Journal of Network Intelligence ©2016 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **10**, Number **2**, August **2025**

Research on Abnormal Toll Path Identification Method for ETC System Combining "Two-Passengers and One-Risk" Vehicle Track Data

Fu-Min Zou, Yi-Heng Su*, Lv-Chao Liao

Key Laboratory of Automotive Electronics and Electric Drive
Fujian University of Technology, Fuzhou 350118, China
fmzou@fjut.edu.cn, suyiheng1219@163.com, achao@fjut.edu.cn

*Corresponding author: Yi-Heng Su

Received May 2, 2023, revised April 6, 2024, accepted August 17, 2024.

Abstract. *Since the cancellation of toll stations at provincial boundaries of highways, the whole country has entered an era of network operation. The cancellation of provincial toll stations has resulted in a significant increase in single-trip costs. Some drivers have the idea of using illegal means to evade payment of tolls. Various new ways of evading tolls have also emerged. Various and hidden behaviors of evading tolls have brought great challenges to the auditing work of highway operation departments, among which the evasion phenomenon of changing the route type is the most frequent. ETC system has a large amount of vehicle toll data, how to extract valuable information from the large amount of data, so as to achieve efficient identification of vehicles with abnormal toll paths, has become the focus of current research. The rise of big data provides convenience for traffic management departments to audit evasion fees. Data mining technology can be used to analyze the hidden evasion behavior characteristics behind abnormal traffic data, and to build a vehicle identification model for abnormal toll paths. The vehicle identification model of abnormal routes can improve the auditing efficiency of management departments.*

Keywords: Expressway, ETC, GPS trajectory, Machine learning

1. **Introduction.** As of the end of 2021, the total mileage of toll roads in China is 18.76 kilometers, accounting for 3.55% of the total mileage of 5.2807 million kilometers [1]. Among them, highways have reached 161200 kilometers, accounting for 85.92% of the total mileage of toll roads. The rapid expansion of highways has promoted the development of transportation and provided great convenience for people's travel, but at the same time, it has also brought some problems. Due to the increase in mileage, the cost of a single trip has also significantly increased. Some drivers take risks to seek maximum benefits and adopt various illegal means to achieve the goal of evading or underpaying tolls [2]. At present, the main method of fee audit is mainly manual, and the efficiency is not high. Seeking an efficient and feasible inspection method has become the main goal of this study.

Fukazawa Kazuo provided a detailed introduction to Japan's toll collection system, stating that Japan's handling of fee evasion is through the installation of identification devices on key sections of highways [3]. In studying the issue of toll loss, D Burgess proposed that in order to reduce the possibility of toll evasion, it is necessary to upgrade the current technology [4]. With the gradual improvement of the ETC charging system and the maturity of data mining technology in practical applications, data mining research based on massive driving data has brought new possibilities for the identification of abnormal vehicles. In [5], the authors categorized the abnormal behavior of vehicles into five categories. In [6], the authors used abnormal vehicle data for transporting poultry meat to construct an audit model for counterfeit green traffic. In [7], the authors conducted mining and analysis on massive toll data, using clustering methods for classification and outlier detection to identify abnormal data. In [8], the authors are based on green traffic detection data, statistical analysis of data attributes, and combined with SMOTE sampling to construct a logistic model. In [9], the authors first constructed a data warehouse based on the highway toll system, applied decision tree algorithm and Bayesian algorithm to the classification of overweight vehicles, and used linear regression algorithm to predict the toll amount. In [10], the authors have improved the neural network algorithm and proposed the IGA-IBP prediction model based on the big data of highway toll evasion, providing theoretical preparation for accurate inspection of toll evasion customers. In [11], the authors use Storm recognition technology to identify vehicles on highways, mainly targeting abnormal vehicles that evade fees through counterfeit free vehicles. In [12-16], the authors focus on the analysis of three types of abnormal data, including repeated transactions, missed transactions, and false transactions, related to the interaction between the passing medium and the gantry during vehicle operation.

In order to more efficiently identify vehicles with abnormal charges, this article first integrates the trajectory data of "Two-Passengers and One-Risk" vehicles and ETC charging data to extract the ETC charging path corresponding to the "Two-Passengers and One-Risk" vehicles. Secondly, this paper proposes an anomaly path recognition algorithm combining GeoHash encoding, and analyze and study the identified abnormal data to extract feature commonalities between the data. Finally, an anomaly vehicle recognition algorithm is constructed based on anomaly feature vectors to achieve efficient anomaly recognition of massive data.

2. **Research on abnormal path recognition.**

2.1. **ETC system data preprocessing.** The research data in this article comes from vehicle toll data collected by the ETC system from September 3-5, 2020, which includes toll station entrance and exit data and ETC gantry toll data. The entrance data includes 54 variables, including transaction identification, entrance toll station name, entrance time, entrance vehicle type, number of entrance axles, entrance weighing, etc; Export data includes 117 variables. ETC gantry charging data includes 103 variables such as transaction identifier, transaction time, gantry name, gantry type, etc. From this, it can be seen that there are indeed many variables in the vehicle traffic information, but not all of them have a role in the study of abnormal path recognition. Redundant variables even interfere with the recognition of abnormal paths, reducing recognition efficiency. In order to establish a more accurate anomaly path recognition model, it is necessary to delete feature attributes unrelated to the research content in the original data structure, filter out relevant available field information, and the filtered data can more clearly reflect the behavioral characteristics of abnormal vehicles. Based on the exploration of the data structure of the ETC system, this article extracts the feature attribute information fields related to it for the convenience of later research on anomaly path recognition, as shown in Table 1 and Table 2.

The previous article introduced the information exchange between vehicles and the gantry through traffic media, uploading transaction data to the provincial highway toll management center and departmental networking center. The ETC system forms the

Table 1. Processed ETC charging data feature attributes

Name	*Description*	*Example*
Passid	Driving ID	02350*******************200903201110
Tradetime	Transaction Time	2020/9/3 21:51:11
Flagid	Gantry number	350265
Obuid	OBU ID	********
Obuplate	OBU License plate	******
Vehclass	Vehicle model	16

Table 2. Processed toll station feature attributes

Name	*Description*	*Example*
Passid	Driving ID	02350*******************200903201110
Entime	Entry time	2020/9/3 20:11:10
Extime	Exit time	2020/9/3 23:57:16
Enstation	Entrance toll station number	3105
Exstation	Exit toll station number	6706
Envehclass	Vehicle model of entry	16
Exvehclass	Vehicle model of exit	16
Entotalweight	Weight of entry	49300
Extolweight	Weight of exit	49300
Envehplate	License plate of entry	******
Exvehplate	License plate of exit	******

ETC transaction gantry trajectory by arranging the obtained transaction data in chronological order. In order to display the trajectory more intuitively and analyze its abnormal phenomena, the following definitions are made here:

Transaction nodes Node: The ETC gantry and toll stations on the highway are collectively referred to as nodes, and the nodes where vehicles interact with the ETC gantry are referred to as transaction nodes. Transaction nodes include travel ID, transaction gantry, and transaction time, and their expression is shown in Formula (1):

$$Node = \{Passid, Flagid, Tradetime\} \tag{1}$$

Vehicle Trading Gantry trajectory Traj: The transaction data generated by the interaction between the passing medium and the gantry during the vehicle's movement, sorted according to the transaction time, which is called the vehicle transaction gantry trajectory $Traj$. The trajectory of the vehicle trading gantry is composed of trading nodes, and its expression is shown in Formula (2):

$$Traj =< Node_1, Node_2, ..., Node_{n-1}, Node_n > \tag{2}$$

Among them, $Node_1$ is the starting transaction node of the vehicle's current highway driving, and $Node_n$ is the ending transaction node of the vehicle's current highway driving.

Passage Section: The passage section is the gantry corresponding to the two adjacent transaction nodes on the track of the vehicle transaction gantry, which is combined in pairs according to the transaction time sequence. Its representation is shown in Formula (3):

$$Section =< Node_n, Node_m > \tag{3}$$

Among them, $Node_n$ and $Node_m$ are the nodes with adjacent transaction times in the vehicle trading gantry trajectory. However, due to ETC gantry equipment issues or vehicle fee evasion behavior, $Node_n$ and $Node_m$ may not necessarily be connected gantry in terms of topology. There is an abnormal situation in the interaction between the vehicle traffic medium and the ETC gantry, which is manifested by incorrect topological relationships of the section gantry. The gantry has the phenomenon of false detection, missed detection, or repeated detection of the traffic medium, as shown in Figure 1. Due to the existence of abnormal situations and fee evasion behaviors within the passage section, it is necessary to further define the section. The topological relationship between $Node_n$ and $Node_m$ in the section, which is connected and adjacent, is called Adjacent section. Classify the topological relationships between $Node_n$ and $Node_m$ that are connected but not adjacent as Missed detected section. The incorrect transaction section where the topological relationship between $Node_n$ and $Node_m$ is not connected is called Unconnected section. By using the unique travel ID of the vehicle traveling on the highway, integrating ETC toll data and entrance/exit toll station data, the vehicle transaction gantry trajectory $Traj$ can be obtained. The transaction node nodes within the vehicle transaction gantry trajectory $Traj$ are combined according to the transaction time to obtain the passage section, which forms the vehicle travel trajectory section table, as shown in Table 3.

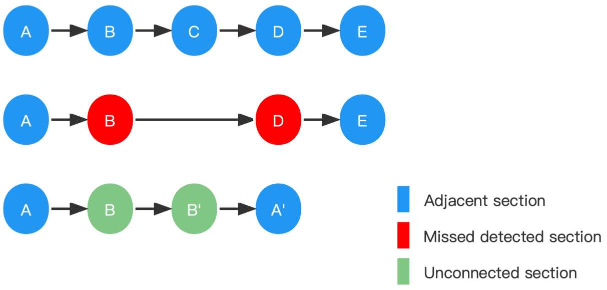


Figure 1. Abnormal trading gantry trajectory

Table 3. Table of vehicle traffic sections

Name	*Description*	*Example*
Passid	Driving ID	02350*******************200903201110
Section	Trading gantry section	$[350267, 350265]$
Section_Time	Section trading time	[2020/9/3 21:48:39,2020/9/3 21:51:11]
From_Flag	Starting gantry number	350267
To_Flag	Arrival gantry number	350265
Distance	Distance(m)	2816
Timedelta	Time difference(s)	152
Vehspeed	Average speed(km/h)	66.69

2.2. **"Two-Passengers and One-Risk" preprocessing.** "Two-Passengers and One-Risk" refer to chartered buses, Class III or above buses engaged in tourism, and road specialized vehicles transporting hazardous chemicals, fireworks, and civilian explosives [17]. Transportation enterprises must comply with relevant requirements and install satellite positioning devices that meet the requirements of "Two-Passengers and One-Risk" vehicles, and connect them to the national key operational vehicle networking and control system to ensure accurate, real-time, and complete transmission of vehicle monitoring data, and to ensure the normal operation, accurate data, and effective monitoring of on-board satellite positioning devices. By collecting the trajectory points of "Two-Passengers and One-Risk" vehicles, the real driving trajectory of the vehicles on the highway is restored, and the real driving trajectory is compared with the vehicle transaction trajectory collected by the ETC system to achieve the purpose of identifying abnormal paths.

The "Two-Passengers and One-Risk" data used in this article comes from the trajectory data provided by Fujian Expressway Futong Logistics Co., Ltd. from September 3, 2020 to September 5, 2020, mainly including basic data such as vehicle identification, positioning time, longitude, dimension, speed, direction, etc. The basic attributes of the "Two-Passengers and One-Risk" trajectory data are shown in Table 4.

Table 4. "Two-Passengers and One-Risk" vehicle trajectory data attributes

Name	*Description*	*Example*
Plate	License plate number	******
Loctime	Positioning time	2020-09-03 20:49:39
Lon	Longitude	119.59
Lat	Latitude	25.79
Direction	Direction	331
Speed	Speed(km/h)	29
Mileage	Mileage(km)	0
Altitude	Altitude(m)	11

From the research on fee evasion behavior mentioned above, it can be seen that the most common behavior on highways is to change the toll path for fee evasion. The method is to use the shortest OD path to charge if no other gantry is detected within the OD section. According to the definition above, the section where no other gantry frames are detected within the OD is called the missed detection section. Therefore, the premise of abnormal path recognition research is to extract the missed detection section of "Two-Passengers and One-Risk" vehicles in the ETC trading gantry track. This chapter uses the fusion method of "Two-Passengers and One-Risk" trajectory data and ETC gantry trading data to obtain the ETC trading segments corresponding to the trajectory points.

Due to the time attribute of both the "Two-Passengers and One-Risk" trajectory data and the ETC gantry trading data, the ETC trading section corresponding to the trajectory point can be obtained by matching the trajectory point positioning time with the gantry trading section time. The extraction process is as follows:

Step 1: Extract the "Two-Passengers and One-Risk" vehicle trajectory points and ETC gantry transaction data of the same vehicle based on the vehicle identification.

Step 2: Traverse the trajectory points of "Two-Passengers and One-Risk" vehicles, and store the data of the trajectory points located within the corresponding ETC gantry trading section in different sections.

Step 3: Process the trajectory points of "Two-Passengers and One-Risk" vehicles in the same section in a row expansion to obtain the "Two-Passengers and One-Risk" vehicle section trajectory table which is shown in Table 5.

2.3. Abnormal path recognition algorithm based on fusion of multiple source data.

Due to the topological relationship of the highway network structure, identifying the gantry can restore the true driving trajectory of the vehicle. The trajectory points of

Table 5. Track table of ”Two-Passengers and One-Risk” Vehicle Sections

Name	*Description*	*Example*
Plate	License plate number	******
Loctime	Positioning time	2020-09-03 20:47:28
Section_Time	Section trading time	[2020-09-03 20:42:13, 2020-09-03 20:56:20]
Section	Trading gantry section	[350139, 350137]
Lon	Longitude	119.59
Lat	Latitude	25.79
Speed	Speed(km/h)	29

“Two-Passengers and One-Risk” vehicles are dense. By identifying ETC gantry frames within a certain range of trajectory points and forming a set of gantry frames, based on the topological relationship within the ETC gantry frame set, the shortest path corresponding to the missed detection section OD in the gantry frame set is found, which is the true driving path corresponding to the vehicle trajectory points [18], Identify abnormal paths by comparing the topological distance of the gantry corresponding to the actual driving path with the topological distance of the toll path gantry. Due to the large amount of computation required to identify whether there is a gantry within the range of trajectory points through distance, the algorithm combines GeoHash encoding to achieve the recognition of gantry points, in order to quickly identify abnormal paths of “Two-Passengers and One-Risk” vehicles.

2.3.1. **GeoHash encoding.** The GeoHash encoding can convert 2D latitude and longitude coordinates into one-dimensional string encoding similar to URLs. The algorithm uses class dichotomy to divide and label geographical regions [19]. It should be noted that GeoHash encoding does not represent a point, but a rectangular region, and the width and height of the rectangular region are related to the length of the string, as shown in Table 6. When the GeoHash encoding length is 5, the width and height of the rectangular area are both 4.9km, while when the length is 8, the width and height of the rectangular area are about 19 meters.

Table 6. GeoHash encoding length and matrix region size correspondence table

Code length	1	2	3	4	5	6	7	8
Width(km)	5009.4	1252.3	156.5	39.1	4.9	1.2	0.153	0.038
Length(km)	4992.4	624.1	156	19.5	4.9	0.609	0.152	0.019

Therefore, the selection of encoding length is quite important for the recognition of gantry frames. The appropriate encoding length ensures that there are not too many redundant gantry frames in the gantry set, and at the same time, the trajectory point recognition process does not have the situation of missing gantry frames, effectively improving the efficiency of abnormal path recognition. This chapter sets the encoding length to 6 to encode the portal frames. The proportion of portal frames corresponding to GeoHash encoding is shown in Figure 2. It can be seen that 31.33% of portal frames have unique GeoHash encoding, and 68.67% of portal frames share GeoHash encoding with other portal frames. 68% of the gantry pairs that share GeoHash encoding are composed of opposite gantry frames. Due to the close installation distance of opposite gantry frames, it is normal for these gantry frames to share GeoHash encoding. In summary, a code length of 6 is basically able to distinguish the position of the gantry on the highway.

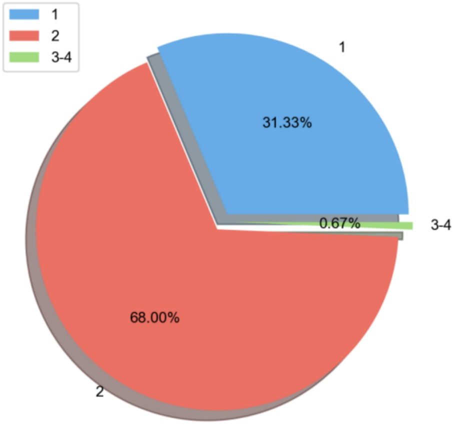


Figure 2. Proportion of GeoHash encoded gantry

2.3.2. **Abnormal path identification process.** Based on the comparison of trajectory points and gantry frames one by one, the gantry frame set is obtained. The Dijkstra algorithm is used to calculate the shortest path of the start and end OD of the missed detection section in the gantry frame set, and the distance is compared with the toll path. If the distance is not equal, the path is an abnormal path. The process of identifying abnormal paths is as follows:

Step 1: Identify missing detection segments in the ETC trading trajectory.

Step 2: Specify the OD of the missed detection section as the starting and ending gantry frames, and extract the corresponding "Two-Passengers and One-Risk" trajectory points within the ETC transaction time of the missed detection gantry frame as the time range.

Step 3: Calculate the GeoHash code corresponding to the trajectory points and compare it with the gantry table. Place the gantry with the same GeoHash code into the gantry set, and repeat this operation until all trajectory points are compared.

Step 4: Perform a deduplication operation on the gantry set, and only use the topological relationship of the gantry within the gantry set to find the shortest path from the starting gantry to the ending gantry, which is the true driving path corresponding to the trajectory point.

Step 5: If there is no shortest path, calculate the shortest distance between the starting gantry and each gantry in the gantry set, take the node with the shortest distance as the endpoint gantry, and calculate its shortest path as the true driving path corresponding to the trajectory point.

Step 6: Compare the distance of the actual driving path with the distance of the toll path. If it is the same, it is normal to miss detection. If it is not the same, there is an anomaly, and this path is the identified abnormal path.

As shown in Figure 3, the shortest path from A to C is A->B->C. However, according to the GeoHash encoding comparison of the trajectory points, the B gantry is not within the specified range. Therefore, the identified shortest path is A->D->E->C, which does not match the shortest path, so this path is an abnormal path.

The pseudocode of the anomaly path recognition algorithm integrating multi-source data is shown in Table 7.

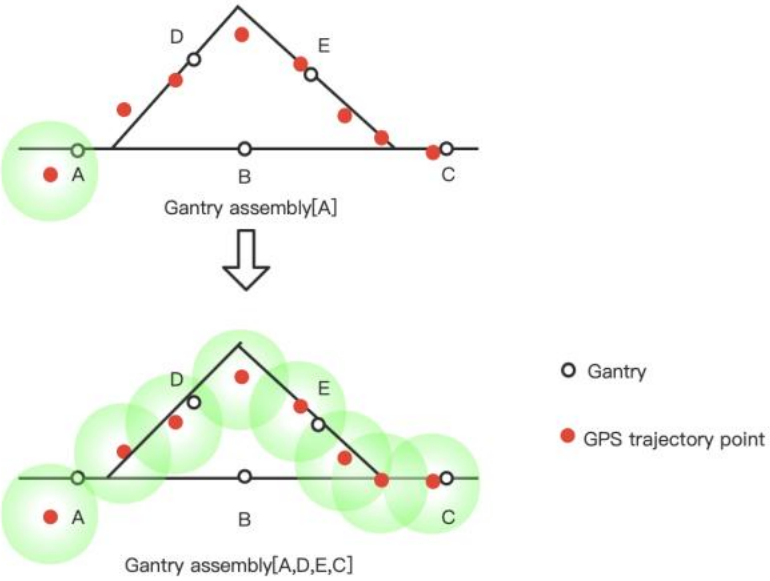


Figure 3. Schematic diagram of GPS trajectory point matching

Table 7. Abnormal path recognition algorithm table

```
Algorithm: Abnormal Path Recognition Algorithm
Input: $DG$, $GPS\_Traj$, $Node$ # Road network model $DG$, vehicle positioning trajectory $GPS\_Traj$, gantry position $Flag\_Point$
Output: $Result$ # Normal = 0, Abnormal = 1
$Flag\_GeoHash$ = getGeoHash ($Flag\_Point$, 6) # Obtain the GeoHash of the gantry based on the encoding length of six
While $Traj$ in $GPS\_Traj$: # Traverse each vehicle trajectory
    While $point$ in $Traj$:
        Check $point.GeoHash$ in $Flag\_GeoHash$ # View positional relationship between trajectory points and gantry
$Flag\_List$ = get $Node$ # Intersecting gantry frames form a set of gantry frames
$NodePath$ = Dijkstra ($CarTraj.start, CarTraj.end, DG, Flag\_List$) # Use the Dijkstra algorithm to find the shortest path from the start point to the end point of the trajectory within the range of the gantry set
If Get_Shorest_Distance($NodePath$) - $PayPath$ != 0: $Result$ = 1
 Else: $Result$ = 0  Determine whether the path distance of trajectory point is equal to toll distance
Return $Result$
```

2.4. **Analysis of results.** The identification results are shown in Table 8. After comparison, there are 172 abnormal paths in the "Two-Passengers and One-Risk" vehicle data. Among them, only 13 abnormal paths with less than 8 missed detection gantry frames in the section are found. This is because the vast majority of OD pairs do not have ambiguous paths between them; When the number of leakage detection gantry frames in a section exceeds 8, there are 159 abnormal paths, accounting for 92.44% of the abnormal data. This is because as the distance between sections increases, the number of ambiguous paths between OD pairs gradually increases, making it easier for the actual path to deviate from the charged path. These vehicles with abnormal paths have caused the loss of tolls. Based on these abnormal data, the following will identify vehicles with abnormal paths.

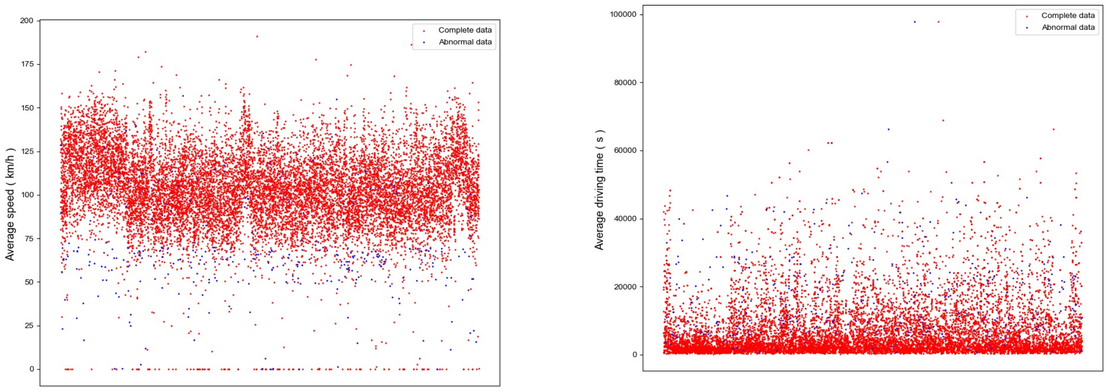
3. **Research on the behaviors of toll evasion.**

Table 8. Abnormal path identification results

Total number of abnormal sections	*Number of abnormal gantry frames*	*Number of abnormal paths*	*Abnormal rate*
18732	1-8	13	0.06%
4734	8-20	159	3.35%
23466	\	172	1.71%

3.1. **Analysis of driving characteristics.** Due to the fact that abnormal behavior is often different from normal driving thinking and has a certain degree of induction, by mining the commonalities between data, the characteristics of vehicle evasion on abnormal paths can be summarized.

3.1.1. **Average speed and average driving time.** From Figure 4(A), it can be seen that the average time distribution of vehicles is concentrated within 0-10000 seconds, while the average travel time of abnormal data mostly exceeds 10000 seconds. Vehicles that shield traffic media from evading fees are more willing to run longer distances to achieve greater benefits. Therefore, data with long travel time needs to be given special attention.From Figure 4(B), it can be seen that the average driving speed of vehicles is basically within the speed limit range of 60-120km/h on highways, with only a small portion of data showing slow speeds, which may be caused by traffic congestion or vehicles entering service areas. The average speed of the abnormal data is mostly below 60km/h, and the abnormal driving time is mainly caused by the inconsistency between the toll path fitted by the ETC gantry and the actual path. Therefore, attention should be paid to vehicle data with abnormal speed.

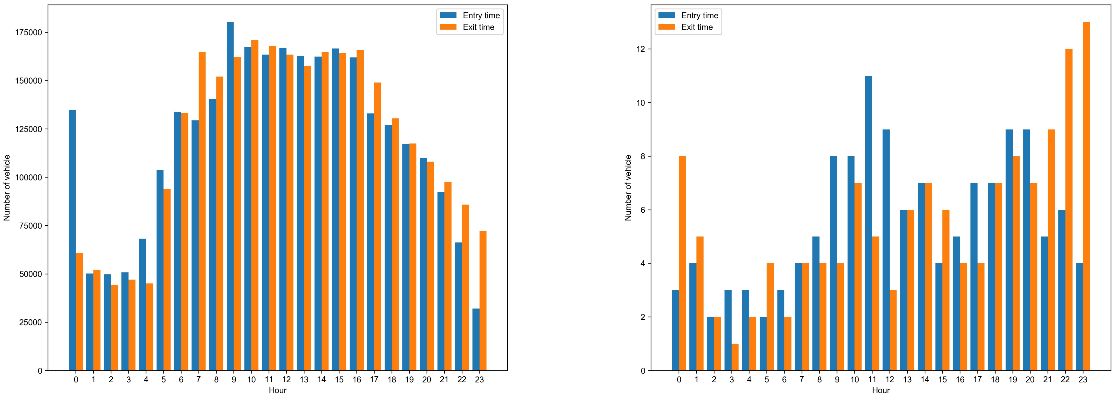


(A)Distribution diagram of average speed

(B)Distribution diagram of average driving time

Figure 4. Schematic diagram of vehicle driving attributes

3.1.2. **Entry and exit time.** From Figure 5(A), it can be seen that the number of vehicles corresponding to the time of vehicle entry and exit is basically consistent with the distribution of traffic flow during peak hours. During peak hours from 9:00 to 17:00, there is a large amount of traffic on the highway, and there are many vehicles entering and exiting the highway. From 17:00 to 4:00 the next day, the number of vehicles entering and exiting the highway gradually decreases. From Figure 5(B), it can be seen that the number of vehicles corresponding to the entry time of abnormal data is basically consistent with the changes in peak traffic flow. However, there are fewer vehicles with exit times between 6:00 and 17:00, while there is a gradual increase in exit vehicles between 17:00 and 0:00. There is a possibility that drivers may evade fees while staff are slack. Therefore, this article includes the exit time in the abnormal feature vector.



(A)Complete data (B)Abnormal data

FIGURE 5. Distribution map of vehicle entry and exit time points

3.1.3. **Vehicle type.** Classify the vehicle models with abnormal data into passenger cars, trucks, and special operation vehicles, as shown in Figure 6. The number of abnormal trucks in the abnormal data is much higher than the other two categories, while the number of passenger cars is the least. This also conforms to the logic that truck payment rates are higher and are prone to fee evasion.

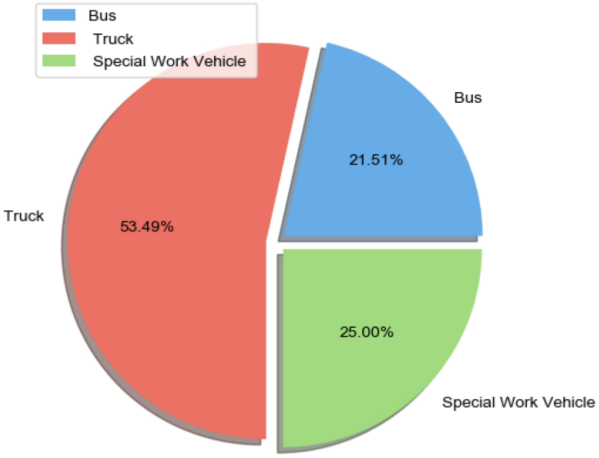


FIGURE 6. Statistical chart of abnormal data vehicle models

3.2. **Load characteristic analysis.** Due to the current charging mode of charging based on vehicle type and the fact that trucks need to be weighed and tested by overload control equipment when entering the highway, the previous method of reducing the weight of the weight is basically unable to evade fees. Instead, more purposeful groups and vehicle teams are organizing to evade fees. This evasion method is often highly covert and difficult to detect through data visualization. However, due to the different driving logic of fee evasion behavior from normal behavior, abnormal data can be detected through relevant feature combinations. Abnormal driving behavior of trucks is mainly divided into long-distance light load and short-term heavy load. This article will analyze abnormal behavior in two situations. In order to more intuitively express the phenomena of light load and heavy load, this article introduces the load rate as shown in Formula (4).

$$r = \frac{w}{w_L} \times 100\% \tag{4}$$

Where r is the load capacity, w is the actual load capacity of the vehicle, and w_L is the maximum load capacity corresponding to the vehicle type (axle). According to the standard overload rate, the normal range of values is 0-1.05, and some vehicles holding large transportation related documents may have a load rate exceeding 1.05.

According to statistics, the load capacity of trucks with a driving distance of over 30 kilometers in the province is shown in Figure 7. Except for the emergency vehicles corresponding to model 26, it can be seen that the load rate also increases with the increase of the model. It can be seen that large vehicles are more concerned about whether they are running empty or not.

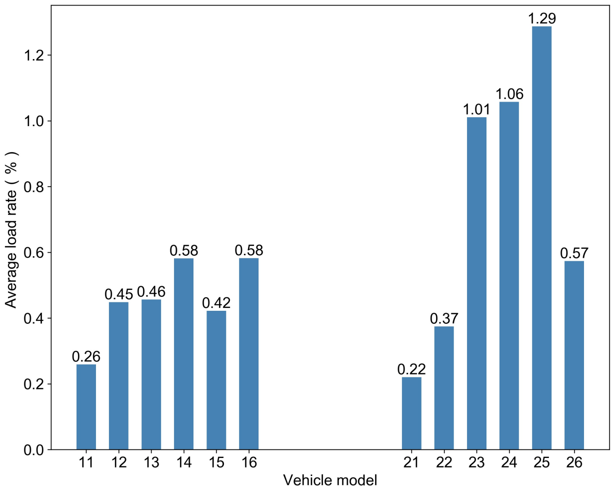
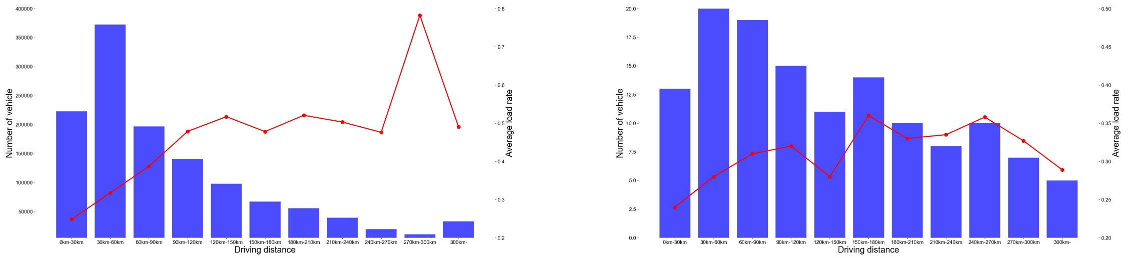


Figure 7. Load rate and vehicle model statistics chart

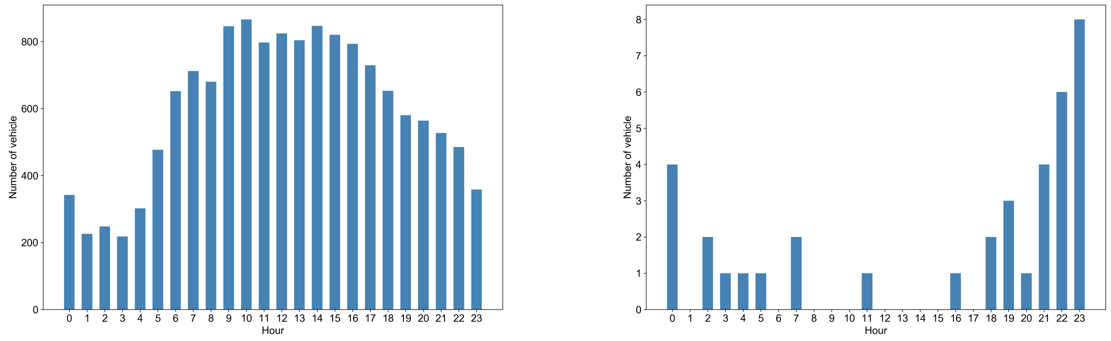
By analyzing the load rate and driving distance of trucks in the complete data, as shown in Figure 8(A), the vast majority of vehicles in the complete data have a driving distance of less than 100km, and the load rate of trucks is basically positively correlated with the driving distance. The larger the driving distance, the higher the load rate, and the overall average load rate is higher than 50%. The load rate and driving distance of trucks in the abnormal data are statistically analyzed, as shown in Figure 8(B). As the driving distance increases, the overall number of vehicles shows a downward trend, which is consistent with the trend of complete data changes. However, the overall load capacity of abnormal data vehicles is lower than that of complete data, and when the driving distance exceeds 150km, there is a downward trend in load capacity, which does not conform to the logic of normal truck loading. Therefore, data with long driving distance but low load capacity needs to be given special attention.



(A)Complete data (B)Abnormal data

FIGURE 8. Load rate and driving data statistical chart

3.3. **Analysis of behavior with entry but without exit.** Extracting complete data and abnormal data, the relevant data of the last gantry in the trading trajectory of the gantry is obtained to obtain the time when this behavior occurred, as shown in Figure 9. The distribution of whether there is an entry or exit in the complete data is basically consistent with the peak period of traffic flow, and the peak period of whether there is an entry or not is 9-16 o'clock. At this time, it is normal for vehicles to exit the province through the provincial toll station; In abnormal data, the frequency of entry and exit phenomena at night is much higher than during the day. The period from 21:00 to 2:00 the next day is the most frequent time period for abnormal phenomena to occur, which conforms to the behavioral logic of vehicles evading fees while staff are tired and relaxed.



(A)Complete data (B)Abnormal data

FIGURE 9. Abnormal time distribution with entrance but without exit

3.4. Analysis of behavior with entry but without exit.

3.4.1. **Feature selection for anomaly recognition models.** In order to build an anomaly recognition model with higher recognition accuracy, it is necessary to select and construct attributes that have stronger correlation with anomaly data as feature vectors. Previous research on fee evasion behavior lacks analysis of vehicle driving conditions, resulting in low audit accuracy. Based on previous research on fee evasion auditing, this chapter establishes vehicle traffic section data in conjunction with ETC toll data. Due to the fact that the most obvious feature of fee evasion behavior in driving trajectories is the abnormal appearance of traffic sections, but the formation of abnormal sections involves multiple factors, it is necessary to construct a new feature to measure the degree of trajectory abnormalities. Due to the direct correlation between travel distance and toll, distance is used as a measure of anomaly, and the section anomaly rate is constructed as shown in Formula (5).

$$Section\_ErrorRate(traj) = \begin{cases} Dis\_1/(Dis\_1 + Dis\_2) & \text{if } no\ disconnect\ sections \\ 1 & \text{if } exist\ disconnected\ sections \end{cases} \tag{5}$$

Among them, Dis_1 is the total distance of the abnormal section, Dis_2 is the total distance of the normal section, and the range of section abnormality rate is 0-1.

This article judges abnormal sections based on the speed within the section, and the judgment process is shown in Figure 10. However, the misdetection section after data cleaning belongs to the disconnected section, which has a great possibility of evading fees. Therefore, the section abnormality rate of this data is directly set to 1. By comparing the speed of a section with the average speed of that section in historical data, identify whether there are abnormalities in the section, and to some extent eliminate the impact of traffic congestion and entering the service area on travel time.

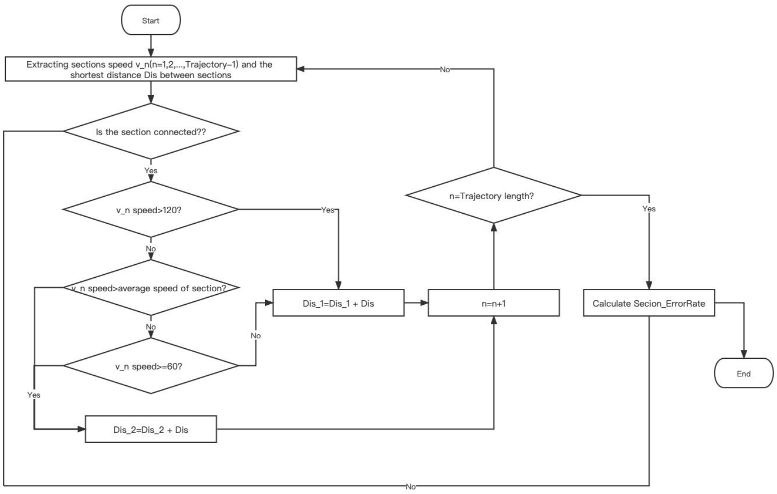

Start
Extracting sections speed v_n(n=1,2,...,Trajectory−1) and the shortest distance Dis between sections
No
Is the section connected??
Yes
n=Trajectory length?
v_n speed>120?
Yes
Yes
No
v_n speed>average speed of section?
Dis_1=Dis_1 + Dis
n=n+1
Calculate Secion_ErrorRate
No
No
v_n speed>=60?
End
Yes
Dis_2=Dis_2 + Dis
No

Figure 10. Flow chart for judging abnormal distance in a section

It can be seen that the segment anomaly rate of abnormal data is generally higher than 0.4 from Figure 11, while the average segment anomaly rate of the overall data is calculated to be only 0.17. It can be seen that when there are anomalies in the path, the value of the segment anomaly rate will be higher than the normal data. Therefore, the segment anomaly rate can to some extent represent the abnormal situation of the vehicle during driving.

In summary, this article extracts and constructs corresponding feature variables based on the commonality of anomalies in abnormal data analysis, and ultimately identifies eight feature variables for the model, as shown in Table 9.

3.4.2. **Analysis and validation of experimental results.** This paper extracts and constructs relevant evasion behavior features and use random forest algorithm to recognize vehicles with abnormal paths [20]. To verify the performance of the model prediction, the test set was introduced into the recognition model, and the recognition results are shown in Table 10. From the recognition results, it can be seen that the model has a

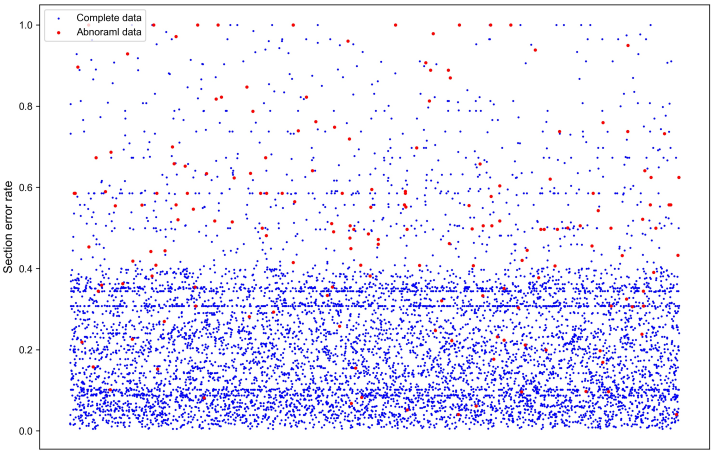


Figure 11. Distribution map of section anomaly rate

Table 9. Meaning and interpretation of characteristic variables

Feature Name	*Feature meaning*	*Feature Description and value range*
Car_Type	Vehicle type	{1=Bus,2=Truck,3=Special operation vehicle}
Is_Same_Plate	License plate consistency	{1=Consistent,0=Inconsistent}
Weight_Rate	Load rating	The actual load/vehicle type corresponds to the maximum load, with 0 for passenger cars
Ex_Hour	Exit hour	The time point of leaving the provincial highway, with a value range of [0,23]
Section_ErrorRate	Section anomaly rate	Abnormal section distance/driving distance
Travel_Time	Driving time	Vehicle driving time
Ave_Speed	Aveage speed(km/h)	Time driving distance/driving time
Wholeness	Entrance and exit integrity	{0=Incomplete,1=Complete or exit provincial boundary gantry}

recognition rate of 96.78% for normal vehicles and 83.38% for abnormal vehicles. Overall, the accuracy of the validation set data recognition reaches 90.17%.

Table 10. Prediction results of random forest algorithm

Sample	*Forecast*		*Total*	*Recognition accuracy%*
	Normal	*Abnormal*		
Normal	607884	20225	628109	96.78
Abnormal	103262	524848	628110	83.55
Total	711146	545073	1256219	**90.17**

In order to verify the feasibility of abnormal path vehicle recognition model and prevent overfitting phenomenon in model training, this paper first puts the data set from September 3 to 4, 2020 into the recognition model for training after sampling, and then puts the data set from September 5, 2020 into the recognition model for recognition. After identification, 36 abnormal data items were identified by the recognition model in the

"Two-Passengers and One-Risk" vehicle dataset on September 5, 2020, with a recognition rate of 91.67%.The recognition results are shown in Table 11.

Table 11. Identification results of data for "Two-Passengers and One-Risk"

Total amount of abnormal data	*Recognition anomaly total*	*Model recognition anomaly total*
36	33	**91.67%**

4. **Conclusions.** This article is based on the massive traffic data of the ETC system and the trajectory data of "Two-Passengers and One-Risk", fully utilizing the characteristics of the highway network, and proposing a trajectory point path recognition algorithm based on GeoHash encoding, which is conducive to efficiently identifying abnormal toll data. And combined with the random forest algorithm, construct a vehicle recognition model based on abnormal features. After verification, the model recognition effect is good and has certain practical significance. However, the algorithm still has shortcomings. The recognition effect of abnormal paths for long-distance vehicles is not accurate enough, and further research is needed in the future.

Acknowledgment. This work is supported by Special project for central guidance of local scientific and technological development: Fujian Key Laboratory of Automotive Electronics and Electric Drive Technology (project No.: 2020L3014, fund No.: KY030167), 2020 Fujian "the Belt and Road" Science and Technology Innovation Platform Project: Beidou Open Laboratory Southeast Asia International Branch (Project No.: 2020D002); Provincial Candidate for Fujian Hundred Thousand Talents Project (Project No.: GY-Z19113) and National Natural Science Foundation of China: Research on Location Information Mining Information and Recommended Methods for Big Data (Project No.: 41971340).

REFERENCES

[1] Ministry of Transport of the People's Republic of China, "Statisitcal Summary of National Toll Roads in 2021," 2022.

[2] H. Wang, "From labor to big data+AI: "One Web" audit is accelerating," *China ITS Journal*, vol. 11, pp. 18–22, 2021.

[3] K. Fukazawa, K. Naito, "Efficiently Managed Electronic Toll Collection System," *Professional Translation,* vol. 6, pp. 54–57, 2002.

[4] D. Burgess, "Caught in the act," *Traffic Technology lnternational,* vol. 2, pp. 37–39, 2008.

[5] C. Li and L. Wu, "Analysis of typical abnormal data behavior on highways," *China ITS Journal,* vol. 7, pp. 94–97, 2016.

[6] H. Chen, "Research on Expressway Green Traffic Inspection Based on Logistic Regression," Chang'an University, 2017.

[7] L. Chu, X. Guo and G. Song, "Identification and analysis of highway toll evasion based on clustering method," *China Intelligent Transportation Annual Conference,* 2013.

[8] F. Zhang, Y. Wang and W. Chen, "Inspection and Analysis of Green Channel Vehicles for Fresh Agricultural Products on Freeways Based on SMOTE-Logistics," *Journal of Highway and Transportation Research and Development,* vol. 15, no. 10, pp. 303–307, 2019.

[9] Z. Hao, "Application of Data Mining Technology in Expressway Toll Collection System," Chang'an University, 2016.

[10] S. Li, Z. Zhou, Y. Li, Y. Wang, X. Song and P. Wang, "Prediction of highway toll evasion based on IGA-IBP algorithm," *Computer Engineering and Design,* vol. 39, no. 12, pp. 3840–3845, 2018.

[11] Q. Ma, Y. Xu, W. Ding and Z. Jie, "A Real Time Detection Method for Expressway Evasion Vehicle Identification," *Computer Technology and Development,* vol. 29, no. 07, pp. 184–189, 2019.

[12] Z. Zhou, “Research on Abnormal Data Detection Methods for Freeways,” Changchun University of Science and Technology, 2018.

[13] Z. Zhang, “Research and application of highway road condition prediction based on vehicle traffic big data,” Donghua University, 2018.

[14] H. Zhang, D. Fu and Z. Wang, “Discussion on the Interference of Adjacent Lane in ETC Toll Lane,” *Hunan Transportation Technology,* vol. 47, no. 02, pp. 41–44+62, 2021.

[15] S. Zou, “Analysis and Application of Abnormal Behavior in Expressway Toll Collection Data,” University of Electronic Science and technology, 2020.

[16] L. Zhou, K. Long and J. Gan, “Intelligent operation and maintenance solutions for highways and their applications,” *China ITS Journal,* vol. 2, pp. 136–138, 2018.

[17] W. Zhou, “Problems and Countermeasures for the Safety Management of Tourist Charter Cars,” *Automobile and Safety,* vol. 12, pp. 61–65, 2021.

[18] Z. Liu, X. Wang and X. Wang, “Map-Matching Algorithm for GPS Trajectory in Complex Urban Road Networks,” *Journal of University of Electronic Science and Technology of China,* vol. 45, no. 06, pp. 1008–1013, 2016.

[19] S. Zhao, W. Tian and Q. Yu, “Research on Ship Entry Detection and Warning Method Based on GeoHash,” *Computer Programming Skills and Maintenance*, vol. 4, pp. 151–153, 2020.

[20] H. Xiang, P. Yang and J. Yi, “State Prediction Model of Expressway Escaping Vehicles Based on RF-LR,” *Journal of Chongqing Normal University(Natural Science),* vol. 37, no. 01, pp. 75–78, 2020.