Spatial temporal analysis of vehicle routing problem from online car-hailing trajectories

Xue Yang (yangxue@cug.edu.cn)
China University of Geosciences

Jianhua Yu
China University of Geosciences

Zihan Kan
Chinese University of Hong Kong

Lin Zhou
China University of Geosciences

Qingfen Guan
China University of Geosciences

Luliang Tang
Wuhan University

Research Article

Keywords: car-hailing trajectories, vehicle routing problem, popular roads avoided by drivers, map-matching, spatial temporal analysis

Posted Date: May 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2830741/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.
Spatial temporal analysis of vehicle routing problem from online car-hailing trajectories

Xue Yang1,2; Jianhua Yu1; Zihan Kan3; Lin Zhou1; Qingfen Guan1,2; Luliang Tang4

1. School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; 2. National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430074, China; 3. Department of Geography and Resource Management and Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong, China; 4. State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: A range of vehicle routing problems, from routing planning that vehicles will apply to the actual route that drivers selected in their environment, depend on many factors including travel length, traffic condition, or personalized experience, etc., raising a fundamental question: To what degree is planned route align with the actual route. Here we explore the spatial temporal differences between the planned route and actual route by studying the popular roads which are avoided by drivers (denoted as: PRAD) from car hailing trajectories. By matching the raw trajectories based on an improved HMM map-matching algorithm, we obtain the OD (origin-destination) matrix and their corresponding actual route that vehicles traveled, and planned route generated by A* routing algorithm. We used the Jaccard index to quantify the similarity between the actual route and the planned route of the same OD pair. The PRAD is detected and further analyzed from the aspects of traffic condition. By using car-hailing trajectories provided by DiDi company, we analyzed drivers’ routing behavior in workday and weekend in Wuhan. The relation of PRAD with traffic condition in workday and weekend is discussed and results shown that about 65% PRAD are occurred with a serious traffic jam especially in workday.

Key words: car-hailing trajectories, vehicle routing problem, popular roads avoided by drivers, map-matching, spatial temporal analysis

1. Introduction

With the increasing level of urban motorization, the urban traffic problems become more and more serious. Based on the statistical data, the number of motor vehicles in China has reached to 395 million and exceeded 2 million in 35 cities by the end of 20211, which caused a lot of traffic problems, especially traffic jams in some popular roads during certain period of time. To alleviate the increasing congestion and improve traffic efficiency, the topic of analyzing drivers’ route choice behavior in both the spatial and temporal dimensions is crucial (Jing et al. 2018; Li et al., 2018; Xu et al., 2020). With the advancements in mobile technologies and location-based services (LBS), GNSS-enabled navigation systems play an important role in people’s route choice behaviors. A key function of such systems is route search/planning to the destination, based on the assumption that the planned route is the optimal. However, the routes
searched by algorithms do not always match people’s actual choices of routes. Therefore, it is important to understand how well the planned routes match the actual routes and the factors that may contributing to the differences between them. By detecting the popular roads avoided by drivers through comparing the planned route and actual route between the same OD pairs becomes a vital step to grasp the characteristics of drivers’ routing behaviors in both the spatial and temporal dimensions. Meanwhile, it can also be used to optimize vehicle routing algorithms to further reduce drivers’ travel costs and alleviate traffic congestion.

At present, most of research on drivers’ routing behavior mainly depends on the stated preference (SP) surveys or data collected by small-scale experiments, which is limited by the number of participants involved (Kroes, 1988; Hensher, 1994). Approaches on mining routing behavior focuses on discrete choice model, e.g., multinomial logit (MNL) model (Dial et al., 1971), CNL (Cross-nested logit) model (C.-H.Wen et al., 2001), and GEV (Generalized extreme value) model (McFadden, 1978), etc. The differences between these models are manifested in the characteristics of datasets, explained variables, and model structures (Deng et al. 2020). For example, Dial et al. (1971) proposed a discrete multinomial logit (MNL) model for multimode selection. To address the independence of irrelevant alternatives (IIA) problem of the MNL model, many studies developed some new models based on the MNL model by adding the modification section to represent the interactions between different routes, such as C-logit model (E.Cascetta et al., 1996) and PS-logit model (M.S.Ramming et al., 2002). Apart from this, some literature demonstrated the use of the developed CNL (Cross-nested logit) model (C.-H.Wen et al., 2001) and PCL (the paired combinatorial logit) model (F.S.Koppelman et al., 2000) to avoid the IIA issue of MNL model based on the GEV principle (McFadden, 1978). However, these previous studies mainly analyze drivers’ routing behaviors depending on a small amount of survey data, which is time-consuming and biased due to the limited data collected.

With the rapid development of information and communication technologies, the positioning technologies and collection or storage capabilities of massive data advanced the application of GNSS trajectories in the field of transportation, such as travel time estimation (Tang et al. 2020), risk assessment of driving behaviors (Zhu et al., 2017; Hu et al., 2015), departure time modeling (Hu et al., 2020), route choice behavior analysis (Lu et al. 2015; Deng et al., 2020; Qi et al., 2020; Deng et al., 2020; Deng et al., 2020), etc. Among them, the vehicle routing problem driven by large tracking datasets has been improved in both effectiveness and accuracy. For instance, Kim et al. (2015) established a framework for clustering and categorizing vehicle trajectories to analyze vehicles’ travel pattern in space and time. Lu et al. (2015) developed a visualization system to help users deal with the massive trajectories and discover the causes of route selections. Based on their study, the contributing factors of routing problems included route-related elements (e.g., route length, traffic light number, route importance, time cost distribution) and trajectory-related elements (e.g., departure time in a day, departure day, trajectory length). Li et al. (2016) discovered the effect of heterogeneity of route selection in from the aspects of drivers’ ages and genders, engine capacity, and characteristics of OD matrix by using the vehicle trajectories collected by
private cars of Toyota City. Deng et al. (2020) applied taxi trajectories to explore the route selection behavior based on the heterogeneous travel distances. They first used the DBSCAN (Density-based spatial clustering of application with noise) algorithm and AIC (Akaike information criterion) to categorize the travel distance into several types and built the PS-logit model by defining 9 explanatory variables to analyze the heterogeneity of trips with varying distance. In summary, most of studies on drivers’ routing behavior analysis by using trajectories focused on exploring the causes of route selection, while lacking an understanding of the differences between planned routes and actual routes.

In this study, we applied car-hailing trajectories to explore drivers’ routing behaviors and further explored why drivers tend to avoid going to some roads that are suggested by routing algorithms (e.g., A* method) by identifying popular dodging roads. To acquire the popular dodging roads, we first match the raw car-hailing trajectories to the motor vehicle road network based on an improved HMM map-matching algorithm. Then, the OD matrix is extracted according to the matched trajectories and their corresponding actual route that vehicles traveled is further obtained. By using A* searching algorithm, we generate the planned route between each OD pair. We use the Jaccard index to quantify the similarity between the actual route and the planned route between a same OD pair and visualize the similarity. The PRAD is detected based on a clustering method and its causes are further analyzed from the aspects of traffic jams and accidents. By using the massive car-hailing trajectories provided by DiDi company in the city of Wuhan, we find about 65% PRAD with a serious traffic jam in workday. The main contributions of this study include:

1) We improved the original version of HMM algorithm by optimizing the computations of angle feature in observation probability and velocity in transition probability to increase the accuracy of map-matching and provide accurate results for the sub-sequent analysis of PRADs.

2) We proposed to detect the PRAD by using the NEAT (a road NEtwork Aware Trajectory clustering) approach from real car-hailing trajectories provided by DiDi company. And we further explore the reason hidden behind the PRADs from the aspect of road condition including traffic congestion and accident.

2. Methodology

The methodological framework of the spatial temporal analysis for vehicle routing from the aspect of PRAD detection is conducted by using the car-hailing trajectories, as shown in Figure 1.
problem from online car-hailing trajectories

In this study, two kinds of spatial data were used to analyze the PRAD. The first is car-hailing trajectory data collected by residents who worked as part-time drivers. Specifically, a trajectory is comprised of a set of corresponding tracking points and can be denoted as $\text{Tra} = (p_1, \ldots, p_n)$, where $n$ is the number of tracking points belonging to the trajectory. Each tracking point records the location (e.g., longitude and latitude), time, speed, heading direction of the moving objects, denoted as $p_i (x_i, y_i, t_i, s_i, a_i)$, $i=1, 2, \ldots, n$. The second type of spatial data used in this study is road network of the study area with road segments, nodes, topology information, and the direction of traffic flow.

2.1 Map-matching based on the improved HMM algorithm

Map matching is the first step for exploring driving behavior from car-hailing trajectories. Its detailed mechanisms have significant effects on the results of drivers’ routing behaviors. Qudus et al. (2007) summarized the existing map-matching algorithm into four categories, including geometry-related methods (B. P. Phuyal et al. 2002), topology-based methods (Y. Meng 2006), probability-based methods (Paul Newson et al. 2009), and mathematical methods (Syed et al. 2004; Li et al. 2014; Dai et al. 2016; Zhao et al. 2017). Among these kinds of methods, HMM-based map-matching has been widely applied because it does not need to train data and considers the features of trajectories and road network both in the geometry and topology. Hu et al. (2019) added the driving direction to the computation of observation probability based on the original model of HMM to improve the accuracy of map-matching. However, their method was limited by the complexity of the direction angle probability calculation and poor performance for low-frequency tracking points matching. To address these issues, we improved the computing model of observation probability and transition probability based on the work conducted by Hu et al. (2019) to enhance the accuracy of map-matching especially for tracking points collected at the complex road intersections.

For a trajectory $\text{Tra} = (p_1, \ldots, p_{i-1}, p_i, p_{i+1}, \ldots, p_n)$, assuming $p_{i-1}$, $p_i$, and $p_{i+1}$ are candidates for map-matching and $s_{i-1}^k$, $s_i^k$, $s_{i+1}^k$ correspond to their state points. The key idea for HMM-based map-matching is to compute the observation probability and transition probability of tracking points based on their corresponding state points. The computation process of observation probability and transition probability of tracking points is conducted in two layers including observation layer and state layer. Specifically, the observation probability quantifies the possibility of tracking points matched with the state of candidates. For most of HMM-based map-matching algorithm, the observation probability was computed based on the distance from the candidate tracking point to the road network (Hu et al. 2019; Liu et al., 2017; Hansson et al., 2020). In this study, we improved the computation method of observation probability for angle feature by adding the angle between tracking points and directed road segments (see Equation 1). So, the observation probability from the aspect of angle feature can be calculated based on Equation 2.


\[ a = \begin{cases} 
\beta - \gamma & |\beta - \gamma| < 180^\circ \\
360 - |\beta - \gamma| & |\beta - \gamma| \geq 180^\circ 
\end{cases} \quad (1) \]

\[ P_{\text{angle}}(o|s^k_i) = \frac{\cos(a) + 1}{2} \quad (2) \]

where \( \beta \) indicates the heading angle of tracking point \( p_i \), \( \gamma \) represents the angle of the candidate matching road segment with the direction of north, and \( a \) is the difference between \( \beta \) and \( \gamma \). The parameter \( P_{\text{angle}}(o|s^k_i) \) represents the observation probability of the observation point \( o_i \) and its corresponding candidate state point \( s^k_i \) from the perspective of angle.

The computation method for observation probability in the aspect of distance is the same with the original version of HMM-based map-matching algorithm (Paul and John, 2009). The comprehensive observation probability (denoted as \( P_{o,\text{dis,ang}}(o|s^k_i) \)) both in the aspects of angle and distance is calculated based on Equation 3, where \( P_{\text{dis}}(o|s^k_i) \) represents the observation probability of the observation point \( o_i \) and its corresponding candidate state point \( s^k_i \) from the perspective of distance, \( \omega_d \) and \( \omega_a \) are their weight respectively, and \( \omega_d + \omega_a = 1 \).

\[ P_{o,\text{dis,ang}}(o|s^k_i) = \omega_d P_{\text{dis}}(o|s^k_i) \ast \omega_a P_{\text{angle}}(o|s^k_i) \quad (3) \]

The transition probability quantifies the possibility of the state point of the previous tracking point changing to the state of the current tracking point. The existing research of HMM-based map matching algorithm mainly consider the distance feature of tracking points (Paul and John, 2009; Goh et al., 2012; George et al., 2017). In this study, we added the speed of tracking point to the computation method of transition probability based on the observation that the speed restrictions of different kinds of roads are different. For example, the driving speed of a ramp in China is limited within 40 km/h which is lower than its adjacent main roads 60km/h. The transition probability in the perspective of speed can be calculated according to Equation 4, where \( v_{i-1} \) and \( v_i \) denote the speed of the previous tracking point \( p_{i-1} \) and the current tracking point \( p_i \), respectively. The parameter in Equation 4 represents the average speed from the candidate state point \( s^k_{i-1} \) of the tracking point \( p_{i-1} \) to the candidate state point \( s^k_i \) of the tracking point \( p_i \). Here, the distance from \( s^k_{i-1} \) to \( s^k_i \) is the network distance which is obtained based on the shortest routing algorithm A* (Candra et al. 2020). The transition probability in the perspective of distance is same with the original version of HMM-based map matching algorithm (Paul and John, 2009). Also, the comprehensive transition probability both in the aspects of speed and distance can be calculated based on Equation 5, where \( P_{t,\text{dis}}(s_{i-1}^k|s_i^k) \) denotes the transition probability of the candidate state point \( s_{i-1}^k \) of tracking point \( p_{i-1} \) and the candidate state point \( s_i^k \) of tracking point \( p_i \), \( \omega_{t,\text{dis}} \) and \( \omega_{t,\text{speed}} \) are their weight respectively, and \( \omega_{t,\text{dis}} + \omega_{t,\text{speed}} = 1 \).

\[ P_{t,\text{dis}}(s_{i-1}^k \rightarrow s_i^k) = \frac{(v_{i-1} + v_i)/2}{v_{(i-1,i)-j(i,r)}} \quad (4) \]

\[ P_{t,\text{dis, speed}}(s_{i-1}^k \rightarrow s_i^k) = \omega_{t,\text{dis}} P_{t,\text{dis}}(s_{i-1}^k \rightarrow s_i^k) \ast \omega_{t,\text{speed}} P_{t,\text{speed}}(s_{i-1}^k \rightarrow s_i^k) \quad (5) \]
2.2 Identification for PRAD

The trajectory data used in this study was collected by car-hailing company DiDi. Each trajectory records the actual route of a vehicle with passengers between an OD pair. Based on the improved map matching algorithm, we can obtain the OD pairs of trajectories that are matched to the road network. We calculate an OD matrix of travel time and distance based on the OD pairs. Then, the actual routes of all OD pairs are extracted based on the map-matching results. The corresponding planned routes between the OD pairs are obtained based on A* routing algorithm because of its performance (Candra et al. 2020). Meanwhile, the travel cost of an OD pair during routing planning by using A* routing algorithm is decided by travel distance and time. It should be noted that the travel distance and time are obtained based on the network distance and speed restrictions of different roads.

For analyzing the differences between actual routes and planned routes of OD pairs, we apply the Jaccard index to quantify their similarity. The Jaccard index (JI), also known as Jaccard similarity coefficient, is mainly used for comparing the differences or similarity of two finite sample sets (Vijay verma & Rajesh, 2020) and calculated based on Equation 5, where \( A \) and \( B \) represent two finite sample sets, respectively. The larger JI is, the higher the sample similarity is, and the smaller JI is, the lower the sample similarity is. In this study, the actual routes and its corresponding planned routes are regarded as the sample sets \( A \) and \( B \). The spatial distribution and OD clustering are further visualized by dividing the interval of JI values.

\[
JI(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5}
\]

We define the PRAD as road segments which are included in the planned route but do not appear in the corresponding actual route. Thus, the PRAD is obtained by comparing the actual route and the planned route of a same OD pair. As shown in Figure 2, ‘road_1’ and ‘road_2’ are two-way road, and ‘road_3’ is a one-way road. Based on the map-matching results, the actual route of a trajectory \( tra_i = (\ldots, p_i, p_{i+1}, \ldots) \) includes ‘road_1’ and ‘road_2’ (see Figure 2). However, the planned route of trajectory \( tra_i \) contains ‘road_1’ and ‘road_3’. Thus, for \( tra_i \), ‘road_3’ is its dodging road. For different trajectories with different OD pairs, their PRAD may be also different. To explore the spatiotemporal pattern of drivers traveling, we need to identify the PRAD first.

![Figure 2. Definition of the PRAD](image-url)

To identify the PRAD, we propose a new clustering method rather than directly
count the frequency of road segments avoided by drivers and estimate the popular one based on descending order. This is because some PRADs appearing dozens of times does not mean that they are popular road segments avoided by drivers due to the existence of contingency. Apart from that, we need to face the problem of how to set a suitable threshold to define the PRAD based on their appearance frequency. And some PRADs will be missing because they have the same traffic direction. To address these issues, we first group the PRADs into a set of clusters based on their location. As shown in Figure 3, assuming $RS_i$ represents the road network segment, $i = 1, 2, \ldots, 5$, among them $RS_5$ is a two-way lane, $RS_1$ to $RS_4$ is a one-way lane. The parameter $DS_i$ represents a set of clusters of PRADs on the corresponding $RS_i$ segment. That is, for two-way lanes, PRADs in two directions on the same road is grouped into the same cluster. And PRADs in the same direction on the same section of the one-way lane is grouped into a cluster. For example, $RS_1$ and $RS_3$, $RS_2$ and $RS_4$, are all connected with the same intersections. In this study, we group $DS_i$ in this case into the same cluster. Then, we improve the clustering method in the third stage of NEAT (road NEtwork Aware Trajectory clustering) proposed by Han et al. (2015) to cluster all $DS$ groups and detect the avoided road segments by drivers. Specifically, the improvement mainly includes: 1) the clustering unit is road segment with the same direction of traffic flow; 2) the distance between two clustering unit is calculated by using Hausdorff distance; 3) the threshold of clustering is adaptively acquired based on the input dataset by using the method proposed by Lee et al. (2007). Based on the clustering results, the clusters of $DS_i$ shown in Figure 3 will be identified as PRADs if they satisfy the clustering threshold and vice versa.

Figure 3. PRAD clustering

3. Case study: vehicle routing behavior analysis in the city of Wuhan

3.1 Data collection and map-matching

3.1.1 Car-hailing trajectories collection and preprocessing

Taking Wuhan as the experimental area, trajectories collected by 300,000 of car-hailing
taxis belonging to the DiDi company from August 8 to 16, 2017 were used to analyze the spatiotemporal pattern of vehicle routing problems. The sampling interval and positional accuracy of these trajectories range from 30 s to 120 s and 5m to 20m, respectively. Each tracking point records the information of the current moving object including latitude, longitude, time, speed, and heading angle. The average amount of tracking records every day was about 600,000 in the experimental region. And each trajectory is composed of about 200 tracking points. These car-hailing trajectories covers the main districts of the city of Wuhan, as shown in Figure 4. In addition, there are outliers in the trajectory data caused by signal drift or irregular driving behavior. Here, we applied the method proposed by Yang et al. (2018) to remove outliers from the raw trajectories. The experimental results showed that about 35.95% tracking points were removed. Apart from trajectory data, road networks of motor vehicle used in this study were acquired from the platform of OSM (OpenStreetMap). Based on the statistics, there are about 77,086 road segments located in the experimental region. Because road networks obtained from the OSM platform were provided by volunteers, there also existed issues such as topological errors. This study applied the method proposed by Hu et al. (2019) to revise these topological errors. Based on the statistics, about 315 road segments had topological errors and were corrected.

![Study area and road network](image1.png) ![GPS trajectory points](image2.png)

**Figure 4.** Study region and data diagram

### 3.1.2 Map-matching based on the improved HMM algorithm

The processed trajectories were matched to the road network by using the improved HMM algorithm. Based on the principle of the map-matching algorithm proposed in this study, we need to set the value of weights of \( \omega_d \), \( \omega_a \), \( \omega_{t\_dis} \), and \( \omega_{t\_speed} \). To get the optimum value of these weights, we randomly selected 200 road segments and manually estimated the matching accuracy of tracking points by tuning the value of them from 0 to 1, as shown in Table 1. In Table 1, we find that the value of \( \omega_d \), \( \omega_a \),
ωt_dis, and ωt_speed was respectively set as 0.7, 0.3, 0.7, and 0.3 with maximum accuracy of map-matching. This accuracy results in Table 1 illustrate that the accuracy of map-matching for vehicle trajectories with low sampling rate was closely related to the distance from the observation point to the candidate state point.

**Table 1.** Matching accuracy of the improved HMM algorithm with different values of parameters

<table>
<thead>
<tr>
<th></th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>ωd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ωa</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>ωt_dis</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>ωt_speed</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Matching accuracy (%)</td>
<td>69.8</td>
<td>70.6</td>
<td>75.2</td>
<td>77.7</td>
<td>80.3</td>
<td>83.5</td>
<td>88.3</td>
<td>83.8</td>
<td>79.5</td>
</tr>
</tbody>
</table>

To verify the effectiveness of the improved HMM algorithm. We randomly selected 35 trajectories and compared the map-matching results with the method proposed by Hu et al. (2019). Specifically, we converted the original trajectories into several kinds of trajectories with fixed sampling rate such 30s, 30s-60s, and more than 60s by using cubic spline interpolation and Douglas-Peuker compress method. Then, these processed trajectories were matched to the road network and manually estimated the average value, variance, and standard deviation of the accuracy of map-matching. As shown in Figure 5, the left panel of box plots indicates the results by using the method proposed by Hu et al. (2019). And the right panel of box plots demonstrates the matching results using the improved HMM method which we optimized based on the work conducted by Hu et al. (2019). The experimental results show that the average value, variance, and standard deviation of the accuracy of map-matching of three kinds of sampling rate trajectories by using the method proposed by Hu et al. (2019) are less than the work based on the improved HMM algorithm in this study. It also means that the distribution of correct matching results based on the improved HMM algorithm is more concentrated.

(a) 30s sample interval (left: mean: 0.676; var: 0.045; std: 0.212, right: mean: 0.888; var: 0.002; std: 0.043)
Beyond that, experimental results illustrate that the matching results by using the algorithm proposed in this paper is more suitable for road segments located on the complex road intersections. As we can see in Figure 6, trajectories collected in a roundabout were matched based on the improved HMM method in this study (Figure 6b) and HMM algorithm proposed by Hu et al., (2019) (Figure 6a). Based on the manual inspection, the raw tracking points were collected in Luoyu road, Lumo Road, and the Guanggu Roundabout which is connected these two roads. After map-matching, these tracking points should be matched to these road segment. However, only 3 tracking points were correctly matched to the road where they were collected by using the method proposed by Hu et al., (2019), as shown in Figure 6a. The other tracking points of the raw trajectory shown in Figure 6a were regarded as the matching failure and abandoned. By comparing with Hu’s method, our algorithm can correctly match all raw tracking points to the right places (Figure 6b).
3.3 PRAD visualization and analysis

3.3.1 JI value categorization and visualization

Based on the map-matching result, we computed the JI value of an OD pair to estimate the similarity between its planned routes and the corresponding actual routes. Since massive trajectories will bring the problem of poor expression in visualization of JI value, we applied kernel density statistics to divide the JI value into several classes to facilitate the visualization. Then, we visualized the JI value of all OD pairs according to its categorization. Figure 7 shows the kernel density distribution of JI value of car-hailing trajectories which were respectively collected on August 12th, and 15th, 2017.

![Figure 7. Kernel density distribution of JI values for GNSS trajectories](image)

In Figure 7, we can find the trends of JI value of car-hailing trajectories obtained in two different dates are roughly the same, and most of JI values concentrate around 0.1. Further, there is almost no JI value in the range between 0.8 and 0.9, which means that most actual routes are significantly different with the corresponding planned routes. Meanwhile, we can see several peaks of the distribution of JI values at 0.1, 0.2, 0.4, 0.6, and 1.0. The number of JI values distributed within the interval of [0, 0.1] is highest, and the densities of other ranges including (0.1, 0.2], (0.2, 0.4], (0.4, 0.6], and (0.6,1.0]
show a descending order. Based on the kernel distribution pattern of JI value of all OD pairs, we divided the JI value into five intervals of [0, 0.1], (0.1, 0.2], (0.2, 0.4], (0.4, 0.6], and (0.6,1.0), respectively. Based on this, we define the similarity between the actual routes and planned routes of five kinds of JI value as ‘totally different’, ‘different’, ‘slightly different’, ‘similar’, and ‘no difference’. These different categories of JI values were visualized by the platform of Kepler. Figure 8 shows the visualization results of JI value of trajectories collected on August 12th, 2017. The statistics show that the proportion of JI values within the interval of [0, 0.1], (0.1, 0.2], (0.2, 0.4], (0.4, 0.6], and (0.6,1.0) of Figure 8 is 45.54%, 24.85%, 21.00%, 6.21%, and 2.40%, respectively. This result indicates that the proportion of actual routes which are totally same with the planned routes is very small.

![Spatial distribution pattern of JI value](image)

**Figure 8.** Spatial distribution pattern of JI value of trajectories collected on August 12th, 2017. (a) JI values with the type of ‘totally different’, (b) JI values with the type of ‘different’, (c) JI values with the type of ‘slightly different’, (d) JI values with the type of ‘similar’, (e) JI values with the type of ‘no difference’

To further analyze the pattern of each type of OD pairs of JI values, we clustered their own OD pairs by using FlowmapBlue method which is a free tool for illustrating aggregated numbers of movements between geographic locations as flow maps. Figure 9 illustrates the clustering results of OD pairs which belongs to each category of JI values. Specifically, the size of the dot shown in Figure 9 represents the number of OD pairs connected to it. The larger the dot, the greater number of OD pairs connected to it. Also, the arrow indicates the direction from the origin point to the destination. Similarly, the size and brightness of these arrows also display the number of traffic flow from the origin point to the destination. The volume of traffic flow is proportionate to
the size or brightness of arrows. As we can see from Figure 9, the distribution of OD pairs becomes gradually disperse with the JI values decreasing, although most of OD pairs still gather at the central area of the experimental region (see Figure 9a). It indicates that the larger the travel distance, the greater the difference between the actual travel route and the planned route. When the travel distance is large, there are more alternative routes to avoid certain risks (such as traffic congestion), thus the possibility of inconsistency with the planned route is higher. So, the route searching algorithm should consider or improve the accuracy of long-distance trip. In contrast, when the driving distance is smaller, avoiding certain risks may lead to higher costs of travel distance or travel time, thus it is less likely to be inconsistent with the planned route. Figures shown in the Appendix also display the pattern of JI values and its own OD pairs of trajectories collected on August 15th, 2017. The features of JI values and OD pairs clustering results shown in the Appendix are similar with the results of Figure 8 and 9. That means, traveling patterns of car-hailing drivers in every day are similar.

![Figure 9. OD clustering results on August 12th, 2017, (a) OD pairs of JI values within [0, 0.1], (b) OD pairs of JI values within (0.1, 0.2], (c) OD pairs of JI values within (0.2, 0.4], (d) OD pairs of JI values within (0.4, 0.6], (e) OD pairs of JI values within (0.6,1.0)](image-url)

Figure 10 shows the relationship between JI value and distance of each OD pairs. In Figure 10, we can find that the JI values of OD pairs decrease as the traveling distance increase. It means traveling pattern with a higher JI value mainly exist in the short distance trips. For long distance trips, car-hailing drivers tend to select the route which is totally different or different with the planned routes.
To further explore the distribution of JI values of all routes, we analyze the time distribution of OD pairs in workday and weekend respectively, as shown in Figure 11. Based on the experimental results, travel activities mainly concentrated on the period of 8:00 am - 10:00 pm, regardless in weekend or workday. But in the weekend, residents’ travel activities usually occurred in three time periods: 8:00 am - 9:00 am, 1:00 pm - 2:00 pm, and 5:00 pm - 6:00 pm (see Figure 11a). In workday, the traveling activities mainly occurred in two time periods, am 8:00 – am 9:00 and pm 5:00 – pm 6:00, as shown in Figure 11b. That is, the traveling activities with slightly different or even more are concentrated on morning peak and evening peak, especially in workday. That means, in workday, drivers tend to select driving routes from original point to the destination based on their experience or real-time situation. These routes may not be the shortest in time or distance for drivers.

3.3.2 Detection of hottest road segments avoided by drivers
Through analyzing the JI value of all OD pairs, we found that most of travel routes of online car-hailing drivers were entirely or partly different with the corresponding planned routes. In this study, the road segments in planned routes but not in actual routes
is defined as the PRAD. To investigate the possible reasons why the drivers did not select these routes, we detected and analyzed the hottest PRADs in a temporal and spatial context. Since these PRADs are more likely to occur during the morning and evening peak hours when the traffic is congested, we detected the hottest one from traveling activities occurred in morning and evening peak hours on workdays and weekends respectively based on the NEAT clustering method. As we can see from Figure 12, the hottest PRADs are shown based on the heat map and grey lines represent the road network. Based on the distribution of the hottest PRADs on workday, we can find that some road segments have always been the PRAD no matter in morning peak hours or evening peak hours, such as Wulu road, Zhongbei road, etc., (see Figure 12a and Figure 12b).
The number of road segments which are avoided by drivers (also named as PRAD) during morning peak hours on August 15, 2017 (on a workday), was about 46. The total length of these PRADs was about 42.03 km. On a workday, the number of PRADs in evening peak was about 21 with 15.4 km total length. Compared with that on a workday, the number of PRADs in morning peak and evening peak on a weekend day (August 12, 2017), was 22 and 17, respectively. The total length of these PRADs were 14.9 km and 16.47 km, respectively. These statistics indicate the number of hottest PRADs on
weekend is obviously less than that on workday in morning peak. And some of them are very similar with PRADs appeared in a workday such as Wuhan Yangtse River tunnel (see Figure 12c and 12d). Table 2 summarizes the road name and type of all hottest PRADs appeared both in workday and weekend. In table 2, we can find most of PRADs are the main road.

<table>
<thead>
<tr>
<th>Road name</th>
<th>Road type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wuluo road</td>
<td>trunk</td>
</tr>
<tr>
<td>Luoyu road</td>
<td>trunk</td>
</tr>
<tr>
<td>Zhongbei road</td>
<td>primary</td>
</tr>
<tr>
<td>Wuhan Yangtse river tunnel</td>
<td>primary</td>
</tr>
<tr>
<td>Huanle avenue</td>
<td>primary</td>
</tr>
<tr>
<td>Jianshe avenue</td>
<td>primary</td>
</tr>
<tr>
<td>Qingnian road</td>
<td>primary</td>
</tr>
</tbody>
</table>

Drivers avoided these road segments may be because of two reasons. First, the traffic congestion of these roads is serious. Most of drivers avoided these roads during the operating time to save time and get a higher income. To validate this assumption, we obtained the traffic monitoring data from Gaode Map traffic monitoring platform\(^3\). Based on the data derived from this platform, we found that about 65% PRADs with a serious traffic jam coincided in workday. In weekend, the traffic congestion happened on about 30% PRADs. In addition, based on the public information provided by the website of WPCOM\(^4\), about 20 road segments often occurred traffic jam in the city of Wuhan. Among of them, about 13 road segments are identified as congested road with a cyclical pattern and 9 of them are considered as PRADs including ‘Air Road interchange’, ‘Fazhan Avenue’, ‘Gusaoshu road’, ‘Zhongshan road’, ‘Qingtai road’, ‘Wutaizha road’, ‘Wuluo road’, ‘Zhongbei road’, ‘Guanggu roundabout on Luoyu Road’, ‘Wuhan Yangtze River bridge and tunnel’. The main reason of these road segments often occur traffic jam is because road networks around these road segments are inadequate which causes them cannot timely alleviate the enormous transportation pressure coming from neighboring commercial places, e.g., large shopping malls, restaurants, and other amenities. Apart from this, about 7 road segments belong to congested road in stages because of construction occupying the road surface and 5 of them are identified as the PRADS including ‘Zhongnan road’, ‘Jiefang road’, ‘the starting part of Development Avenue’, ‘Hanzheng road’ and its surrounding roads.

We also investigated the traffic accident happened in the experimental area based on the information collected from the paper published by Fan et al. (2018). The result shown that about the incidence of traffic accidents occurred in PRADs ranged from 0 to 8.52%. To further quantify the relation between PRADs and traffic jam and accident, we estimated their correlation by using Spearman’s correlation. In Appendix, we counted the rate of avoidance of all PRADs identified based on the trajectories collected on a workday (August 15\(^{th}\), 2017) and a weekend day (August 12\(^{th}\), 2017). Here, the rate of avoidance of the PRAD was computed based on its occurrence frequency. That is the rate of avoidance of one PRAD is equal to divide its occurrence frequency by the total number of occurrence frequency of all PRADs. Meanwhile, we also obtained
traffic jam index and traffic accident rate of these PRADs through Gaode Map traffic monitoring platform and the investigate data provided by Fan et al. (2018). The correlation between these PRADs and traffic jam and accident both in workday and weekend was computed, see Table 3.

Table 3. Spearman’s correlation between PRADs and traffic jam and traffic accident

<table>
<thead>
<tr>
<th></th>
<th>Traffic jam</th>
<th>Traffic accident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning peak of workday</td>
<td>0.472**(0.001)</td>
<td>0.779**(0.000)</td>
</tr>
<tr>
<td>Evening peak of workday</td>
<td>0.801**(0.000)</td>
<td>0.737**(0.000)</td>
</tr>
<tr>
<td>Morning peak of weekend</td>
<td>0.878**(0.000)</td>
<td>0.772**(0.000)</td>
</tr>
<tr>
<td>Evening peak of weekend</td>
<td>0.604*(0.010)</td>
<td>0.596*(0.012)</td>
</tr>
</tbody>
</table>

Note: ** denotes significance at 0.01 level; * denotes significance at 0.05 level.

In Table 3, we can find that traffic jam and accident are associated with PRADs no matter in workday or weekend. Specifically, traffic jams and traffic accident occurred in peak of workday are significantly associated with PRADs. In peak of weekend, traffic jams and traffic accident are also associated with PRADs but is not significantly in the evening peak. This result indicates that the occurrence of traffic jam and accident is dynamic in the evening peak of weekend. In general, most of drivers do avoid these roads during the operating time to save time and get a higher income.

4. Conclusion

Addressing vehicle routing problem needs to consider many factors including travel distance, road condition (e.g., traffic jam and accident), personalized preference, etc. The basis of weighting these factors during routing planning is to figure out the differences between planned route based on these factors with the actual route selected by drivers. In this study, we answer this question by studying the popular dodging roads from a large number of car-hailing trajectories. Specifically, we optimized the HMM map-matching algorithm by improving the computations of angle feature in observation probability and velocity in transition probability to increase the accuracy of map-matching and provide accurate matched results for the subsequent analysis of PRADs. The actual route of the OD matrix was generated based on the map-matching results. The planned route of the corresponding OD pair was generated by A* routing algorithm. By using the Jaccard index, we quantified and visualized the similarity between the actual route and the planned route between the same OD pair. The most popular road segments avoided by drivers are detected based on the clustering method of NEAT and its causes are further analyzed from the aspects of traffic condition including traffic jam and accident.

Taking online car-hailing in Wuhan as a case study, we explored the spatiotemporal patterns of PRADs. The experimental result shewed that the hottest PRADs on a weekend day are significantly fewer than that on a workday. In general, a planned route is selected based on several principles including minimum travel distance, shortest travel time, or the comprehensive optimum scheme from the aspects of travel time, distance, number of traffic light, road speed limits, etc. Although a routing planning algorithm has considered many factors to get an optimal route for drivers, the actual traffic condition is very complicated. There are two main reasons why drivers avoided
the planned route and selected other roads to arrive at their destination. First, some roads of planned route may not be optimal due to serious traffic congestion. The statistics obtained from Gaode Map validated that there are about 65% dodging routes with a serious traffic jam coincided in workday. Apart from that, drivers want to select a safe way to arrive at their destinations. However, the traffic accident rate of some planned roads is very high. Based on the correlation analysis, the accident rate of PRADs was significantly associated with the road segments which are identified as the PRADs. These findings by analyzing PRADs can be used for optimizing routing planning strategies, which means users can avoid these road segments in the specific time such as the peak of workday to reduce their travel time cost.

Reference


the kdd.


of the 2004 National Technical Meeting of the Institute of Navigation.


Yaxing, F. (2019). *Spatial-temporal analysis of urban road traffic accidents and multi-constraint spatial partition optimization research.* (PhD), Wuhan University.


1 http://www.gov.cn/xinwen/2022-01/12/content_5667715.htm
2 https://flowmap.blue/
4 https://www.sanyefengji.cn/qichezatan/414998.html
Figure 1. Spatial distribution pattern of JI value of trajectories collected on August 15th, 2017, (a) JI values with the type of ‘totally different’, (b) JI values with the type of ‘different’, (c) JI values with the type of ‘slightly different’, (d) JI values with the type of ‘similar’, (d) JI values with the type of ‘no difference’
Figure 2. OD clustering results on August 15\textsuperscript{th}, 2017, (a) OD pairs of JI values within [0, 0.1], (b) OD pairs of JI values within (0.1, 0.2], (c) OD pairs of JI values within (0.2, 0.4], (d) OD pairs of JI values within (0.4, 0.6], (e) OD pairs of JI values within (0.6, 1.0)

Figure 3. Relationship between JI values and the distances between OD pairs on August 12\textsuperscript{th}, 2017

Table 1. Details of PRADs including its name, the rate of avoidance, traffic jam index, and traffic accident rate

<table>
<thead>
<tr>
<th>Road name of PRADs</th>
<th>The rate of avoidance on the PRAD</th>
<th>Traffic jam index</th>
<th>Traffic accident rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luoyu Road</td>
<td>2.27</td>
<td>3.6</td>
<td>1.35</td>
</tr>
<tr>
<td>Jianshe Avenue</td>
<td>2.17</td>
<td>2.3</td>
<td>1.13</td>
</tr>
<tr>
<td>Zhongbei Road</td>
<td>4.17</td>
<td>4.2</td>
<td>2.46</td>
</tr>
<tr>
<td>Huanle Avenue</td>
<td>1.17</td>
<td>2.1</td>
<td>0.52</td>
</tr>
<tr>
<td>Jiefang Avenue</td>
<td>3.17</td>
<td>3.6</td>
<td>2.32</td>
</tr>
<tr>
<td>Youyi Avenue</td>
<td>2.03</td>
<td>2.2</td>
<td>1.08</td>
</tr>
<tr>
<td>Zhongshan Road</td>
<td>2.38</td>
<td>3.3</td>
<td>0.28</td>
</tr>
<tr>
<td>Xinhua Road</td>
<td>4.32</td>
<td>2</td>
<td>2.27</td>
</tr>
<tr>
<td>Fazhan Avenue</td>
<td>4.2</td>
<td>3.9</td>
<td>2.26</td>
</tr>
<tr>
<td>Road Name</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>------------------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Wuluo Road</td>
<td>3.59</td>
<td>4.1</td>
<td>2.09</td>
</tr>
<tr>
<td>Wusheng Road</td>
<td>2.95</td>
<td>1.6</td>
<td>0.25</td>
</tr>
<tr>
<td>Hongkong Road</td>
<td>2.87</td>
<td>2.3</td>
<td>0.54</td>
</tr>
<tr>
<td>Huangpu Avenue</td>
<td>2.49</td>
<td>2.5</td>
<td>0.37</td>
</tr>
<tr>
<td>Air Road interchange</td>
<td>2.14</td>
<td>4.3</td>
<td>0.52</td>
</tr>
<tr>
<td>Youyi Road</td>
<td>1.82</td>
<td>1.3</td>
<td>1.05</td>
</tr>
<tr>
<td>Xiongchu Avenue</td>
<td>1.64</td>
<td>1</td>
<td>0.54</td>
</tr>
<tr>
<td>Luoshi Road</td>
<td>1.58</td>
<td>1.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Xudong Avenue</td>
<td>1.54</td>
<td>1.9</td>
<td>0.11</td>
</tr>
<tr>
<td>Qingnian Road</td>
<td>1.35</td>
<td>1.5</td>
<td>0.08</td>
</tr>
<tr>
<td>Qiaochang Street</td>
<td>1.14</td>
<td>1.3</td>
<td>0.07</td>
</tr>
<tr>
<td>Dazhi Road</td>
<td>1.13</td>
<td>1.6</td>
<td>0.05</td>
</tr>
<tr>
<td>South Luoshi Road</td>
<td>0.94</td>
<td>1.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Yinwu Avenue</td>
<td>0.86</td>
<td>0.9</td>
<td>0.54</td>
</tr>
<tr>
<td>Wuhan Yangtze River tunnel</td>
<td>0.79</td>
<td>3.6</td>
<td>0</td>
</tr>
<tr>
<td>Bayi Road</td>
<td>0.78</td>
<td>0</td>
<td>1.04</td>
</tr>
<tr>
<td>Heping Avenue</td>
<td>0.6</td>
<td>1.1</td>
<td>0</td>
</tr>
<tr>
<td>Qiaokou Road</td>
<td>0.54</td>
<td>1.3</td>
<td>0.28</td>
</tr>
<tr>
<td>Zhongshan Avenue</td>
<td>0.54</td>
<td>3.5</td>
<td>0</td>
</tr>
<tr>
<td>Jinqiao Avenue</td>
<td>0.53</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Renhe Road</td>
<td>0.52</td>
<td>0</td>
<td>0.28</td>
</tr>
<tr>
<td>Huquan Street</td>
<td>0.52</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tangjiadun Road</td>
<td>0.51</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>Sanyanqiao Road</td>
<td>0.5</td>
<td>1.8</td>
<td>0.52</td>
</tr>
<tr>
<td>Guanggu roundabout</td>
<td>0.37</td>
<td>2.1</td>
<td>0.31</td>
</tr>
<tr>
<td>Jinghan Avenue</td>
<td>0.31</td>
<td>0</td>
<td>0.95</td>
</tr>
<tr>
<td>Hanxi Road</td>
<td>0.29</td>
<td>1.6</td>
<td>0</td>
</tr>
<tr>
<td>Yangtze River tunnel</td>
<td>0.28</td>
<td>3.6</td>
<td>0</td>
</tr>
<tr>
<td>Shengli Street</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nanjing Road</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zhongnan Road</td>
<td>0.26</td>
<td>3.5</td>
<td>0.52</td>
</tr>
<tr>
<td>Jinghan Road</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mingzu Avenue</td>
<td>0.25</td>
<td>2.4</td>
<td>0.94</td>
</tr>
<tr>
<td>Xingye Road</td>
<td>0.25</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>Second ring road</td>
<td>0.24</td>
<td>0.6</td>
<td>0</td>
</tr>
</tbody>
</table>