STM JOURNALS

International Journal of
Algorithms Design and Analysis Review

Volume 1, Issue 2, 2023
July–December
DOI (Journal): 10.37591/IJADAR

https://journals.stmjournals.com/ijadar

Review IJADAR

An Approach for Travel Pattern Analysis Using HDBSCAN and Apriori Algorithms

P.V.S. Anil Kumar[1], Anuradha Purohit[2,*]

Abstract

Most mega-city regions around the world are suffering from an ongoing increase in the number of commuting trips. Understanding commuting patterns is crucial for both public and authority planners. The understanding of travel patterns helps passengers to know about the places and the time where they could get vacant transport and also helps authority planners in laying out a new transport service. The traditional way of understanding travel patterns includes conducting surveys, which is time consuming process. Therefore, authority planners are exploring another way of acquiring knowledge about the travel patterns within less time and effort. Using the concept of clustering and sequential pattern mining helps in understanding the travel patterns by generating clusters and frequent patterns. In this paper, a new approach for trajectory pattern analysis using hierarchical density–based clustering algorithm and Apriori sequential pattern mining is proposed. The use of a hierarchical density–based clustering (HDBSCAN) algorithm instead of traditional density-based clustering (DBSCAN) algorithm gave better results in terms of quality of clusters. Performing clustering on an hourly basis helped in getting deeper insights of the travel patterns. We utilized the Porto taxi trajectory dataset from the UCI Machine Learning Repository for our experimentation.

Keywords: hierarchical density–based clustering (HDBSCAN), density-based clustering (DBSCAN), Apriori, pattern mining, global positioning system (GPS)

INTRODUCTION

Vehicles can now convey their positions and conditions thanks to contemporary technologies like the global positioning system (GPS), global system for mobile communications (GSM), and Wi-Fi. This new method of communication allows collecting valuable spatiotemporal information like the time-series of locations [1]. These data hold valuable information about the routes and movements of a fleet of vehicles and humans. GPS devices can be used to detect the movements of people [2]. They can then be mined for important data like the mode of transportation and significant sites. In addition to GPS, radio frequency identification (RFID) has been extensively utilized in numerous industrial applications. It is an RFID device that can be used to monitor and control the operations of a warehouse [3]. There are several applications for this information. Some of these include planning and traffic management, security, animal protection, and marketing.

Frequently, substantial volumes of collected data undergo analysis through the utilization of data mining techniques. They can assist city planners in locating abnormalities and identifying road networks [4]. Authorities can better understand what influences people's wellbeing by having this information at their disposal.

*Author for Correspondence
Anuradha Purohit
E-mail: anuradhapurohit78@gmail.com

[1]Student, Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore, Madhya Pradesh, India
[2]Professor, Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore, Madhya Pradesh, India

Received Date: July 20, 2023
Accepted Date: September 20, 2023
Published Date: October 25, 2023

Citation: P.V.S. Anil Kumar, Anuradha Purohit. An Approach for Travel Pattern Analysis Using HDBSCAN and Apriori Algorithms. International Journal of Algorithms Design and Analysis Review. 2023; 1(2): 1–9p.

Therefore, in situations where travel patterns have to be analyzed, systems have been designed using data mining techniques like DBSCAN, K-means clustering algorithms for understanding the passengers travel movements in a much easier way.

To gain deeper insights from trip data, there are a few restrictions that need be taken into consideration. These limitations include fine-tuning the parameters in algorithms like DBSCAN, K-means [5], and it has also been observed that clustering the trajectory movements has not been done at a higher granularity level, that is, hourly.

Given these limitations, this study aims to provide an approach for trajectory analysis using hierarchical density–based clustering (HDBSCAN) and Apriori pattern mining to generate clusters and frequent itemset from the travel data. Unlike previous studies, in order to get deeper insights clustering is performed on the travel data on hourly basis.

The paper is structured as follows: *Background study* provides information about HDBSCAN and the K-means method. *Literature review* focuses on trajectory clustering and sequential pattern mining. *Research methodology* discusses the proposed approach for trajectory pattern analysis. *Experiments and results* presents the experimental findings.

BACKGROUND STUDY

The background study of HDBSCAN algorithm and Apriori sequential pattern mining algorithm is discussed in this section.

Hierarchical Density Based Clustering Algorithm

The clustering algorithm known as HDBSCAN was created by Campello, Moulavi, and Sander in 2013 [6]. It is a density-based clustering technique that only creates clusters using the core points. This algorithm is easy to implement and is fast when compared to traditional density-based clustering algorithm (DBSCAN), because it takes as an input only one parameter, which is minpts, and performs clustering. The distance metric it uses for performing clustering is mutual reachability which is given as follows:

$d_{mreach-k}(a, b) = \max\{core_k(a), core_k(b), d(a, b)\}$

where,

$d(a, b)$ = Euclidian distance between points a and b.

core (a) = core distance of point a.

core (b) = core distance of point b.

Apriori Algorithm

The Apriori sequential pattern mining algorithm, introduced by Agrawal *et al*. in 1996 [7], was designed to identify frequent item sets within a dataset. The technique is known as Apriori since it depends on prior knowledge of frequently used itemset. Data mining frequently item sets is a commonly utilized method for uncovering patterns and performing exploratory data mining. It is a very effective strategy for mining frequently occurring item sets in databases. The Apriori algorithm follows a level-wise approach, where it leverages k-frequent item sets to identify k+1 item sets. In other words, it uses frequent 1 item sets to generate frequent 2 item sets, and so forth. This algorithm takes in two parameters as input, they are support and confidence. Where support represents the item's frequency of occurrence and confidence represents the conditional probability [8].

For itemset x and y support and confidence are given as:

$$\text{Support (x)} = \frac{Number\ of\ transactions\ in\ which\ x\ appears}{Total\ number\ of\ transactions}$$

$$\text{Confidence (x}\rightarrow\text{y)} = \frac{\text{Support(x U y)}}{\text{Support(x)}}$$

LITERATURE REVIEW

In this section various approaches proposed by the researchers for performing travel pattern analysis are presented.

Palma *et al.* [9] proposed an approach for finding major places on a trajectory data, based on its speed. They extracted the knowledge of travel behaviour from trajectory data by combining geographical characteristics with a semantic description for each trajectory stop. They utilized the density-based spatial clustering of applications with noise (DBSCAN) algorithm in their approach to identify stops with the shortest trip distances. Bermingham and Lee [10] performed spatiotemporal trajectory sequential pattern mining. To illustrate each point in a trajectory, they used images from Flickr taken by tourists in the Queensland region. They mined the dataset using a sequential pattern mining framework and discovered interesting trends in the east coast and Brisbane districts.

Bhaskar and Cheng *et al.* [11] proposed an approach to detect travel patterns of passengers. They used DBSCAN algorithm to segment passengers into four different classes, namely transit passengers, regular passengers, habitual passengers and irregular passengers. They were able to segment passengers into respective classes and their work helped bus operators in providing specific service to specific class of passengers. Shen *et al.* [12] used DBSCAN algorithm to detect hot spots for passenger's pick-up and drop-off locations. They enhanced the algorithm parameters sensitivity by applying clustering on each grid and initializing the density threshold for each grid. Also, they created a weighted tree with multiple factors such as speed, time, and distance to recommend the most suitable routes.

Saptawati *et al.* [13] partitioned trajectory data into small grids, and then applied spatiotemporal clustering. As they clustered each grid for each time interval, the authors were able to and congestion areas with the high- density population in a specific time interval. They used the DBSCAN algorithm for clustering in their approach. Mao *et al.* [14] proposed a novel approach to detect spatiotemporal patterns for household travel. They first clustered trip origin and destination stops to identify attractive places, then they visualized the identified clusters to analyze the behavior for different areas based on the spatial distribution and temporal trends.

Qiu *et al.* [15] proposed a novel approach to detect potential passengers out of regular passengers. They applied pairwise DBSCAN algorithm to cluster trip origin and destination stops, to identify attractive places, then they visualized the identified clusters to analyze the travel behavior in urban areas. Their work helped bus operators successfully designing initial bus lines. Zhao and colleagues [16] employed a hierarchical clustering method, leveraging the DBSCAN algorithm, to identify similarities within ship trajectories. They were able to discover the ships' traffic flow into the sea by identifying clusters with various densities.

Kieu *et al.* [17] introduced a weighted stop density-based clustering algorithm (WS-DBSCAN) for performing travel pattern analysis. This algorithm performs a two-level analysis process. In the first level of analysis, it makes use of existing knowledge of individual travel patterns to perform clustering. And in the second level, it performs neighborhood search operation only when required. This way the algorithm showed great results in terms of complexity, by bringing the quadratic level complexity of DBSCAN to a sub-quadratic level in WS-DBSCAN.

RESEARCH METHODOLOGY

In this paper, an approach for travel pattern analysis using HDBSCAN algorithm and Apriori sequential pattern mining algorithm is proposed. The approach that is proposed includes two major stages. In the first stage, HDBSCAN algorithm is executed for searching top clusters in passenger travel data. In the next stage, frequent item sets from the passenger travel data are generated using Apriori sequential pattern mining algorithm. The major steps carried out for the travel pattern analysis are shown in block diagram in Figure 1.

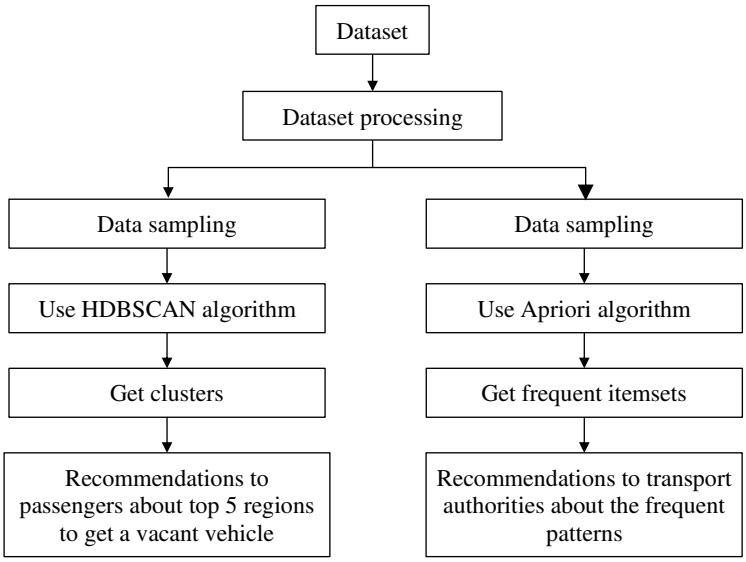

Figure 1. Block diagram of proposed approach.

The following is a detailed explanation of the steps identified:

1. *Data preprocessing:* Preprocessing is a critical step in clustering the travel data which is used to extract interesting knowledge from the data set. Two methods are applied for preprocessing of dataset:
   i. *Remove missing values:* In this step, the whole dataset is scanned in search of missing values and if such values are present, the entire tuple is removed from the dataset. In the approach the dataset was scanned and tuple with missing values were removed.
   ii. *Feature engineering:* This stage entails converting raw data into features in order to more accurately depict the underlying issue. In the approach pickup_date_time attribute was delimited into separate columns for easy analysis and also on polyline attribute feature engineering was applied to get pickup coordinates.
2. *Data sampling:* After preprocessing the data, data sampling is performed on the pre-processed dataset for the application of HDBSCAN and Apriori algorithms. For the approach, one-month trip samples from the original dataset for clustering were taken into account. These trips were from the month of March and consisted of around 1,38,000 trips. Further this sampled dataset is divided into 24 subsets, where each subset consists of trips recorded in that time interval. For sequential pattern mining, a hundred trip samples were taken into consideration for sequential pattern mining.
3. *Algorithms used in the proposed approach:* In order to get deeper insights from the passenger's trip data clustering the travel data on an hourly basis is proposed in the approach. Therefore, instead of clustering the whole sampled data, data is divided into 24 subsets, where each subset consists of trips recorded in that time interval and on each individual subset following clustering algorithm is applied:
   i. *Algorithm 1: Hierarchical density based clustering (HDBSCAN) algorithm*
   Input: Number of minpts
   Output: Clusters process:
   Step 1: Select one subset.
   Step 2: Select the coordinates of that subset.
   Step 3: For that subset set a variable i and iterate i from 1 to for all pair of coordinates in that subset.
   Step 4: Set a variable count = number of nonempty records in that subset.
   Step 5: Then set minpoints to round of Log (Count)
   i.e. MinPts = Log (Count).
   Step 6: If MinPts = 1
   // The cluster cannot contain one point only.

Break;
Else
//For this pair of coordinates, use clustering.
B = HDBSCAN (Coordinates, MinPts)

Step 7: After clustering add B attribute to the subset.
Step 8: Select next subset.
After applying the clustering algorithm on all the subsets, only top clusters are taken into consideration, and the first stage of the proposed approach ends here. Before entering the second stage of the proposed approach, from the hundred samples taken during sampling for second stage, coordinates in each sample are mapped to their names with the help of internet and finally a sequence database consisting of a set of ordered elements is created and the following sequential pattern algorithm is applied. This algorithm takes as an input a support value.

ii. *Algorithm 2: Apriori sequential pattern mining algorithm*
Input: Support = 0.04
Output: Frequent 1,2,3 Itemsets Process:
Step 1: To start, search the sequence database for the support 'S' for each 1-itemset, compare that support to the minimum support, and then get the support of 1-itemsets.
Step 2: Use join to create a collection of potential k-item sets. Use the apriori property to remove the infrequently used k-item sets from this list.
Step 3: Check the support 'S' of each possible k-item set in the supplied set in the sequence database, compare 'S' to the minimal support, and you'll get a list of often occurring k-item sets.
Step 4: In the event that the candidate set is empty, proceed by generating all nonempty subsets of frequent item set 1.
Step 5: For all nonempty subsets identified in 1, output the rule as "s=>(1-s)". If the confidence (C) of the "s=>(1-s)" rule is below the minimum confidence threshold.
Step 6: If the candidate set is not empty, proceed to step 2.
After applying the sequential pattern mining algorithm on all the mapped samples, frequent one, two and three item sets are generated and out of which only frequent two and three item sets are taken into consideration and with this second and the final stage of the proposed approach ends.

EXPERIMENTS AND RESULTS

Dataset

The UCI Machine Learning Repository's Port taxi trajectory dataset can be downloaded [18]. The cab trips in this dataset were taken between June 2013 and June 2014. The initial CSV file was about 2 GB in size and contained 1.7 million travels. Each trip was a separate tuple with numerous properties, with trip id serving as the unique identifier. The call type attribute specified the location of the taxi request, such as the downtown area, a random street, or a taxi station. The "origin stand" attribute indicated the starting location in case the cab originated from a station, while the "origin call" attribute contained the phone number used to order the cab. Each trip's start time was recorded by a Unix timestamp attribute, and the taxi driver could be recognized using the taxi id attribute. The day-type element reflected the day type (weekday, holiday, or weekend), while the missing attribute indicated the condition of the GPS reading.

If the reading was not successfully captured, the attribute was assigned a value of true; conversely, if the reading was successfully captured, the attribute was assigned a value of false. A polyline was the attribute that characterized each trip's route, it stored a sequence of longitudinal coordinates where each set of this sequence represented a specific location's longitude and latitude coordinates.

Results

In this section, results obtained after applying the HDBSCAN clustering algorithm on all the 24 subsets and frequent itemsets obtained after applying the Apriori pattern mining algorithm as proposed in the approach are presented.

A table is used to present the clustering findings. The table contains information of regions identified as top clusters. Table 1 shows the results in the morning 6-6:59 am interval. Table 2 shows the results in afternoon 1-1:59 pm interval. Table 3 shows the results in evening 5-5:59 pm interval and Table 4 shows the results in late night 11-11:59 pm interval.

The results of pattern mining obtained after applying the Apriori sequential pattern mining algorithm on sequence database consisting of a set of ordered elements for trajectory dataset is shown in table. Frequent 2 itemsets along with their support value is shown in Table 5. Frequent 3 itemsets along with their support value is shown in Table 6.

Table 1. Results in morning 6:00-6:59 am.

S.N.	Region cluster ID	Region description	Number of pickups
1	90	Cooperativa Dos Pedreiros	353
2	148	Café Santiago, Rau de Passos Manuel 198	346
3	146	Hospedaria Novo Mundo	213
4	145	Marques Soares	152
5	24	Survifor Porto Surf Hostel	123

Table 2. Results in afternoon 1:00-1:59 pm.

S.N.	Region cluster ID	Region description	Number of pickups
1	212	Jardim de Carrilho Videira	199
2	218	Praça Dom Joao I	154
3	146	Praça de Mouzinho de Albuquerque 3220	152
4	216	Labmed. Rua Alexandre Herculano	132
5	132	ERA Boavista, 269 R. de Gonçalo Sampaio	129

Table 3. Results in evening 5:00-5:59 pm.

S.N.	Region cluster ID	Region description	Number of pickups
1	20	Campanhã Railway Station	198
2	5	Hospital de São João	131
3	135	novo banco	127
4	133	Praça da Liberdade and Rua dos Clérigos	88
5	93	*Rua* Monsenhor *Salazar rua de diogo botelho*	82

Table 4. Results in late night 11:00-11:59 pm.

S.N.	Region cluster ID	Region description	Number of pickups
1	126	Campanhã Railway Station	355
2	60	Campus Sao Joao Rua Doutor Plácido Da Costa Porto	156
3	12	Novo Banco	142
4	236	ERA Boavista, 269	140
5	228	Mercado do Bolhão	134

Table 5. Frequent two itemsets.

S.N.	Itemsets	Support
1	Se, Sao Nicolau	0.20
2	Se, Bonfin & Campha	0.10
3	Santo Iildefonso, Bonfin and Campha	0.07
4	Miaragaia, Se	0.07
5	Santo Iildefonso, Se	0.07

Table 6. Frequent three itemsets.

S.N.	Itemsets	Support
1	Se, Vitoria, Sao Nicolau	0.04
2	Miaragaia, Se, Sao Nicolau	0.04

Metric Used for Evaluating Clustering

To measure the efficiency of clustering used in the proposed approach, the Silhouette coefficient/score has been used. Silhouette A clustering technique's quality can be determined using a coefficient or score [19]. Its value is between –1 and 1. Here 1 denotes clusters that are well separated and distinct from one another, 0 denotes clusters are unimportant, or distance between clusters that is not significant and a value of –1 indicates that clusters have been assigned incorrectly.

$$\text{Silhouette score} = \frac{b(o)-a(o)}{\max\{a(o),b(o)\}}$$

Here, "s(o)" represents the silhouette coefficient of a data point, "a(o)" signifies the average distance of the data point from all other data points within its cluster, and "b(o)" denotes the smallest average distance to any cluster to which the data point does not belong. Figure 2 shows the silhouette score for DBSCAN and HDBSCAN generated after applying clustering on all the 24 subsets.

Comparison with DBSCAN Algorithm

In this section, we compare the HDBSCAN algorithm used in the proposed approach with DBSCAN algorithm using silhouette score. Figure 2 displays the silhouette scores for the HDBSCAn and DBSCAN. From Figure 2, the quality of clusters is measured. For DBSCAN algorithm silhouette score ranges from –0.06 to 0.52, whereas the silhouette score for HDBSCAN algorithm ranges from 0.46 to 0.60.

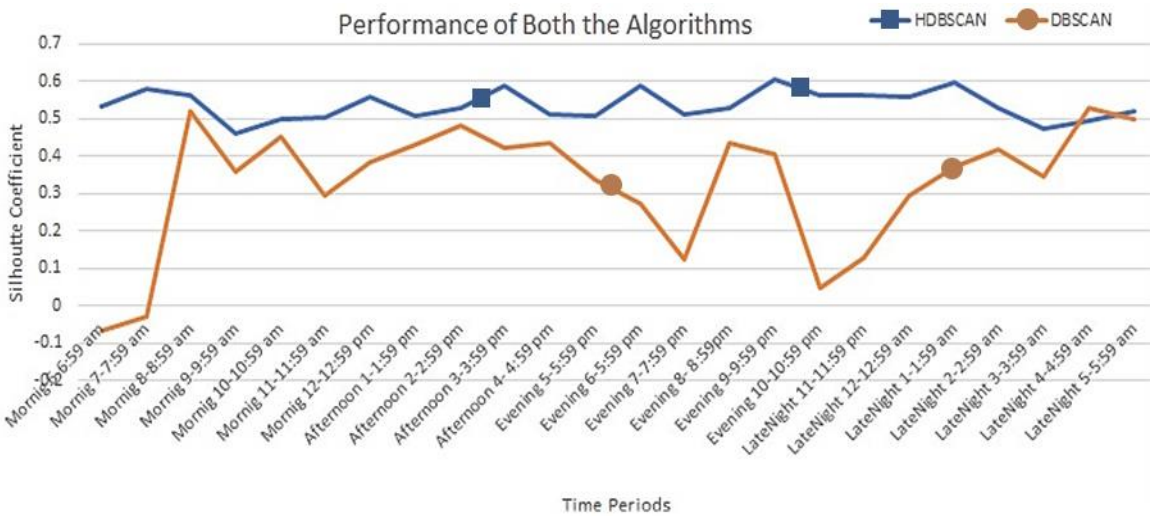

Figure 2. Silhouette scores for HDBSCAN and DBSCAN algorithms on the proposed approach.

From the plotted results, it is noted that use of HDBSCAN, instead of traditional DBSCAN algorithm for travel pattern analysis gave better results in terms of quality of clusters.

CONCLUSION AND FUTURE WORK

In this paper, a new approach for trajectory analysis is proposed where a HDBSCAN algorithm and Apriori sequential pattern mining algorithm is used to identify travel patterns of the passengers. Clustering on an hourly basis helped in getting deeper insights of travel patterns. The use of a HDBSCAN instead of DBSCAN algorithm helped in achieving better results in terms of quality of clusters. Along with clustering an Apriori sequential pattern mining algorithm is applied on dataset to extract frequent patterns. The clusters obtained can be used as recommendations for passengers. This

recommendation can increase a passenger's chances of getting a vacant vehicle. The frequent itemsets obtained from pattern mining can be used as recommendations to transport authorities. These recommendations can help them in laying out a new transport service. Understanding passenger travel patterns is considered as the biggest problem in trajectory analysis domain. The use of combination of clustering and sequential pattern mining algorithms as proposed in the approach, provides a better pavement for the travel pattern analysis.

Future Work

The proposed approach can be further amended by considering the use of efficient hybrid and ensemble clustering methods for the analysis of a passenger's travel data and the work can further be extended by performing analysis at a higher granularity level, that is, by performing clustering on minute basis.

Acknowledgments

Our thanks to Dr. Anuradha Purohit who contributed towards the development of the paper.

REFERENCES

1. Zheng Y. Trajectory data mining: an overview. ACM Trans Intell Syst Technol. 2015; 6 (3): 1–41.
2. Lin M, Hsu WJ. Mining GPS data for mobility patterns: a survey. Pervasive Mobile Comput. 2014; 12: 1–6.
3. Zhong RY, Huang GQ, Lan S, Dai QY, Chen X, Zhang T. A big data approach for logistics trajectory discovery from RFID-enabled production data. Int J Prod Econ. 2015; 165: 260–272.
4. Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: concepts, methodologies, and applications. ACM Trans Intell Syst Technol. 2014; 5 (3): 1–55.
5. Dang S, Ahmad PH. Text mining: techniques and its application. Int J Eng Technol Innov. 2014; 1 (4): 22–25.
6. Campello RJ, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Germany: Springer; 2013. pp. 160–172.
7. Agrawal R, Mehta M, Shafer JC, Srikant R, Arning A, Bollinger T. The Quest data mining system. In: Simoudis E, Han J, Fayyad UM, editors. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press; 1996. pp. 244–249.
8. Wei YQ, Yang RH, Liu PY. An improved Apriori algorithm for association rules of mining. In: 2009 IEEE International Symposium on IT in Medicine & Education, Jinan, China, August 14–16, 2009. pp. 942–946.
9. Palma AT, Bogorny V, Kuijpers B, Alvares LO. A clustering-based approach for discovering interesting places in trajectories. In: Proceedings of the 2008 ACM Symposium on Applied Computing, Fortaleza, Ceara Brazil, March 16–20, 2008. pp. 863–868.
10. Bermingham L, Lee I. Spatio-temporal sequential pattern mining for tourism sciences. Procedia Computer Sci. 2014; 29: 379–389.
11. Bhaskar A, Chung E. Passenger segmentation using smart card data. IEEE Trans Intell Transport Syst. 2014; 16 (3): 1537–1548.
12. Shen Y, Zhao L, Fan J. Analysis and visualization for hot spot based route recommendation using short-dated taxi GPS traces. Information. 2015; 6 (2): 134–151.
13. Saptawati GAP. Spatio-temporal mining to identify potential traffic congestion based on transportation mode. In: 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, Indonesia, November 1–2, 2017. pp. 1–6.
14. Mao F, Ji M, Liu T. Mining spatiotemporal patterns of urban dwellers from taxi trajectory data. Front Earth Sci. 2016; 10: 205–221.
15. Qiu G, Song R, He S, Xu W, Jiang M. Clustering passenger trip data for the potential passenger investigation and line design of customized commuter bus. IEEE Trans Intell Transport Syst. 2018; 20 (9): 3351–3360.

16. Zhao L, Shi G, Yang J. An adaptive hierarchical clustering method for ship trajectory data based on DBSCAN algorithm. In: 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, March 10–12, 2017. pp. 329–336.
17. Kieu LM, Bhaskar A, Chung E. A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. Transport Res Part C Emerg Technol. 2015; 58: 193–207.
18. Frank A. UCI Machine Learning Repository. [Online]. 2010. Available at http://archive.ics.uci.edu/ml
19. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20: 53–65.

