. Thus, we herein analyzed a two-week car-hailing dataset collected from a major car-hailing operator in China, which includes millions of car-hailing order records. Our aim was to unravel hidden mobility patterns (combined spatiotemporal topic distributions) at a traffic analysis zone (TAZ) scale. A hierarchical mixture model based on a two-dimensional latent Dirichlet allocation (LDA) model was used to handle the large, high-dimensional spatiotemporal dataset efficiently and effectively. More specifically, we incorporated regional properties into the LDA model to reconstruct regional mobility patterns using a linear combination of derived spatial and temporal hidden patterns. The interaction between spatial and temporal dimensions in each TAZ was captured using only a few hidden patterns. Moreover, Gibbs sampling was employed for efficient inference. The topics uncovered by our model were regarded as routines; therefore, we were able to investigate the uncertainties and similarities in regional mobility patterns. We then employed the degree of predictive perplexity to determine whether TAZs are prone to abnormal or unpredictable mobility patterns. From a methodology perspective, we proposed a workflow to reveal TAZ-based mobility patterns, which combined the intrinsic properties of a TAZ, the temporal travel intensity, and trip departures from and arrivals at TAZs. In addition, the probability mixture model was able to overcome the high-dimensional and sparsity limitations of the car-hailing order dataset. From the perspective of our numerical results, we found several TAZ-based mobility patterns of car-hailing services in Beijing compared with other public transit modes. Thus, it was found that a greater number of travel intensity peaks were observed on weekdays, with the four key peaks being observed at 9:00–10:00, 14:00–15:00, 21:00–22:00, and 18:00–19:00, respectively. Furthermore, travel demands were more concentrated on the weekend, with only one salient high intensity peak being observed, i.e., at 18:00–19:00. Moreover, our results indicated that it is difficult to hail a car in transportation hubs and scenic areas, and the predictive perplexity tends to be high in these two kinds of areas. Our results therefore revealed the efficient detection of hidden patterns and anomalies in TAZs in the study areas. However, our research has some limitations. Firstly, the car-hailing demand changes substantially with time. Thus, when creating the temporal corpus, the interval between adjacent time windows may have a large influence on the uncovered latent patterns. Secondly, the "bag-of-words" assumption is a weakness of the LDA model, and so the sequence in the corpus does not influence the uncovered latent patterns. However, the high predictability of the car-hailing demand suggests that we could develop a unified framework for both pattern identification and prediction.

Data Availability

The partial discretized car-hailing order data used to support the findings of this study have been deposited in the "Beijing car-hailing order dataset" repository, http://doi.org/10.21227/7mss-w794.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was partially supported by the National Key and Development Program of China (2017YFC0803903) and National Natural Science Foundation of China (no. 71601006).

References

[1] Z. Bai, W. Liu, and Y. Xing, "Evolutionary Game Theory-based choice behavior analysis of order dispatching modes for online car-booking service," in *Proceedings of the CICTP 2017*, pp. 2417–2425, Shanghai, China, July 2018.

[2] X. Chen, H. Zheng, Z. Wang, and X. Chen, "Exploring on-demand ridesplitting behavior and impact on mobility: a case study in Hangzhou, Hangzhou, China," in *Proceedings of the Transportation Research Board 97th Annual Meeting*, Washington, DC, USA, 2018.

[3] D. Sun, K. Zhang, and S. Shen, "Analyzing spatiotemporal traffic line source emissions based on massive didi online car-hailing service data," *Transportation Research Part D: Transport and Environment*, vol. 62, pp. 699–714, 2018.

[4] Beijing Transport Institute, *Beijing Transport Annual Report*, 2019.

[5] E. Thuillier, L. Moalic, S. Lamrous, and A. Caminada, "Clustering weekly patterns of human mobility through

mobile phone data," *IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 817–830, 2018.

[6] G. Zhong, J. Zhang, L. Li, X. Chen, F. Yang, and B. Ran, "Analyzing passenger travel demand related to the transportation hub inside a city area using mobile phone data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 50, pp. 23–34, 2018.

[7] Z. He, G. Qi, L. Lu, and Y. Chen, "Network-wide identification of turn-level intersection congestion using only low-frequency probe vehicle data," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 320–339, 2019.

[8] Z. He, L. Zheng, P. Chen, and W. Guan, "Mapping to cells: a simple method to extract traffic dynamics from probe vehicle data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 3, pp. 252–267, 2017.

[9] X. Ma, C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.

[10] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[11] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: segmenting space through digital signatures," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 78–84, 2010.

[12] C. Kang and K. Qin, "Understanding operation behaviors of taxicabs in cities by matrix factorization," *Computers, Environment and Urban Systems*, vol. 60, pp. 79–88, 2016.

[13] M. G. Demissie, G. Correia, and C. Bento, "Analysis of the pattern and intensity of urban activities through aggregate cellphone usage," *Transportmetrica A: Transport Science*, vol. 11, no. 6, pp. 502–524, 2015.

[14] N. Yong, S. Ni, S. Shen, P. Chen, and X. Ji, "Uncovering stable and occasional human mobility patterns: a case study of the Beijing subway," *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 28–38, 2018.

[15] H. F. Yu, N. Rao, and I. S. J. C. S. Dhillon, *High-dimensional time series prediction with missing values*, 2015, http://arxiv.org/abs/1509.08333.

[16] G. Qi, A. Huang, W. Guan, and L. Fan, "Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1197–1214, 2019.

[17] Y. Gong, Y. Lin, and Z. Duan, "Exploring the spatiotemporal structure of dynamic urban space using metro smart card records," *Computers, Environment and Urban Systems*, vol. 64, pp. 169–183, 2017.

[18] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.

[19] 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.

[20] Z. Fan, A. Arai, X. Song, A. Witayangkurn, H. Kanasugi, and R. Shibasaki, "A Collaborative filtering approach to Citywide human mobility completion from sparse call records," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, NY, USA, July 2016.

[21] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa, *Fast Mining and Forecasting of Complex Time-Stamped Events*, Association for Computing Machinery, Beijing, China, 2012.

[22] J. B. Sun, J. Yuan, Y. Wang, H. B. Si, and X. M. Shan, "Exploring space–time structure of human mobility in urban space," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 5, pp. 929–942, 2011.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet Allocation*, pp. 993–1022, University of California, Berkely, CA, USA, 2003.

[24] 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.

[25] T. L. Griffiths and M. J. P. Steyvers, "Finding scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5228–5235, 2004.

[26] G. Qi, J. Wu, Y. Zhou et al., "Recognizing driving styles based on topic models," *Transportation Research Part D: Transport and Environment*, vol. 66, pp. 13–22, 2019.

[27] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, Alberta, Canada, August 2002.

[28] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer Science & Business Media, Berlin, Germany, 2008.

[29] Z. Gan, M. Yang, T. Feng, and H. J. T. Timmermans, "Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations," *Transportation*, vol. 45, no. 3, pp. 1–22, 2018.

[30] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 712–725, 2015.

[31] M. Young and S. Farber, "The who, why, and when of Uber and other ride-hailing trips: an examination of a large sample household travel survey," *Transportation Research Part A: Policy and Practice*, vol. 119, pp. 383–392, 2019.

[32] C. Zhong, E. Manley, S. Müller Arisona, M. Batty, and G. Schmitt, "Measuring variability of mobility patterns from multiday smart-card data," *Journal of Computational Science*, vol. 9, pp. 125–130, 2015.