Extracting potential bus lines of Customized City Bus Service based on public transport big data

To cite this article: Yibin Ren et al 2016 IOP Conf. Ser.: Earth Environ. Sci. 46 012017

View the article online for updates and enhancements.
Extracting potential bus lines of Customized City Bus Service based on public transport big data

Yibin Ren¹, Ge Chen¹,², Yong Han¹,², Huangcheng Zheng¹
¹Qingdao Collaborative Innovation Center of Marine Science and Technology, College of Information Science and Engineering, Ocean University of China, China
²Laboratory for Regional Oceanography and Numerical Modeling, Qingdao National Laboratory for Marine Science and Technology, China
E-mail: yonghan@ouc.edu.cn

Abstract. Customized City Bus Service (CCBS) can reduce the traffic congestion and environmental pollution that caused by the increasing in private cars, effectively. This study aims to extract the potential bus lines and each line’s passenger density of CCBS by mining the public transport big data. The datasets used in this study are mainly Smart Card Data (SCD) and bus GPS data of Qingdao, China, from October 11th and November 7th 2015. Firstly, we compute the temporal-origin-destination (TOD) of passengers by mining SCD and bus GPS data. Compared with the traditional OD, TOD not only has the spatial location, but also contains the trip’s boarding time. Secondly, based on the traditional DBSCAN algorithm, we put forwards an algorithm, named TOD-DBSCAN, combined with the spatial-temporal features of TOD. TOD-DBSCAN is used to cluster the TOD trajectories in peak hours of all working days. Then, we define two variables P and N to describe the possibility and passenger destiny of a potential CCBS line. P is the probability of the CCBS line. And N represents the potential passenger destiny of the line. Lastly, we visualize the potential CCBS lines extracted by our procedure on the map and analyse relationship between potential CCBS lines and the urban spatial structure.

1. Introduction

Public transport is one of the most common ways of daily travel [¹]. In recent years, Customized City Bus Service (CCBS) becomes a new popular mode of public transportation service. It not only improves the comfort of public bus, but also improves the efficiency of commuting. It is conducive to easing traffic congestion and environmental pollution caused by increasing of private cars. So, the promotion of using CCBS is very significant for urban management. In China, CCBS was first
introduced and implemented in Qingdao in August 2013 and it spread to other cities quickly. By 2015, there have been more than 30 cities across the country running CCBS systems. Currently, the way of setting a CCBS line is “from bottom to top”: passengers send the origins and destinations of their trips to public transportation management department through APP or Web page; bus department collects all requests and decides which lines to be set up. This way is limited by network and mobile devices. Thus, if the bus management department takes the initiative to analyze the potential CCBS lines, there will be an effective complementary way to improve the use of CCBS.

The way to discover potential CCBS lines is finding the similarities of transits between different passengers. So, we need to grasp the daily travel characteristics of passengers. Smart Card Data (SCD) records the travel process of the cardholder, widely covered and quickly accessed. We can mine the passengers’ origin and destination (OD) from consumption records of SCD and further extract the potential CCBS lines by analyzing spatio-temporal characteristics of OD trajectories. Mostly, SCD doesn’t record the boarding station when cardholder brushing the card. We need to combine SCD with bus GPS data to get passenger’s boarding station. Then, we can deduct the alighting station based on origins by trip chain model. A series of study work have been done to extract OD information and analyze the spatio-temporal characteristics of passengers. Nishiuchi H et al analyze variations in trip patterns to understand how passengers’ daily travel patterns vary temporally and spatially among one month. Ma X et al use the K-Means++ clustering algorithm and the rough-set theory to cluster and classify travel pattern regularities. And they develop a data-driven platform for online transit performance monitoring. Tao S et al analyze the travel differences between BRT and non-BRT passengers. Zhao J et al study the spatio-temporal travel patterns of individual passengers in metro system of Shenzhen, China. Recently, some researchers focus on the travel characteristics of specific groups. Long et al analyzes the travel characteristics including travel frequency, commute time, as well as jobs-housing location and their change of underprivileged residents in Beijing by SCD. And they also study the travel characteristics of 4 kinds extreme public transit riders (early birds, night owls, and tireless/recurring itinerants).

Existing studies mentioned above are focusing on OD deduction, commuting analysis and travel regularity discovering etc. The studies about discovering similarities between passengers public travel behaviors which is aiming to extract potential CCBS lines are a little fewer. Therefore, this study aims to extract the potential CCBS lines and each line’s passenger density by mining the public transport big data. We propose temporal-origin-destination (TOD) data structure and extent the traditional DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm in both spatial and temporal dimensions, forming TOD-DBSCAN algorithm. Then, we use the TOD-DBSCAN to find the similarly travel patterns and extract the potential CCBS lines. At last, we visualize the potential CCBS lines on the map by a Web GIS system and make some discussion.

The remainder of the paper is structured as follow: Section 2 introduces the public transport big data used in this study. Section 3 expounds the procedure to extract potential CCBS lines, including construction of TOD, TOD-DBSCAN algorithm, and extraction and visualize of potential CCBS lines. In Section 4, we come to the conclusion of this study.

2. Public transport big data
The dataset used in this study is provided by the Qingdao Public Transportation Group, which has
around 5000 buses on the core roads of Qingdao. The dataset includes SCD data, bus GPS data and scheduling table of drivers. The SCD data contains more than 50 million transactions, collected from 1.2 million smart cards between October 11th and November 7th 2015. The format of SCD record is shown in table 1. The bus GPS data is an important auxiliary data to extract boarding station of cardholders, because it records the time when the bus arriving and leaving the station, table 2. And the scheduling table of drivers records the relationship between bus and driver, table 3. It is the bridge to link SCD and bus GPS Data.

### Table 1. Data structure of SCD

| Field Name | Description |
|------------|-------------|
| CardID     | Unique number for each card. |
| RouteName  | Line name of this transaction. |
| DriverID   | Driver’s ID number. |
| ActTime    | Detial time of the transaction. |

### Table 2. Data structure of bus GPS data

| Field Name | Description |
|------------|-------------|
| BusID      | Unique number of bus. |
| RouteName  | Line name of this bus. |
| StationNum | Station number of this line. |
| ActTime    | Detial time of the action. |
| Flag       | Arriving or leaving. |

### Table 3. Data structure of schedule table

| Field Name | Description |
|------------|-------------|
| BusID      | Unique number of bus. |
| RouteName  | Line name of this bus. |
| DriverID   | Number of driver on this bus. |

3. Extracting potential CCBS lines by TOD-DBSCAN algorithm

3.1. Construction of TOD

In order to construct the TOD, we should extract OD trajectory firstly. The Automatic Fare Collection System of Qingdao is flat fares: passengers pay a fixed price and they brush the card only when they boarding on the bus. So, we need to extract the passenger’s origin (boarding station) firstly and then deduce the destination (alighting station) based on the origin. As shown in table 1, the SCD record doesn’t contain boarding station. The bus GPS data records arriving time and leaving time of each station. So, we can calculate the boarding station by matching card time with arriving and leaving time of the bus. The scheduling table contains the corresponding relationship between the driver and the bus. Thus, we can match SCD record with bus GPS data through schedule table (Step 1 in figure1.) to get the corresponding relationship between SCD and bus GPS. Actually, some passengers brush the card after the bus leaving station. So, we use the arriving times of two adjacent stations as time threshold, figure1. Further, for the deduction of alighting station, we use the trip chain model which is based on two assumptions\cite{14}: ① The previous trip destination is nearby the bus stop of next trip origin; ② Passengers return to the first boarding station of the day at the end of the day. They are widely adopted by most studies.

![Figure 1. Steps to extract boarding station](image)

Extracting boarding station by matching card time with bus arriving and leaving time.

After above processes, we got the passengers’ OD trajectories. But a complete OD maybe divided into...
several ODs by transfer stations. We need to find the transfer behavior and merge them to a complete one. We compute the time interval between current boarding behavior and last alighting behavior. If the time interval is less than 30 minutes, we merge two ODs to one, (figure 2). Since we have obtained the corresponding relationships between transactions and bus, and the alighting station is known, so the alighting time is the time when the bus leaving from the alighting station.

Based on three datasets mentioned in Section 2, we extract OD trajectories which including boarding and alighting station, boarding and alighting time. In order to analyze the spatial and temporal similarity of OD trajectories, we propose TOD (temporal-origin-destination) data structure: {ID, \(T_B\), \((X_O, Y_O)\), \((X_D, Y_D)\)}. ID is the card ID, \(T_B\) is the boarding time at origin station, \((X_O, Y_O)\) is the coordinate of origin station, and \((X_D, Y_D)\) is the coordinate of destination station. Taking card “2***26” (figure 2) for example, its TOD model is {2***26, 2015-10-12 07:13:15, (120.34783, 36.0822), (120.42315, 36.07415)}. Because the alighting time is impacted by various factors, we don’t take it as a factor to analyze the similarity between OD trajectories. And it’s not included in TOD. Then, we will detail the algorithm to cluster TOD trajectories.

| Card ID | Date       | Boarding time | Alighting time | OD                        |
|--------|------------|---------------|----------------|---------------------------|
| 2***26 | 2015.10.12 | 07:13:15      | 08:17:17       | Line 125:Taidong -> Qingdao University |
| 2***26 | 2015.10.12 | 08:24:32      | 08:37:56       | Line 227:Qingdao University -> Qingda Yilu |

\[(08:24:32 - 08:17:17) = 7.25\text{ minutes} < 30\text{ minutes},\text{ merging two OD trajectories.}\]

| Card ID | Date       | Boarding time | Alighting time | OD                        |
|--------|------------|---------------|----------------|---------------------------|
| 2***26 | 2015.10.12 | 07:13:15      | 08:37:56       | Line 125:Taidong -> Line 227: Qingda Yilu |

**Figure 2.** Modifying raw OD trajectories

3.1 **TOD-DBSCAN algorithm**

We analyze the spatial and temporal similarities of OD trajectories by clustering operation. DBSCAN algorithm is not required to define the number of clusters or identify arbitrarily shaped clusters because higher-density records are more likely to be grouped into a cluster \(^{[15]}\). It needs two parameters: distance threshold (\(\varepsilon\)) and the minimum number of points (\(\text{MinPts}\)). With two parameters, it calculates the connected relation between points iteratively, forming result clusters. In 2014, Kieu L M \(^{[16]}\) used DBSCAN algorithm to cluster OD trajectories, including 3 steps: clustering alighting points, clustering boarding points and clustering boarding time. He extended the traditional DBSCAN algorithm from one dimension (point) to three dimensions (line segment with time information). But, this method clusters identical OD trajectories repeatedly, increasing the clustering time, especially for long time series public transport big data. To solve this problem, we analyze the characteristics of public transport OD trajectories, designing TOD-DBSCAN algorithm. The new algorithm contains two parts: spatial clustering and temporal clustering. The \(\varepsilon\) is no longer just a space distance. It becomes a combination of spatial distance (\(l\)) and temporal distance (\(t\)).

**Figure 3.** OD trajectories which are similar in spatial
We illustrate the spatial clustering process firstly. OD trajectories which are similar in spatial are show in figure 4: complete overlap (figure 3(a)), only one point overlap (figure 3(b)) and complete under lap but both terminals are within limited distance (figure 3(c)). So, in the process of OD clustering, we should judge whether origins and destinations between different OD trajectories are within the distance threshold respectively. The bus stations in the city form a fixed points set. And both the origin and the destination are all belong to this points set. So, we take the unique origins and unique destination of all OD trajectories instead of all origins and destinations as the cluster objects to find the spatial similarities. Take the situation in figure 5 for example. There are 10 OD trajectories in figure 4(a). Using the method of Kieu L M, we have to clustering 20 points, even if there are some duplicate points. But in our procedure, firstly, we extract the unique points collection of origins and destinations: \( \{O_1,O_2,O_3,O_4,O_5,O_6\} \) and \( \{D_1,D_2,D_3,D_4,D_5\} \) (figure 4(b)). Then, we cluster two point collections respectively by DBSCAN, froming 2 origin clusters (OC1, OC2) and 3 destination clusters (DC1, DC2, DC3) (figure 4(c)). And the number of clustering objects is reduced from 20 to 11. Lastly, for each group in origin clustered results, we select OD trajectories whose origins belong to the group and cluster the OD trajectories based on result clusters in step 2, forming the final spatial clustering results (figure 4(d)). Taking OC1 for example, there are 5 OD trajectories (OD-ID:1,2,3,7,10) whose origins are included in OC1. And the 5 destinations (D1,D2,D3,D4,D5) are clustered into 3 groups(DC1,DC2,DC3) in step 2. So, combining the origin clusters with destination clusters, these 5 OD trajectories are clustered into 3 groups (Cluster ID: 1, 2, and 3). In this way, all the 10 OD trajectories are clustered into 5 groups by spatial clustering process.

**Figure 4.** Spatial clustering process of TOD-DBSCAN algorithm

Temporal clustering is relatively simple. After above processing, OD trajectories which are similar in spatial are clustered into one spatial group. Temporal clustering is clustering all trajectories from one group in time dimension. As mentioned above, we use the boarding time to represent the temporal information of OD trajectory. And the boarding time is recorded in TOD, so we just need to set the boarding time as cluster variable, clustering TOD trajectories belong to one spatial group using DBSCAN algorithm with the time distance. For example, OD trajectories whose ID numbers are 4, 5 and 6 are clustered into the 4\(^{th}\) group (cluster ID is 4 in figure 4(d)). After temporal clustering, they are clustered into 2 groups (cluster1: OD4 and OD5; cluster2: OD3) with the time distance 15mins finally (figure 5).
3.2 Extracting potential CCBS lines using TOD-DBSCAN

The main idea to extract potential CCBS lines is to discover the similarities between OD trajectories of working day. We use TOD-DBSCAN algorithm to cluster TOD trajectories. Our analysis objects are TODs whose \( T_B \) values are in morning peak period (06:30am ~ 08:30am). In order to determine the spatial distance threshold, we change the threshold value to cluster the unique boarding stations collection (a total of 2673 stations), setting \( l \) as 100m, 250m and 500m respectively. The \( MinPts \) in this stage is 0, because a single station should also become one group. The results are shown in Figure 6. The number of clusters are 1759 (\( l=100m \)), 1089 (\( l=250m \)) and 352 (\( l=500m \)). The spatial clustering achieves a better result with 250m. When \( l \) is 100m, the clustering results are too scattered. And when it increases to 500m, the results are too concentrated. Both two values can’t be very good to characterize the spatial proximity between stations. So, we use the 250m as spatial distance threshold.

In temporal clustering stage, we set the value of \( t \) as 15 minutes. And the weekends of four weeks aren’t included. The \( MinPts \) is 300, because if the passenger destiny less than 15 per day on average, there is no need to set the line.

![Figure 5. Temporal clustering process](image)

**Figure 5.** Temporal clustering process

After spatial and temporal clustering, we define two variables \( P \) and \( N \). \( P \) is the probability of the CCBS line, and it can calculate by equation 1: \( P_i \) is the probability of line \( i \), \( M_i \) is the number of days in group \( i \), \( M_a \) is the number of all working days during the dataset’s time span (it’s 20 in this study). If \( P_i \) is less than 0.6 (3 days a week), there will be no need to set the potential CCBS line. \( N \) is the number of potential passengers of the line. It is calculated by equation 2, \( S_i \) is all the passengers in group \( i \), and \( M_i \) is the number of days in group \( i \). Generally, most groups contain multi origins and destinations, and we

![Figure 6. Clustering results of unique boarding stations with different values of l](image)

**Figure 6.** Clustering results of unique boarding stations with different values of \( l \)
select the one with the highest frequency as the origin or destination of CCBS line. The departure time of the line is also determined in the same way and we set 5mins as time unit to half-adjust it.

$$P_i = \frac{M_i}{M_n}$$  

$$N_i = \frac{S_i}{M_i}$$

The details of the results are shown in table 4. There are 49 potential CCBS lines extracting by our procedure. And they mainly contain 3 directions: from north to south (① in figure 7 (a)), from east to south and west (② in figure 7 (a)) and from Taidong to Qingdao North Station (③ in figure 7 (a)). This phenomenon is related to the urban structure of Qingdao. The southern coastal line is the most prosperous region in Qingdao, containing a series of companies, shopping malls and tourist areas. And the northeast part of the city is a new residential area with a relatively low housing price, attracting a lot of young people. So, they have to across the city for work. And the Taidong is a representative residential of the old city. Traditional heavy industry is distribution near the north railway station and it is the main work place of the older generation. The potential CCBS lines are corresponding to the urban structure of Qingdao.

### 4. Conclusion

This study aims to extract potential CCBS lines by mining public transport big data. For this goal, we propose the TOD data structure and extend the traditional DBSCAN algorithm to TOD-DBSCAN. We cluster all the TOD trajectories from October 11th to November 7th 2015 by TOD-DBSCAN algorithm. A total of 49 potential CCBS lines are extracted from the clustering results. They are corresponding to the urban structure and visualized in a Web GIS system. Next step, we will cooperate with the public transportation department to validate the key lines.

### Acknowledgement

This work is supported by Science and Technology Development Plan of Qingdao; National High Technology Research and Development Program of China (863 Program): System of Ocean Observation Based on Shipboard UAV (Grant NO. 2014AA09A511).
References

[1] Carmien S, Dawe M, Fischer G, Gorman A and Kintsch A 2005 Socio-technical environments supporting people with cognitive disabilities using public transportation ACM.TOCHI 12 233-262
[2] Liu T and Ceder A 2015 Analysis of a new public-transport-service concept: Customized City Bus Service in China Transport. Policy 39 63-76
[3] Pelletier M, Trépanier M and Morency C 2011 Smart card data use in public transit: A literature review Transport. Res. C. 19 557-568
[4] Ma X, Wang Y, Chen F and Liu J 2012 Transit smart card data mining for passenger origin information extraction J. Zhejiang Univ. - Sci. C 13 750-760
[5] Barry J, Newhouser R and Rahbee A 2002 Origin and destination estimation in New York City with automated fare system data Transport.Res.Rec. 1817 183-187
[6] Trépanier M, Tranchant N and Chapleau R 2007 Individual trip destination estimation in a transit smart card automated fare collection system J INTELL TRANSPORT S 11 1-14
[7] Nishiuchi H, King J and Todoroki T 2013 Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data International Journal of Intelligent Transportation Systems Research 11 1-10
[8] Ma X, Wu Y J, Wang Y and Liu J 2013 Mining smart card data for transit riders’ travel patterns Transport.Res.C 36 1-12
[9] Ma X, Wang Y 2014 Development of a data-driven platform for transit performance measures using smart card and GPS data J.Transp.Eng, 140( 04014063)
[10] Tao S, Corcoran J, Mateo-Babiano I and Rohde D 2014 Exploring Bus Rapid Transit passenger travel behaviour using big data Appl Geogr 53 90-104
[11] Zhao J, Tian C, Zhang F and Xu C 2014 Understanding temporal and spatial travel patterns of individual passengers by mining smart card data IEEE 17th International Conference on Intelligent Transportation Systems (ITSC) p 2991-2997
[12] Long Y, Shen Z 2015 Profiling underprivileged residents with mid-term public transit smartcard data of Beijing Computer Science 2015 169-192
[13] Long Y, Liu X, Zhou J and Chai Y 2015 Early Birds, Night Owls, and Tireless/Recurring Itinerants: An Exploratory Analysis of Extreme Transit Behaviors in Beijing, China Computer Science 2015(1502.02056)
[14] Li D, Lin Y, Zhao X and Zou N 2011 Estimating a transit passenger trip origin-destination matrix using automatic fare collection system International Conference on Database Systems for Advanced Applications (Springer Berlin Heidelberg) p 502-513
[15] Ester M, Kriegel H P, Sander J and Xu X 1996 A density-based algorithm for discovering clusters in large spatial databases with noise Kdd. 96 226-231
[16] Kieu L M, Bhaskar A, and Chung E 2014 Transit passenger segmentation using travel regularity mined from Smart Card transactions data Engineering 2014