Journal of Computers Vol. 35 No. 2, April 2024, pp. 251-267
doi: 10.53106/199115992024043502016

Identification Method of High Incidence Areas of Bad Driving Behaviors Based on DBSCAN-Graham Combination

Luo-Chen Liu[1,2], Ying-Ji Liu[1,2*], Wei Zhou[1,2], Hai-Ying Xia[1,2], Wen-Jie Cai[1,2]

[1] Key Laboratory of Operation Safety Technology on Transport Vehicles, Ministry of Transport,
[2] Research Institute of Highway Ministry of Transport,
Beijing 100088.China

{lc.liu,yj.liu,w.zhou,hy.xia,wj.cai}@rioh.cn

Received 27 March 2024, Revised 7 April 2024, Accepted 15 April 2024

Abstract. This study utilizes data obtained from vehicle positioning, speed, and engine status through satellite positioning system monitoring terminals. The data is processed using spherical interpolation of latitude and longitude to handle missing values. The study establishes indicators for identifying bad driving behaviors, including speeding, rapid acceleration and deceleration, prolonged vehicle idling, and fatigue driving. The traditional DBSCAN algorithm is improved by incorporating both temporal and spatial search ranges into the algorithm, enabling cluster analysis of bad driving behaviors from both temporal and geographical perspectives. Finally, the Graham algorithm is utilized to calculate cluster boundaries, determining the specific locations and boundaries of high incidence areas of bad driving behaviors. This research aids regulatory authorities in more precisely supervising risk areas during transportation processes in terms of time and location.

Keywords: satellite positioning system, bad driving behavior, DBSCAN algorithm, cluster analysis, Graham algorithm

1 Introduction

With the development of information technology, various data such as vehicle data, video data, and driving trajectories recorded by devices such as in-vehicle satellite positioning systems, CAN bus, and cameras have become available. These data are used for identifying drivers' bad driving behaviors [1] and evaluating the classification and severity of driver behaviors [2-3]. Bad driving behaviors not only reduce the safety of operational vehicle transportation but also are a major cause of increased fuel consumption and vehicle wear and tear. Therefore, monitoring drivers based on bad driving behaviors helps reduce the risks of vehicle operation on roads [4] and also aids in reducing vehicle emissions, contributing to achieving national carbon neutrality goals.

The human-vehicle-road system is a complex coupled system. From a road perspective, identifying high incidence areas of bad driving behaviors is beneficial for regulatory authorities to take targeted measures and focus on supervising risk areas and time periods. Several domestic studies have addressed this issue. Some studies have analyzed the impact of urban prosperity and travel times on speeding by calculating the speeding rate on each road segment [5]. Jiang analyzed the association between road observation section data and speeding behavior, finding that the probability of speeding on steep slopes with longitudinal slopes exceeding 5% and 3% was significantly higher than other types of slope sections [6]. Gao [7] utilized the DSBCAN method to divide the two-dimensional data plane into grids and cluster based on the grids. These studies have been proven effective in specific road sections and specific time periods. However, there is currently limited research and application related to identifying high incidence areas and time periods of bad driving behaviors using widely available satellite positioning data.

To address this issue, this paper proposes a method for identifying high incidence areas of bad driving behaviors based on DBSCAN-Graham. The method automatically selects the number of clusters using the characteristics of the DBSCAN algorithm, avoiding errors caused by subjective parameter settings. Additionally, to address the shortcomings of the classical DBSCAN algorithm, a low-cost DBSCAN clustering method considering the spatial and temporal attributes of bad driving behavior points is introduced, and the region range is optimized using the Graham algorithm. The results obtained from this algorithm provide an effective technical means for

[*] Corresponding Author

road transportation enterprises and traffic safety management departments to strengthen operational safety management.

The technical route of this paper is shown in Fig. 1.

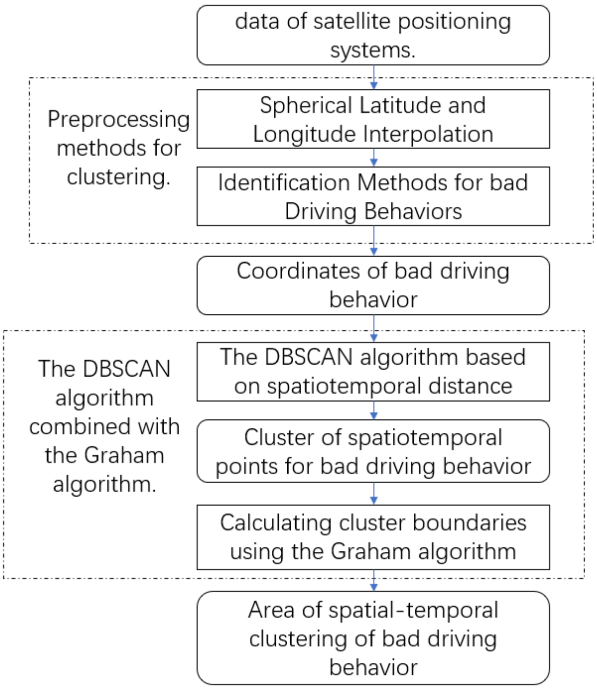
data of satellite positioning systems.
Preprocessing methods for clustering.
Spherical Latitude and Longitude Interpolation
Identification Methods for bad Driving Behaviors
Coordinates of bad driving behavior
The DBSCAN algorithm combined with the Graham algorithm.
The DBSCAN algorithm based on spatiotemporal distance
Cluster of spatiotemporal points for bad driving behavior
Calculating cluster boundaries using the Graham algorithm
Area of spatial-temporal clustering of bad driving behavior

Fig. 1. Technical route

2 Preprocessing Methods for Clustering

In the current road transportation industry, operating vehicles have largely achieved the installation of satellite positioning system monitoring terminals in accordance with relevant regulations and standards. Through these terminals, partial information such as the direction angle, longitude, latitude, and engine ignition status of the vehicles during the driving process can be obtained. However, these raw data suffer from issues such as missing values, inconsistencies, noise, and similar duplicate records, which can impact further data analysis. Therefore, it is necessary to preprocess the raw data to ensure accuracy and completeness. For noisy data and duplicate records, given the relatively clean nature of the data under study, these can be simply addressed using methods such as boxplot and search-delete. As for the issue of missing values, this study employs a method of latitude and longitude interpolation based on spherical coordinates.

Upon completion of data preprocessing, it is essential to extract indicators related to driving safety and environmental protection from the data fields themselves. Based on the extracted indicators, the data is processed to identify points of bad driving behavior that affect driving safety and the environment. This study has designed a set of bad driving behavior indicators related to safe and environmentally friendly driving behavior using existing information such as direction angle, longitude, latitude, and engine ignition status.

2.1 Spherical Latitude and Longitude Interpolation

The existing research methods for interpolating satellite positioning data mostly start from the perspective of plane geometry, without considering the characteristics of the Earth's own sphere, which leads to errors in the position of interpolation points compared to reality. In this context, the spherical latitude and longitude interpolation method is crucial in processing vehicle satellite positioning data, especially when facing the missing data. It is necessary to accurately fill in the position information of blank data points based on the vehicle's driving direction, speed, and known latitude and longitude information. The core of this method is to consider the spherical geometric characteristics of the Earth's surface, rather than simple plane coordinate system interpolation.

Firstly, we need to recognize that the latitude and longitude coordinate system is a projection representation on a spherical surface. Therefore, when a vehicle moves on the surface of the Earth, the changes in latitude and longitude are not uniform, especially in the longitude direction. With different latitudes, the actual length of the ground represented by each longitude will also change. Specifically, the closer to the equator, the longer the arc length corresponding to longitude, while near the pole, all longitude lines converge at one point.

When performing spherical latitude and longitude interpolation, we adopted the following steps:

(1) Coordinate Forward Calculation Formula

This involves calculating the coordinates of the desired point based on the coordinates of known points, the azimuth from the known points to the desired point, and the side length. This type of calculation is referred to as coordinate forward calculation in measurements (Fig. 2).

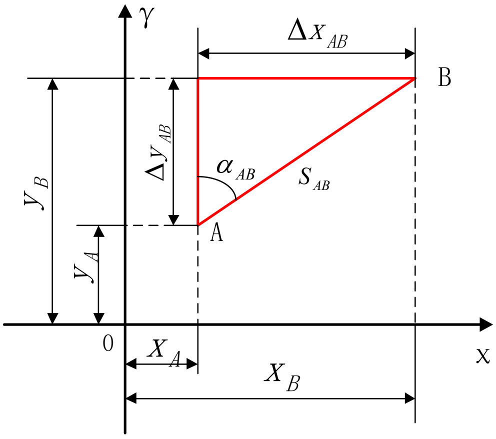

Fig. 2. Coordinate forward calculation

As depicted in the diagram, given the coordinates of point A as (x_A, y_A), and the length and polar angle of the edge from A to B as S_{AB} and α_{AB}, respectively, the coordinates of the point B can be determined as

$$\begin{aligned} x_B &= x_A + \Delta x_{AB} \\ y_B &= y_A + \Delta y_{AB} \end{aligned}. \tag{1}$$

In the given equation, the value of Δx_{AB}, Δy_{AB} represents the increment in coordinates. Based on the illustration, it can be inferred that

$$\begin{aligned} \Delta x_{AB} &= S_{AB} sin\alpha_{AB} \\ \Delta y_{AB} &= S_{AB} cos\alpha_{AB} \end{aligned}. \tag{2}$$

In the given equation, S_{AB} represents the horizontal side length, and α_{AB} denotes the coordinate increment. Substituting the second equation into the first yields the following result.

$$\begin{aligned} x_B &= x_A + S_{AB} sin\alpha_{AB} \\ y_B &= y_A + S_{AB} cos\alpha_{AB} \end{aligned}. \tag{3}$$

Given the coordinates of point A as (x_A, y_A), the length of the side from A to B, and the polar angle in the coordinate system S_{AB}, α_{AB}, the coordinates of point B can be calculated using the aforementioned formula.

(2) Coordinate System Conversion and Spherical Distance Calculation

If the longitude and latitude changes of a set of data are not significant, the characteristics of the Earth's ellipse can be ignored, and the Earth can be regarded as a perfect sphere, and the spherical coordinates of this set of data can be calculated based on this perfect sphere. Here we use the average radius of the Earth to construct three-dimensional coordinate points.

In a spherical coordinate system, the position of a point is defined by three parameters: radius r, inclination angle (or zenith angle) $\boldsymbol{\theta}$, and azimuth angle $\boldsymbol{\phi}$. Inclination angle $\boldsymbol{\theta}$) is the angle between the straight line measured from the positive z-axis downwards to the point and the z-axis, while the azimuth angle $\boldsymbol{\phi}$ is the angle of a line projected on the xy plane from the positive x-axis to a point. In a coordinate system with the Earth as the sphere, azimuth and tilt angles can be converted through latitude and longitude coordinates.

$$\begin{gathered} \theta = \frac{\pi}{2} - rad\left(long\right) \\ \phi = rad\left(lat\right) \end{gathered}. \tag{4}$$

For a point A (long, lat), its representation in the spherical coordinate system is as follows:

$$\begin{cases} x = r \cdot cos\left(rad\left(lat\right)\right) \cdot cos\left(rad\left(long\right)\right) \\ y = r \cdot cos\left(rad\left(lat\right)\right) \cdot sin\left(rad\left(long\right)\right). \\ z = r \cdot sin\left(rad\left(lat\right)\right) \end{cases} \tag{5}$$

Through the above formula, we can convert the longitude and latitude coordinates of the vehicle into spherical spatial coordinates, and calculate the spherical coordinates of the new position of the vehicle after moving a certain distance through spherical trigonometry. In the actual calculation process, in order to simplify the calculation and take into account the characteristics of the Earth's sphere, we consider a simplified situation: taking the spherical coordinates of point A as an example, its change in 1km in the longitude and latitude direction can be expressed by the following formula:

$$\begin{gathered} \Delta\phi = \frac{1}{R_{mean}\cos(\phi)} \times 1\text{km} \times \text{conversion factor} \\ \Delta\lambda = \frac{1}{\mathrm{R}_{\mathrm{eq}}\cos(\phi)} \times 1 \sim \text{km} \times \text{conversion factor.} \end{gathered} \tag{6}$$

R_{eq} refers to standard Earth radius, R_{mean} refers to average Earth radius, $\boldsymbol{\phi}$ refers to dimensions, λ refers to longitude;

The *conversion factor* depends on the proportion of the Earth's circumference at that latitude, which is a rough reference parameter in this step;

Due to the fact that the Earth is not a perfect sphere, but rather presents an elliptical shape, it is necessary to introduce the flattening of the Earth in the actual calculation process, and use the ellipsoidal model in the WGS-84 coordinate system to improve the accuracy of interpolation positions.

Considering that the Earth is a sphere, we can hypothetically extend the geographic coordinates to a flat plane, with longitude corresponding to the Y-axis and latitude to the X-axis. If the latitude and longitude of the initial point, azimuth, and distance are known, the coordinates of the target location can be calculated using the direct calculation formula. However, as the divisions of latitude and longitude are not evenly spaced like those on a flat coordinate axis, the formula must be first transformed.

Latitudes are parallel, so the distance between each degree remains almost constant. However, longitudes, which are farthest apart at the equator and converge at the poles, vary significantly in their distances. Each degree of latitude equates to approximately 111 kilometers. In fact, due to the Earth's elliptical shape, with R_{eq} being 6,378.137 km and R_{mean} being 6,371 km, each degree of latitude ranges from 110.567 km at the equator to

111.699 km at the poles. Even with extreme errors, these remain within 1%. Hence, we take the average value for the unit latitude distance, approximating one degree of latitude to 111km.

The distance corresponding to a 1° longitude difference in the East-West direction is closely related to its latitude. While a 1° difference in longitude equates to approximately 111 km over the equator, the distance varies per degree in different regions.

Therefore, a change of one kilometer on the Earth's surface corresponds to a change in latitude and longitude. Depending on the direction of the "one kilometer" extension, three scenarios can be distinguished:

1. A change of 1 kilometer along a meridian corresponds to a change of $\frac{1}{111}$ degrees in latitude.

2. A change of 1 kilometer along a parallel does not change the latitude. Let the latitude value at this point be *lat*, then the change in longitude is $\frac{1}{111 * cos\left(lat\right)}$ degrees.

3. A change of 1 kilometer in a direction with an angle of *θ* degrees from true north, let the latitude value at this point be *lat*, then the change in longitude is $\frac{sin\left(\theta\right)}{111 * cos\left(lat\right)}$ degrees and the change in latitude is $\frac{cos\left(\theta\right)}{111}$ degrees.

Based on the aforementioned steps, the latitude and longitude of the interpolation point B, denoted as (lng_B, lat_B), can be calculated using the following steps:

Given the latitude and longitude of point A, denoted as (lng_A, lat_A), bearing *θ*, and distance *distance*, the calculation can be performed using the known speed. Since the data collection frequency for satellite positioning is once per second, the speed of point A can be used to infer $distance = \frac{speed}{3600}$.

$$\begin{aligned} Ing_B &= lng_A + \tfrac{speed}{3600} * \frac{\sin(\theta)}{111 * \cos(\text{lat})} \\ \text{lat}_B &= \text{lat}_B + \tfrac{speed}{3600} * \frac{\cos(\theta)}{111} \end{aligned} . \quad (7)$$

(3) Evaluation of Interpolation Results

Five points are randomly selected from the dataset as known points. The point following the known points is designated as the reference points. The interpolation method described above is applied to interpolate between the preceding and following points. Subsequently, the interpolation results are compared with the reference point. The process is shown in Table 1 and Table 2:

Table 1. Known points

id	Velocity	Longitude	Latitude	Azimuth angle
1392	62.0	115.602541	24.910296	223
7274	25.0	115.645428	24.952875	13
8150	51.0	115.625121	24.929896	225
10350	30.0	115.454318	24.702125	100
11527	0.0	115.649593	24.960056	27

Table 2. Reference points

id	Velocity	Longitude	Latitude	Azimuth angle
1393	61.0	115.602435	24.910178	219
7275	24.0	115.645443	24.952933	11
8151	51.0	115.625015	24.929811	228
10351	29.0	115.454393	24.702111	98
11528	0.0	115.649593	24.960056	27

The results point is calculated by using known points as examples, as shown in Table 3.

Table 3. Interpolation results

Longitude	Latitude
115.602424	24.910183
115.645443	24.952936
115.625022	24.929806
115.454399	24.702112
115.649593	24.960056

Comparing the resulting points and reference points for comparison, as shown in Table 4.

Table 4. Result comparison

Reference points		Interpolation results		Error (in meters)
Longitude	Latitude	Longitude	Latitude	
115.602435	24.910178	115.602424	24.910183	1.189469
115.645443	24.952933	115.645443	24.952936	0.337372
115.625015	24.929811	115.625022	24.929806	0.874021
115.454393	24.702111	115.454399	24.702112	0.656020
115.649593	24.960056	115.649593	24.960056	0.000000

After conducting multiple comparative experiments, the calculated position error of the results remains within a range of 3 meters from the actual position.

Through this approach, we have successfully solved the problem of missing values in satellite positioning data, achieved continuous tracking of the vehicle's driving path on a spherical coordinate system, and provided accurate positional information for the identification of subsequent bad driving behaviors.

2.2 Extraction and Identification of Indicators for Bad Driving Behavior

In recent years, numerous studies have been conducted to analyze the impact of unsafe driving behavior on traffic accidents and exhaust emissions. For instance, Peng [8] utilized parameter learning and Bayesian network modeling to discover a significant correlation between fatigue driving, frequent speeding, and accident rates. Similarly, Cheng [9] collected data on instantaneous fuel consumption of trucks and identified prolonged idling and abrupt deceleration as significant factors contributing to increased fuel consumption. Building on these studies [10], this paper proposes a set of typical impact indicators covering hazardous driving behavior and energy-saving emissions reduction. These indicators include abrupt acceleration/deceleration, speeding, duration of fatigue driving, and prolonged idling. The specific indicators and identification criteria are presented in Table 5.

Table 5. Indicators and identification criteria for bad driving behavior

Bad driving behavior indicators		Identification criteria	
Rapid acceleration and deceleration	Acceleration, a	$\lvert a\rvert > a_m$, a_m represents the threshold for acceleration.	
Overspeed	Velocity, v	The velocity of three consecutive points satisfies the following conditions v_{i-1} and v_i and $v_{i+1} > v_m$ where v_m represents the velocity threshold.	
Fatigue driving	Cumulative duration of fatigue driving, t	Daytime fatigue driving: Fatigue driving within one day: Night fatigue driving:	t > 4h t > 8h t > 2h
Prolonged idle	Cumulative idling time of vehicles, t	Prolonged idle:	t > 2min

(1) Rapid acceleration (deceleration) identification.

The main process steps of the rapid acceleration behavior recognition algorithm are explained as follows:

① Input the trajectory data of the vehicle T and initialize time step t = 0.

② The length of the trajectory data is denoted as L. If t = L − 1, the computation ends; otherwise, proceed to step 3.

③ Calculate the acceleration $a_t = \frac{v_{t+1} - v_t}{3600}$ and compare the magnitudes of $|a_t|$, a_m.

④ If a $|a_t| < a_m$, t = t + 1, return to step 2; if $|a_t| > a_m$, consider T(t) as a point of rapid acceleration (deceleration), increment t by 1, and return to step 2.

(2) Overspeed

To determine when the speeding behavior will be terminated, a threshold for judging the violation of speeding behavior will be introduced α, based on industry experience preset: $\alpha \geq 200m$.

To determine whether speeding behavior is a violation of speeding regulations, a time limit for defining speeding violations is introduced δ, according to industry experience, the preset time is around 30 seconds.

Define the starting time $t1$ for speeding violations, the deadline $t2$ for speeding violations, and the duration Δt for speeding violations, where $\Delta t = t_2 - t_1$.

A set of one or more consecutive overspeed records, regardless of whether Δt is greater than δ, all are defined as "speeding behavior". When $\Delta t > \delta$, the definition of speeding behavior is "illegal speeding behavior.". The following is the identification process for speeding behavior points:

① Input the vehicle's trajectory data T, initialize t = 0, and set the sliding window *win* = [t, t + 1, t + 2].

② The length of the trajectory data is L. If t = t = L − 2, then end the computation; otherwise, proceed to step 3.

③ Calculate the speeds v_t, v_{t+1}, v_{t+2} of the points within the sliding window and compare v_t, v_{t+1}, v_{t+2} with v_m.

④ If v_t, v_{t+1}, v_{t+2} are greater than v_m, then the sub-trajectory $T'_t = [T_t, T_{t+1}, T_{t+2}]$ is a speeding trajectory, where T_t, T_{t+1}, T_{t+2} are the speeding points. Increment t by 1 and proceed to step 2.

This article proposes an improved speeding behavior recognition algorithm to address the existing problems of the original algorithm. The purpose is to avoid the occurrence of noise in the vehicle mounted satellite positioning system, which is caused by a single or extremely short period of coordinate drift noise point, resulting in intermittent recognition of speeding behavior.

If the time interval between adjacent speeding data is less than the threshold, the improved algorithm will merge the two speeding behaviors into one speeding behavior. That is, regardless of whether the speeding behavior occurs beyond the threshold or not, the improved algorithm will merge the two speeding behaviors to improve the algorithm's anti-interference ability against noise points and avoid the following situations: 1) Cut off a speeding behavior that exceeds the total time limit of the violation behavior into a speeding behavior that has a very short interval between segments a and b, but the duration of a single segment does not exceed the violation behavior limit, resulting in the violation behavior not being correctly identified; 2) Cutting off a period of speeding behavior that exceeds the total time limit of the violation behavior into two segments, A and B, with extremely short intervals, but each segment lasting longer than the violation behavior limit, resulting in the identification of two segments of speeding behavior; 3) Cutting off a period of speeding behavior that exceeds the prescribed time limit into two extremely short intervals, a and b, with one segment lasting longer than the prescribed time limit and the other segment not exceeding the prescribed time limit, resulting in the identification of one segment of speeding behavior.

To merge two intermittent speeding behaviors caused by noise points, the positioning time interval T between adjacent speeding records is introduced, and a threshold for merging the duration of illegal speeding behaviors is introduced β, based on industry experience preset: $0 \leq \beta \leq 4s$.

① Set the obtained driving record number to i and the violation and speeding behavior record number to j. Let i =1 and j = 0. When $i < n$, take the respective positioning coordinates of the i and i + 1 speeding records, and calculate the spherical distance d between the two coordinate points based on the spherical distance calculation formula;

② Determine if d is greater than α, if so, then determine if T is greater than β, if so, let t_2 be the i record time, with the sequence number $i = i + 1$. If $\Delta t > \delta$, the serial number of illegal speeding behaviors is $j = j + 1$, the number of illegal speeding behaviors increases by 1, and the speeding information is recorded; If the starting time t_1 of this violation of speeding behavior is the same as the starting time of the data in the output violation of speeding records, keep the vehicle information, the number of violations of speeding behavior, t_1, and the starting latitude and longitude, update and output other speeding information, t_1, set $\Delta t = 0$, and proceed to step 5. If T ≤ $T \leq \beta$, then proceed to step 4; If $d \leq \alpha$, then proceed to step 4;

③ Determine if t_1 is 0. If so, set t_1 to be the location time of the i overspeed data, and proceed to step 5;

④ Check if i is less than n. If so, return to step 1;

(3) Fatigue driving

In the process of identifying fatigue driving, it should be noted that if the vehicle is between a continuous trajectory and the time interval between two points where v ≠ 0 is greater than 20 minutes, it can be considered that the driver has stopped to rest, and the cumulative driving time for fatigue driving recognition is initialized to 0. The fatigue driving recognition process is as follows

① Input vehicle driving trajectory data: Set the obtained driving record number to i, initialize i = 1, set the overspeed record number to j, initialize j = 0. Input vehicle driving data T, which includes the positioning coordinates, speed, and time of each record. Initialize the number of violations and speeding behaviors to 0, set the vehicle running time win_r = [], and the parking rest time win_b = [].

② Determine the speed of the current record: Check the speed v_i corresponding to the current record point T_i If v_i is greater than 0, it indicates that the vehicle is in motion, execute ③ . If $v_i = 0$, it indicates that the vehicle may be stopping to rest, execute ④ .

③ Record vehicle travel time: Place the current time point T_i into the vehicle travel time win_r, where $i = i + 1$, reset win_b = [] to an empty list, return to execution ② , and continue checking the next record.

④ Record the vehicle's parking rest time: Place the current time point T_i into the parking rest time win_b, where $i = i + 1$. Calculate the time interval Δtb between the first and last points in win_b. if Δtb is less than 20 minutes, it is considered a brief stop, and return to execution ② to continue monitoring. if Δtb is greater than 20 minutes, it may be a break time. Execute ⑤ for further judgment.

⑤ Judging fatigue driving behavior: calculating the time interval Δtb between the first and last points in win_r. Based on the value of Δtb, select the processing method:

If the duration of Δtr during the daytime exceeds 4 hours, it is considered that the trajectory with a part $\Delta tr = 4 + \Delta tdr'$ greater than 4 hours represents a fatigue driving trajectory.

If the duration of Δtr during the nighttime exceeds 2 hours, it is considered that the trajectory with a part $\Delta tr = 2 + \Delta tnr'$ greater than 2 hours represents a fatigue driving trajectory. Afterwards, reset win_r= [] to an empty list and execute ② to continue monitoring new driving data.

(4) Prolonged idling

Vehicle idle refers to the state in which the vehicle is still in the starting state, i.e. the engine is in the ignition state, even when the vehicle is not in motion, i.e. $v = 0$. Being in this state for a long time can lead to vehicle wear and tear, as well as an increase in fuel consumption and energy waste. At the same time, it emits additional vehicle exhaust, which has a negative impact on the environment. The process for identifying excessive idle is as follows.

① Input the vehicle driving trajectory data T: initialize t = 0, identify the Prolonged idling idle window
win = [], and prepare to use it to identify the vehicle's idle behavior.

② Check vehicle speed and ignition status: For each time point T_t, check the vehicle speed v_t and ignition status ac. If the vehicle speed v_t is 0 and the ignition status ac == 1 (i.e., the vehicle is stationary but the engine is running), then proceed to step ③ . Otherwise (i.e. when the vehicle is moving or the engine is not running), reset the Prolonged idling recognition window win = [], as this indicates an interruption or absence of idle behavior.

③ Record idle status: Add the current time point Tt to the Prolonged idling recognition window win. Subsequently, check the time interval Δt between the first and last time points in win. If Δt exceeds 2 minutes, the recorded points during this period are recognized as the vehicle's idle point, indicating that the vehicle has been in an excessively long idle state during this period.

This process identifies Prolonged idling behavior by monitoring the duration of the vehicle's stationary and engine running state, which helps manage the vehicle's energy consumption and maintenance status.

2.3 Identification Results of Bad Driving Behavior

The data used in this article comes from the driving trajectories of thirty freight vehicles within a natural month, and the bad driving behavior recognition method is used to process and identify the driving trajectories of these vehicles. The license plate numbers of the top ten vehicles are shown in Table 6.

Table 6. Test vehicle license plate number

Number	License plate number
1	皖 A5C63*
2	皖 A5D03*
3	皖 A6A89*
4	皖 A1928*
5	皖 A2767*
6	皖 A2952*
7	皖 A3217*
8	皖 A3419*
9	皖 A3504*
10	皖 A3689*

Each data interval is the default data transmission interval of the system, which is about 30 to 60 seconds. The satellite positioning data size is a total of 6.75GB, and each data record contains fields such as license plate number, time, longitude and latitude, vehicle speed, and other information.

(1) Fatigue driving behavior recognition

By analyzing the data through fatigue driving behavior recognition algorithms, the recognition results can be obtained. The algorithm calculates the effective continuous driving time of the driver and the effective cumulative driving time of the day, and statistically analyzes the fatigue driving behavior that exceeds the time judgment threshold. Different types of alarms and the degree of violation are obtained, and relevant parameters such as the start and end times of alarms are recorded. As shown in Table 7, after the algorithm is implemented by the program, a dedicated algorithm runtime tool is used to calculate the algorithm completion analysis time of 2.612 seconds. There are a total of 427 identified fatigue driving behaviors.

Table 7. Fatigue driving vehicles

Number	License plate number	Current continuous driving time (hours)	Accumulated driving time of the day (hours)
1	皖 A5C63*	7.11	
2	皖 A5D03*	5.49	
3	皖 A6A89*		9.55
……	……	……	……
425	皖 A3217*		9.01
426	皖 A3419*	8.23	
427	皖 A3504*		9.00

(2) Identification of speeding behavior

This project selects partial driving data of operating vehicles on a certain route from Hefei to Jiangxi on a highway section, and the data transmission frequency is pre-set to 1 second. Preliminary screening was conducted on the overspeed data exceeding the speed threshold in the satellite positioning data transmitted by the vehicle terminal. An alarm was issued and stored in the real-time overspeed database of the dynamic monitoring platform. Some overspeed vehicle information is shown in Table 8, with a total of 87858 instances of overspeed behavior.

Table 8. Partial satellite positioning data

Number	Positioning time	Longitude	Latitude	Speed (km/h)
……	……	……	……	……
623	01-1617:53:00	117.11187	32.26702	105.6
……	……	……	……	……
647	01-1617:53:24	117.11278	32.27349	109
648	01-1617:53:25	117.11284	32.27376	109.4
649	01-1617:53:26	0	0	109.6
650	01-1617:53:27	117.11296	32.2743	109.1
651	01-1617:53:28	117.11302	32.27457	108.8

……	……	……	……	……
659	01-1617:53:36	117.11354	32.27668	106.9
……	……	……	……	……

(3) Recognition of rapid acceleration and deceleration behavior

The recognition of rapid acceleration and deceleration on the driving trajectory resulted in a total of 43001 results, as shown in Table 9:

Table 9. Results of speeding behavior detection

Number	Longitude	Latitude	Duration of rapid acceleration and deceleration	Maximum acceleration	Average acceleration
1	117.24834	31.785243	1	4.157	4.157
2	117.243853	31.804173	1	-3.056	-3.056
……	……	……	……	……	……
716	117.271906	31.815153	2	3.170	3.082
717	116.59953	35.027855	1	-4.319	-4.319
……	……	……	……	……	……

(4) Identification of prolonged idling

The identification of prolonged idling behavior on the driving trajectory resulted in a total of 11951 results, as shown in Table 10:

Table 10. Results of prolonged idling detection

Number	Longitude	Latitude	Engine ignition status	Speed	Duration
1	117.2531	31.8210	1	0	0:41:11
2	117.2724	30.91978	1	0	0:12:51
……	……	……	……	……	……
10000	117.2897	31.8635	1	0	1:02:36
10001	117.2578	31.8499	1	0	0:27:10
……	……	……	……	……	……

3 The Clustering Area Calculation Method of DBSCAN-Graham Combination

In traditional clustering algorithms, once the clustering results are computed, a visualization method is usually employed to assign different colors to points of different categories in order to distinguish points between different categories. However, this method is difficult to provide accurate ranges when identifying areas of bad driving behavior. To address this issue, we introduce the DBSCAN algorithm with the Graham algorithm to process the DBSCAN clustering results, thereby accurately delineating the boundaries of areas with frequent occurrences of bad driving behavior. This method assists management departments in more precisely monitoring and managing bad driving behavior and frequent occurrence areas in terms of time and location.

Based on the characteristics of satellite positioning data, we optimized the DBSCAN algorithm to balance the dual dimensions of time and space. Then, we used the Graham algorithm to delineate the boundaries of areas with high incidence of bad driving behavior, ultimately forming a complete method for identifying areas with high incidence of bad driving behavior based on the DBSCAN Graham combination. This method not only reveals the spatial distribution characteristics of bad driving behavior, but also reveals its dynamic evolution through time series analysis, greatly promoting the advocacy and implementation of road transportation safety management and environmentally friendly driving behavior.

3.1 Improved DBSCAN Algorithm Considering Spatiotemporal Conditions

Taking speeding behavior as an example, in order to extract the spatial distribution characteristics of this type of behavior, a point feature-based spatial clustering algorithm can be used to classify points of speeding behavior in order to achieve a similar effect within the same category, while ensuring differences between different categories. The widely used algorithm for this purpose is the density-based spatial clustering of applications with noise (DBSCAN) algorithm. This algorithm separates dense clusters from sparse noise by using a specified distance. The DBSCAN algorithm designates regions with a distance less than or equal to the search distance as clustering categories, while those with a distance greater than the search distance are considered noise points.

This algorithm has several advantages in the clustering scenario of speeding behavior points:

1. Non-convex shape clustering. DBSCAN is suitable for discovering non-convex shaped clustering structures. In geographical spatial data, especially points distributed along roads, they are often distributed along the direction of the road and may exhibit non-convex distributions, such as at bends, which may not conform to regular geometric shapes. The adjustable density parameter of DBSCAN enables it to adaptively find clusters of different shapes.

2. Robustness against noise and outliers. DBSCAN is robust against noise and outliers because it defines clusters based on density, unaffected by global averages. In speeding behavior data, there may be some outliers or exceptional points, and DBSCAN can effectively handle these cases.

3. Handling non-uniform density. Speeding behavior points exhibit non-uniform distribution characteristics in space and time, with density varying with geographical location. Some classic clustering algorithms, such as K-Means, may not perform well for clusters with non-uniform density, while DBSCAN divides clusters by adapting to the local density of data points, making it more suitable for areas with different densities.

In the process of identifying speeding behavior points, in addition to the latitude and longitude coordinates of speeding points, the time at which bad driving behavior occurs is also an attribute. When using the classic DBSCAN algorithm for computation, the data for both time and space attributes need to be used for distance calculation, i.e., a speeding behavior point includes distance (m) and time difference (min or s). The resulting Euclidean distance is influenced by dimensional effects, making it difficult to represent actual meaning. Therefore, a proposed improvement method separates the calculation of time and space attributes to address this issue. Only when the minimum number of points threshold is met within both the spatial and temporal neighborhoods, will the point be collected in a loop as a directly spatiotemporal density-reachable object.

The algorithm flow is as follows:

1. Given parameter Eps_{time}, Eps_{space}, $Minpts$, traverse the dataset $X, x_i\left(time, space\right) \in X$;

2. If x_i is not in any cluster, conduct a search on the two neighborhoods of x_i, denoted as $Eps = Eps_{time}$ and $Eps = Eps_{space}$. If both neighborhoods satisfy $|Npts(x_i)| > Minpts$, then x_i is a core point, and the corresponding cluster for x_i is denoted as $cluster_k$, with the set of points in the neighborhood denoted as $C, c_j \in C$. If $|Npts(x_i)| < Minpts$, then x_i is a noise point.

3. For any $c_j \in C$, calculate the number of points in the neighborhoods $Eps = Eps_{time}$ and $Eps = Eps_{space}$, denoted as $|Npts(x_i)|$. If there exists $|Npts(x_i)| > Minpts$, then c is density-reachable from x, and c is a core point belonging to c. Repeat steps 2 and 3 with c as the input.

4. If the calculation result in step 3 yields $|Npts(x_i)| < Minpts$, then c_j is a border point. It still belongs to $cluster_k$, but it is not considered a core point for subsequent calculations.

3.2 Graham's Algorithm

In the real vector space V, for a given set H, the intersection S of all convex sets containing X is referred to as the convex hull of X. In two-dimensional Euclidean space, the convex hull can be visualized as a rubber band containing all points, as shown in Fig. 3:

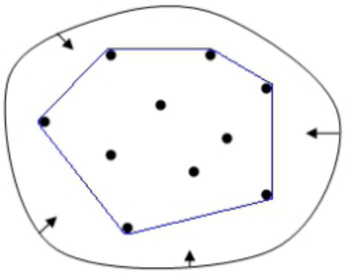

Fig. 3. Convex hull in two-dimensional Euclidean space

The two-dimensional convex hull can address problems such as fencing, urban planning, and cluster analysis. In this paper, the Graham's algorithm is utilized to obtain a more intuitive representation of areas with frequent occurrences of speeding behavior. Fig. 4 outlines the process of Graham's Scan algorithm for computing

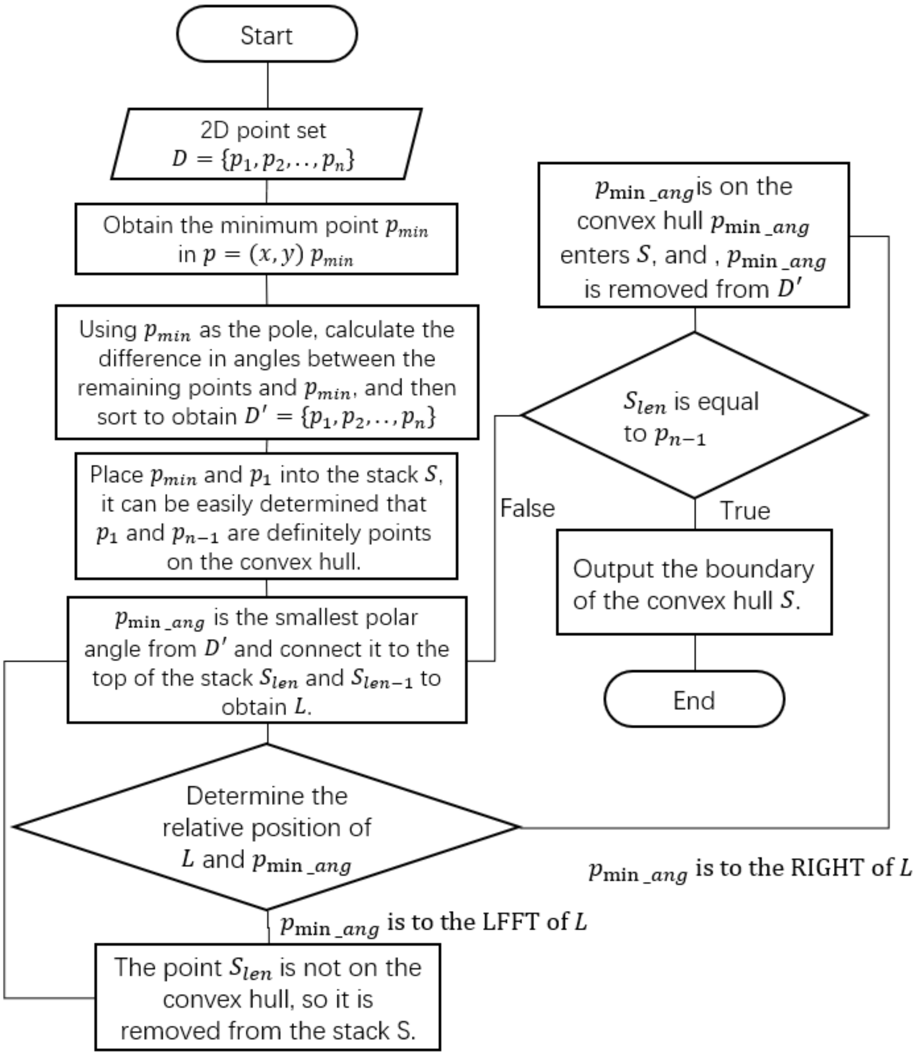
Start
2D point set $D = \{p_1, p_2, \ldots, p_n\}$
Obtain the minimum point p_{min} in $p = (x, y)$ p_{min}
Using p_{min} as the pole, calculate the difference in angles between the remaining points and p_{min}, and then sort to obtain $D' = \{p_1, p_2, \ldots, p_n\}$
Place p_{min} and p_1 into the stack S, it can be easily determined that p_1 and p_{n-1} are definitely points on the convex hull.
$p_{\min\_ang}$ is the smallest polar angle from D' and connect it to the top of the stack S_{len} and S_{len-1} to obtain L.
Determine the relative position of L and $p_{\min\_ang}$
$p_{\min\_ang}$ is to the LFFT of L
The point S_{len} is not on the convex hull, so it is removed from the stack S.
$p_{\min\_ang}$ is to the RIGHT of L
$p_{\min_ang}$is on the convex hull $p_{\min\_ang}$ enters S, and , $p_{\min\_ang}$ is removed from D'
S_{len} is equal to p_{n-1}
False
True
Output the boundary of the convex hull S.
End

Fig. 4. Procedure of the Graham's Scan algorithm

Graham’s Scan algorithm possesses the following characteristics:

(1) It is simple and efficient. The algorithm is relatively straightforward and performs well in most cases, with a time complexity of O(nlogn)

(2) It is stable. Through polar angle sorting and stack maintenance, the algorithm ensures the stability of the convex hull. Even if the input point set contains duplicate points, the output convex hull is uniquely determined.

(3) It is not dependent on specific distance metrics. The implementation of the algorithm does not rely on specific distance metrics, but rather on the polar angle relationships of the points, making it applicable to any point set for which polar angles can be defined.

4 Experimental Analysis

4.1 Identification of Bad Driving Behaviors

The data utilized in this study was obtained from the driving trajectories of thirty freight vehicles over the course of one calendar month. The identification of unsafe driving behaviors was conducted on these vehicle trajectories using a specific method. Due to the large amount of data reaching 6.75GB, this article uses random sampling to extract some data for identification and clustering at this stage, as shown in Table 11:

Table 11. Identification results of bad driving behavior

Behavior Car number	Fatigue driving	Prolonged idling	Overspeed	Rapid acceleration and deceleration
皖 D81***9	0	5	46	12
皖 D82***6	5	6	50	9
皖 AB***8	2	8	21	5
		……		
皖 M0***2	4	3	35	8
Total	27	717	5271	2580

Through data analysis, it has been determined that speeding and prolonged idling are the two most prevalent bad driving behaviors identified in the data, accounting for over 90% of all such behaviors. Specifically, speeding constitutes 61% of all bad driving behaviors. According to existing research findings [11], speeding is a major contributing factor to traffic accidents and significantly impacts the severity of such incidents. Additionally, prolonged idling is a primary cause of excessive vehicle emissions, leading to environmental pollution. Therefore, in order to enhance the safety and environmental standards within the road transportation industry, it is imperative to implement effective measures to control both speeding and prolonged idling behaviors.

4.2 Analysis of Clustering Results

1. Clustering of prolonged idling behavior

After multiple experimental analyses and considering social industry management experience, and combining silhouette coefficient and visual evaluation of clustering results, this study adopts $Eps_{time} = 60min$; $Eps_{space} = 500m$, $Minpts = 5$ as the input parameters for the DBSCAN algorithm based on spatiotemporal distance to cluster prolonged idling behavior points. After computation, a total of 38 normal clusters and 2133 individual noise points were obtained. Some of the clusters are displayed in descending order of quantity in Table 12.

Table 12. Clustering results for prolonged idling behavior

Group/ Cluster	Number of points in cluster	Time
Cluster I	105	00:53:41 to 11:58:56
Cluster II	103	00:07:56 to 09:19:56
Cluster III	79	04:47:17 to 10:50:41
	……	
Noise	1133	-

The distribution of the first and second clusters is shown in the following Fig. 5.

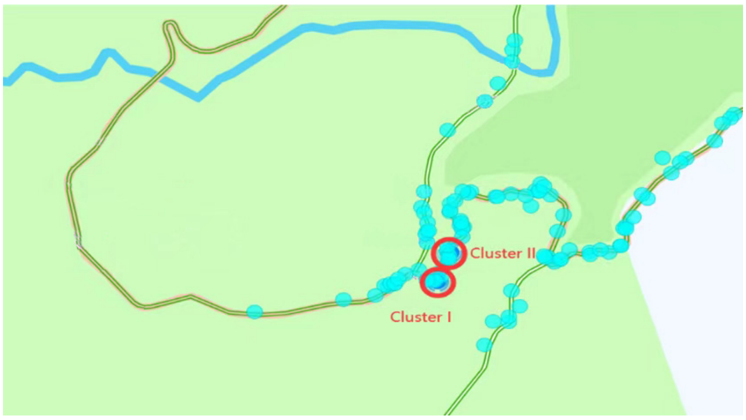

Fig. 5. The distribution of the first and second types of clusters on the map

The analysis of the first and second clusters reveals that these clusters are geographically close in proximity and have similar quantities. The examination of these two clusters yields the following insights:

(1) From a geographical distance perspective, the intersection of Provincial Road S230 and Township Road Y964 is an area with frequent occurrence of prolonged idling. In this area, the prolonged idling behavior is mainly concentrated on Provincial Highway S230.

(2) From the perspective of time distribution, the ultra long idle behavior of both clusters occurs during the time period from early morning to morning. There are fewer vehicles on the road during this time period, but driving behavior is mainly concentrated in this time period, indicating that the driver may be in two bad driving states: fatigue driving and prolonged idle. In this situation, the operational safety of the vehicle will be seriously affected.

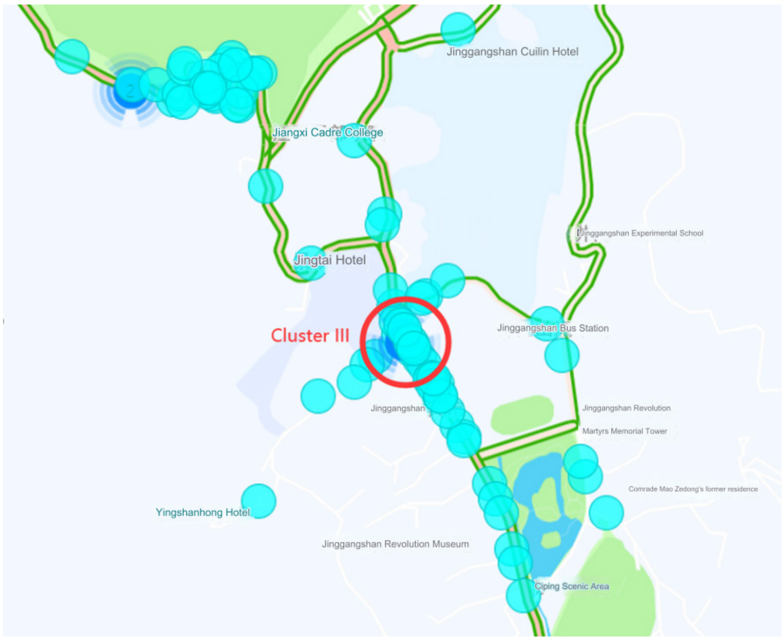

Fig. 6. The distribution of the second types of clusters on the map

Furthermore, the analysis of the third cluster, exemplified by local visualization shown in Fig. 6, with deep blue points representing clustered areas, reveals the following:

(1) The north-south direction road in the graph is national highway G220, with the northeast direction road on the south side being a city road named Changkeng Road, and the northeast direction road on the north side being a city road named Wujing Road. Instances of speeding behavior are prevalent at the intersections of G220 with the two city roads.

(2) Instances of speeding behavior predominantly occur during the late night to early morning period, a time when the roads are sparsely populated, making it easier for drivers to engage in speeding behavior when transitioning from city roads to national highways.

In summary, the spatial-temporal distribution of the main clustering clusters obtained through the consideration of spatiotemporal attributes using the DBSCAN clustering algorithm indicates that:

1) Transitioning from township roads, city roads, and other relatively narrow roads to national and provincial roads constitutes high-incidence areas for speeding behavior.

2) The late night to early morning period is a high-incidence timeframe for speeding behavior among commercial vehicles.

2. Overspeed behavior clustering

Due to the fact that speeding behavior mainly occurs on highways, its frequency is significantly higher than other bad driving behaviors. In order to accurately identify it, parameters have been reset in this section, Eps_time = 30min; Eps space = 100m, Minpts = 5, and using the inverse geocoding method, the speeding coordinate points located on the highway were screened. The spatiotemporal DBSCAN algorithm was used to cluster this type of data, and the local clustering results were plotted on the map. The cluster with the most matching number in the class was labeled in the map. After calculation, a total of 76 normal populations and 2626 individual noise points were obtained. Some ethnic groups are shown in descending order of number in Table 13.

Table 13. Clustering results for overspeed behavior

Group/Cluster	Number of points in cluster	Time
first kind	304	11:12:00-12:47:00
second kind	286	12:01:00-12:38:00
third kind	284	12:59:00-13:50:41
	……	
Noise	2626	-

The area with the most obvious concentration of speeding behavior is shown in Fig. 7.

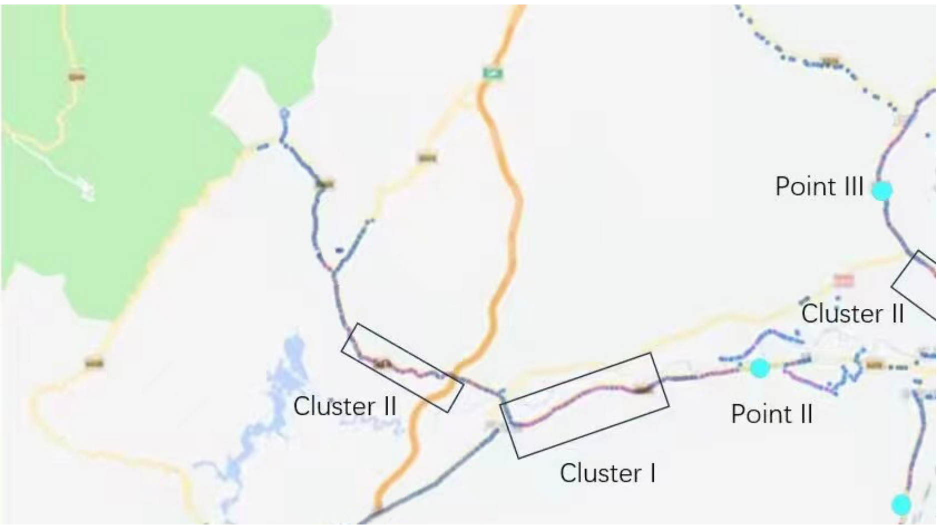

Fig. 7. The distribution of the overspeed behavior on the map

Based on time, it is found that the high incidence of speeding behavior is mostly concentrated in the middle of the day. Within the four hours from 11:00 to 14:00, there are 23 high incidence sub sections, accounting for nearly 47% of all high incidence sub sections. From the 21st hour to 1am the next day, within five hours, there were only two high speed sub sections. In terms of time dimension, the high speed sub section reached its peak at noon, indicating that this batch of operating vehicles had the highest possibility of speeding behavior occurring at noon

In terms of spatial evaluation, as speeding behavior often occurs on a longer road section, this article selects the cluster with the highest local density, Point I, as the analysis case. As shown in the following example, there are 38 speeding behavior points in a 200 meter road section. These speeding behavior points have a serious impact on road transportation safety due to a considerable number of speeding behaviors occurring within the short circuit section.

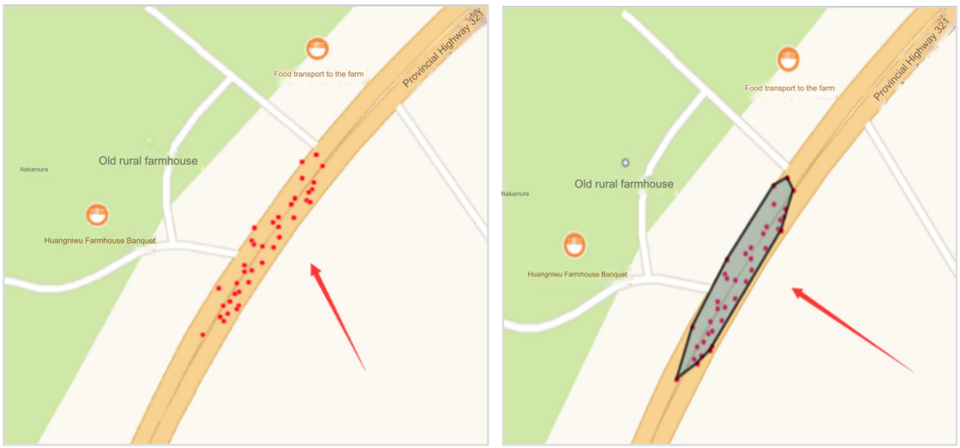
Food transport to the farm
Provincial Highway 321
Old rural farmhouse
Nakamura
Huangniwu Farmhouse Banquet
Food transport to the farm
Provincial Highway 321
Old rural farmhouse
Nakamura
Huangniwu Farmhouse Banquet

Fig. 8. Comparison of Graham algorithm performance

Utilizing the previously described Graham algorithm, the boundary of this category is computed and plotted on the map, as illustrated in Fig. 8. The black line area indicated by the arrow represents the boundary line of this category. It can be observed from the figure that the boundary line encompasses all the located points and effectively delineates the area of this category.

5 Conclusion

This article designs a spherical latitude and longitude interpolation method to reduce errors in data interpolation steps. Multiple indicators of bad driving behavior were calculated based on vehicle satellite positioning data. For the two indicators that occur more frequently and have a more concentrated distribution: excessive idle and speeding behavior, the DBSCNA algorithm considering time and space distance was used to cluster the designated points of bad driving behavior, and several areas of bad driving behavior were obtained. The clustering results show that speeding: ① prolonged idling driving behavior mostly occurs at night, and for vehicles driving on national roads from urban and rural roads, the likelihood of excessive idle behavior is higher. ② Overspeed and poor driving behavior often occur in the midday, in hot areas on certain road sections, and local law enforcement agencies need to improve law enforcement targeting. Poor driving behavior points are highly concentrated.

In addressing the clustering outcomes, particular attention was given to the challenge of high-density clusters that lack distinct boundaries. To tackle this issue, we developed a clustering region detection algorithm grounded in the Graham scan technique, which effectively delineates the clustering area using a two-dimensional spatial linear shape. This method allows for the precise identification of regions with a prevalence of poor driving behaviors, based on specific geographic locations and time frames. Moreover, recognizing the existence of clustering boundaries enables regulatory authorities to carry out more precise and targeted vehicle supervision within the defined area ranges. Such an approach not only enhances the efficiency of supervisory efforts by the management authorities but also contributes to a safer road traffic environment. Ultimately, these improvements lead to a reduction in carbon emissions produced by vehicles in road transport.

References

[1] J. Lu, K. Wang, Y.-M. Jiang, Real-time identification method of abnormal road driving behavior based on vehicle driving trajectory, Journal of Traffic and Transportation Engineering 20(6)(2020) 227-235.

[2] H.-X. Wang, X.-Y. Wang, Z.-X. Wang, X.-D. Li, Dangerous Driving Behavior Clustering Analysis for Hazardous Materials Transportation Based on Data Mining, Journal of Transportation Systems Engineering and Information Technology 20(1)(2020) 183-189.

[3] K. Wang, J. Lu, Y.-M. Jiang, Abnormal road driving behavior spectrum establishment and characteristic value calculation method based on vehicle driving trajectory, Journal of Traffic and Transportation Engineering 20(6)(2020) 236-249.

[4] Q. Bao, Q.-K. Qu, H.-R. Tang, J.-M. Chen, Y.-J. Shen, Multi-drivers Risk Evaluation Based Proactive Intervention of Drivers' Risky Behavior Under Connected Transportation Contexts. Journal of Transportation Systems Engineering and Information Technology 22(4)(2022) 283-292.

[5] Z.-L. He, Analysis of driving behavior of truck driver and risk of overspeed assessment based on big data, [Master's dissertation] Guangdong: Guangdong University of Technology, 2022. DOI: 10.27029/d.cnki.ggdgu.2021.001868

[6] J. Zhong, W.-X. Zhang, W. Li, K. Chen, X.P. Chen, J.-M. Li, Analysis of operating speed characteristics and overspeed situation on curved slope sections of highways, Highway Transportation Technology (Applied Technology Edition) 16(7)(2020) 302-305+319. < https://kns.cnki.net/kcms2/article/abstract?v=YoFA4grnCX5rWu6gc1R5A2x-Il0l9fqZrJMFMEltR0SVmbTJn0T2bu88-v1ufPZvdwe-fnSJ4PDNzRSo1H74ma6Dj1oQOjjKweyqVvXPP7SP49Wn-qlqhUcIUd35qgrD06aPyplyegC5w=&uniplatform=NZKPT&language=CHS >

[7] W.-X. Gao, Analysis and Research on Speeding Section Based on DBSCAN Algorithm, [dissertation] Harbin:Harbin Engineering University, 2022. DOI: 10.27060/d.cnki.ghbcu.2021.001515

[8] Z.-P. Peng, H.-Y. Pan, Y.-G. Wang, Analyzing the Causes of Traffic Accidents of Online Ride-Hailing Cars Using the Bayesian Network, Journal of Northeastern University (Natural Science) 44(1)(2023) 145-152.

[9] Y. Cheng, J.-L. Zhang, S.-J. Zhang, J.-F. Guo, D. Zhang, Evaluation of Eco-driving Behavior and Fuel-saving Potential of Large Freight Vehicles, Journal of Transportation Systems Engineering and Information Technology 20(6)(2020) 253-258.

[10] Y.-J. Liu, C. Zeng, S.-J. Wang, Y. Yao, M.-Y. Liu, An Evaluation Method of Safety and Energy-saving Driving Behavior Based on Satellite Positioning Data, Journal of Highway and Transportation Research and Development 35(1) (2018) 121-128+158.

[11] L. Zheng, P. Gu, J. Lu, A Cause Analysis of Extraordinarily Severe Traffic Crashes Based on T-S Fuzzy Fault Tree and Bayesian Network, Journal of Transport Information and Safety 39(4)(2021) 43-51+59.