Electronic Research Archive

ERA, 30(12): 4507–4529.
DOI: 10.3934/era.2022229
Received: 19 July 2022
Revised: 10 September 2022
Accepted: 20 September 2022
Published: 17 October 2022

http://www.aimspress.com/journal/ERA

Theory article

A taxi detour trajectory detection model based on iBAT and DTW algorithm

Jian Wan[1,2], Peiyun Yang[1,3,*], Wenbo Zhang[1], Yaxing Cheng[1], Runlin Cai[3] and Zhiyuan Liu[1,*]

[1] School of Transportation, Southeast University, China
[2] Research and Development Center on ITS Technology and Equipment, China Design Group, China
[3] China Urban Planning and Design Institute Shanghai Branch, China

* **Correspondence:** Email: 942164919@qq.com, zhiyuanl@seu.edu.cn.

Abstract: Taxi detour is a chronic problem in urban transport systems, which largely undermines passengers' riding experience and the city's image while unnecessarily worsening traffic congestion. Tourists unfamiliar with city roads often encounter detour problems. Therefore, it is important for regulatory authorities to develop a tool for detour behavior detection in order to discover or identify detours. This study proposes a detour trajectory detection model framework based on the trajectory data of taxis that can identify taxi driving detour fraud at the microscopic level and analyze the characteristics of detouring trajectories from the perspective of microscopic motion traits. The deviation from normal driving trajectories provides a framework for the automatic detection of detour trajectories for the off-site supervision platform of the taxis. Considering drawbacks of the isolation-Based Anomalous Trajectory (iBAT) algorithm, this paper made further improvements in trajectory anomaly detection. In this study, three methods including the iBAT, iBAT + Dynamic Time Warping (DTW), and iBAT + DTW algorithms considering the driving distance and time are compared using the relevant experimental data. The case studies verify that the proposed method outperforms the other methods. Verified by the experiments based on the trajectory data coming from Nanjing, the false positive rate of this framework is only 1.64%.

Keywords: taxi trajectory data; iBAT; DTW; anomaly detection

1. Introduction

Currently, taxis in major cities are generally equipped with a global positioning system (GPS). The GPS used by taxis has high data accuracy, extensive coverage, and the ability to obtain dynamic positioning and timing information in real time. The on-board GPS equipment of taxis can not only assist the taxi supervision and dispatch system in knowing the status of taxis (including vacancy, busy, and rest) and the traffic flow condition of the road network [1], but it can also help the taxi management center in supervising the operational behavior and efficiently managing taxis by analyzing vehicle operating data [2]. Taking Nanjing city as an example, which has an average of approximately 15,000 taxis put into operation every day, each taxi uploads its location information in real time through the on-board GPS, which is uploaded once every 60 s on an average, generating approximately 300 million trajectories every day. These data are analyzed to offer a decision-making basis for vehicle operation supervision by analyzing taxi operation characteristics and indicators.

Scholars worldwide have conducted various studies based on taxi travel trajectory data, which are rich in traffic information. The majority of research focuses on discovering the travel characteristics of urban transportation [3–6], identifying areas of interest [7], discovering behavior patterns [8], exploring the laws of mobility interactions [9], and analyzing urban road conditions and traffic accessibility [10]. However, when compared with regular information, the data contains some interesting information that usually involves anomalous behavior patterns associated with events [11,12]. For example, a previous study identified hacking behaviors from abnormal network data, discovered drunk or reckless driving behaviors from abnormal traffic flow data, detected bank fraud incidents from abnormal credit card transaction data, and identified hazards from abnormal medical images.

In the domain of transportation, a small fraction of outliers detected in taxi datasets may contain abnormal behavior patterns [13], which may be due to detours, events (such as concert, fair, gatherings), unlawful pricing, or even taxi hijacking. Among them, the detour was the most worthy of our attention. It is a serious problem of urban transportation systems that has negatively impacted passengers' perceptions of the city and their sentiments. Some unscrupulous taxi drivers profit by purposefully detouring to extend their driving distance [14,15]. The illegal operation of taxis not only violates the legitimate rights of passengers but also disrupts the normal transport order, damages the operating interests of passenger transport enterprises, and reduces the standard of service provided by the taxi system.

According to studies, the detection problem of abnormal taxi behaviors (such as detours for no reason and illegal gatherings) can be viewed as a specialization of the general problem of identifying data with different patterns [16–20]. The main task of detecting abnormal data is to identify data containing behavioral differences. More specifically, detecting abnormal data involves finding datasets that are distinct from most of the datasets [21]. Most studies on abnormal taxi behavior detection have focused on a series of points representing taxi trajectories. Generally, there are two ways to identify abnormal datasets: identifying datasets that differ from the global behavior (i.e., anomalous behavior detection for a single trajectory), and looking for data groups that differ from neighborhoods (especially generalized local neighborhoods) after grouping all datasets.

The feature-based abnormal behavior detection method is used to establish a global feature model based on the existing trajectory data [22], and the trajectory data that deviates significantly from the global feature model is classified as abnormal behavior. For example, by combining geographic features with semantic descriptions of each trajectory site, Palma et al. extracted anomalous

information from trajectory data [23]. Grady and Schwartz used vehicle position, speed, and corner direction as discriminative features and used comprehensive discriminant indicators to identify abnormal vehicle behaviors [24]. Combining mathematical and statistical methods, Zhao et al. established a feature matrix of target trajectories and regarded trajectories with unsatisfactory confidence as abnormal [25].

Meanwhile, time-series analysis can be used to analyze behavioral changes in trajectories over time. By visualizing the trajectory data, the shape of the trajectories can be recognized and their similarity can be identified [26]. For example, Ibrahim used Euclidean distance and Dynamic Time Warping (DTW) to perform hierarchical division and similarity discrimination of taxi trips, group similar trips, and study their trends over time [27]. Clustering methods are one of the most popular techniques for grouping data into homogeneous clusters and aim to minimize the distance between individual data groups within a cluster [23,28,29].

In general, the detour behavior detection of taxis based on trajectory data has achieved several advances. However, there are several problems that persist. First, taxi detour behavior detection is commonly conducted by experienced staff based on the feedback from passengers. However, this method is inefficient, time-consuming, and labor-intensive [30]. Meanwhile, it is usually difficult for passengers who are not familiar with local roads to identify detours, and abnormal data are usually eliminated as "noisy data" in practical applications. Second, existing anomaly detection systems have shortcomings, such as single discrimination criteria and high false alarm ratios in complex urban road network environments. There is a lack of research on the combination of anomaly detection algorithms and the joint use of spatio-temporal data to identify detours. Therefore, appropriate detour behavior detection algorithms that meet the needs of intelligent supervision services are urgently required. With the current advancement of smart and off-site law enforcement, this study proposes a detour behavior detection model based on Nanjing taxi trajectory data. This method uses grid-based abnormal trajectory detection algorithms to detect abnormal data and analyze the microscopic characteristics of detour trajectories.

2. Problem description

2.1. Dataset

The raw taxi trajectory data were obtained from nearly 20,000 taxis in Nanjing in 2021. The dataset includes the location, speed, direction, status, and other attributes. The data features included SERIAL, ORDER_ID, VEHICLE_NO, POSITION_TIME, LON, LAT, SPEED, DIRECTION, MILEAGE, vehicle VEH_STATUS, ACC, and operating status. These features were recorded at least every 60 s. The description of the individual features is as follows. The ACC feature represents the status of the engine, with values 1 and 0. When the value of the feature ACC equals 1, it indicates that the engine is working. In contrast, when the value of the feature ACC is 0, it indicates that the engine is off. For the feature representing the order number, the value of 0 represents the non-operational status. The value range of the feature representing the direction angle is 0–359 in the clockwise direction. The unit of MILEAGE is the KM. There are four operating statuses (1: passenger, 2: order, 3: empty, and 4: out-of-service). Table 1 lists the features and examples included in the car-hailing positioning information data.

Table 1. Example of car-hailing location information data.

Number	Feature Name	Type of data	Description
1	SERIAL	NUMBER (18)	serial number
2	ORDER_ID	VARCHAR2 (64)	order number
3	VEHICLE_NO	VARCHAR2 (32)	vehicle number plate
4	POSITION_TIME	VARCHAR2 (14)	positioning time
5	LON	VARCHAR2 (10)	longitude
6	LAT	VARCHAR2 (10)	latitude
7	SPEED	VARCHAR2 (10)	instantaneous speed
8	DIRECTION	VARCHAR2 (10)	direction angle
9	MILEAGE	VARCHAR2 (10)	mileage
10	VEH_STATUS	VARCHAR2 (10)	vehicle status
11	ACC	CHAR (1)	engine status
12	Operating Status	CHAR (1)	operating status

2.2 Data preprocessing

2.2.1 Grid meshing

Taxi trajectory data are used as spatiotemporal data, and grid meshing is a basic method for spatiotemporal big-data analysis. In this study, the advantages are mainly reflected in the following points: First, the trajectory data were originally continuous points scattered on the map. Meshing can discretize the geographic space into small areas one by one, making the analysis of point data easier. Second, the grids had the same size, which ensured that their properties could be compared. This method can also control the grid size and the accuracy of the analysis. The last advantage is that the GPS point trajectory is transformed into a grid sequence that can achieve rapid data correspondence and effectively improve the efficiency of anomaly detection.

As shown in Figure 1, the trajectory points, grid cells, and map data were first connected and corresponded. After the GPS data were gridded, each data point contained the corresponding grid information. When the grid is used to express the distribution of the data, the distribution situation represented by it is close to the real situation. Meanwhile, the GPS trajectory is transformed into a grid sequence, which can realize rapid data correspondence to effectively improve the efficiency.

The grid-based trajectory representation method includes the following steps. First, consider the geographic boundary of Nanjing as the boundary of the research scope and divide it into separate square grids of 500 m * 500 m. As shown in Table 2, the grid number corresponds to the latitude and longitude of the grid center point. After obtaining the gridding parameters and corresponding GPS data for the grid, a grid can be jointly specified by the columns "LONCOL" and "LATCOL", which can be regarded as the horizontal and vertical coordinates of the grid. Thus, the grid was recorded as an $G_i = (\mathrm{LONCOL}, \mathrm{LATCOL}) = (x_i, y_i)$. At the same time, for the convenience of calculation, the trajectory was recorded as $t_i = (G_1, G_2, \ldots, G_i) = \left[(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\right]$. Each trajectory consisted of a series of continuous trajectory points or a continuous sequence of one or more grids.

Finally, the trajectory set T was recorded as $T=(t_0,t_1,\ldots t_n)$, and each trajectory set consisted of sub-tracks t_n with the same start and end points.

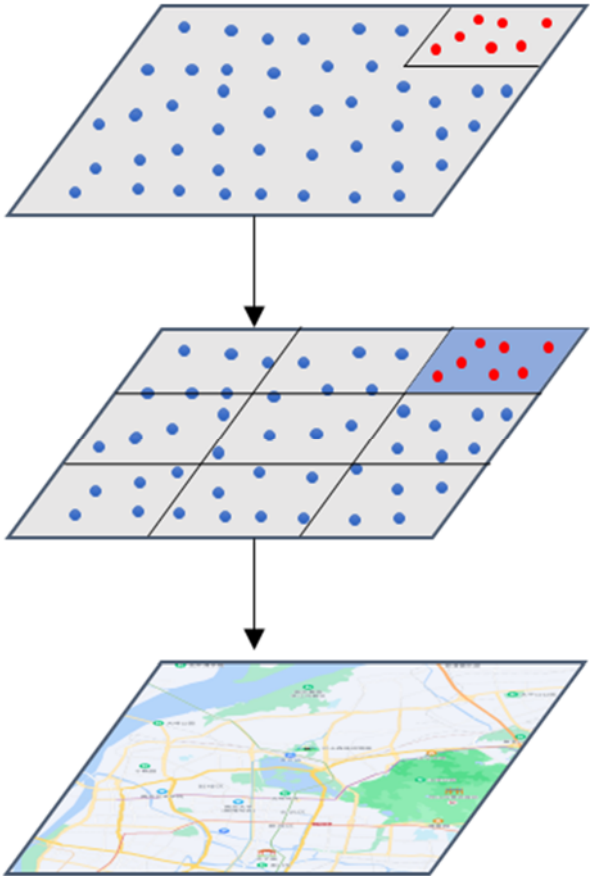


Figure 1. Schematic diagram of the meshing of trajectory points.

Table 2. Example of grid output corresponding to trajectory point.

VEHICLE_NO	POSITON_TIME	LAT	LON	ACC	SPEED	LON COL	LAT COL
***	2021-03-04 02:31:00	118.71	32.12	1	60	71	184

Figure 2 depicts the trajectory after meshing. In addition, the number of trajectory points in each grid was counted. A heat map can also be used to reflect the degree of taxi aggregation in different areas while generating grid geographic graphics. Areas with a larger degree of aggregation will be the key research areas for abnormal behavior detection. The numbers in the legend of Figure 2 represent the order number, that is, the number of orders corresponding to the trajectories of each color. Taking Figure 2 as an example, the purple dot correspond to numbers (1.00, 2.00), indicating that the purple trajectories in the Figure 2 are visualizations of trajectories in order 1 and order 2. The meaning of the numbers after other color points can be deduced in the same way.

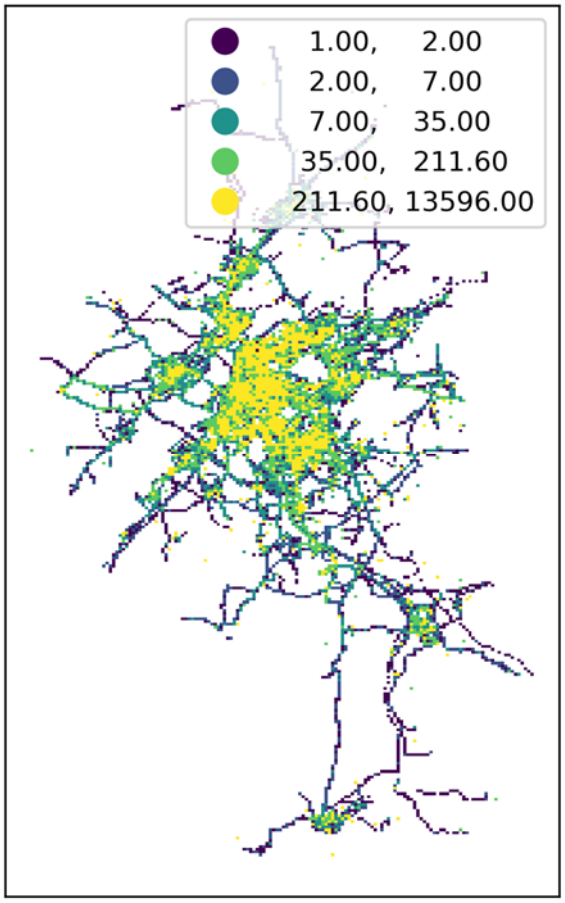


Figure 2. Heatmap of gridded trajectories.

2.2.2. OD extraction and grouping statistics

After completing the meshing and correlating trajectory points with grids, as shown in Table 3, the OD (origin-destination) as well as the operating status of the taxi can be extracted based on where the trajectory points are located on the grid. In Table 3, the columns "SLONCOL" and "SLATCOL" represent the grid coordinates of the starting point, and the columns "ELONCOL" and "ELATCOL" represent the grid coordinates of the ending point.

Table 3. Example of OD extraction output of taxi trip.

SLONCOL	SLATCOL	ELONCOL	ELATCOL
102	97	56	78

GPS devices typically report data at a low frequency of approximately one record per minute. This results in a less detailed representation of taxi trajectories, as a taxi may traverse multiple consecutive cells without recording its GPS points. Therefore, before OD grouping, trajectory point densification is carried out with a time interval of 1 s, ensuring that there a trajectory point is generated every 1 s. After augmenting the trajectory points between each OD pair, all taxi ODs that passed through the same pair of destination cells were grouped. The classification process of the start and end

points of the same trajectory is illustrated in Figure 3. Therefore, the problem of abnormal driving trajectory detection was transformed into the problem of finding abnormal trajectories from all trajectories with the same start-end-point unit pair.

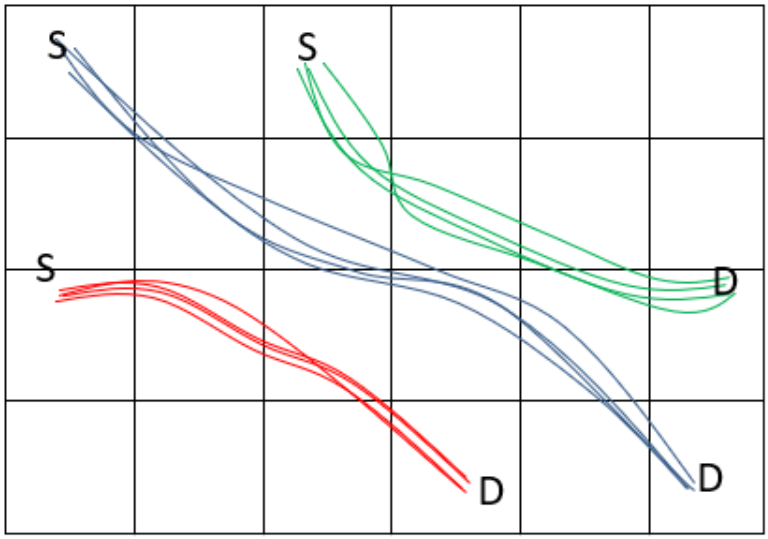


Figure 3. Schematic diagram of taxi trajectories grouped by OD.

2.3. Definition of detour trajectory

Taxi trajectory data is a type of spatio-temporal trajectory data that is generated by moving objects in geographic space and is usually represented by spatial points with temporal order. The formal expression is $Trajectory\, t_k = p_1 -> p_2 -> \ldots\ldots -> p_n$, where P_i indicates the location and other properties of the target in t_i. Usually, the elements of P_i include the ID of the positioning point, the ID of the trajectory, longitude, latitude, altitude, speed, and time. Trajectory dataset $T = \{t_1, t_2, \cdots, t_m\}$, where t_i represents the i_{th} trajectory. This study found that not all spatially anomalous trajectories are detour trajectories and may be the well-intentioned choices of drivers, such as taking shortcuts to save time or avoiding road congestion. Therefore, as shown in Figure 4, this study judges whether the trajectory is abnormal in space or time based on spatial characteristics and travel time, and categorizes the abnormal trajectories into four categories: normal, temporal anomaly, spatial anomaly, and spatiotemporal anomaly. As shown in Figure 4, t_2 is a normal trajectory. t_1 is similar to a normal trajectory in terms of time but different in terms of space characteristics. t_3 has spatial characteristics similar to normal trajectories, but abnormal temporal characteristics, which are usually caused by road congestion. Since both the temporal and spatial features of t_4 exhibit anomalies, such trajectories are classified as spatiotemporal anomalous trajectories.

The study object is a spatiotemporal anomalous trajectory based on which detour behaviors can be identified. The detour behavior presents abnormal distributions in both time and space, which complicates the determination process. Detour behavior is a manifestation of trajectory abnormality, but it is not the same as trajectory abnormality; it is a subset of or a special type of trajectory

abnormality. Combining the actual situation and the spatiotemporal characteristics of the trajectory, as shown in Figure 5, four conditions are supposed to be satisfied: small quantity and deviation from the normal trajectory; graphically abnormal; travel time larger than the threshold value; and locally generated abnormal driving distance. Trajectories satisfying the above conditions are classified as detour trajectories.

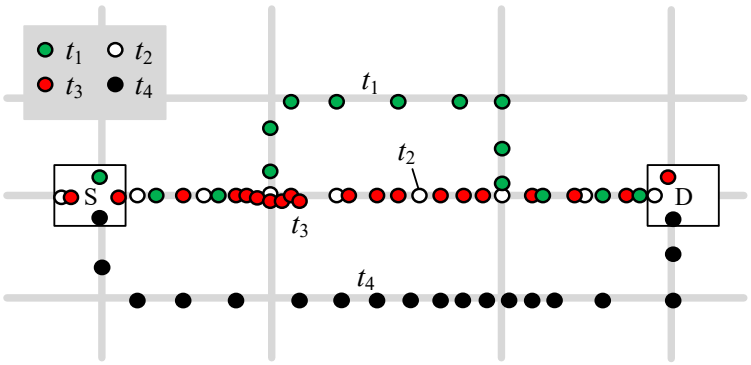


Figure 4. Four patterns of abnormal trajectories.

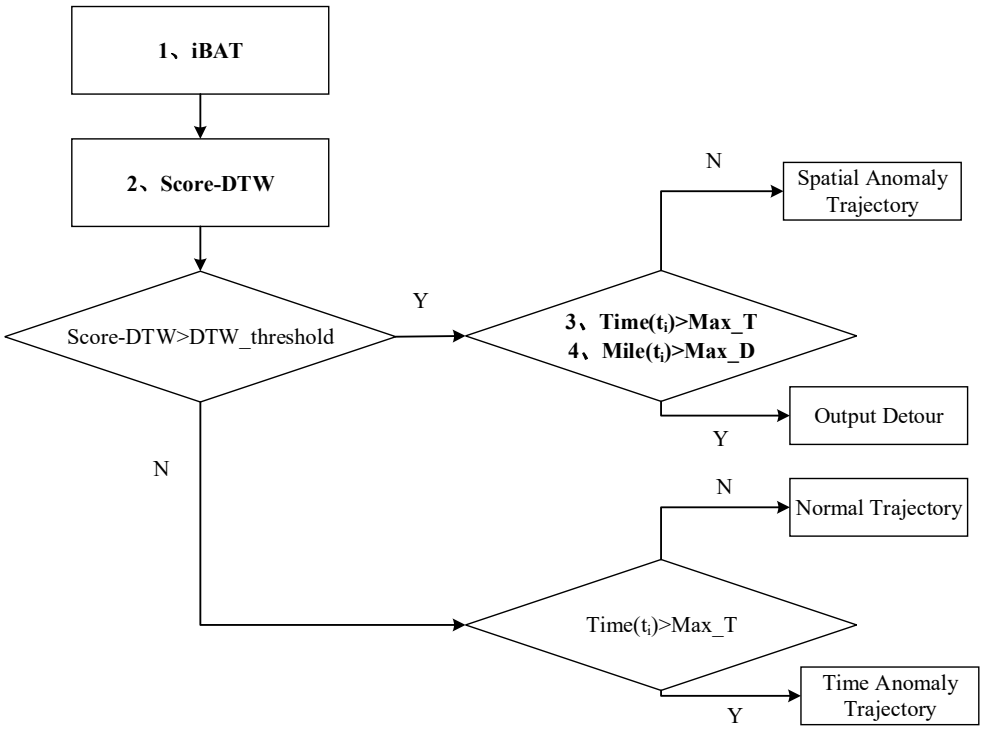


Figure 5. Four conditions to identify the detour trajectory.

3. Technical framework of taxi detour trajectory detection

After preliminary preparations, this study divides the detour trajectory detection technology framework into two parts: preprocessing and detection. This framework was designed according to the definition and characteristics of detour trajectories.

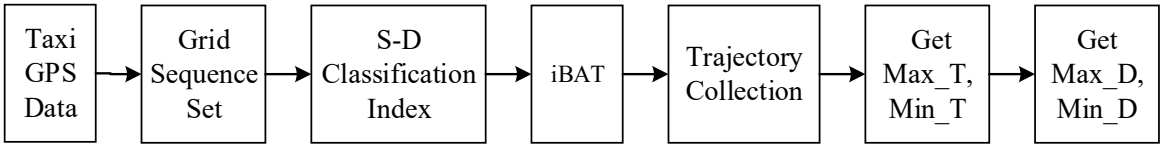


Figure 6. Process of preprocessing stage.

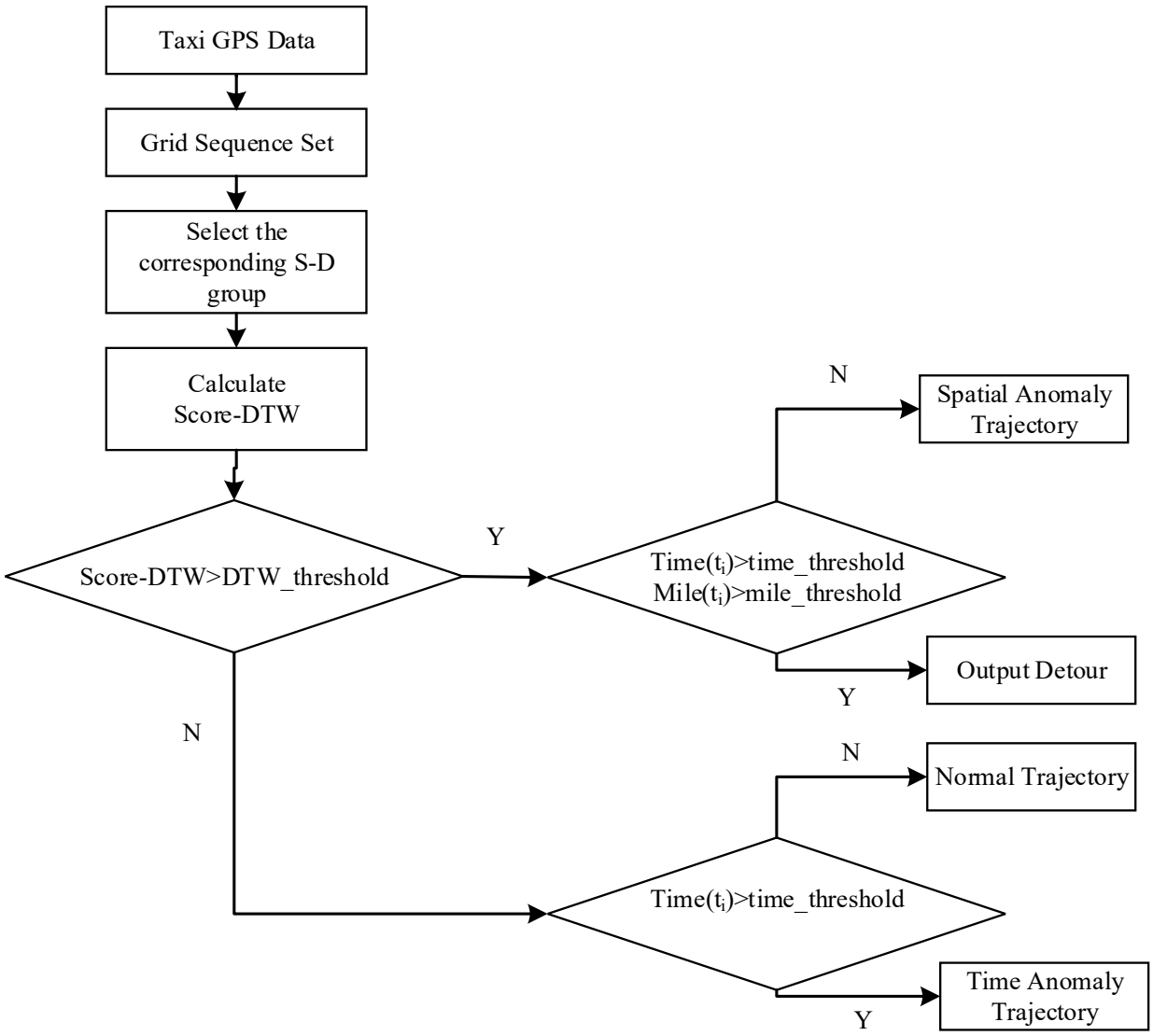


Figure 7. Detection stage process.

As shown in Figure 6, in the preprocessing stage, the trajectory data are first converted into a grid sequence, and their origin and destination are classified and indexed. The trajectories that passed through the same origin-destination grid pair were separated into groups, and the question was transformed into finding anomalous trajectories for the same origin-destination grid pair. Second, after the classification index, anomaly detection based on the isolation-Based Anomalous Trajectory (iBAT)

algorithm is performed on the trajectory grid sequence, that is, to obtain a set of trajectories, including normal and abnormal, and calculate the mean and variance of the normal trajectory travel time to obtain the corresponding threshold.

As shown in Figure 7, in the detection stage, four steps are performed: conversion of trajectory points to grid sequences; selection of corresponding groupings according to OD pairs; similarity measurement with normal trajectories; calculation of the distance matrix; and DTW score. The last step is to compare the mean and variance of the normal trajectory travel time and output the detour trajectory.

3.1. Abnormal trajectory detection based on iBAT algorithm

In principle, the detection methods for abnormal trajectories can be divided into four categories based on category, historical trajectory similarity, distance index, and grid. The principles, advantages, and disadvantages of these four categories are as follows:

1) Classification-based detection methods.

The classification-based trajectory anomaly detection method generally involves two steps: one is to label the training set to learn to build a classifier, and the other is to use the classifier to distinguish noise, normal values, and outliers in the test set [31,32]. In practical applications, to obtain higher detection accuracy, it is necessary to manually attach labels to the training set data, which consumes manpower and material resources and hence does not meet the requirements of online anomaly detection [33].

2) Detection methods based on historical similarities.

As it is difficult to obtain a relatively complete dataset, historical data can be used to ensure the integrity of the acquired abnormal trajectory types [34,35]. The trajectory anomaly detection method based on the similarity with historical trajectories is used to extract abnormal features based on the existing trajectory database and assess if it is an abnormal trajectory by calculating the matching degree between the target trajectory and the abnormal features [36].

3) Distance-based detection method.

The trajectory anomaly detection method based on the distance index measures the difference between trajectories based on the distance between trajectory points or between trajectories (particularly between trajectories) [37], identifying abnormal trajectories that are significantly off away from usual trajectories [38,39]. However, the distance-based detection method is aimed at the trajectory characteristics at a certain time point and ignores the trajectory changes in the entire process [40], that is, the difference in the spatial position characteristics at a specific instant.

4) Detection method based on grid divisions.

The main idea of the grid-based trajectory anomaly detection method is to divide the area into grids, attach a series of sequences to the trajectory, and identify anomalies according to the sequences [41]. Anomaly detection methods are mainly based on the likelihood ratio and isolation mechanism. The abnormal trajectory detection method based on the likelihood ratio was used to predict the abnormal trajectory by constructing the likelihood ratio detection statistic [42]. When this method is applied to abnormal trajectory recognition, it is necessary to count the maximum deviation from the expected situation at the vehicle level. The main idea of the abnormal trajectory detection algorithm of the isolation mechanism is that the abnormal trajectory distribution is sparse and unique [43]. An anomaly detection algorithm based on the isolation forest is a commonly used algorithm for this type of detection method [44–46]. The basic idea of the algorithm is that the outliers are sparser than the

normal points; therefore, they are easier to divide, which is completely consistent with the above-mentioned isolation mechanism.

Considering the characteristics of several anomaly detections, this study finally selected the iBAT detection method among the grid-based detection methods for preliminary anomaly trajectory screening. The iBAT is a mining algorithm for abnormal trajectories based on the idea of isolation, which is a decision tree based on trajectory gridding. According to the algorithm, trajectories can be divided into two types: normal and abnormal. The normal trajectories are "many and approximate", and the abnormal trajectories are "few and special". The latter is the focus of the present study. Compared to the complexity and intractability of normal trajectories, abnormal trajectories are easier to isolate. It utilizes the inherent "few and different" characteristics of abnormal trajectories and applies a data-induced random tree to divide all trajectories until they are isolated. The specific algorithm steps of iBAT randomly select a grid, dividing the trajectory set into two trees according to the presence or absence of this grid, recursively processing the subtrees, obtaining a complete decision tree, and determining abnormal data according to the iForest algorithm. The iBAT algorithm used in this study is Algorithm 1, which is summarized as follows:

Algorithm 1. The iBAT algorithm.

Function of algorithm: Taking advantage of the "few and different" inherent characteristics of outliers, the anomaly trajectories are mined based on the idea of isolation.
Input: x-selected track to be detected, T-track set composed of n sub-tracks, c(n)-number of binary trees, that is, the average number of track points that need to be selected to separate the track in the track set from other tracks, n-sample size, the number of tracks in the same group as the start and end points.
Output: abnormal score, $Score(x,n)$.

For example, Figure 8 shows seven trajectories in a set of trajectories. All the trajectories begin at 1 and end at 28. t_0 and t_6 are significantly different from other trajectories, and the path length of abnormal trajectories in t_0 is significantly shorter than that of other trajectories.

t_0: 1→2→3→4→5→6→7→14→21→28

t_1: 1→8→15→22→23→24→25→26→27→28

t_2: 1→8→15→23→24→25→26→27→28

t_3: 1→8→15→16→23→24→25→26→27→28

t_4: 1→8→15→16→24→25→26→27→28

t_5: 1→8→16→23→24→25→26→27→28

t_6: 1→8→15→22→23→16→15→22→23→24→25→26→27→28

1（S）	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28(D)

Figure 8. Schematic diagram of possible paths.

In addition, because each trajectory point selection is random, the final abnormal score should be obtained by comprehensively considering the results of multiple judgments. The calculation formula for the abnormal score is given in Eqs (1) and (2).

$$Score(x,n) = 2^{-\frac{E(t(x))}{c(n)}} \tag{1}$$

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{2}$$

where x is the track to be detected and n represents the size of the sub-sampling sample, that is, the number of tracks in the same group as x. $E(t(x))$ represents the number of track points that should be selected to separate the track from other tracks in the same group. $c(n)$ represents the average number of trajectory points that must be selected in a group of n trajectories to separate the trajectories in this group from other trajectories. H is the total number of harmonics that equals $ln(n) + 0.57721566(Euler\ Constant)$, and when $E(t(x)) \to 0$ and $\mathrm{Score}(x,n) \to 1$, the trajectory is viewed as an abnormal trajectory. When $E(t(x)) > c(n)$ and $\mathrm{Score}(x,n) < 0.5$, the trajectory can be viewed as a normal trajectory.

3.2. Abnormal trajectory detection based on iBAT and DTW improved algorithm

This study considers the flaws of iBAT; specifically, the anomaly detection algorithm based on iBAT lacks an accurate judgment of partial trajectory detours. Therefore, in view of the shortcomings of the iBAT algorithm, this study provides several improvements. The trajectory similarity measurement algorithm was used for further anomaly detection on abnormal trajectories processed by the iBAT algorithm, lowering the misjudgment rate.

Commonly used metrics for calculating the spatiotemporal similarity between two trajectories are Euclidean distance, Hausdorff distance, Edit distance, and DTW. The Euclidean distance refers to the

distance between two points in Euclidean space. When using Euclidean distance to measure the similarity between trajectories, it is necessary to first convert the two trajectories into sequences of the same dimension and length. The Hausdroff distance describes the distance between two point sets. The distance metric between them can be determined using the exhaustive method at the algorithm level.

Edit distance is a calculation index used to measure the difference between two character sequences. The fundamental idea of using edit distance is to change one string into another by adding, deleting, and replacing. The premise of the change is that the lengths of the strings are equal, and after the lengths of the two strings are the same, they are assessed by measuring the similarity of the longest common sequence between the two character sequences. It is difficult to calculate the distance between sequences of different lengths using the traditional Euclidean distance calculation method, whereas DTW is commonly used to determine the similarity between two time series. Thus, the DTW algorithm was used to calculate the distance between two time series of different lengths. In addition, the DTW algorithm can be executed without reference to time.

Therefore, in this study, the abnormal trajectories initially processed by the iBAT algorithm were further processed by DTW to measure the similarity between the trajectories. The application principle of the DTW algorithm used in this study is as follows: To find the similarity between two trajectories, a distance matrix dG, which can be considered as a multidimensional array, needs to be calculated, and each point of sub-trajectory t_1 needs to be mapped with the actual distance of each point of sub-trajectory t_2. To determine the distance between two trajectory points, this study used the Harvesine formula shown in the following equation:

$$\Delta\sigma = 2arcsin\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_1 \cos\phi_2 \sin^2\left(\frac{\Delta\lambda}{2}\right)} \tag{3}$$

$$\Delta\phi = (\phi_1 - \phi_2)^2 \tag{4}$$

$$\Delta\lambda = (\lambda_1 - \lambda_2)^2 \tag{5}$$

where λ_1, ϕ_1 and λ_2, ϕ_2 represent the radians of the geographic longitude and latitude of points 1 and 2, respectively.

To use DTW to calculate the distance between the subtrajectories t_1 and t_2, this study defines a function to calculate the ground distance between two points. Subsequently, using the principles of dynamic programming, the matrix is recursively traversed until the final score representing the DTW between the two trajectories is obtained. The formula is as follows:

$$dtw(i,j) = \begin{cases} \infty & if\ i = 0 \text{ or } j = 0 \\ 0 & if\ i = j = 0 \\ d(t_1^i, t_2^j) + \min\left(dtw(i-1,j), dtw(i,j-1), dtw(i-1,j-1)\right), & otherwise \end{cases} \tag{6}$$

In this study, the trajectories were initially divided into two categories through iBAT: normal trajectories were marked with label 0, and abnormal trajectories were marked with label 1. The DTW similarity measurement process is illustrated in Figure 9.

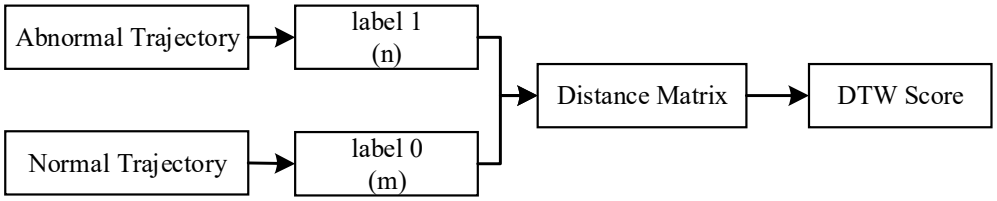


Figure 9. The process of DTW similarity measurement.

3.3. Abnormal trajectory detection based on iBAT algorithm

In the preceding two sections, this study detects abnormal trajectories using the iBAT and DTW algorithms, and the main goal of this study is to detect detour behavior. Through related research, it was found that the driving distance and time are used to establish if a taxi's trajectory is a detour behavior. Therefore, in this section, we perform a statistical analysis on the spatiotemporal characteristics of the detected abnormal trajectories and investigate the reasons behind the abnormal trajectories. The intentional detour of the taxi can be more accurately identified if evidence to rule out the driver's detour for special reasons such as traffic accidents is provided.

This study assumes that detour behavior will lead to longer route lengths than normal, increasing both time and cost. The mileage and travel time of the normal trajectory between point S and point D can be determined using the historical trajectory database and previous abnormal detection results. If the mileage and travel time of the trajectory are greater than the corresponding thresholds, the abnormal trajectory is deemed to be a detour trajectory. This study defines Max_D and Min_D as the maximum and minimum travelling lengths in a normal trip, respectively, and Max_T and Min_T as the maximum and minimum travelling times in a normal trip, respectively. These values exhibit great variability considering the different traffic conditions at different time periods.

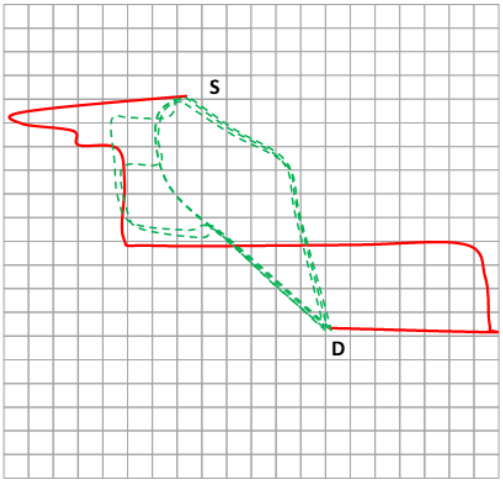


Figure 10. Schematic diagram of the trajectory between S-D pairs over the same time period.

Therefore, for each S-D pair, the mean travel time μ_T of the normal trajectory and its standard deviation σ_T were calculated, and the threshold was defined as $Max\_T = \mu_T + \sigma_T$. In addition, in

order to account for the influence of different time periods on the same road stretch, this study analyzed all trips between S-D pairs in the same time period to exclude taxi drivers detouring for special reasons. Special reasons for this include road traffic accidents and other special events. An illustrative example is shown in Figure 10. The green dotted line trajectories indicate that traffic interference in this time period is unlikely; hence, the red trajectory may be correctly identified as an intentional detour trajectory.

4. Experimental results and analysis

4.1. Experimental data

In this study, the detour trajectory data from the Nanjing 2021 case list were screened out, indexed by order ID and marked in the experimental data (see Table 4), and then corresponded to the grid sequence.

To verify the algorithm, this study selected three sets of grid trajectory sequences with the same S-D, where the detour trajectory data was obtained from the case list data. Comparing the order ID to the data of the order, the experimental dataset obtained is shown in Table 5.

Table 4. Case list data.

ORDER_ID	VEHICLE_NO	POSITION_TIME	Location Stamp	Illegal behavior
1411158572	***	2021/03/15 14:22	Nanjing South Railway Station	Detour without passenger's consent
1411161425	***	2021/02/17 16:39	Nanjing Railway Station	Detour without passenger's consent
1411161340	***	2021/12/20 13:46	Nanjing Railway Station	Detour without passenger's consent
1411160321	***	2021/09/22 15:58	China Pharmaceutical University	Detour without passenger's consent

Table 5. Experimental dataset.

Set of trajectories	Amount of trajectories	Amount of detour	Proportion
T-1	1011	23	2.3%
T-2	1409	43	3.1%
T-3	1318	36	2.7%

4.2. Experimental results

4.2.1. Experimental results of iBAT

To improve the effect of the iBAT algorithm, this study investigates the influence of the values of

the parameters $c(n)$ and n on the AUC (Area Under Curve) value. $c(n)$ is the number of binary trees, that is, the average number of trajectory points that need to be selected to separate from other trajectories. n is the sample size, which is the number of trajectories in the same group as the trajectory x. In this study, the experiment has been carried out on the T-2 trajectory set with the largest amount of data, where the value of $c(n)$ is $[1,150]$ and the value of n is $\{2,4,8,16,\ldots,1024\}$. As shown in Figure 11, the AUC value converges at a smaller $c(n)$.

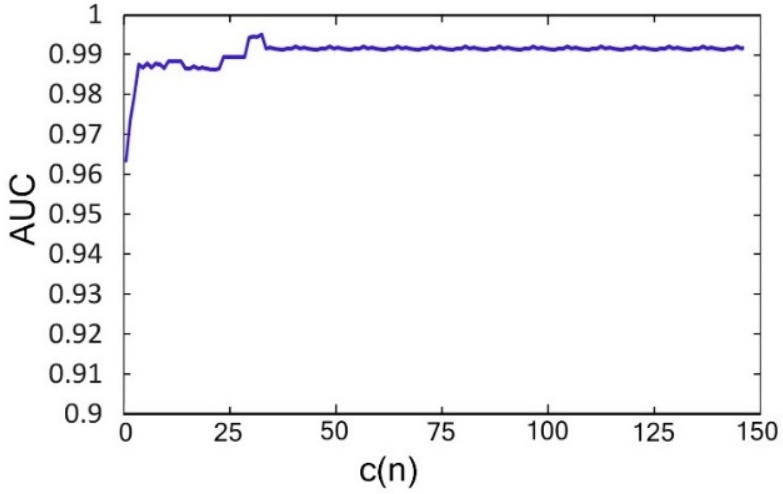


Figure 11. iBAT parameter c (n) experiment.

As shown in Figure 12, we set $c(n)=100$ and observed the impact of changes in n on performance. If the value of n is too small, more trajectories are isolated. On the other hand, if the amount of data is large, the larger is the value of n, and the longer is the average running time. $n=256$ was selected in this experiment.

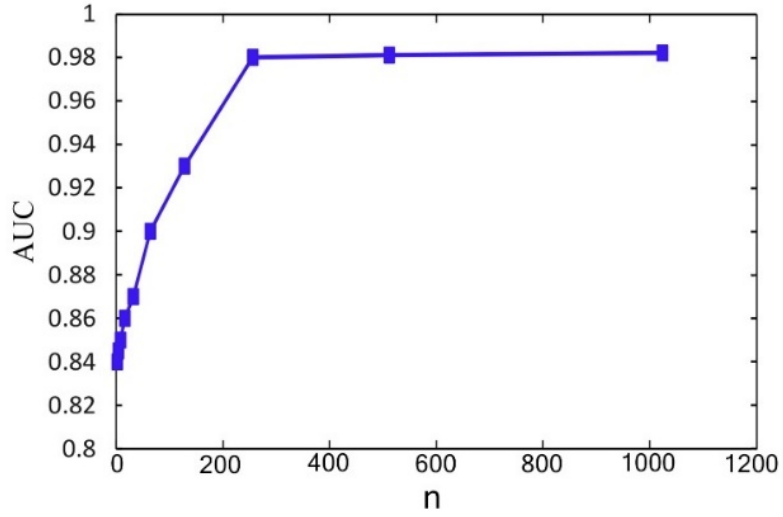


Figure 12. iBAT parameter n experiment.

In the preprocessing stage of the detour detection technology framework, this study preliminarily divides all trajectories into abnormal and normal trajectories using the iBAT algorithm and outputs the maximum travel lengths Max_D, minimum travel lengths Min_D, maximum travel times Max_T, and minimum travel times Min_T. The output data are listed in Table 6.

Table 6. Output data of iBAT.

Set of trajectories	Min_T of normal trajectory	Max_T of normal trajectory	Min_D of normal trajectory	Max_D of normal trajectory
T-1	41	102	40	68
T-2	26	32	13	20
T-3	35	71	26	35

4.2.2. Experimental results of DTW

As shown in Figure 13, DTW threshold experiments were conducted in this study, and the results showed that when the DTW threshold was 0.0006, the false detection rate was the lowest.

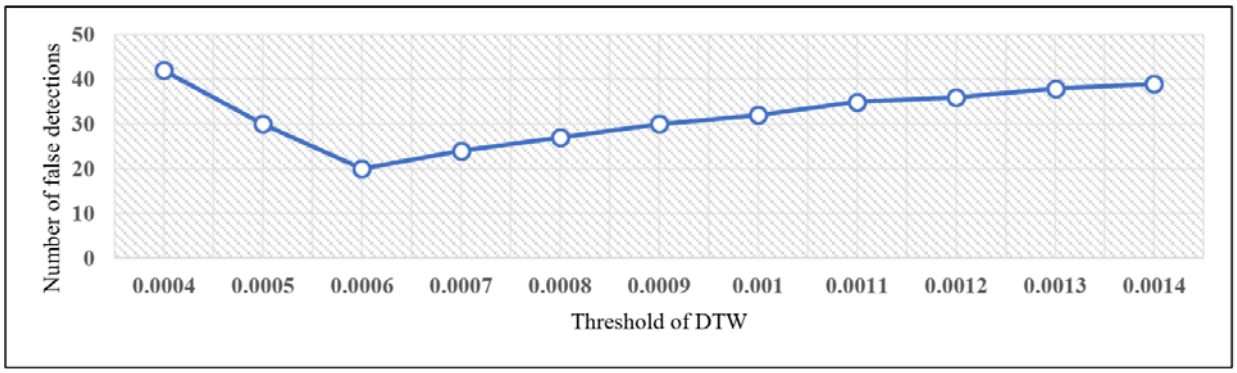


Figure 13. Experiments on DTW Thresholds.

4.2.3. Detour trajectory recognition results considering spatiotemporal features

Table 7 shows the distribution of the abnormal trajectories in terms of driving distance and time. It can be seen that more than 60% of abnormal trajectories take longer time and cover more distance than normal trajectories, indicating that intentional detour behavior is one of the main motivations behind abnormal taxi driving behavior.

Each set of trajectories was divided into four categories. The first category is TP (True Positive), indicating that abnormal trajectories are correctly classified as abnormal. The second category is FP (False Positive), indicating that normal trajectories are incorrectly classified as abnormal. The third category is FN (False Negative), indicating that abnormal trajectories are incorrectly classified as normal. The last category is TN (True Negative), indicating that normal trajectories are correctly classified as normal.

Table 7. Distribution of abnormal trajectories over travel distance and time.

	$[0, Min\_T)$	$[Min\_T, Max\_T)$	$[Max\_T, \infty)$
$[0, Max\_D)$	0.0011	0.0139	0.0112
$[Min\_D, Max\_D)$	0.0060	0.1065	0.0821
$[Max\_D, \infty)$	0.0042	0.1525	0.6225

In this study, the DR (detection rate) and the FAR (false alarm rate) are used to determine the accuracy of the anomaly detection results. DR indicates that the abnormal trajectory is correctly classified as abnormal, that is, the proportion of anomalous trajectories that were successfully detected. FAR refers to the proportion of normal trajectories that are incorrectly classified as abnormal. DR and FAR can be defined as follows:

$$DR = \frac{TP}{TP + FN} \tag{7}$$

$$FAR = \frac{FP}{FP + TN} \tag{8}$$

The anomaly detection algorithm is more effective when DR is closer to 1 and FAR is closer to 0. This study describes the degree of balance between these two indicators by plotting the ROC curves. This study quantified the trade-off by plotting FAR on the x-axis and DR on the y-axis and measuring AUC value.

This study compares the misjudgment rate of the iBAT, iBAT + DTW, and iBAT + DTW algorithm considering the driving distance and time based on 8 million taxi trajectory datasets. The specific steps are to randomly divide the total data into three data sets, and experiment three methods three times based on each data set. The final misjudgment rate value of each method is the average value of three experiments. The results are presented in Table 8. The second method adds the DTW algorithm to the iBAT algorithm to measure trajectory similarity. The abnormal trajectory is further distinguished based on the difference between the abnormal and normal trajectories, thereby effectively reducing the misjudgment rate. Compared with the first two methods, the third method can more accurately identify the detour fraud trajectory owing to the additional consideration of the travel time and mileage.

Table 8. Comparison of the results based on three algorithms.

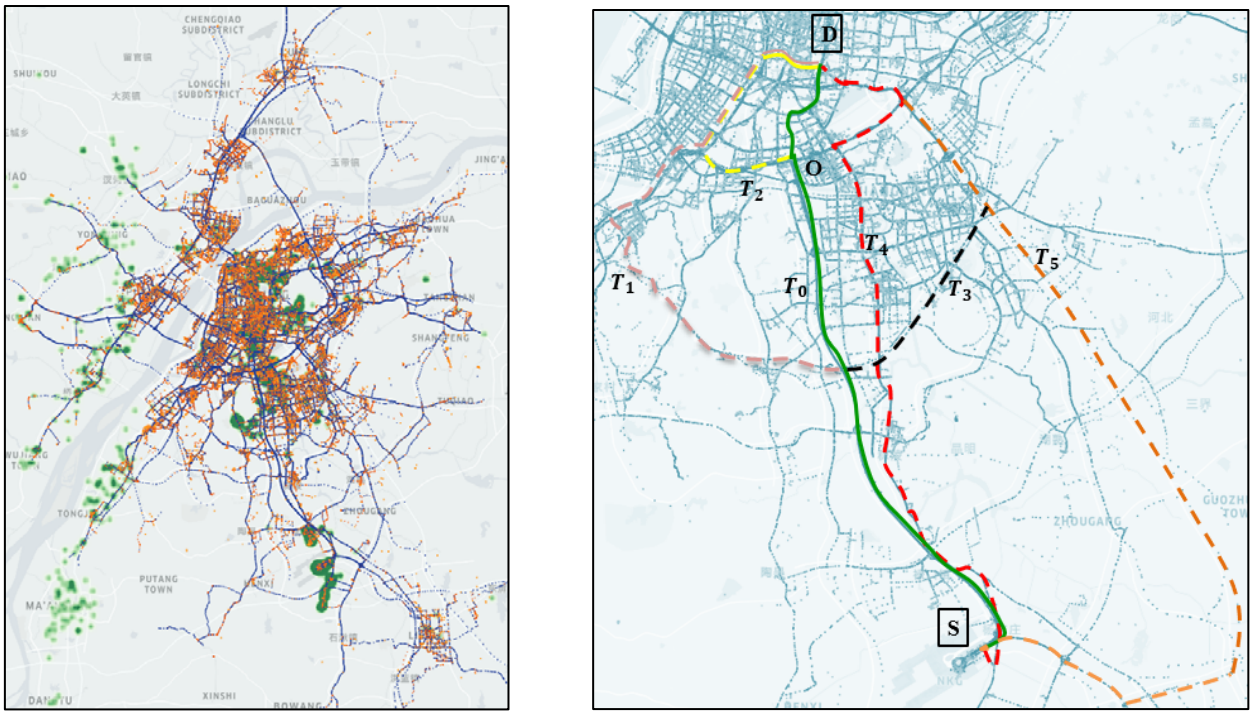
Alogorithm	Misjudgment rate %
iBAT	19.7
iBAT + DTW	3.41
iBAT + DTW + Max_D, Max_T	1.64

This study analyzed the taxi trajectory in Nanjing for one week. The color of the trajectory line represents the speed of the vehicle (blue indicates fast speed, and yellow indicates low speed), as shown

in Figure 14(a). Based on spatiotemporal data, this study identified the similarities and trends of abnormal trajectories. The green heat map represents the area from where the abnormal trajectory originates. It has been found that the origin of abnormal trajectories is mainly concentrated in Nanjing Railway Station, Lukou Airport, Nanjing South Railway Station, and other large transportation hubs, important urban subway stations, areas with a large number of tourists or remote areas with fast moving vehicles and scattered roads.

For example, the trajectories from trajectory set T-1 in the same period were selected and visualized to further analyze the microscopic characteristics of the detour trajectory. Figure 14(b) shows the test trajectory from T-1, where points S and D correspond to the starting and ending points of the trajectory set, respectively. T_0 is a normal trajectory, and the detected detour trajectories are represented by dotted lines, which include the partial detour trajectories T_1, T_2, T_3, and the global detour trajectories T_4, T_5.

Matching the T-1 trajectory set with the map, it was found that the trajectory set included trajectories from Nanjing Lukou Airport to Confucius Temple. As shown in Figure 14(a), the areas where the starting and ending points of these trajectories are located have high incidences of detours. Meanwhile, in this normal road section, the average speed of the road from point S to point O was high, and the traffic conditions were good.



(a) Visualization of overall trajectory. (b) Visualization of T-1 trajectory set.

Figure 14. Visualization of the trajectory.

Therefore, it can be ruled out that the taxi driver detoured owing to road congestion or a traffic accident in this stretch during this time period, and detours T_1, T_3, T_5 are judged to be intentional detours. In the normal track T_0, the section between points O and D experienced traffic congestion, so

some of the detour trajectories T_2 and T_4 might be regarded as detours to avoid congested stretches. In addition, this study generates order feature information corresponding to the detected detour trajectories and provides them to off-site law enforcement managers. The order feature information provided includes mileage, travel time, driving costs, vehicle information, and driver information, which were compared with those of normal trajectories. For comparison, as shown in Table 9, one order is selected for each group of detour trajectories.

Table 9. Comparison of the results based on three algorithms.

Trajctory	VEHICLE_NO	Mileage	Travel time	Travel cost
T_0	***	39.2	35	98
T_1	***	57.9	61	189
T_2	***	46.4	52	146
T_3	***	49	56	152
T_4	***	37.9	64	125
T_5	***	67.6	102	237

5. Conclusions

This study proposes a technical framework for detour trajectory anomaly detection, which is divided into preprocessing and detection stages. In the preprocessing stage, taxi historical trajectory data were first converted into a grid sequence, and trajectories that passed through the same start-end grid pair were grouped together. Second, anomaly detection based on the iBAT on the trajectory was performed, a trajectory set including normal and abnormal trajectories was obtained, and the travel time of all normal trajectories was analyzed to determine the threshold. During the detection stage, the trajectory grid sequence selected the corresponding groups according to the S-D point pairs and measured the similarity with the normal trajectory. The distance matrix was then calculated, the DTW score was obtained, and the mean and variance of the normal trajectory travel time were compared. Finally, the output detours the trajectory. The innovation of this study is adding distance metrics (DTW) and parameters of normal trajectories (average driving distance and time) to iBAT algorithm to achieve accurate identification of detour trajectories.

To verify the actual effect, this study marked the detour trajectory data in the Nanjing 2021 case list, corresponding to the grid sequence as the experimental data, and used the model proposed in this study to analyze the test data. The experimental results show that the method proposed in this study has a low misjudgment rate for taxi detours. The analysis results of the model test data in this study showed that there is a certain correlation between the choice of detours and geographical location, and that it is highly correlated with time and space factors.

In the future, based on the existing research results, an online detection framework for detecting the abnormal behavior of taxis can be proposed, which can not only detect road segments with abnormal behavior, but also update the route behavior model through newly added trajectories to detect the detour behavior of taxi drivers in real time.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Q. Cheng, Z. Liu, Y. Lin, X. S. Zhou, An s-shaped three-parameter (S3) traffic stream model with consistent car following relationship, *Transp. Res. Part B Methodol.*, **153** (2021), 246–271. https://doi.org/10.1016/j.trb.2021.09.004
2. H. Wang, Transportation-enabled urban services: A brief discussion, *Mutimodal Transp.*, **1** (2022), 100007. https://doi.org/10.1016/j.multra.2022.100007
3. M. Ester, H. P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Kdd*, AAAI, **96** (1996), 226–231.
4. Q. Cheng, Z. Liu, J. Guo, X. Wu, R. Pendyala, B. Belezamo, et al., Estimating key traffic state parameters through parsimonious spatial queue models, *Transp. Res. Part C Emerging Technol.*, **137** (2022), 103596. https://doi.org/10.1016/j.trc.2022.103596
5. Z. Liu, Y. Wang, Q. Cheng, H. Yang, Analysis of the information entropy on traffic flows, *IEEE Trans. Intell. Transp. Syst.*, **2022** (2022), 1–12. https://doi.org/10.1109/TITS.2022.3155933
6. D. Huang, J. Xing, Z. Liu, Q. An, A multi-stage stochastic optimization approach to the stop-skipping and bus lane reservation schemes, *Transportmetrica A Transp. Sci.*, **17** (2021), 1272–1304. https://doi.org/10.1080/23249935.2020.1858206
7. F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory pattern mining, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Jose, USA, (2007), 330–339. https://doi.org/10.1145/1281192.1281230
8. I. Syarif, A. Prugel-Bennett, G. Wills, Data mining approaches for network intrusion detection: from dimensionality reduction to misuse and anomaly detection, *J. Inf. Technol. Rev.*, **3** (2012), 70–83.
9. Y. Yue, H. D. Wang, B. Hu, Q. Li, Y. G. Li, A. G. Yeh, Exploratory calibration of a spatial interaction model using taxi GPS trajectories, *Comput. Environ. Urban Syst.*, **36** (2012), 140–153. https://doi.org/10.1016/j.compenvurbsys.2011.09.002
10. Q. Cheng, Y. Chen, Z. Liu, A bi-level programming model for the optimal lane reservation problem, *Expert Syst. Appl.*, **189** (2022), 116147. https://doi.org/10.1016/j.eswa.2021.116147
11. G. Münz, S. Li, G. Carle, Traffic anomaly detection using k-means clustering, in *GI/ITG Workshop MMBnet*, **7** (2007), 9.
12. I. N. Junejo, O. Javed, M. Shah, Multi feature path modeling for video surveillance, in *Proceedings of the 17th International Conference on Pattern Recognition*, IEEE, Cambridge, UK，**2** (2004), 716–719. https://doi.org/10.1109/ICPR.2004.1334359
13. Q. Meng, P. Liu, Z. Liu, Integrating multimodal transportation research, *J. Multimodal Transport.*, **1** (2022), 100001. https://doi.org/10.1016/j.multra.2022.100001
14. Y. Zheng, Trajectory data mining: an overview, *ACM Trans. Intell. Syst. Technol. (TIST)*, **6** (2015), 1–41. https://doi.org/10.1145/2743025
15. Z. Feng, Y. Zhu, A survey on trajectory data mining: techniques and applications, *IEEE Access*, **4** (2016), 2056–2067. https://doi.org/10.1109/ACCESS.2016.2553681
16. N. Paragios, R. Deriche, Geodesic active regions: a new framework to deal with frame partition problems in computer vision, *J. Visual Commun. Image Represent.*, **13** (2002), 249–268. https://doi.org/10.1006/jvci.2001.0475

17. C. Stauffer, W. E. Grimson, Adaptive background mixture models for real-time tracking, in *1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, IEEE, Collins, USA, **2** (1999), 246–252. https://doi.org/10.1109/CVPR.1999.784637
18. S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, F. Bremond, Toward abnormal trajectory and event detection in video surveillance, *IEEE Trans. Circuits Syst. Video Technol.*, **27** (2016), 683–695. https://doi.org/10.1109/TCSVT.2016.2589859
19. J. Huo, X. Fu, Z. Liu, Q. Zhang, Short-term estimation and prediction of pedestrian density in urban hot spots based on mobile phone data, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 10827–10838. https://doi.org/10.1109/TITS.2021.3096274
20. D. Huang, Y. Wang, S. Jia, Z. Liu, A Lagrangian relaxation approach for the electric bus charging scheduling optimisation problem, *Transportmetrica A Transp. Sci.*, **2022** (2022), 1–24. https://doi.org/10.1080/23249935.2021.2023690
21. J. Simon, Remote supply revisited: the jeep problem with costly transfer points, *Multimodal Transp.*, **1** (2022), 100019. https://doi.org/10.1016/j.multra.2022.100019
22. J. Qiu, K. Huang, J. Hawkins, The taxi sharing practices: matching, routing and pricing methods, *Multimodal Transp.*, **1** (2022), 100003. https://doi.org/10.1016/j.multra.2022.100003
23. A. T. Palma, V. Bogorny, B. Kuijpers, L. O. Alvares, A clustering-based approach for discovering interesting places in trajectories, in *Proceedings of the 2008 ACM Symposium on Applied Computing*, ACM, Fortaleza, Brazil, (2008), 863–868. https://doi.org/10.1145/1363686.1363886
24. L. Grady, E. L. Schwartz, Isoperimetric graph partitioning for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **28** (2006), 469–475. https://doi.org/10.1109/TPAMI.2006.57
25. L. Zhao, G. Shi, J. Yang, An adaptive hierarchical clustering method for ship trajectory data based on DBSCAN algorithm, in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, IEEE, Beijing, China, (2017), 329–336. https://doi.org/10.1109/ICBDA.2017.8078834
26. Y. Xi, D. Huang, Y. Yuan, Z. Liu, K. Anish, N. Zheng, Improved dynamic time warping algorithm for bus route trajectory curve fitting, *J. Transp. Eng.*, **147** (2021), 04021044. https://doi.org/10.1061/JTEPBS.0000544
27. R. F. Ibrahim, *A recommendation system based on clustering and classification for optimal trajectory analysis*, PhD thesis, Carleton University, 2019. https://doi.org/10.22215/etd/2019-13400
28. M. Khashei, M. Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, *Appl. Soft Comput.*, **11** (2011), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015
29. Y. Yuan, W. Zhang, X. Yang, Y. Liu, Z. Liu, W. Wang, Traffic state classification and prediction based on trajectory data, *J. Intell. Transp. Syst.*, **2021** (2021), 1–15. https://doi.org/10.1080/15472450.2021.1955210
30. V. Hodge, J. Austin, A survey of outlier detection methodologies, *Artif. Intell. Rev.*, **22** (2004), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9
31. S. Y. Huang, Y. N. Huang, Network traffic anomaly detection based on growing hierarchical SOM, in *2013 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, IEEE, Budapest, Hungary, (2013), 1–2. https://doi.org/10.1109/DSN.2013.6575338
32. A. S. da Silva, J. A. Wickboldt, L. Z. Granville, A. Schaeffer-Filho, ATLANTIC: A framework for anomaly traffic detection, classification, and mitigation in SDN, in *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*, IEEE, Istanbul, Turkey, (2016), 27–35. https://doi.org/10.1109/NOMS.2016.7502793

33. K. K. Santhosh, D. P. Dogra, P. P. Roy, Anomaly detection in road traffic using visual surveillance: A survey, *ACM Comput. Surv. (CSUR)*, **53** (2020), 1–26. https://doi.org/10.1145/3417989
34. S. Chawla, Y. Zheng, J. Hu, Inferring the root cause in road traffic anomalies, in *2012 IEEE 12th International Conference on Data Mining*, IEEE, Brussels, Belgium, (2012), 141–150. https://doi.org/10.1109/ICDM.2012.104
35. P. R. Lei, A framework for anomaly detection in maritime trajectory behavior, *Knowl. Inf. Syst.*, **47** (2016), 189–214. https://doi.org/10.1007/s10115-015-0845-4
36. J. Wang, I. C. Paschalidis, Statistical traffic anomaly detection in time-varying communication networks, *IEEE Trans. Control Network Syst.*, **2** (2014), 100–111. https://doi.org/10.1109/TCNS.2014.2378631
37. E. M. Knorr, R. T. Ng, V. Tucakov, Distance-based outliers: Algorithms and applications, *VLDB J.*, **8** (2000), 237–253. https://doi.org/10.1007/s007780050006
38. E. M. Knorr, R.T. Ng, Finding intensional knowledge of distance-based outliers, in *Vldb*, **99** (1999), 211–222.
39. J. G. Lee, J. Han, X. Li, Trajectory outlier detection: A partition-and-detect framework, in *2008 IEEE 24th International Conference on Data Engineering*, IEEE, Cancun, Mexico, (2008), 140–149. https://doi.org/10.1109/ICDE.2008.4497422
40. S. A. Ahmed, D. P. Dogra, S. Kar, P. P. Roy, Surveillance scene representation and trajectory abnormality detection using aggregation of multiple concepts, *Expert Syst. Appl.*, **101** (2018), 43–55. https://doi.org/10.1016/j.eswa.2018.02.013
41. Y. Ge, H. Xiong, Z. Zhou, H. Ozdemir, J. Yu, K. C. Lee, Top-eye: top-k evolving trajectory outlier detection, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, Toronto, Canada, (2010), 1733–1736. https://doi.org/10.1145/1871437.1871716
42. W. Qin, J. Tang, C. Lu, S. Lao, A trajectory abnormal detection method based on segmentation and clustering, in *Journal of Physics: Conference Series*, **2010** (2021), 012188. https://doi.org/10.1088/1742-6596/2010/1/012188
43. X. Zhao, Y. Rao, J. Cai, W. Ma, Abnormal trajectory detection based on a sparse subgraph, *IEEE Access*, **8** (2020), 29987–30000. https://doi.org/10.1109/ACCESS.2020.2972299
44. Z. Ding, M. Fei, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window, *IFAC Proc. Vol.*, **46** (2013), 12–17. https://doi.org/10.3182/20130902-3-CN-3020.00044
45. D. Xu, Y. Wang, Y. Meng, Z. Zhang, An improved data anomaly detection method based on isolation forest, in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, IEEE, Hangzhou, China, **2** (2017), 287–291. https://doi.org/10.1109/ISCID.2017.202
46. Z. Cheng, C. Zou, J. Dong, Outlier detection using isolation forest and local outlier factor, in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, ACM, Chongqing, China, (2019), 161–168. https://doi.org/10.1145/3338840.3355641

AIMS Press