Traffic & Transportation

Loading and Unloading Points Identification Based on Freight Trajectory Big Data and Clustering Method

Si-Yuan SUN[1], Rong-Hui BI[2], Zong-Yao WANG[3], Yu JI[4]

Original Scientific Paper
Submitted: 12 Oct. 2022
Accepted: 7 Feb. 2023

[1] SIYUANsemail@163.com, Collaborative Innovation Center for Transport Studies, Dalian Maritime University
[2] 14251275@bjtu.edu.cn, Collaborative Innovation Center for Transport Studies, Dalian Maritime University
[3] wzy@dlmu.edu.cn, Collaborative Innovation Center for Transport Studies, Dalian Maritime University
[4] 592347943@qq.com, Collaborative Innovation Center for Transport Studies, Dalian Maritime University

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT

Based on the GPS trajectory data of a freight enterprise in Dalian, China, this paper studies the identification of loading and unloading points by a clustering algorithm. Firstly, by analysing the characteristics of freight loading and unloading behaviour, combined with the spatial and temporal distribution characteristics of truck GPS trajectory data, three characteristic variables of the number of trucks passing through a certain place, the average speed of trucks and the average stay time of trucks in the place are extracted. Then, the clustering algorithm and visual analysis are used to obtain the target cluster, and the POI language of the geographic information is obtained according to the points in the target cluster. The meaning information is crawled to accurately identify the result of the freight loading point. Finally, two classical clustering algorithms, K-means and GMM, are evaluated and compared. The results show that the identification method designed in this paper finally identifies 2,320 freight loading and unloading points from 11,406,000 trajectory data, which can realise the accurate extraction of freight loading and unloading points.

KEYWORDS

loading and unloading points identification; cluster analysis; GPS truck tracking; K-means algorithm; GMM algorithm; data mining.

1. INTRODUCTION

With the development of technology and the intensification of market competition, traditional freight enterprises begin to provide more scientific and reasonable theoretical guidance for enterprise decision-making by mining and analysing various data, such as freight track data, and are gradually transformed in the direction of "informatisation".

Loading and unloading points identification is an important information basis for freight route planning, logistics facility location and other research. The identification results of loading and unloading points can provide data support for traditional freight enterprises to divide customer requirements and determine key cargo distribution areas. It can help enterprises grasp the demand of freight markets, promote the digital development of the logistics industry, assist urban planning decision-makers in logistics service point planning and contribute to future freight management and planning.

In recent years, many scholars have also carried out in-depth research on freight track big data mining. Among them, as one of the research highlights in this field, the stopping points stay recognition of freight cars has received extensive attention [1–5]. However, because there are many kinds of truck stops, and the stop data generated due to truck failures, drivers' physiological needs, traffic congestion and other reasons are irrelevant to freight operations, it is inaccurate to simply identify the stop.

The key problem in identifying loading and unloading points is distinguishing ordinary stopping points and loading and unloading points. Ordinary stopping spots usually include those caused by truck drivers that eat,

rest and refuel, as well as by traffic congestion. The truck parking time of an ordinary parking point is long, and the extraction of the loading and unloading point is strongly interfered with, which needs to be excluded.

This paper takes the trajectory data collected by truck GPS equipment as the research object, analyses the universal characteristics of freight loading and unloading behaviour and solves the practical engineering problems of space stay, time consumption, and loading and unloading demand in the process of truck information logistics. Through the establishment of the loading and unloading point identification process, the extraction of characteristic data such as spatial aggregation, stay time and vehicle speed when the vehicle is loading and unloading at the loading and unloading point. The clustering results are visualised, the truck behaviour is described and compared, and the problem of screening the target cluster of the loading and unloading point is solved.

Additionally, combined with POI (point of interest) semantic information, the accuracy of loading and unloading point identification is improved. The clustering quality of the K-means algorithm and GMM algorithm is compared, the performance of the two algorithms in processing trajectory data is compared, and the application effect of the K-means algorithm in trajectory data is obtained.

The identification of loading and unloading points is of great significance to the long-distance freight industry, and it can be used to segment the journey and obtain information such as popular routes and traffic hotspots. It can help traditional freight companies reduce the sales transmission chain, maintain and expand customer relationships, improve the income of freight companies and provide theoretical support for traffic management departments to further optimise the freight system.

The paper is organised as follows. Section 2 gives a brief review of the related works. Section 3 introduces the data source and feature selection used in this study. Sections 4 and 5 are the key parts that present the methodology and testing in this study and show the results. Finally, Section 6 concludes the study with contributions, limitations and future work.

2. LITERATURE REVIEW

A large number of studies have shown that it is feasible to study freight characteristics based on trajectory data, but in different freight tasks, time and space ranges, the specific characteristics of freight need to be determined according to the specific objects identified [6–8]. Therefore, the key to distinguishing is to grasp the key characteristics of freight behaviour.

In the analysis of the key characteristics of freight-related behaviour, the most commonly used indicators of freight trajectory characteristics are truck stay time, average speed, distance between road networks and heading variation. Yang et al. determined the loading and unloading points of trucks in New York City using three characteristic variables. Although the accuracy has been improved, there are regional limitations, the application scenarios are limited to the road sections studied in the literature, and they do not have universality [9]. Du et al. used a combination of truck dwell time, distance to the road network and heading variation to identify the endpoint of the truck's journey, reflecting information about the nature of the parking incident [10]. Under the condition of big data, the course change may be affected by the road topology, resulting in incorrect classification [8]. Stephane et al. set the dwell time interval to distinguish the parking types of freight trucks. The study found that the dwell time interval of freight trucks is affected by many factors, and other methods are still needed to further distinguish loading and unloading points and other stop points [11]. Therefore, the selection of feature variables needs to be considered comprehensively from the aspects of improving the recognition accuracy and the generalisation ability of the algorithm to mine the typical characteristics of the loading and unloading points.

At present, there are three methods to realise the identification of stay points. The first method is to identify auxiliary information such as truck drivers and land use information through field investigation [12, 13]. The second method determines the stay point by setting the stay time threshold [7, 8, 14]. The third is through the machine learning algorithm, combined with the behaviour characteristics of the research target to achieve [15, 16]. Referring to the existing stay point identification method, loading and unloading point identification can be realised by analysing the characteristics of freight loading and unloading behaviour, combined with the information mining technology and geographic information technology of trajectory big data [17–21].

Clustering analysis is one of the core methods of machine learning and the key technology for mining internal information on big data. In related research based on trajectory big data, the application of clustering algorithms is increasingly in-depth, and the K-means is one of the most widely used algorithms [22]. Yan Xuedong et al. used the improved K-means algorithm to divide and obtain 19 major functional areas of the Beijing metropolitan area based on tailwind car data. Based on crowd motion trajectory data [23], You Feng et al. extracted features by analysing the spatial and temporal distribution characteristics of the trajectory data and used the K-means algorithm to analyse the trajectory source points and vanishing points [24]. This paper finds that the K-means algorithm is simple and easy to implement and belongs to the 'hard' clustering algorithm [25].

Under specific constraints, the K-means algorithm can be regarded as a special form of the GMM (Gaussian mixture model, GMM). However, the GMM algorithm is widely used in pattern recognition and machine learning and is less used in the field of transportation, especially for related research on key point recognition. The affiliation of data points in clustering is not only related to the neighbour points but also depends on the shape of the cluster. Compared with the K-means of 'hard clustering', GMM is a 'soft clustering' statistical model based on probabilistic clustering, which is not affected by the shape of the cluster. GMM can obtain the probability that each sample belongs to each category and has a stronger description ability [26].

In summary, based on big data, to realise the identification of loading and unloading points, it is necessary to grasp the typical characteristics of freight behaviour and apply appropriate identification methods. Compared to previous similar studies in the field of the stopping points recognition of freight cars, this paper has the following innovation: combined with the typical characteristics of freight behaviour, a freight loading and unloading point identification method is proposed, which is not limited by data sources and geographical regions, and the accurate extraction of loading and unloading points is realised. The K-means algorithm and GMM algorithm are compared in clustering quality, and the performance of the two algorithms in processing track data is compared.

3. DATA AND THEIR CHARACTERISTICS

3.1 Data description

The data in this paper are from a freight enterprise in Dalian, China, and the trajectory data generated by the truck-mounted GPS device from 28 August to 31 August 2020 are extracted nationwide. In total, there are 11,406,388 pieces of data in the format shown in *Table 1*.

GPS trajectory data contain information such as the license plate number, latitude and longitude coordinates, recorded time and instant speed of the moving vehicle. The latitude and longitude coordinates are accurate to 4 decimal places.

In this paper, Python web crawler technology is used to extract POI semantic information of longitude and latitude coordinate points collected by truck GPS equipment with the help of AMAP API, including location attributes, area, detailed address and so on, as shown in *Table 2*.

3.2 Selection of loading dock features

After preprocessing the truck trajectory data, a total of 257,652 GPS points were obtained, and the time was recorded in 10 minutes as the counting unit. This paper uses three characteristic variables: the number of trucks passing through GPS point *j*, the average speed of trucks and the average stay time of trucks at GPS point *j*.

Table 1 – Original data presentation

Number	License plate number	Time	Longitude	Latitude	Speed [km·h^{-1}]
1	Liao B****5	2020-08-28 7:40:57	121.8176	121.8176	0
2	Liao B****5	2020-08-28 7:40:54	121.8434	121.8434	0
3	Liao B****5	2020-08-28 7:40:37	111.9405	111.9405	0
...	...	...	...	...	...

Table 2 – Clusters of freight-related POIs

Base category	Enterprises	Daily life service	Transportation service
Subcategory	(1) advertisement and decoration (2) construction company (3) medical company (4) machinery and electronics (5) chemical and metallurgy (6) network science and technology (7) commercial trade (8) telecommunication company (9) mining company (10) factory (11) farming, forestry, animal, and fishery base (12) comprehensive market (13) home building materials market (14) stationary store	(1) logistics service (2) logistics warehouse space (3) gas station (4) service centre	(1) airport related (2) port & marina (3) railway station (4) border crossing

The number of trucks passing through the place X_j:

$$X_j = \sum (x_{j,t})^0 \tag{1}$$

where $x_{j,t}$ represents the vehicle appearing at time t in j.

The average speed of trucks through the place $\overline{V}_j$, km/h:

$$\overline{V} = \frac{1}{X_j} \sum_{t=1}^{X_j} v_{j,t} \tag{2}$$

where $v_{j,t}$ represents the speed value collected at time t in j.

Average stay time of trucks at the place $\overline{T}_j$, min:

$$\overline{T_j} = \frac{10 \sum_{r=1}^{X_j} m_r}{X_j} \tag{3}$$

where m_r represents the frequency of the r-th car in j.

4. IDENTIFICATION OF MAJOR LOADING AND UNLOADING POINTS

4.1 Clustering algorithm

The principle of the K-means algorithm is to randomly select points as the initial clustering centre, calculate the Euclidean distance between each data point and the initial clustering centre, and use it as the similarity evaluation standard to assign data points to the cluster represented by the clustering centre with the largest similarity. Finally, according to the similarity between the dataset and the clustering centre, the position of the clustering centre is constantly updated until the clustering centre does not change.

As a statistical model based on probability clustering, the GMM algorithm is determined by two parameters: the mean vector and the covariance matrix. The model combines the advantages of nonparametric and parameter estimation methods, which can well express the distribution characteristics of objects in the parameter space and obtain more flexible allocation results.

4.2 Selection of cluster number value

In the application research of the K-means and GMM algorithms, it is necessary to assign the number of clusters. The commonly used algorithms to determine the optimal number of clusters include the elbow method and silhouette coefficient method. The complexity of the silhouette coefficient method is greater, while that of the elbow method is relatively simple. The research object of this paper is the vehicle parking point. The number of feature types is small, and the required clustering value is not large. The elbow method is more effective in estimating the number of clusters.

The SSE (sum of squared error) from the sample point to the centroid of the cluster is used as the core index to measure. The smaller the value is, the more convergent the clusters.

First, a possible maximum number of clusters is randomly specified. Then the number of clusters is increased from 1, and SSE is calculated. In the process of setting the number of clusters to approach the real number of clusters, the initial SSE shows a rapid downward trend. When the number of clusters increases to a reasonable value, the decreasing range of SSE decreases and tend to be gentle.

By drawing the k-SSE curve, the inflection point on the way down is found, that is, the determined value. The specific calculation method is as follows:

$$SSE = \sum_{f=}^{k} \sum_{q \in Df} \left| q - u_f \right| \tag{4}$$

where D_f is the f cluster, q is the sample point in the D_f cluster, and u_f is the centroid of the D_f cluster.

Through the analysis of the characteristics of the driving behaviour of the truck, it can be found that the driving state of the truck can be divided into six categories according to the related activities of the truck: one is the truck driving at high speed in the highway section, the second is the truck driving normally in the urban road section, the third is the parking point caused by the truck parking loading and unloading, the fourth is the truck driver eating and resting on the way, as well as truck refuelling, the fifth is the parking point caused by traffic congestion, the sixth is the parking point caused by the traffic charge, the traffic light and so on.

Through the analysis of the characteristics of truck driving behaviour, combined with *Figure 1* of elbow method results, it is found that the slope of the image changes greatly before and after k=7, and the SSE is much smaller than k=2 and k=3. Therefore, this paper selects the k value of k=7.

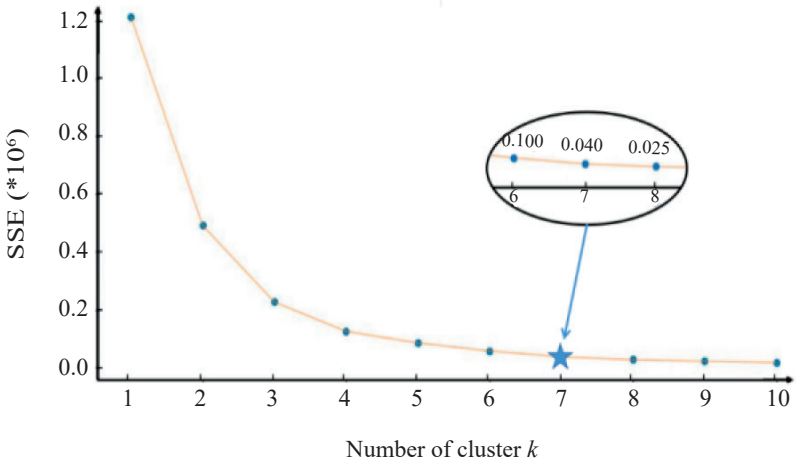


Figure 1 – The elbow method determines the k value

4.3 Primary loading and unloading points identification process

In this paper, K-means and GMM algorithms are used to study the trajectory data, compare the quality of clustering results and obtain information on loading and unloading points. The specific steps of identification are as follows:

Step 1: The original truck GPS trajectory data (latitude and longitude, time and other related data) are used for data preprocessing.

Step 2: Feature extraction and quantitative statistics of the data. The trajectory data are quantified and counted based on three characteristic variables (the number of trucks passing through the GPS point, the average speed, and the average stay time at the GPS point). The quantitative results are shown in *Table 3*.

Step 3: Reduce the data results of Step 2 to reduce the amount of data processing. The principal component analysis method is used to reduce the dimension of the data twice. The first is to reduce the dimension of the three high-dimensional feature data to two-dimensional data and merge them. The second is to reduce the dimension of the combined six-dimensional data to three-dimensional data. To eliminate the dimensional influence between the indicators, the data are normalised. The specific process of data reduction is shown in *Figure 2*.

Table 3 – Quantification of the results

Longitude	Latitude	Number of vehicles	Average vehicle speed [km·h^{-1}]	Average stay time [min]
115.563	37.344	1	0	10.6667
118.627	35.088	5	4.92	30
124.046	41.844	15	0.8267	10.1552
117.455	39.08	4	3.075	40
111.409	34.712	1	44.2	10
121.938	39.222	23	3.8913	20.8934
...	...	...	...	...

Step 4: Use cluster analysis for the three-dimensional data obtained by Step 3. The elbow method is used to estimate the number of clusters based on the K-means and GMM algorithms for data clustering, which are divided into k clusters.

Step 5: Select the target clusters. Each cluster is visualised and numerically counted according to the three characteristics of trucks. Through the comparative analysis between the clusters, the target clusters that meet the characteristics of the loading and unloading points are retained.

Step 6: Consider crawling POI semantic information. The retained target clusters are subjected to AMAP geographic inverse coding and XML parsing, and the POI semantic information data of the facility closest to the latitude and longitude data in the target cluster are extracted, such as specific location information and location attribute data.

Step 7: Remove unreasonable results and output. The POI semantic information obtained by Step 6 accurately identifies the results and improves the accuracy of loading and unloading points identification. Use sentence keyword filtering, excluding non-loading points, such as scenic spots, residential areas, financial services and other places.

Step 8: Obtain the result of loading and unloading points information and store it in the file.

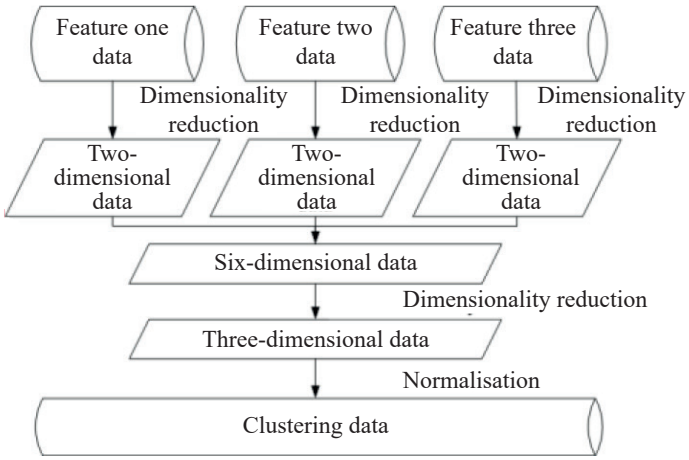


Figure 2 – Data Reduction

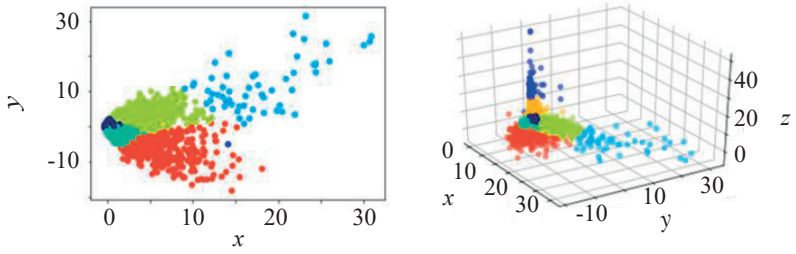
In summary, this method first performs feature extraction and quantitative statistics of the data according to the three feature variables of freight behaviour. Statistical results were reduced to reduce the amount of data. Then, the clustering algorithm is used for the clustering analysis. Finally, the cluster data are statistically analysed, and the loading and unloading point information results are obtained by combining the POI semantic information.

5. TEST RESULTS AND ANALYSIS

5.1 Clustering results and evaluations

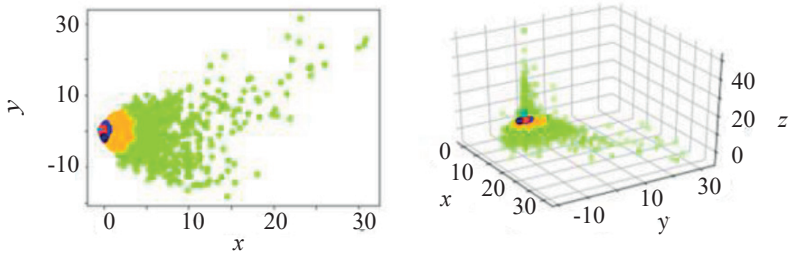
In this paper, K-means and GMM algorithms are used to analyse the data. In the case of unknown actual cluster information, the following four commonly used evaluation indicators are selected to evaluate the clustering results: SC (silhouette coefficient), DBI (Davies–Bouldin index, DBI), CH (Calinski-Harabaz Index, CH) index and running time [14, 27, 28].

The clustering results are visualised and drawn into a two-dimensional plan view and a three-dimensional plan view, as shown in *Figures 3 and 4*. *Figures 3 and 4* show that both clustering algorithms can divide the data into different categories. The clustering results of the GMM algorithm partially overlap in the figure, and due to the different clustering principles, the obtained morphological distribution of clusters varies widely.



a) Two-dimensional plan view *b) Three-dimensional plan view*

Figure 3 – K-means clustering results



a) Two-dimensional plan view *b) Three-dimensional plan view*

Figure 4 – GMM clustering results

Through the four evaluation indices, the quality of the clustering results of the two algorithms is compared and evaluated, the results of which are shown in *Table 4*. *Table 4* shows that the K-means is superior to the GMM in terms of the SC, CH and DBI.

Table 4 – Quantification of the results

Clustering algorithm	SC	CH	DBI	Running time [s]
K-means	0.55	108352.005076	0.50	14.60
GMM	-0.15	12295.155725	11.82	13.30

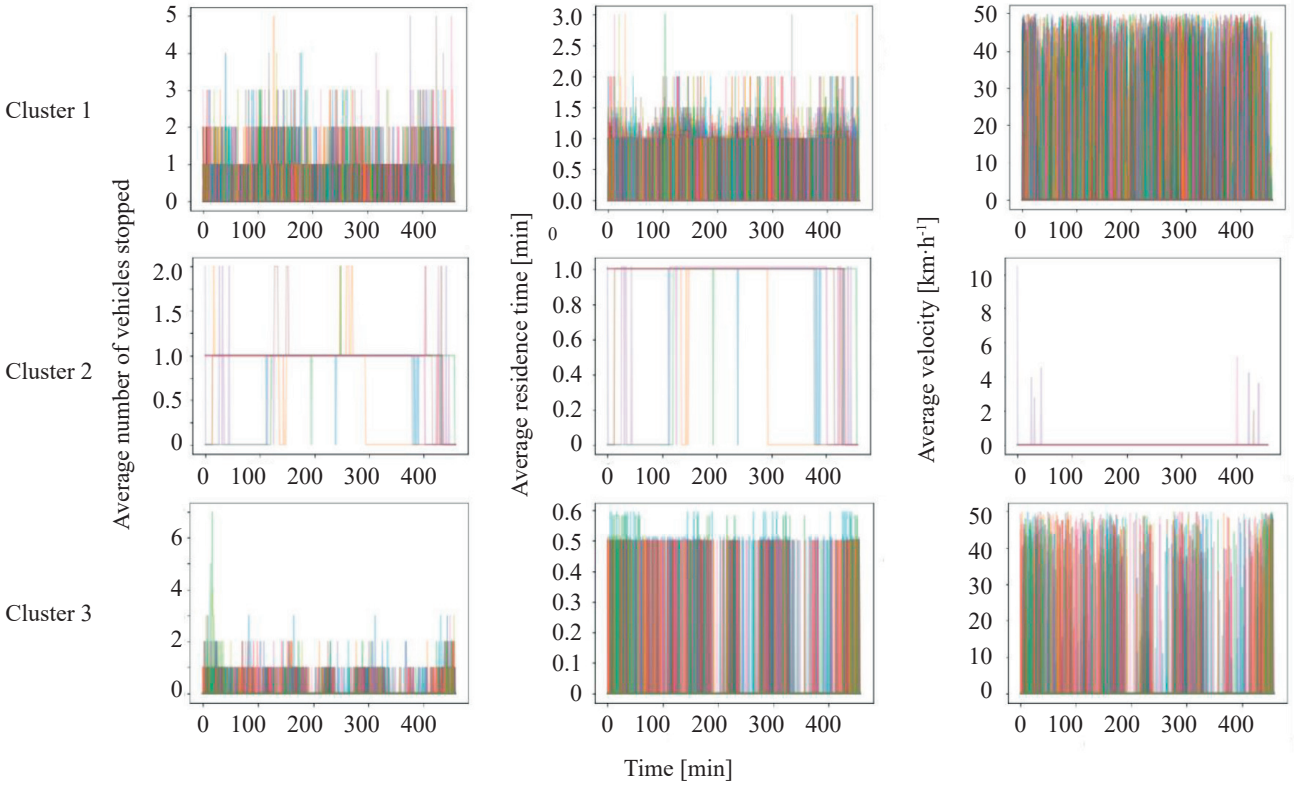
The SC is the average of the silhouette coefficient of all samples. The range of its value is [-1,1]. The closer it is to 1 means clustering of the samples is more reasonable. It is found that the SC of the K-means is larger than that of the GMM.

The larger the CH is, the closer the class itself is. The more dispersed the class is, the better clustering results can be obtained. For the CH index, the CH of the K-means is larger.

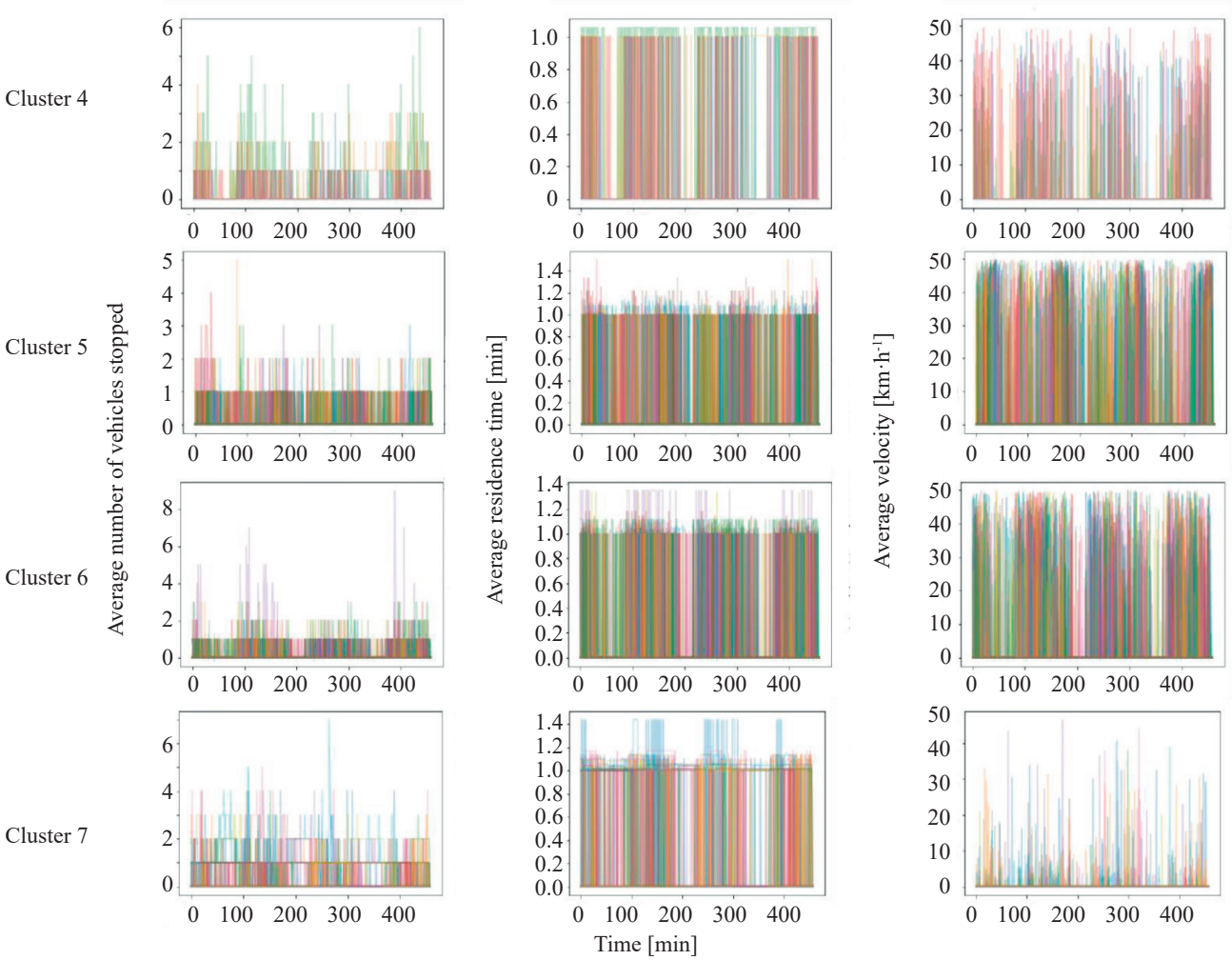
The DBI is the sum of the average distance of the intraclass distance of any two categories divided by the centre distance. The smaller the DBI is, the smaller the intraclass distance, the larger the interclass distance, and the better the clustering effect. The DBI of K-means is smaller.

Comparing the running time of the two clustering algorithms, it can be found that the GMM runs faster.

In general, in terms of the clustering effect, the K-means algorithm has a better clustering effect on truck trajectory data than the GMM. In terms of running speed, the GMM algorithm processes data faster.



a) Characteristic map of clusters 1, 2, 3



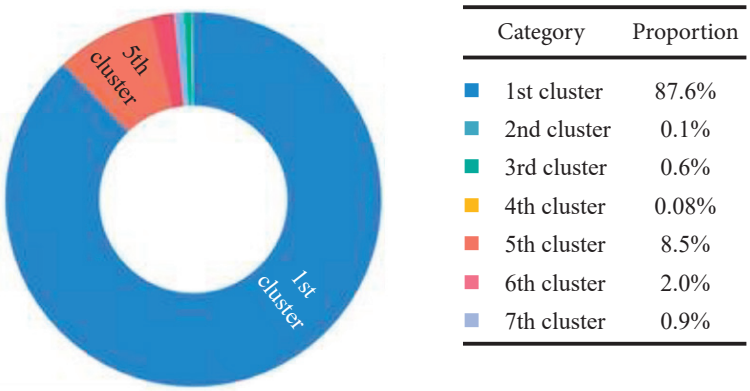
b) Characteristic map of clusters 4, 5, 6, 7

Figure 5 – Characteristic map of truck behaviour

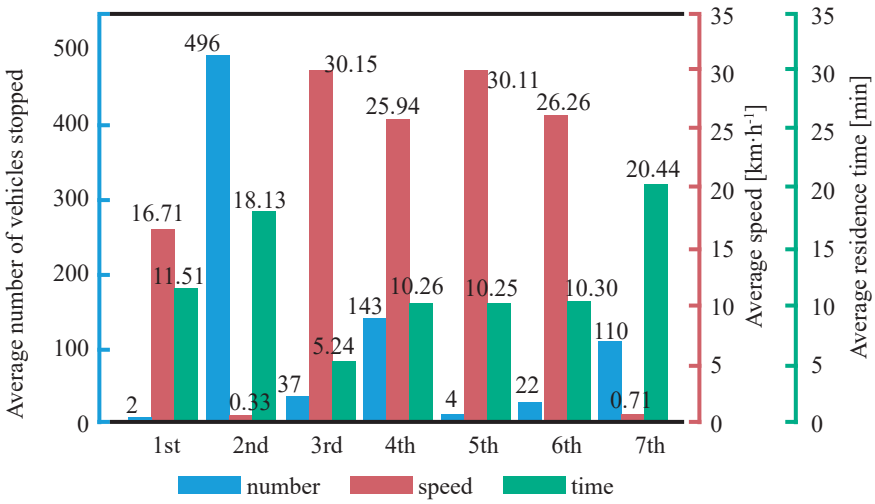
5.2 Depiction and comparative analysis of truck behaviour

Freight loading and unloading points are usually characterised by a high number of passing trucks and slow speeds, with long truck stay times. At the same time, according to the strength of the freight demand we can divide the loading and unloading points into popular loading and unloading points (such as industrial parks and logistics parks etc.) and general loading and unloading points (such as distribution centres and commercial outlets etc.).

To eliminate non-loading point clusters in the seven clusters, the data such as the average number of trucks passing through the GPS points, the average speed, and the average stay time are analysed as shown in *Figure 5*, where each vertical line represents the eigenvalue of a truck at a certain moment. The denser the vertical lines, the more points in the cluster. The height value of the vertical lines represents the characteristic value of the points. The average eigenvalues of each cluster are calculated, and the specific values are shown in *Figure 6*.



a) Percentage of points in the cluster



b) Numerical details

Figure 6 – Target class cluster details

The following can be obtained through *Figure 6*.

Cluster #1 has the highest number of points, which shows the lowest average number of trucks stopping, lower average speed and longer average stay time. After actual investigation, it is found that Cluster #1 is the result of actions such as drivers eating and resting on the way, trucks refuelling, or severe traffic congestion.

Cluster #2 shows that the number of trucks is 1, the average speed tends to be 0, and the average stay time is long at most points in time. After actual investigation, it is found that most of the points in Cluster #2 are urban freight nodes, such as distribution centres, small enterprises, factories and companies, and other sites with frequent freight demand but little single transport volume.

Clusters #3, #4, #5 and #6 have high average speeds and do not meet the characteristics of loading and unloading points and were excluded first.

Cluster #7 has more vehicles, lower average speed and the longest average stay time compared to other clusters, and after actual investigation, it was found that most of the points in cluster #7 belong to popular loading and unloading points such as industrial parks, professional markets, logistics parks and large enterprises.

Comprehensive comparative analysis of image and data statistics excluded clusters #1, #3, #4, #5 and #6 that did not meet the characteristics of loading and unloading points and retained clusters #2 and #7.

5.3 Loading and unloading points identification results

After clustering analysis, the retained Clusters #2 and #7 contain a large amount of geographical location data, which includes loading and unloading points information and some non-loading points information interference.

According to the POI information in the urban land use attributes and road facilities information in the statement, keywords were removed, excluding all non-loading points, such as scenic spots, residential areas, financial services and other places. Finally, the geographic information of the 2,320 loading and unloading points is shown in *Table 5*.

Table 5 – Loading and unloading points extraction results

Location attribute	Specific location	Longitude	Latitude
commercial residence; industrial park	Shanghai Malu Industrial City, Malu Town, Jiading District, Shanghai	121.33	31.351
company; enterprise	Shanghai Automobile Transmission Co., Ltd. (Joint Factory), Zhongjing 1st Road, Xuhang Town, Jiading District, Shanghai	121.186	31.392
transportation facilities services; railway stations; freight railway stations	Luoyang Town Station, Yunjing Road, Luoyang Street, Chenggong District, Kunming City, Yunnan Province	102.839	24.936
...	...	...	...

The distribution of the obtained loading and unloading points is visualised by the ArcGIS software, as shown in *Figure 7*. These points include large logistics parks, industrial parks, professional markets, airports, stations, ports etc.

Figure 7 – Distribution of loading and unloading points

Through truck behaviour characterisation and qualitative and quantitative analysis, the 2,320 loading and unloading points belong to two categories: general loading and unloading points and popular loading and unloading points, in line with the actual situation of freight loading and unloading point clusters.

6. CONCLUSION

Identifying the loading and unloading points can reflect the actual freight demand more accurately. The identification results can not only reflect the spatial distribution and hot areas of freight demand but also help enterprises determine the location of logistics centres by analysing the spatial distribution characteristics of loading and unloading points.

This paper quantifies the characteristics of freight behaviour, combines dimension reduction technology and a clustering algorithm, extracts the target cluster of loading and unloading points from freight trajectory data, obtains POI semantic information through geographic information technology, accurately identifies the results and identifies loading and unloading points information.

1) The research shows a method to identify loading and unloading points that are not limited by data sources and geographical regions. A total of 2,320 freight loading and unloading points are effectively identified from the trajectory data of 11,406,388, which realises accurate data extraction.
2) The research shows that there are differences between loading and unloading points and non-loading points. Three characteristic variables related to loading and unloading behaviour are analysed: the number of trucks passing through a place, the average speed, and the average stay time at the place. These features are universal, do not require specific road network support and are not affected by road topology.
3) K-means is superior to GMM in terms of the SC, DBI and CH. The running time of GMM is shorter. In improving the accuracy of recognition, K-means has a better effect than GMM.

The limitations of this study are as follows.

1) According to the method of this paper, the workload required for the later verification of the loading and unloading points is large. Therefore, in future research, the neural network should be used to learn the characteristics of the loading and unloading points to realise the automatic identification and extraction of the loading and unloading points.
2) All the loading and unloading points identified in this paper are not screened according to factors such as traffic volume and freight rate. In future research, factors such as traffic volume and freight rate can be added to classify the identified loading and unloading points. The classification of loading and unloading points can provide data support for the establishment of logistics centres.

ACKNOWLEDGEMENT

This study was supported by the Green Port and Shipping Network Operation Management Optimization Research Fund (71831002).

REFERENCES

[1] Liu S, Chen G, Wei L, Li G. A novel compression approach for truck GPS trajectory data. *IET Intelligent Transport Systems*. 2021;15:74–83. DOI: 10.1049/itr2.12005.

[2] Duan M, Qi G, Guan W, Guo R. Comprehending and analyzing multiday trip-chaining patterns of freight vehicles using a multiscale method with prolonged trajectory data. *Journal of Transportation Engineering, Part A: Systems*. 2020;146:04020070. DOI: 10.1061/JTEPBS.0000392.

[3] Xiao ZP, Zou HX, Sun YH. Using GPS data to visualize the intra-city freight mobility—the case of Shenzhen. *Journal of Human Settlements in West China*. 2017;32:9–15. DOI: 10.13791/j.cnki.hsfwest.20170102.

[4] Csendes B, Albert G, Szander N, Munkácsy A. Where truck drivers stop–application of vehicle tracking data for the identification of rest locations and driving patterns. *Promet – Traffic&Transportation*. 2021;33:821–32. DOI: 10.7307/ptt. v33i6.3962.

[5] Gan M, Nie YM, Liu X, Zhu D. Whereabouts of truckers: An empirical study of predictability. *Transportation Research Part C: Emerging Technologies*. 2019;104:184–95. DOI: 10.1016/j.trc.2019.04.020.

[6] Gingerich K. *Studying regional and cross border freight movement activities with truck GPS big data*. PhD Thesis. University of Windsor (Canada); 2017.

[7] Gingerich K, Maoh H, Anderson W. Classifying the purpose of stopped truck events: An application of entropy to GPS data. *Transportation Research Part C: Emerging Technologies*. 2016;64:17–27. DOI: 10.1016/j.trc.2016.01.002.

[8] Thakur A, et al. Development of algorithms to convert large streams of truck GPS data into truck trips. *Transportation Research Record*. 2015;2529:66–73. DOI: 10.3141/2529-07.

[9] Yang X, Sun Z, Ban X J, Holguín-Veras J. Urban freight delivery stop identification with GPS data. *Transportation Research Record*. 2014;2411:55–61. DOI: 10.3141/2411-07.

[10] Du J, Aultman-Hall L. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transportation Research Part A: Policy and Practice*. 2007;41:220–32. DOI: 10.1016/j.tra.2006.05.001.

[11] Hess S, Quddus M, Rieser-Schüssler N, Daly A. Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review*. 2015;77:29–44. DOI: 10.1016/j.tre.2015.01.010.

[12] Bernardin Jr VL, Steven T, Jeffery S. Expanding truck GPS-based passive origin-destination data in Iowa and Tennessee. *TRB 94th Annual Meeting Compendium of Papers*. 2015.

[13] Bassok A, McCormack ED, Outwater ML, Ta C. Use of truck GPS data for freight forecasting. *TRB 90th Annual Meeting Compendium of Papers*. 2011.

[14] Yang Y, et al. Identifying intracity freight trip ends from heavy truck GPS trajectories. *Transportation Research Part C: Emerging Technologies*. 2022;136:103564. DOI: 10.1016/j.trc.2022.103564.

[15] Cheng Z, Wang W, Lu J, Xing X. Classifying the traffic state of urban expressways: A machine-learning approach. *Transportation Research Part A: Policy and Practice*. 2020;137:411–28. DOI: 10.1016/j.tra.2018.10.035.

[16] Yuan Y, et al. Traffic state classification and prediction based on trajectory data. *Journal of Intelligent Transportation Systems*. 2021:1–15. DOI: 10.1080/15472450.2021.1955210.

[17] Gan M, Qing S-D, Liu X-B, Li D-D. Review on application of truck trajectory data in highway freight system. *Journal of Transportation Systems Engineering and Information Technology*. 2021;21:91. DOI: 10.16097/j.cnki.1009-6744.2021.05.009.

[18] Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*. 2009;17:285–97. DOI: 10.1016/j.trc.2008.11.004.

[19] Zheng Y. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2015;6:1–41. DOI: 10.1145/2743025.

[20] Feng Z, Zhu Y. A survey on trajectory data mining: Techniques and applications. *IEEE Access*. 2016;4:2056–67. DOI: 10.1109/ACCESS.2016.2553681.

[21] Portugal I, Alencar P, Cowan D. A framework for spatial-temporal trajectory cluster analysis based on dynamic relationships. *IEEE Access*. 2020;8:169775–93. DOI: 10.1109/ACCESS.2020.3023376.

[22] MacQueen J. Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*. 1967. p. 281–97.

[23] Yan XD, et al. Regional division and hierarchical structure of metropolitan area based on carpooling data and clustering method. *Journal of Transportation Systems Engineering and Information Technology*. 2021;21:30. DOI: 10.16097/j.cnki.1009-6744.2021.04.004.

[24] You F, et al. The trajectory of densely tracked the trajectory of the multi-target tracking and the semantic perception of sports. *Transportation System Engineering and Information*. 2021;21:42. DOI: 10.16097/j.cnki.1009-6744.2021.06.006.

[25] Liu T, et al. Study on driving style clustering based on K-means and Gaussian mixture model. *China Safety Science Journal*. 2019;29:40. DOI: 10.16265/j.cnki.issn1003-3033.2019.12.007.

[26] Chen Y, et al. Gaussian mixture clustering algorithm combining elbow method and expectation-maximization for power system customer segmentation. *Journal of Computer Applications*. 2020;40:3217. DOI: 10.11772/j.issn.1001-9081.2020050672.

[27] Zhang XH, et al. Research on the evaluation index of duty cycle-based clustering algorithm. *Computer Engineering and Applications*. 2022;58:175–81. DOI: 10.3778/j.issn.1002-8331.2007-0298.

[28] Jin Y, et al. Intelligent on-demand design of phononic metamaterials. *Nanophotonics*. 2022;11(3):439-460. DOI: 10.1515/nanoph-2021-0639.

孙思远，毕容珲， 王宗尧，季禹

基于货运轨迹大数据和聚类方法的装卸点识别研究

摘要

利用中国大连某货运企业货车的GPS轨迹数据，研究基于聚类算法的装卸点识别。首先，通过分析货物装卸行为特征，结合货车GPS轨迹数据的时空分布特征，提取经过某地的货车数量、平均速度、在该地的平均停留时间3个特征变量； 然后，利用聚类算法和可视化分析得到聚类结果，获取不同聚类类簇中各点的具体位置； 利用高德地图API接口对POI语义信息信息进行抓取，以准确识别货物装货点的结果。最后，对两种聚类算法K-means和GMM进行了评估。结果表明，本文设计的识别方法最终从1 140.6万条轨迹数据中识别出2320个货物装卸点，可实现货物装卸点的准确提取。

关键词:

卸点识别; 聚类分析; 货车GPS轨迹; K-means; GMM; 数据挖掘