A novel approach for online car-hailing monitoring using spatiotemporal big data

TONG ZHOU 1,2,3, WENZHONG SHI 3, XINTAO LIU 3, ZHEN QIAN 1, FEI TAO 1,2 and RUIJIA ZHANG 4,
1School of Geographic Science, Nantong University, Nantong 226007, China; 2Key Laboratory of Virtual Geographical Environment, MOE, Nanjing Normal University, Nanjing 210046, China 3Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong 4Institute of College of Foreign Studies, Nanjing Agriculture University, Nanjing 210095, China

ABSTRACT Car-hailing service has increasingly become popular and fundamentally changed the way people travel in the era of sharing economy. Although such service brings convenience to people's lives, it also causes safety and property concerns. Many studies have been conducted to access the efficiency and effectiveness of car-hailing, but little has been done on its safety monitoring. However, with the rapid development of information technologies such as Internet of Things (IoT), Geographical Information Science (GIS) and automatic monitoring, a more advantageous approach than the current simple drivers screening and testing is feasible. A new model including five indexes i.e. region dangerous index, offset distance of the origin-destination, real-time speed under traffic conditions, vehicle travel time and passenger information, is therefore proposed in this paper based on big data mining of the historical vehicle GPS trajectory data. Experiments were conducted to validate the model in the Gangzha District of Nantong City, China. Several other types of data were used in the experiments, e.g. points of interest (POI), road network data and urban image. The results showed that the proposed model effectively monitored the vehicle when it was driving in a “potentially dangerous area”. In addition, the model could accurately identify the driver's abnormal driving behaviors, such as bypass and abnormal stop. The prediction accuracy of the experiments was 92.06%, among which the discrimination accuracy of the abnormal stop was 100% and that of the detour was 90.57%. All these validate the applicability of the model for future management systems for car-hailing services.

INDEX TERMS Car-hailing; Big data; Trajectory data; Internet of Things (IoT); Geographical information science (GIS); points of interest (POI); potentially dangerous area;

I. INTRODUCTION

The online car-hailing industry has been a popular commercial travel mode that combines the Internet and the sharing economy [1]. The online car-hailing integrates the traditional operation mode of Cruise and Carpooling, making it easy for people to call driving service through mobile devices at any time or even in remote areas of the city. It greatly facilitates the travel of production and also solves the imbalance between the demand and supply of vehicles. Many existing studies have focused on the factors that affect sharing efficiency, e.g. [2-4]. At present, it has become one of the main modes of urban public transport travel. However, due to the low entry threshold and fast development speed of the "online car-hailing" industry, especially the imbalanced development in medium and small cities, the comprehensive quality of the workers in the industry is at different levels. The lack of high worker quality makes vehicle safety management more difficult, for example, a series of cases have been reported in which passengers and property safety was significantly infringed [3].

Government departments and management agencies have strengthened the management of laws and regulations, such as via the central-local two-level management mechanism.
To a certain extent, it helps solve the above problems [5, 6]. However, it is limited to regulatory supervision, and only raises the threshold for vehicles to join the online ride-hailing platform. As a result, while it works in the verification stage of drivers, it is not effective in strengthening the supervision and management of the whole transportation process. Meanwhile, the rapid development in mobile communications, sensor networks, and geospatial analysis technology allows sufficient hardware and software support for real-time monitoring [7-9]. This makes real-time network car supervision possible. However, currently available supervision platforms only focus on real-time location query and historical track playback [10, 11], owing to a large number of net cars managed by the platform [12, 13]. There is a lack of technology to automatically find abnormal vehicles from a large number of vehicles. In fact, the conventional monitoring platform is only semi-automatic, and it requires a lot of manual assistance to deal with special situations. This is not conducive to timely prevention of possible security issues, nor to the rapid emergency work in the event of problems. There is an urgent need for automatic supervision technology for car-hailing safety.

Existing studies have shown that there is a certain correlation between the crime rate and the geographical location [14, 15]. In addition, criminal behavior is highly correlated with spatiotemporal factors, often revealing high incidence periods and some hotspot spatial areas. Moreover, population density, economic status, and regional categories can have significant impacts on such areas [16, 17]. Therefore, it is of great significance to predict criminal behavior based on the spatiotemporal distribution principal. In the present study, we intended to select the sparsely populated and relatively remote areas based on the temporal and spatial characteristics of the crime committed by the driver of online car-hailing [18, 19]. To a certain degree, the selected areas are also related to the built-up environment of the city. In the past, there have been researching cases based on Landsat and nocturnal light data. Landsat 8 has a relatively high spatial resolution [20, 21], and the night light data can reflect on the more dynamic part of the building [22, 23]. Combined with the POI (points of interest) data provided by third-party social media websites such as Dazhong Dianping, it can further indicate the prosperous areas of the city [24, 25]. Through the integration of these data, we can extract the areas in the built region that are more prone to criminal behaviors to a certain extent, so as to provide a geographical reference for real-time supervision.

At present, most existing research has been dedicated to online car-hailing issues, for instance, the prediction of possible congestion using vehicle speed and other factors, the prediction of possible hot spots using historical empirical data [28, 29], the discovery of driver bypass fraud using trajectory data [30, 31], and other studies on outlier detection of urban traffic vehicles [32, 33]. Nevertheless, little has been done to monitor the safety of car-hailing considering the multiple driving indexes comprehensively as well as the geographic environment.

A novel approach is proposed in this study, which includes several innovations as following. First, it is a quantitative model that integrates five indicators into a score, namely "risk score", to evaluate the driving state of the vehicle. The higher the score, the more likely the vehicle is in an abnormal state. Base on the method, we can easily find several suspect vehicles from a massive set, and this makes our current management objects more streamlined, thus our work efficiency can be improved. Second, the geographical environment is considered in our model, this is also a highlight of our work. According to the high correlation between driver crime and the geographical environment, the practice of quantifying the geographical environment as an evaluation factor will have a huge contribution to the safety detection of vehicles. Finally, the option setting of this model is flexible, in case of different monitoring purposes, different kinds of indexes combination can be selected. The number of factors and the range of the parameters can be adjusted according to different objects. The greater the amount of data accumulated in the future, the more simulations will be performed, and the settings of these indexes and parameters will become more reasonable.

The rest of this paper is organized as follows: Section 2 and 3 introduces the main ideas and design course of the model; study area, data source, data processing are given in Section 4; Section 5 describes the results and discussion, followed by a summary in Section 6.

II. METHODOLOGY AND MODEL

A. MOTIVATION AND CONCEPTS

Criminal behaviors are closely related to temporal factors and tend to concentrate in a high-incidence time period and some hotspot areas. Based on this, a concept of "potentially dangerous zone" is proposed and implemented in the present study. A potentially dangerous zone can be an area with sparsely populated (low-vitality area), remotely distributed (non-built-up areas), surrounding hollow-ness (mountains and forests, construction land) or an area where passengers have high rate of criminal record, among which specific areas such as sparsely populated, remote and open zones can be regulated according to regulatory needs and actual conditions. The data show that the proportions of crimes committed by online car-hailing drivers differ in different time periods. Accordingly, one day can be divided into different periods: high-incidence section, potential danger section, and safety section [3]. Based on this concept, this paper assigns corresponding risk factors to the dangerous areas in different time periods to achieve reasonable supervision of urban areas.

The vehicle GPS trajectory describes the current driving state and can be used not only for traffic monitoring/forecasting [18], personalized route recommendation [37],...
driving route selection [20] but also for monitoring the driver’s abnormal driving behavior [38]. In the view of the monitoring of passenger safety status, this study will use the big data analytical technology to select appropriate model factors to be combined with the potential danger areas for a comprehensive analysis. When a vehicle is driving in a potential danger zone, the model will calculate the risk score. Combined with the particularity of cases of online ride-hailing drivers of high risk to infringe on passengers, this model contains the processing of passenger gender and quantity information, that is, when the passengers are female and the number is small, the level of ride safety will be reduced. In doing so, the model can effectively provide security assurance.

Based on big data analysis technology, this study aims to mine the origin and destination (OD) frequency law of different orders of vehicles and then selects the model superiority factors. The OD frequency law refers to the frequency law of different order routes of vehicles based on the same starting position and ending position. The route with high frequency is the normal driving route, and the route with low frequency may have abnormal driving behavior e.g. detour of the driver [36, 37].

In this paper, OD frequency analysis of 15.36 million taxi order data of Nantong, China in 2018 is carried out. Figure 1 demonstrates how the space-time law of order trajectory with low frequency is applied in this analytical process of origin-destination frequency. Through analysis, the time and distance factors of this kind of orders are higher than those of normal driving orders and have certain regularity. Considering the particularity of online car-hailing cases, such as parking crime, this paper takes vehicle OD distance, abnormal speed and driving time as model factors.

**FIGURE 1.** Analysis process of Origin-Destination frequency (After the origin trajectory data filtering, the main routes can be found, and the frequency of each route can be counted)
B. CONSTRUCTION OF MODEL

A multi-parameter safety state monitoring model is constructed based on a series of factors. The current driving safety can be judged by calculating the real-time risk score $P_{\text{total}}$ of the vehicle:

$$P_{\text{total}} = W \times D_i \times \sum_{i=1}^{l} \theta_i \times P_i \quad (1)$$

In the formula, $D_i$ is the risk index of the potential danger zone, $P_i$ is the risk score of each factor, $W$ is the passenger safety factor, $P_{\text{total}}$ is the total risk score, and $\theta_i$ is the weight of $P_i$. The structure of the model is shown in Figure 2.

![Modeling structural diagram](image)

III. INDICATORS OF MODEL

In this model, the travel trajectory and passenger information of vehicles in a potentially dangerous zone are analyzed, and the corresponding risk score is given. The safety condition of passengers in vehicles is monitored according to the total risk score $P_{\text{total}}$. The detailed analytical steps are as follows.

A. POTENTIALLY DANGEROUS ZONE

The steps for dividing the potentially dangerous zone are as follows:

1) According to the safety conditions, time period of the city is divided into high-incidence period $T_1$, potential danger period $T_2$, and safety period $T_3$.

2) Based on the POI data as well as the pick-up and drop-off points of urban vehicles at different time periods, make low-thermal regions are classified and established, from which dense areas are assigned with the risk index $D_1$.

3) Extract the urban non-built areas and assign the risk index $D_2$.

4) Subject the result data of step 2 and step 3 to the overlay operation. The overlay result is taken as the remote region $Sp1$, and the danger index of the corresponding area is superimposed.

5) Extracting the construction site and taking the forest area as the open zone $Sp2$, and assigning it the risk index $D_3$.

6) For areas where relevant cases have occurred or have been complained about, the point set is used to describe $Pi$, and a buffer zone with a radius of $r$ is built around it, which is the high-incidence zone $Sp3$, and the hazard index $D_4$ is assigned.

7) Combine the risk index of $Sp1$, $Sp2$, and $Sp3$ and superimpose the risk index accordingly. The area where the risk index is not 0 is the potential danger zone.

8) The risk degree of potential danger zone will change in different time periods. In the high incidence period of the case, the risk index will be increased by $I_1$, while the risk index will be increased by $I_2$ in the potential danger zone. The risk index will remain unchanged in the safety period. The formula is as follows:

$$D_1|T_j = I_j + D_i \quad (2)$$

In which, $D_1|T_j$ means that the risk index of the zone $i$ in the period of $T_j$ is $D_i$, where $j$ is 1 or 2.

"Dangerous degree index", as shown in Table 1, can be assigned to the corresponding index according to the regional characteristics [24], in order to differentiate degrees of dangerous zones in the remote areas, open zones and high-incidence zones.

<table>
<thead>
<tr>
<th>Type</th>
<th>Element</th>
<th>Subdivision of risk index (1-20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote zone</td>
<td>Non-built-up area</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>low-vitality area</td>
<td>5</td>
</tr>
<tr>
<td>Open zone</td>
<td>Mountain and forest area</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Construction areas</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>The area where have been complained</td>
<td>19</td>
</tr>
<tr>
<td>High-incidence zone</td>
<td>The area where cases have occurred</td>
<td>20</td>
</tr>
</tbody>
</table>
B. OD DISTANCE OF VEHICLES
For a long time, there have been widespread cases in which passengers are detoured by some drivers and charged more. Especially when driving in unfamiliar areas, where passengers may not be able to recognize such anomalous behavior. For this phenomenon, the real-time vehicle trajectory can be studied to analyze the current route and compare it with the route with high OD frequency [22, 23], so as to judge whether the driver has detour behavior.

The above is carried out by analyzing the potentially dangerous zones in conjunction with the characteristics of specific cases.

When the vehicle drives in a dangerous area, the distance between the current position and the destination is not reduced but is larger than the distance between the previous track point and the destination. Moreover, when the distance difference exceeds a certain threshold value, the danger assessment is carried out, and the specific method is as follows:

1) Initialize the danger score $P_1$ of the current vehicle order as 0 and the flag $dis\_flag$ in the "distance increasing state" as 0.

2) When the vehicle drives into a potential danger zone, calculate the real-time distance $S(\text{Pos}_i, \text{Des}_i)$ between the vehicle's current location $\text{Pos}_i$ and the destination $\text{Des}_i$:

$$S(\text{Pos}_i, \text{Des}_i) = \sqrt{(\text{Des}_i.x - \text{Pos}_i.x)^2 + (\text{Des}_i.y - \text{Pos}_i.y)^2} \quad (3)$$

In the formula, $n$ is the number of the order, $i$ is the trajectory point number of order $n$.

3) Judge whether $S'$ is greater than 0. If greater than 0, set $dis\_flag$ as 1, and judge distance difference $\Delta S$; when $\Delta S$ exceeds the threshold value X, the risk score $P_1$ is updated to $P_1 + 1$ for every $n$ meters. Among them, $S'$ and $\Delta S$ are calculated as follows:

$$S' = \frac{dS(\text{Pos}_i, \text{Des}_i)}{dt} \quad (4)$$

$$\Delta S = S(\text{Pos}_{i-1}, \text{Des}_{i-1}) - S(\text{Pos}_i, \text{Des}_i) \quad (5)$$

In the formulas, when $S'$ is greater than 0, $S(\text{Pos}(n), \text{Des}(n))$ will increase.

C. ABNORMAL SPEED OF VEHICLES
The speed information in the GPS data can directly reflect the current state of the vehicle. Through the speed variation and speed magnitude, the driving state of the vehicle can be directly identified. When the vehicle speed is high or the speed variation is large, the driver will have abnormal driving behavior [21, 23]. In the past cases of online car-hailing drivers infringing on passengers, drivers would choose to stop to commit crimes [17]. At this time, GPS would record the current vehicle speed state, so it is crucial to judge abnormal parking behavior in the order process.

If the parking time is not too long, it may be because of waiting for the red light or traffic jam. If it exceeds a certain threshold, the behavior can be judged as abnormal behavior, and the risk assessment is carried out. The specific methods are as follows:

1) Initialize the risk score $P_3$ of the current order as 0 and the sign $time\_flag$ which is in "abnormal parking state" as 0.

2) When a vehicle is driving in a dangerous area, if the speed is 0, the real-time traffic congestion and traffic light waiting condition are analyzed to judge whether the vehicle is in a congested or red light state, that means, then the vehicle data either falls into the congested section or in the traffic light buffer zone, $stop\_flag$ will be 1.

3) When the $stop\_flag$ is 1, and the parking time exceeds the threshold $Y$, the risk score $P_3$ is updated to “$P_3 + 1” for every $n$ minutes.

In practical applications, considering that the driver tends to change shifts and take short breaks when they are in "passenger-carrying mode", the order behavior of "stop state" may result in the excessive risk score. To correct this, the data source will be processed. When the number of passengers is 0 (at which state the APP does not transfer passenger data to the server), the current order will not be counted.

FIGURE 3. Figure 3. Road Congestion Diagram

D. DRIVING TIME OF VEHICLES
Abnormal driving behaviors of drivers are not only manifested in OD path but also can be perceived in time. Combining the current position and time of vehicle trajectory, space-time analysis is conducted, and the abnormal behaviors of vehicles can be identified by comparing the travel time of actual path and the travel time of shortest path [25].

In the model, when the vehicle has been driving for longer than the estimated OD time, and the time difference exceeds a certain threshold, the behavior is judged to be a potential abnormal behavior, and the risk is assessed. The specific method is as follows:

1) Initialize the risk score $P_2$ of the current order as 0 and the sign $time\_flag$ which is in “abnormal time state” as 0.
2) Calculate the estimated OD time $T_p$ through Route Matrix API provided by Baidu Map.

3) Calculate the driving time $T_e$ and $\Delta T$. When $\Delta T$ is greater than 0, time_flag is 1. $T_e$ is calculated as follows:

$$T_e = time_e - time_o \quad (6)$$

In the formula, time_e is the current time, time_o is initial departure time.

4) When time_flag is 1, calculate $\Delta T$. If $\Delta T$ is greater than the threshold $T_s$, the score $P_3$ is updated to $P_3+1$ for every $m$ seconds. Where $\Delta T$ is calculated as follows:

$$\Delta T = T_e - T_p \quad (7)$$

In which, $T_p$ is the estimated driving time.

### VI. INFORMATION OF THE PASSENGERS

In the past, most of the victims of online car-hailing cases have been associated with female passengers, and when there were few passengers in the car [17]. There was a strong correlation between the gender and number of passengers and the probability of the crime case [24]. Therefore, it is very important to analyze the case based on the information of passengers. In this model, by analyzing the gender and number of passengers, the appropriate risk assessment coefficient is given. The specific methods are as follows:

1) Initialize the passenger risk coefficient $W_c$ as 0 and the sign of “gender of passenger” sex_flag as 0.

2) Use the self-developed mobile APP to get the number of passengers count (count ≥ 1) and gender. When all the passengers are female, sex_flag is 1 and $W_c$ is 1.

3) When sex_flag is 1, the risk coefficient of each additional passenger is reduced by $w_c / (1 ≤ i ≤ 3)$, and $W_c$ is $W_c - w_c$.

### B. POI DATA

The POI data was downloaded from the Baidu Map API by the web crawler technology, which includes 1052 records with property fields such as residence, school, food, hospital, scenic spot, and community, and other keywords such as coordinates, and names et al, Table 3 shows different parts recorded for of Nantong City.

<table>
<thead>
<tr>
<th>License plate number</th>
<th>Time</th>
<th>Speed</th>
<th>Lon</th>
<th>Lat</th>
<th>Order</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Su FB****</td>
<td>2018-12-23 02:11:46</td>
<td>0</td>
<td>120.8499</td>
<td>32.01197</td>
<td>1</td>
<td>Pick-up point</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23 02:12:51</td>
<td>7.6</td>
<td>120.8498</td>
<td>32.01182</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23 02:18:44</td>
<td>24.9</td>
<td>120.8685</td>
<td>32.01253</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23 02:19:10</td>
<td>18.3</td>
<td>120.8684</td>
<td>32.01493</td>
<td>1</td>
<td>Drop-off point</td>
</tr>
</tbody>
</table>

### TABLE2. Partially Processed Trajectory Data Record

### TABLE3. Partially Processed Trajectory Data Record
C. ROAD NETWORK DATA AND ITS CONGESTION CONDITION

The experimental data was the OSM open-source road network first-level road data [34], whose attributes include road name, geometric type and other characteristics. The real-time road condition query interface provided by the web crawler and Baidu Map and the Ali Cloud server are combined to obtain the real-time road condition from 0 o'clock on the 1st of December 2018 to 0 o'clock on the 1st of January, 2019[35]. As is referred to by the Baidu Map standard, the status of 1 stand for unblocked, which means that the vehicles can run fast; 2 stands for slow, that means the vehicles slow down but do not stop; 3 stands for congestion, which means vehicles are congested. 4 represents serious congestion, in other words, the vehicles are difficult to drive. Because of the delay in the HTTP request and data return, the time in congestion data was not in full hours, so it needs to be corrected. The data is processed in batches, and Table 4 is the road congestion status data after time processing.

<table>
<thead>
<tr>
<th>Time</th>
<th>Road_name</th>
<th>Status</th>
<th>Status_desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018/12/04 15:30</td>
<td>Gongnong Road</td>
<td>2</td>
<td>Slow</td>
</tr>
<tr>
<td>2018/12/04 16:00</td>
<td>Gongnong Road</td>
<td>1</td>
<td>Unblocked</td>
</tr>
<tr>
<td>2018/12/04 16:30</td>
<td>Gongnong Road</td>
<td>1</td>
<td>Unblocked</td>
</tr>
<tr>
<td>2018/12/04 17:00</td>
<td>Gongnong Road</td>
<td>1</td>
<td>Unblocked</td>
</tr>
<tr>
<td>2018/12/04 17:30</td>
<td>Gongnong Road</td>
<td>1</td>
<td>Congested</td>
</tr>
</tbody>
</table>

D. REMOTE SENSING IMAGE DATA

The remote sensing image data of Lujia-01 night-light data of the 29th of October, 2018 was downloaded from the Hubei Data and Application Network of the High-Resolution Earth Observation System. The Landsat-8 OLI image data was the image of December 2017 downloaded from the Geospatial Data Cloud. The original data contained 12 bands, with the cloud volume of 0.46% and the spatial resolution of 30 meters, and Figure 4 shows the sample data within the study area.

V. RESULT AND DISCUSSION

A. PARAMETERS SETTING

The vehicle trajectory data and the POI data respectively describe the dynamic and static urban vitality characteristics. In order to distinguish the vitality of different areas of the city, the nuclear density analysis was carried out on pick-up and drop-off point and POI data [25-26], from which high and low-density areas were identified, and sparsely populated areas were extracted.

In recent years, many scholars have extracted urban built-up areas by means of night-light remote sensing data and conducted studies on urban expansion, GDP development level [27]. Based on the hypothesis that greater light intensity, results in higher possibility of urban built-up of the area, the present paper uses a bisection method to set the threshold which can be continuously adjusted by comparing with the statistical data, and to extract the non-built area [28] and set as a remote zone in the city. Through the digitization of Landsat-8 OLI image, extract mountain forests and building land were also extracted as open areas in the city. The areas extracted above are assigned the risk index and merged. Then, by superimposing the risk index of the corresponding location, the potential danger zone in $T_1$, $T_2$, and $T_3$ periods (in this paper, $T_1$ is 0:00~6:00, $T_2$ is 15:01~24:00, $T_3$ is 6:01~15:00) were extracted (Figure 5).
According to the parameter selection of OD distance, speed anomaly and driving time, we selected the orders that may have abnormal behavior (driver detour and stop time is too long) as the experimental data, and compared the effect of different parameters on selected factors among different groups. These orders involving the passenger complaint were extracted from the Nantong taxi management system.

Figure 6 illustrates how orders with possible detour and abnormal stop behavior were displayed. These orders may correspond with the possible detour and abnormal stop behaviors, and then, we can easily find the strange graph after its trajectory data been displayed on the map.

From these data, risk scores of each factor were calculated for different parameters, and the results are shown in Figure 7.

FIGURE 6. Order trajectory with abnormal behavior (a) detour behavior (b) abnormal stop behavior

FIGURE 7. Comparison of various factors' risk scores (a) $P_1$ risk score (b) $P_2$ risk score (c) $P_3$ risk score
For orders with abnormal behavior, the calculated results of the risk score of each factor in the model better reflect the real-time state of vehicles. To more accurately assess the real-time safety state of passengers, on the premise that the danger integral value is moderate and the real-time trajectory status can be accurately reflected. Taking the $P_1$ risk score as an example, as shown in Fig. 7(a), when $x$ is one-fifth of the $OD$ distance ($OD\_DIS$), abnormal fluctuations in behavior can be captured more easily. Under this condition, $n_1$ is selected as one-sixth of $x$, then the result is more reasonable. Because the simulation curve of $P_2$ and $P_3$ risk score is similar to $P_1$, so the parameter selection is similar as above. The optimal parameters were selected and listed in Table 5, in which $OD\_DIS$ is the required distance for OD.

We first extracted all the order data in the Gangzha district and took different orders as experimental objects to select the optimal parameters. Then, used the model script to perform a batch processing to obtain real-time risk score of the vehicle. The Analytic Hierarchy Process (AHP) was used to calculate the weight, and the results were 0.3, 0.44 and 0.26 respectively. The weights were used to calculate the comprehensive risk score $P_{total}$. When the trajectory points were in dangerous areas, the potential danger index of the current trajectory area was assigned. Since the App is still in the developmental stage, the total risk score $P_{total}$ was calculated based on the case that the passenger is a single woman ($sex\_flag$ is 1) in the present study. Some results are shown in Table 6.

### TABLE5. Optimal parameters of safety analysis

<table>
<thead>
<tr>
<th>Factors</th>
<th>Parameter</th>
<th>The optimal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD distance</td>
<td>X</td>
<td>$OD_DIS/15$</td>
</tr>
<tr>
<td>Abnormal speed</td>
<td>n1</td>
<td>X/6</td>
</tr>
<tr>
<td>Driving time</td>
<td>n2</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>n3</td>
<td>1/6</td>
</tr>
<tr>
<td></td>
<td>Ts</td>
<td>Tp*2</td>
</tr>
</tbody>
</table>

**TABLE6. Parts records of score**

<table>
<thead>
<tr>
<th>License plate number</th>
<th>Time</th>
<th>Type</th>
<th>$P_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>Pick-up point</td>
<td>0</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:02:48</td>
<td>0</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:06:19</td>
<td>0</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:05:19</td>
<td>0</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:05:49</td>
<td>0</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:07:20</td>
<td>5.512</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:07:49</td>
<td>7.956</td>
</tr>
<tr>
<td>Su FB****</td>
<td>2018-12-23</td>
<td>02:08:20</td>
<td>9.945</td>
</tr>
</tbody>
</table>

### B. ANALYSIS AND DISCUSSION

Analysis of the study results showed that there were more orders in the $T_2$ period, and the risk scores in this period were generally higher than those in periods $T_1$ and $T_3$. The risk score of certain individual orders in the $T_2$ period could be dozens of times higher than other orders in other time periods. By tracking orders with higher risk scores, it was found that the $P_1$ and $P_3$ points were high in most moments of this particular high-risk order. The distance and time of the order in the model were found to be abnormal thus the vehicle was judged as having a detour behavior [38]. The same vehicle was found at some time travel in potentially dangerous areas, which ultimately leads to a higher calculated total score in the model (Figure 7). In addition, the risk score during the holiday was found to be generally lower than that of the working days. Due to the increased number of travelers, the number of orders also increased. As a result, the probability of illegal behavior of the driver was actually decreased and the safety of the passengers was significantly improved. This suggests that the monitoring model proposed in the present study to be reliable.

**FIGURE 8.** Maximum score order trajectories in $T_2$
By comparing the risk scores of the orders on the 23rd of December and the 24th of December, it was found that latter were generally lower, and the number of orders in the T1 on the 24th of December was about twice as that of the same period on the 23rd of December. The number of orders increased as a direct result of being near Christmas Eve. Such increase led to a reduced probability of illegal behavior of the drivers, significantly improved the safety of driving, which proved again the reliability of the proposed detection model.

By analyzing the safety situation, it was found that the risk score $P_{\text{total}}$ was generally higher in the $T_2$ on the 23rd and 24th. Further analyses of the trajectory of the order and the $P_1$ and $P_2$ scores of the model, it could be concluded that detours in individual vehicles might have occurred. In all periods, when $P_{\text{total}}$ of the vehicle is not zero, the $P_2$ score of most vehicles is low, meaning that the number of abnormal stops of vehicles is less. Because of the larger weight of $P_2$ in the model, the order is deemed less risky. In the present study, interviews were conducted with drivers with a high $P_{\text{total}}$ and a high $P_2$. It was found that when the red light time was too long, the score tends to be higher. The present model could, therefore, change parameters accordingly. In addition, by comparing data between holidays and workdays, it was concluded that it would be best for female passengers to choose to travel when the traffic volume is large and to avoid choosing the time when the number of passengers is small, so as to improve the safety of driving.

Since there are no examples of passengers being infringed in the experimental area, the "potential danger zone" could not be established by referencing the real crime locations, and the abnormal behaviors of vehicles will be evaluated by comprehensive use of the other factors. The orders were sorted in a descending order based on the total value of $P_{\text{total}}$, and the top 0.5%, which is 63, of the total 12600 orders were selected.

The trajectory data of the selected orders were identified using the taxi management system and OD frequency analysis method based on big data analysis. As a result, 58 orders were identified to have abnormal behaviors, such as detour, abnormal stop, and the experimental accuracy reached 92.06%, among which the discrimination accuracy of the abnormal stop was 100% and that of the detour was 90.57%. Partial results are shown in Table 7. Among them, $P_2$ of most vehicles was lower, which suggest that the number of abnormal stops of vehicles was less, although some orders remained in higher scores. Some of this be due to the abnormally long red light time, therefore by interviewing drivers with higher $P_{\text{total}}$ and higher $P_2$ orders, the proper adjustment was performed for the model parameters to better reflect the characteristics of urban traffic.

Due to the lack of considering the actual situation of urban traffic, such as road congestion and road construction status, the estimated OD time of the current model was calculated based on Baidu Map API under ideal conditions. As a result, some calculated results may be obviously less than the actual required time, which leads to certain errors in the model results.

### TABLE 7. Result of Model Checking

<table>
<thead>
<tr>
<th>License Plate Number</th>
<th>Start Time of Order</th>
<th>End Time of Order</th>
<th>Maximum Value of $P_{\text{total}}$</th>
<th>Type of Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Su 2018-12-02</td>
<td>07:50:30</td>
<td>16:17:20</td>
<td>104.843</td>
<td>detour</td>
</tr>
<tr>
<td>FB**** 2018-12-23</td>
<td>07:00:18</td>
<td>17:37:34</td>
<td>91.702</td>
<td>abnormal stop</td>
</tr>
<tr>
<td>Su 2018-12-23</td>
<td>07:50:30</td>
<td>08:03:30</td>
<td>228.634</td>
<td>detour</td>
</tr>
</tbody>
</table>

### VI. CONCLUSION

In summary, the present study examined the safety monitoring of "online car-hailing" passengers from the perspectives of geographical time and space, and proposed a multi-factor safety state monitoring model based on "potentially dangerous zone". By analyzing different factors and their combined effects, potentially dangerous zones in cities were extracted. Model factors were selected based on big data analysis, and better parameters were selected based on multi-group experimental analysis. The real-time trajectory behavior of vehicle orders was analyzed by using space-time behavior. The experimental results showed that the model has a better monitoring effect than exist APP. However, in practical application, due to the location accuracy of vehicle GPS trajectory points and the calculation accuracy of the Baidu Map API, there is still some inaccuracy in this proposed monitoring model. In the future work, we will focus more on solving the impact of data and will take into consideration the actual complex traffic conditions to reduce the scoring error.

### ACKNOWLEDGEMENT

The authors would like to thank the editor and the anonymous reviewers who provided insightful comments on improving this article and thank The Hong Kong Polytechnic University for providing the opportunity of academic exchange. We also thank Mr. Rong Tsao, Research Scientist of Agriculture and Agriculture-Food Canada, for his valuable advices on the improvement of English writing.
REFERENCES


TONG ZHOU is an Associate Professor of School of Geographic Sciences and Professional Head of Geographic Information Science, Nantong University. He is currently a visiting scholar of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. He received a bachelor's degree from Jiangsu Normal University in 2000, and receive the M.S. degrees from Nanjing Normal University in 2004. His teachings in the areas of remote sensing and GIS algorithm, his research interests include geographic information science, urban informatics, human mobility, and spatiotemporal data analysis and mining.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2939787, IEEE Access

T Zhou et al: A Novel Approach for Online Car-Hailing Monitoring Using Spatiotemporal Big Data

WENZHONG SHI is Head of Department of Land Surveying and Geo-Informatics, Otto Poon Charitable Foundation Professor in Urban Informatics, Chair Professor in GISci and remote sensing, Director of Laboratory for Smart City and Spatial Big Data Analytics, The Hong Kong Polytechnic University. He obtained his doctoral degree from University of Osnabrück in Vechta, Germany in 1994. Prof Shi’s current research interests are in the areas of urban informatics for Smart Cities, GISci and remote sensing with focusing on analytics and quality control for spatial big data, object extraction and change detection from satellite images and LiDAR data, integrated mobile mapping technology, and 3D and dynamic GISci modeling. Prof Shi served as President of Commission II for International Society for Photogrammetry and Remote Sensing (2008-2012), President for Hong Kong Geographic Information System Association (2001-2003). He also serves as an editorial board member for a number of international journals. He has published more than 400 scientific articles (with over 200 SCI papers) and 15 books. He received a number of prestige awards, including an award from International Society of Photogrammetry and Remote Sensing and Natural Science Award from the State Council, China.

XINTAO LIU received the B.Eng. degree in the survey from Hohai University, China, in 1998, the M.Sc. degree in cartography and GIS from Nanjing Normal University, China, in 2003, and the Ph.D. degree in geo-informatics from the Royal Institute of Technology, Sweden, in 2012. In 2012, he joined the Department of Civil Engineering, Ryerson University, Canada, where he was a Postdoctoral Fellow in GIS and transportation until 2016. He was a Sessional Lecturer with Ryerson University, since 2015. He is currently an Assistant Professor with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. He is also a PI and a Co-PI of several national projects funded by Sweden, Canada, and Hong Kong. His research interests include GI services and science, urban computing, and GIS in transportation. His research goal is to use state-of-the-art technologies to advance smart city for a better urban life. He received the Ph.D. Scholarship from Lars Erik Lundebergs.

FEI TAO received the B.S. degree in geography teacher and the M.S. degree in marine geography in 2004 and 2007 from Nanjing Normal University, Nanjing, China. She is currently pursuing a Ph.D. degree in cartography and geographic information system.

From 2007 to 2019, she was a lecturer with the School of Geographical Sciences, Nantong University, Jiangsu, China. She has presided over a project supported by the Young Scientists Fund of the National Natural Science Foundation of China and published more than 10 papers. Her research interests include remote sensing applications and geographic big data analysis.

ZHEN QIAN was born in Zhenjiang, Jiangsu Province in 1997, he entered the School of Geographic Sciences of Nantong University in 2016, Jiangsu Province, and majored in Geographic Information Science. He is mainly engaged in spatiotemporal data mining research, published six papers on spatiotemporal data mining, applied for four national patents of China, and two software copyrights. Mr. Qian has won many national awards and scholarships, including the M Prize of the American Mathematical Modeling Competition and the National Scholarship.

RUIJIA ZHANG was born in Lianyungang, Jiangsu, China in 1997. She received a bachelor's degree from Nantong University in 2019, and has been admitted to Nanjing Agricultural University as a graduate student with the major in English translation. From 2018 to 2019, she was a member of Intelligent Analytical Teacher-student Team for Geographic Big Data, Nantong University, Jiangsu. Her research interest are the development of Chinese and Western translation theories and geographic information technology. Her awards include the second prize in National English Competition for College Students and Preliminary Round of the 2018 "FLTRP Cup" English Writing Contest, the Meritorious Winner award in 2018 Mathematical Contest in Modeling, which she worked as a translator.