Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 9457486, 11 pages
https://doi.org/10.1155/2018/9457486

Research Article

Passenger Travel Regularity Analysis Based on a Large Scale Smart Card Data

Qi Ouyang, Yongbo Lv, Yuan Ren, Jihui Ma, and Jing Li

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Qi Ouyang; 14114203@bjtu.edu.cn

Received 16 June 2018; Revised 25 August 2018; Accepted 4 September 2018; Published 24 September 2018

Guest Editor: Javier Sánchez-Medina

Analysis of passenger travel habits is always an important item in traffic field. However, passenger travel patterns can only be watched through a period time, and a lot of people travel by public transportation in big cities like Beijing daily, which leads to large-scale data and difficult operation. Using SPARK platform, this paper proposes a trip reconstruction algorithm and adopts the density-based spatial clustering of application with noise (DBSCAN) algorithm to mine the travel patterns of each Smart Card (SC) user in Beijing. For the phenomenon that passengers swipe cards before arriving to avoid the crowd caused by the people of the same destination, the algorithm based on passenger travel frequent items is adopted to guarantee the accuracy of spatial regular patterns. At last, this paper puts forward a model based on density and node importance to gather bus stations. The transportation connection between areas formed by these bus stations can be seen with the help of SC data. We hope that this research will contribute to further studies.

1. Introduction

Traditional studies on passenger travel patterns and passenger segmentation solely focus on passenger physical characteristics or the use of transit user surveys. This classification has little help of knowing passenger travel habits. Therefore, we need another method to study the temporal and spatial regularity. This method must be based on actual data with passenger travel information. SC data meets the needs.

SC data gathered by automated fare collection systems records travel details which are very valuable. However, passenger travel patterns can only be watched through a period time, and a lot of people travel by public transportation in big cities like Beijing daily, which leads to large-scale data and difficult operation. This paper adopts SPARK platform to solve this problem. Several computers are used to build the platform and calculate together.

This paper adopts a systematic approach to mine the travel pattern and search the temporal and spatial regularity using SC data. After the literature review in this section, this paper introduces the SC dataset adopted and the method used to drop invalid data. We consider each item in the dataset a transaction. Then, this paper rebuilds the SC dataset by reconstructing the completed transactions of SC users into a trip, which can recognize SC user transfer behaviour. After the reconstruction, the density-based spatial clustering of application with noise (DBSCAN) algorithm is adopted to mine the travel pattern and obtain the temporal and spatial regularity of each SC user in Beijing. In spatial dimension, this paper designs an algorithm based on passenger travel frequent items to handle the phenomenon that passengers swipe their cards before arriving to avoid the crowd. Then, we put the temporal data and the spatial data together to classify SC users by temporal and spatial features. Finally, this paper puts forward a model based on distance and node importance to gather bus stations into areas without any intersection, and then we assign the SC data into every area to investigate transportation connection between them. In Section 4, this paper sums up some conclusions.

Analysis of SC data has attracted research interest and a lot of researches have been done in the past few years. Catherine Morency et al. (2007) measured transit variability with SC data. They built an object model to understand the relationship between different elements within the transit network, and then used k-means cluster to indicate the spatial variability of passengers [1]. Ka Kee Alfred Chu et

al. (2008) detected transfer coincidence based on information of the vehicles and SC data, and then they obtained temporal distribution and cumulative percentage of transfer time [2]. Their group (2010) proposed a methodology to enhance transit trip characterization by adding a multiday dimension to SC transactions. They detected individual, anchor points—precise to an exact address for each SC user. Then, they adopted spatial statistics, spatial analyses with geographic information system, visualizations, and data mining to describe passenger activity space, locations and departure time [3].

In recent years, with the development of the associated modelling methods, solving technology and computing capabilities, the study of SC data has developed rapidly. Jun Liu et al. (2014) presented a traffic monitoring and analysis system for large-scale networks based on Hadoop, an open-source distributed computing platform for big data processing on commodity hardware [4]. Sui Tao et al. (2014) applied a geovisualisation-based method to a large SC database to examine spatial temporal dynamics on BRT systems in Brisbane, Australia. They displayed their analysis result by a thermodynamic chart [5]. Cynthia Chen (2016) introduced how to use big data and small data datasets, concepts, and methods to analyse travel behaviour [6].

With the help of big data, researchers can identify passenger patterns derived by data through some complex models and algorithms. Le Minh Kieu et al. (2015) adopted the DBSCAN algorithm, which can find clusters of arbitrary shapes based on different parameters, to mine the travel patterns based on around 34.8 million transactions made by a million SC users over 15000 transit stops of the bus, city train, and ferry networks. They segmented transit passengers into four identifiable types based on the above research. However, because of the high algorithm complexity, this algorithm takes a long time to cluster convergence when the dataset is very large [7]. Mohamed K. ElMahrsi et al. (2017) proposed two approaches to cluster SC data. The first approach clusters stations based on when their activity occurs; i.e., how trips made at the stations are distributed over time. The second approach makes it possible to identify groups of passengers that have similar boarding times aggregated into weekly profiles [8]. Xiaolei Ma et al. (2017) measured spatial-temporal regularity of individual commuters, including residence, workplace, and departure time, using one-month transit SC data. They divided one day into 48 intervals, which means one interval contains half an hour, to observe temporal regularity [9]. Anne-Sarah Briand et al. (2017) presented a two-level generative model that applied the Gaussian mixture model to regroup passengers based on their temporal habits in their public transportation usage. They observed the year-to-year changes in public transport passenger behaviour [10].

2. Materials and Methods

The Materials and Methods should contain sufficient details so that all procedures can be repeated. It may be divided into headed subsections if several methods are described.

Section 2.1 in this section introduces the dataset used for the case study, as well as the methods for the reconstruction of travel itineraries. This part also introduces some definitions makes a simple statistics analysis. Sections 2.2 and 2.3 analyse passenger travel behaviour in time dimension and spatial dimension. Section 2.4 tries to find the relation between different areas based on station density and node contraction in weighted complex networks.

2.1. Dataset and Reconstruction of Travel Itineraries. The SC data used in this paper come from Beijing, which is one of the largest cities in the world. 6 million SC records are collected by AFC every day in this city. The total dataset contains around 150 million transactions over 7000 transit stops of the bus from October 1, 2015 to October 30, 2015. The dataset includes the following main fields:

(1) CARDID: The unique SC ID, which has been hashed into a unique number to maintain the privacy of the cardholder.

(2) TRADETIME: The time that passengers swipe their cards when they get on the bus.

(3) MARKTIME: The time that passengers swipe their cards when they get off the bus.

(4) LINEID: The transit routes that passengers take.

(5) TRADESTATION: The station that passengers swipe their cards at when they get on the bus.

(6) MARKSTATION: The station that passengers swipe their cards at when they get off the bus.

The SC dataset only contains information of passengers, which can be used when combined with bus operation data. The bus operation data includes the following fields:

(7) XLDM: The same as LINEID.

(8) ZM: The name of bus stops.

(9) ZDXH: The same as TRADESTATION and MARKSTATION.

(10) ZDJD: The longitude of bus stops.

(11) ZDWD: The latitude of bus stops.

In this paper, we call each item collected by AFC a transaction. As we know, one passenger may take buses many times each working day, so there are several transactions for each SC user every day. Sometimes, a bus trip one passenger takes contains two or three transactions. How to construct the travel trip from individual transactions is a fundamental problem before mining the travel patterns.

This paper adopts an algorithm to connect the individual transactions: there are two principles of this algorithm: first, if two transactions can be connected into one trip, the time interval between transactions must be less than 60 mins [11]; then, we sort two transactions by MARKTIME. If two transactions can be connected into one trip, the origin stop of the first transaction must be different from the destination stop of the second.

Here, the first boarding stop and the last alighting stop of a completed trip are defined as the "origin stop" and the "destination stop," respectively. The time interval between the alighting time of a transaction and the boarding time of the next transaction of the same trip is defined as the transferring time. Figure 1 shows the flowchart. The steps of the algorithm are described as follows.

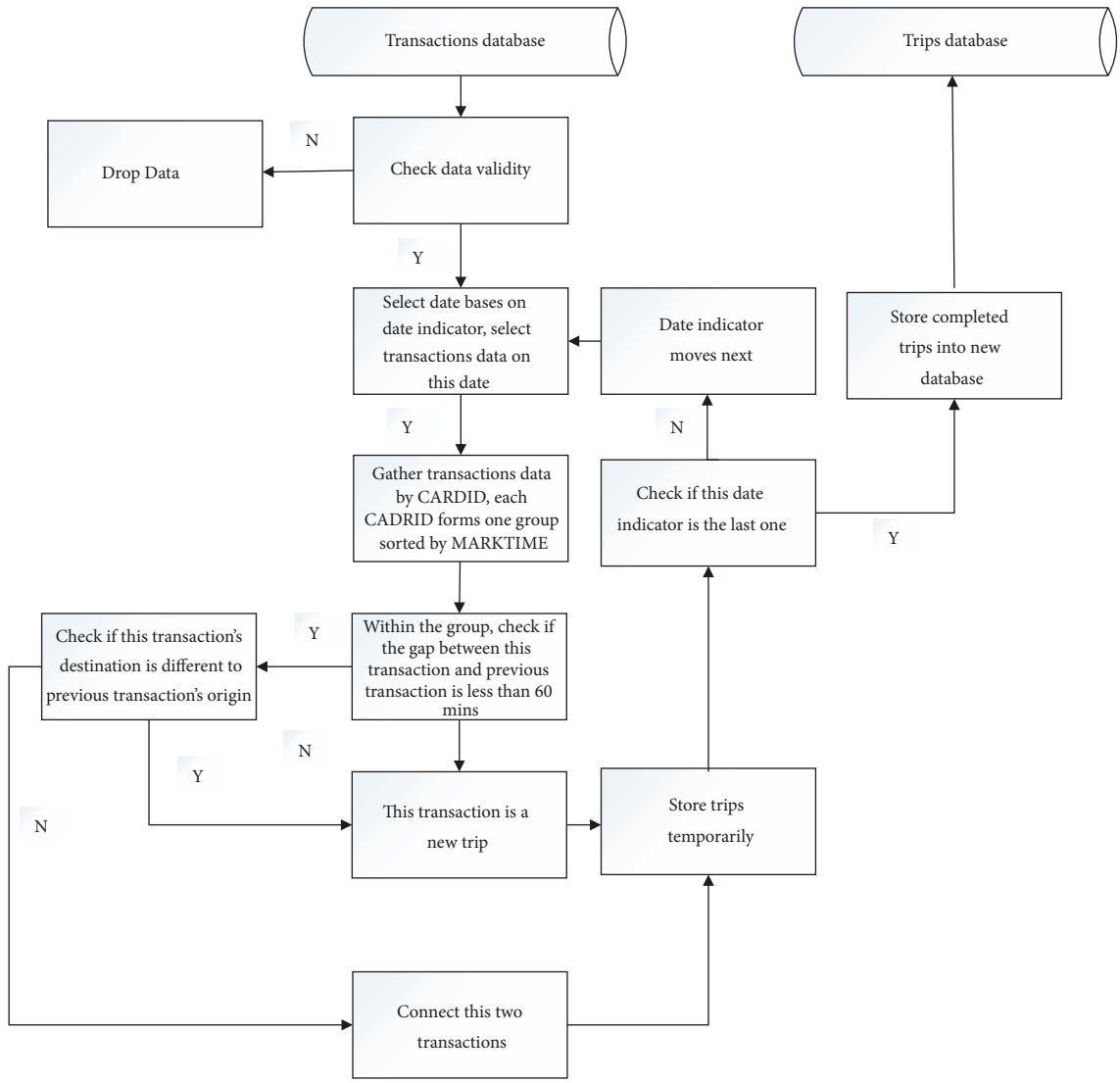

Figure 1: Trip reconstruction flowchart.

Step 1. Check data validity. Because of some hardware problems, the data collected by AFC cannot be used, as two items are exactly the same as each other: any data missed in fields "MARKTIME, TRADETIME, LINEID, MARKSTATION, and MARKSTATION" or the MARKTIME equals TRADETIME in one item. By checking data validity, this paper drops around 1% of the transactions data.

Step 2. Set time indicator. Each item in the time indicator represents a date. At first, the time indicator points to the first item and then we select whole data based on the date.

Step 3. All the data selected are classified by "CARDID" to form different groups. Each group represents all the transactions a SC user made on one working day, and the transactions in each group are sorted by "MARKTIME."

Step 4. The time interval between current transaction and previous transaction is calculated. If the time interval is less than 60 min, the destination of current transaction is compared to the origin of previous transaction. If they are different from each other, these two transactions are connected to one transaction and then we continue to connect two transactions into one trip until two transactions cannot satisfy two principles mentioned above.

Step 5. After calculating all the data of different groups on this date, whether time indicator is in the final position is judged. If not, the time indicator moves to the next and then Steps 2–5 is repeated.

A passenger may take several journeys with the same origin and destination at different time of one day, which has a big influence on the analysis. So, this paper defines that passenger travel times is the number of days which has travel records, namely for each person one travel contains all the different trips recorded on one working day. Based on this principle, the number of travel times a passenger takes cannot be larger than the number of working days within one month.

2.2. Passenger Travel Behaviour Analyses in Temporal Dimension. This part adopts DBSCAN algorithm [12, 13] to mine travel patterns. DBCAN is a clustering algorithm based on density, which has great advantage in the following aspects:

(1) If we consider a passenger travels regularly, he must travel by bus several times within a certain time. The DBSCAN algorithm is proper to this data mining. The number of "a certain time" is the *Eps* in the DBSCAN algorithm, and the number of "several times" is the *minPts* in the DBSCAN algorithm.

(2) As we can see, a passenger may take bus to deal with individual random events. These trips have a lot of differences with regular trip, and we call these trips "noise." This algorithm can find the clusters (regular pattern) and deal with the varying noise effectively.

(3) This algorithm can identify a cluster of any shape and size. It means that we can use this algorithm to obtain various travel regular patterns in consideration of travel frequency.

(4) This algorithm does not require the predetermination of initial cores or the number of clusters. This feature is also essential for travel pattern analysis because the number of patterns from an individual passenger is unknown.

(5) Because of the high complexity of the DBSCAN algorithm, this paper extends this algorithm to a distributed platform, which means that we gather SC data based on CARDID to form a group and then calculate passenger's travel data within each group. After this change, we can use a computer cluster to mine SC data. Each computer in the cluster calculates several group data to increase the speed of calculation.

For a D-dimension dataset containing N points: $X = \{x_1 \dots x_i \dots x_n\}$, where $x_i \in R^d$ and DBSCAN algorithm has related definitions as follows:

(1) *Eps* neighbourhood of point x_j: An area with the center x_i and the radius *Eps*.

(2) Density of point x_i: The total number of points within the *Eps* neighbourhood of point x_i. We mark this *density*(x_i), and the number of *density*(x_i) is $|N_{Eps}(x_i)|$.

(3) Core point x_i: If the density of point x_i is no less than threshold *minPts*, point x_i is a core point.

The goal of DBSCAN algorithm is to find out the whole core points in dataset X, and then for each core point x, x makes up a cluster together with all the other points whose distance to x is less than *Eps*. This paper choses 20 min as *Eps* in temporal dimension.

For SC data analysis, *minPts* has great meaning in two sides. On one side, the absolute number of the *minPts* must reach a certain value. If the times of a passenger travelled by bus are too small, the regular passenger travel patterns will not be very clear. The number chosen in this paper is 4, which means a passenger must have travel regular record in at least 4 days among 18 working days. On the other side, *minPts* is a relative value, namely, if a passenger travels regularly, the proportion of regular-travel-day number to all the travel times is more than 50%.

2.3. Passenger Travel Behaviour Analyses in Spatial Dimension. This part also adopts DBSCAN algorithm to analyse passenger travel behaviour. Because the density of the bus station in Beijing is large and the frequency of buses is high, there is no need for passengers to choose another station to board or alight. So the *Eps* for passenger travel behaviour analysis in spatial dimension is chosen by 0 m, and the principles to *minPts* is the same as the principles in temporal dimension above. However, in reality, some passengers sometimes choose to swipe their cards in advance in order to ensure the efficiency of alighting, or swipe their cards later to avoid the crowd near SC inductors. For this phenomenon, this part proposes an algorithm based on frequent items to identify the advanced or postponed records.

Step 6. We gather SC data by CARDID for each user to form a 3-dimension set containing N points. The 3 dimensions are LINEDID marked l, MARKSTATION marked m, and TRADESTATION marked t. Then the dataset for each user can be expressed as $X = \{x_1(l_1m_1t_1) \dots x_i(l_im_it_i) \dots x_n(l_nm_nt_n\}$.

Step 7. The occurrences number of each element in the set is calculated, and the elements appearing more than four times are selected as a candidate frequent set, and the other elements are put into an infrequent set. A new dimension, the occurrences number, is added to the candidate frequent set, so the candidate frequent set can be expressed as $X_f = \{x_i(l_im_it_i) \dots x_j(l_jm_jt_j) \dots x_k(l_km_kt_k\}$ and the infrequent set can be expressed as $X_{if} = \{x_u(l_um_ut_u) \dots x_v(l_vm_vt_v) \dots x_w(l_wm_wt_w\}$.

Step 8. The distance between different elements in the candidate frequent set is calculated, which can be expressed as follows: $DistanceL_{ij} = |x_i(l_i) - x_j(l_j)|$, $DistanceM_{ij} = |x_i(m_i) - x_j(m_j)|$, and $|x_i(t_i) - x_j(t_j)|$. If $DistanceL_{ij} = 0$, $DistanceM_{ij} = 1$, $DistanceT_{ij} = 1$, and $c_i > c_j$, we combine x_i and x_j together to form a frequent set.

Step 9. The distance between each element in the infrequent set and each element in the frequent set is calculated. If $DistanceL_{ij} = 0$, $DistanceM_{ij} = 1$, $DistanceT_{ij} = 1$, we change $x_i(l_im_it_ic_i)$ into $x_i(l_im_it_ic_i + 1)$, and drop item x_j from the infrequent set. At least, we put the infrequent set and the frequent set together.

2.4. Bus Station Clustering and Connection Analysis. In a complex bus transit network, we define the number of different lines passing through a bus station as the node importance d, namely, the lager the node importance is, the more lines the bus station connects. Each bus stop has different functions as a node in the network. Some nodes have small node importance but they are very close to other nodes. Some of these nodes can be put together because of the similar roles they play in a certain area, and then the relationship can be seen between different areas or between an area and some nodes.

To gather different nodes, this paper puts forward an algorithm based on density and node importance. The flow of the algorithm is as follows:

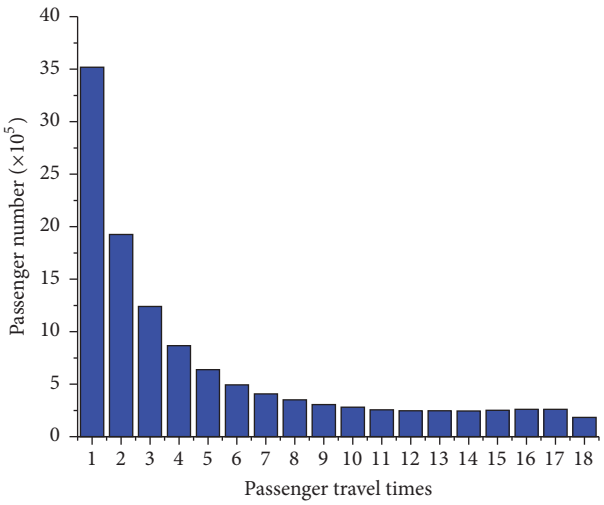

Figure 2: The relationship between passenger number and passenger travel times.

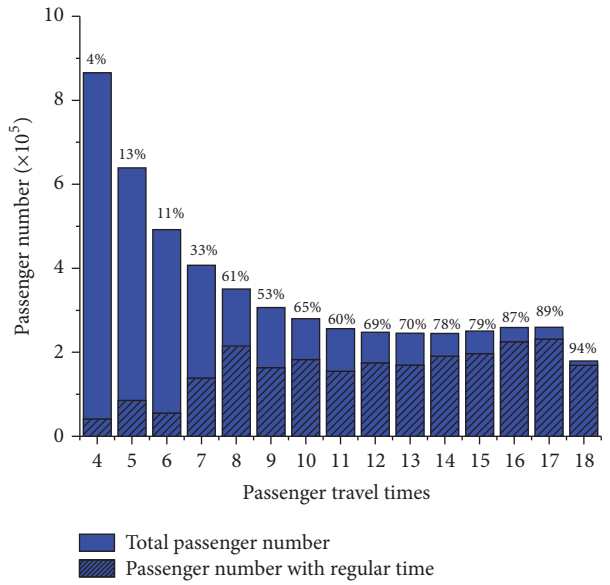

Figure 3: The relationship between passenger number with regular time and travel times.

Step 10. This step is similar with DBSCAN. A 3-dimension set containing all bus stops is built, whose dimensions are node importance, longitude and latitude. The dataset can be expressed as $X = \{x_1(d_1 lon_1 lat_1) \ldots x_i(d_i lon_i lat_i) \ldots x_n(d_n lon_n lat_n)\}$. A threshold *Eps* 1000 m is chosen. For bus stop x_i, if the distance between x_j and x_i, which can be calculated by the latitude and longitude, is less than *Eps*, we define x_j as an appendage to x_i. Then, a new dataset containing each bus stop and its appendages is built. The new dataset can be expressed as $X_e = \{x_1(d_1 x_i \ldots x_n) \ldots x_i(d_1 x_1 \ldots x_n) \ldots x_n(d_j x_1 \ldots x_n)\}$.

Step 11. For bus stop x_i and x_j in dataset X_e, if the node importance d_i is larger than d_j, bus stop x_i will absorb the same elements belong to both x_i and x_j. Then each x_i forms an area and each bus stop appears only once in a certain area.

3. Results and Discussion

After calculating by the trip reconstruction algorithm, this paper finds that the total number of passengers travelled by bus on working days in October is 11966945. Around 30% of passengers travelled by bus on only one working day, around 18% of passengers had travel records on over ten working days. The numbers of passengers show little change when the traveling-working day is from 10 to 17. The details show in Figure 2.

The DBSCAN algorithm is used to analyse the travel behaviour for each passenger and identify whether passengers travel regularly. Passenger number with regular travel time or passenger number with regular travel ODs appears below the sum number of passengers whose trips have a certain *Eps* and *Minpts* (two main factors in DBSCAN algorithm). The travel time or ODs of regular passengers is the core point clustered by DBSCAN algorithm for each passenger.

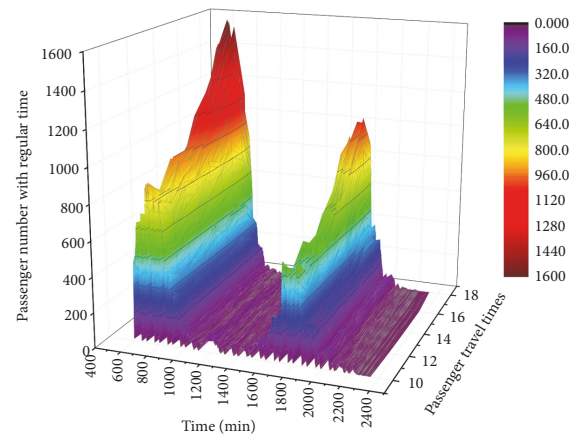

Figure 4: Passenger number variation with regular travel time (X axis) for different time (Y axis) and travel times (Z axis).

According to the DBSCAN algorithm introduced above, passenger travel regularity in temporal dimension is analyzed. Passenger number with regular time for different travel times is shown in Figure 3. As passenger travel times increased, passenger number with regular time decreases, but the proportion of passengers with regular time increases.

Figure 4 shows passenger number variation with regular travel time for different time and travel times (from 10 to 18). The morning peak hour begins at 6:30 am and ends at 9:00 am, and the evening peak begins at 17:00 pm and ends at 18:30 pm. Both passenger numbers with regular time in two peak hours are the largest when the passenger travel times is 17. Figure 5 indicates passenger number variation with regular travel time for different time and regular travel times (from 10 to 18). The two peak hours are the same as Figure 4, but passenger numbers with regular time are the largest when the passenger regular travel times is 10.

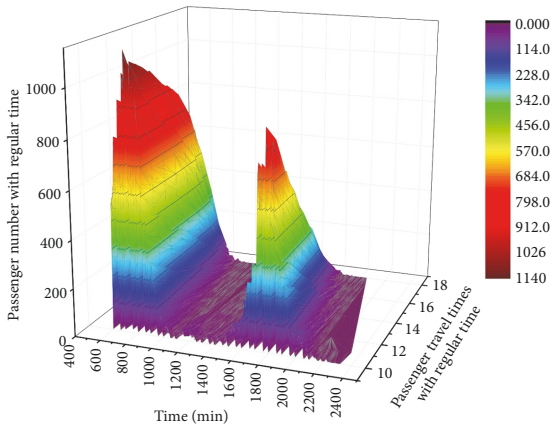

FIGURE 5: Passenger number variation with regular travel time (X axis) for different time (Y axis) and regular travel times (Z axis).

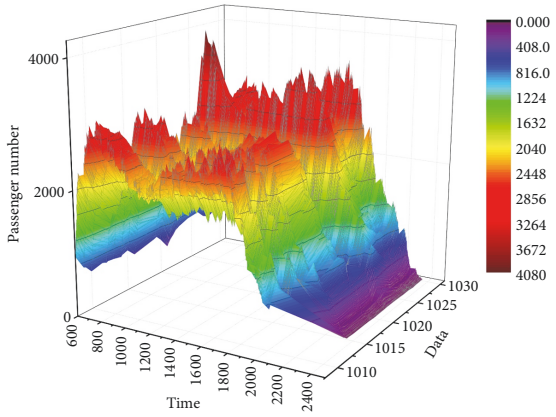

FIGURE 6: Passenger number variation with different time (X axis) and different date (Y axis).

In the temporal dimension, we can see that the morning peak hour begins at 9:00 for irregular passengers, and the number of passenger during a day varies little with travel time (8:00-20:00). This phenomenon is quite different from this of regular passenger. More details show in Figure 6.

According to the DBSCAN and frequent items algorithm introduced above, passenger number with regular travel ODs is calculated. Passenger number with regular travel ODs for different travel times is shown in Figure 7. As passenger travel times increased, passenger number with regular ODs decreases, but the proportion of passengers with regular ODs increases.

Figure 8 shows five origin stops and five destination stops in the morning with the largest passenger number with regular ODs for different travel times (from 10 to 18). Dabeiyao South stop, located in CBD, is the destination stop with the largest alighting passenger number in the morning. Sihuishuniu stop is the destination stop with the largest boarding passenger number in the morning. It is also an origin stop among the top five origin stops. When passenger travel times are over 10, the peak of passenger travel times is 17 and the peak of passenger travel times with regular ODs is 10.

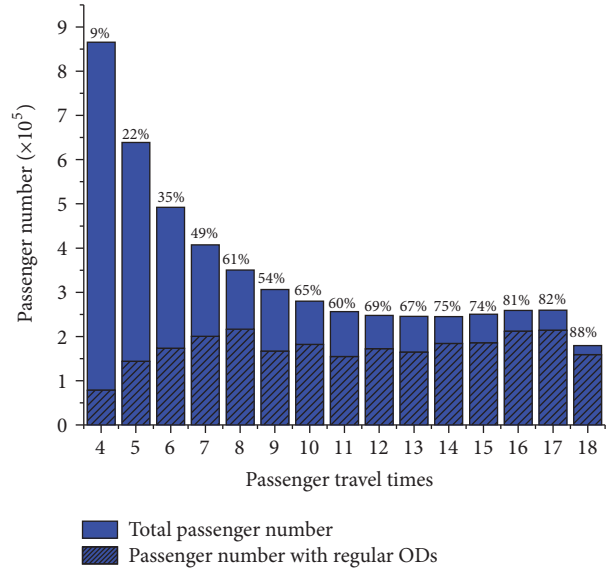

FIGURE 7: Passenger number variation with regular ODs for different travel times.

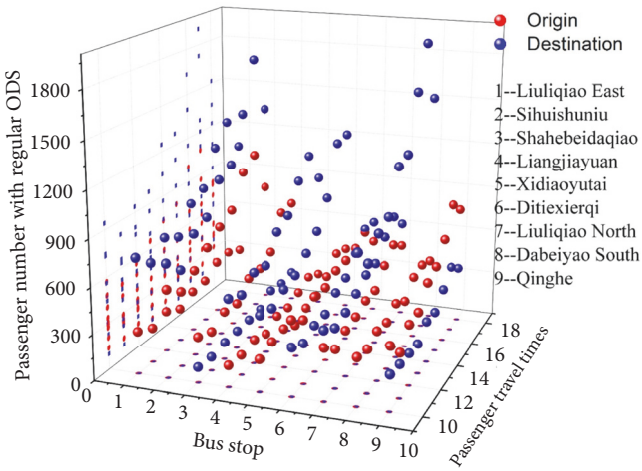

FIGURE 8: Five origin stops and five destination stops (X axis) with the largest passenger number with regular ODs (Z axis) in the morning for different travel times (Y axis).

Figure 9 indicates five origin stops and five destination stops in the evening with the largest passenger number with regular ODs for different travel times (from 10 to 18). The boarding and alighting result in the evening is opposite to that in the morning. The two peaks are the same as that in the morning.

In the spatial dimension, we count the number of irregular passengers at the 5 largest bus stops on ODs per working day. From Figure 10, we can observe that stations1-7 in Figure 10 almost appeared everyday which means that although each passenger traveled irregularly, they had some same destinations. Compared with regular passengers whose target is work area, their destinations are scenic spots, markets, and schools. Station1 (Dongcemen) is close to Tian'anmen Square, and Station2 (Chuangbeixiaoqu) is the

Table 1: Detail data for the top five origin stops and destination stops with the most passengers in the morning.

Stop	Lon[°]	Lat[°]	Passenger number (D)	Passenger number (O)	Stop type
①Sihui	116.4903	39.9052	13228	9295	OD
②Xierqi	116.3014	40.0496	5195	6829	O
③Qinghe	116.3417	40.0290	3630	6601	O
④Shahebeidaqiao	116.2626	40.1290	1820	5178	O
⑤Liuliqiao North	116.3040	39.8887	7078	5022	O
⑥Dabeiyao South	116.4552	39.9038	15842	4755	D
⑦Liuliqiao East	116.3114	39.8865	11415	3837	D
⑧Liangjiayuan	116.4641	39.9071	10051	2405	D
⑨Xidiaoyutai	116.2936	39.9226	9122	3883	D

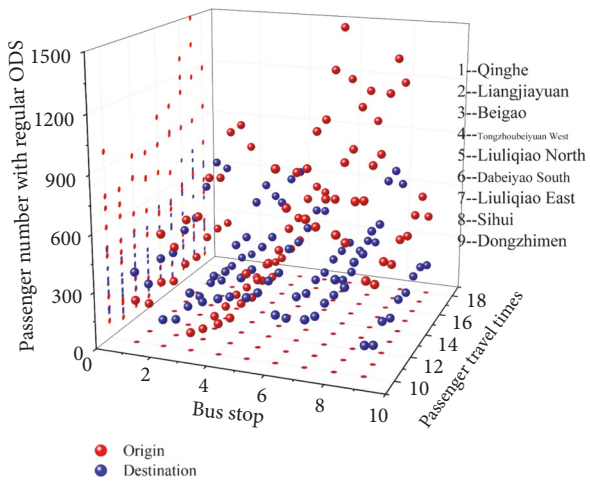

Figure 9: Five origin stops and five destination stops (X axis) with the largest passenger number with regular ODs (Z axis) in the evening for different travel times (Y axis).

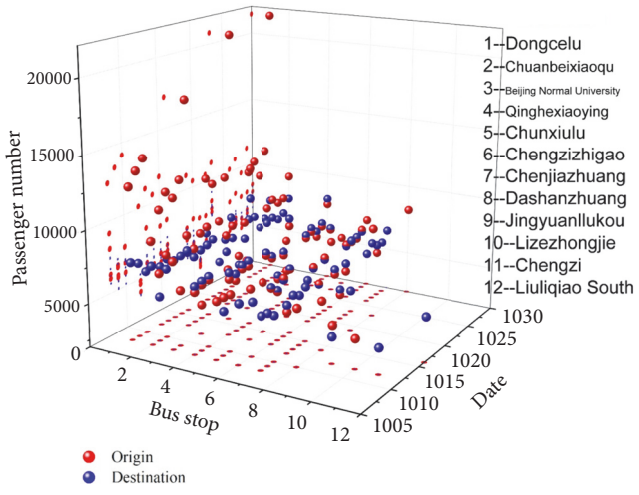

Figure 10: Five origin stops and five destination stops (X axis) with the largest passenger number with regular ODs (Z axis) for different date (Y axis).

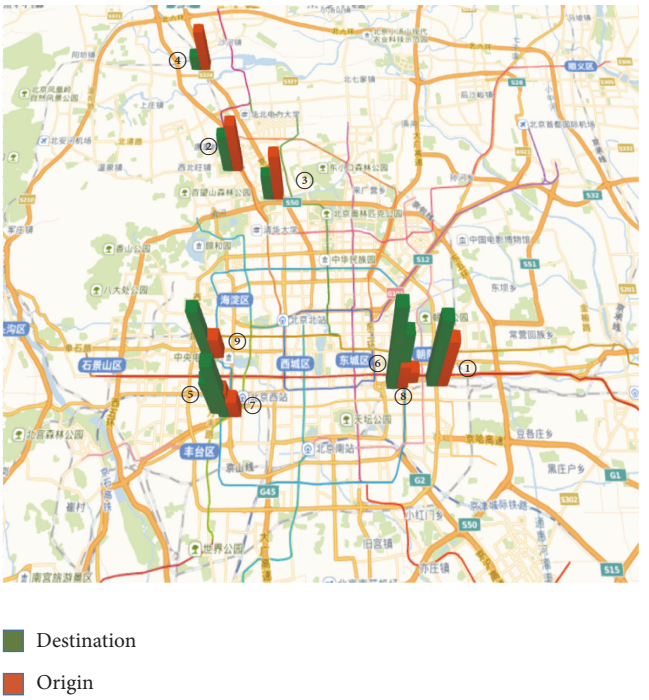

Figure 11: The top five origin stops and destination stops with the most passengers in the morning.

main site for the Great Wall. The number of irregular passengers whose destination is Station1 (Dongcemen) changes a lot during different days. This phenomenon is quite different from that of regular passengers.

This paper distributes passengers with regular ODs according to their travel time (morning or evening) and ODs. Then we chose the top five origin stops and destination stops with the most passengers for a detailed analysis. Sihui station belongs to both the top five origin stops and the destination stops. The distribution and detail data of the top five stops in the morning shows in Figure 11 and Table 1.

From Figure 12 and Table 2 we can see that the top 5 destination stops are located in the 3rd ring and three of the top 5 origin stops are located outside the 5th ring. The

Table 2: detail data for the top five origin stops and destination stops with the most passengers in the evening.

Stop	Lon[°]	Lat[°]	Passenger number (D)	Passenger number (O)	Stop type
①Dabeiyao South	116.4552	39.9038	3703	14671	O
②Liuliqiao East	116.3114	39.8865	4091	12597	O
③Sihui	116.4903	39.9052	7459	12259	OD
④Liangjiayuan	116.4641	39.9071	1982	9707	O
⑤Dongzhimen	116.4302	39.9408	2729	8239	O
⑥Tongzhoubeiyuan	116.6337	39.9051	5118	2426	D
⑦Qinghe	116.3417	40.0290	5457	3022	D
⑧Liuliqiao North	116.3040	39.8887	5101	6730	D
⑨Beigao	116.5063	40.0101	4589	1121	D

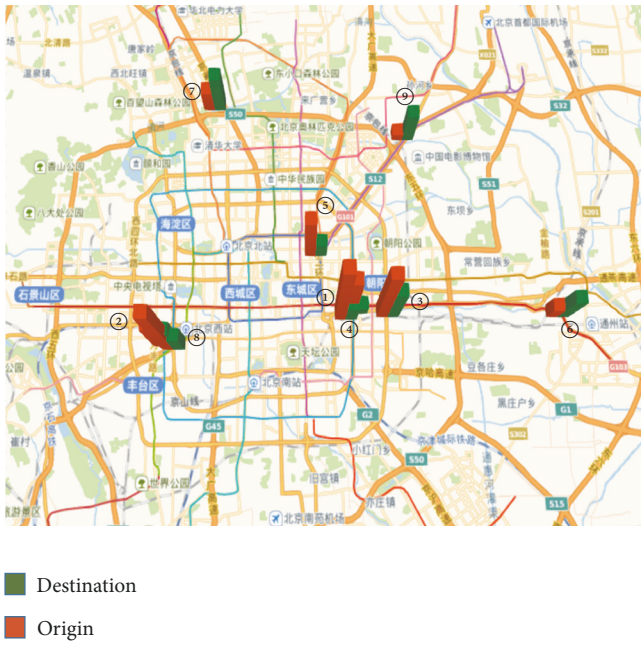

Figure 12: The top five origin stops and destination stops with the most passengers in the evening.

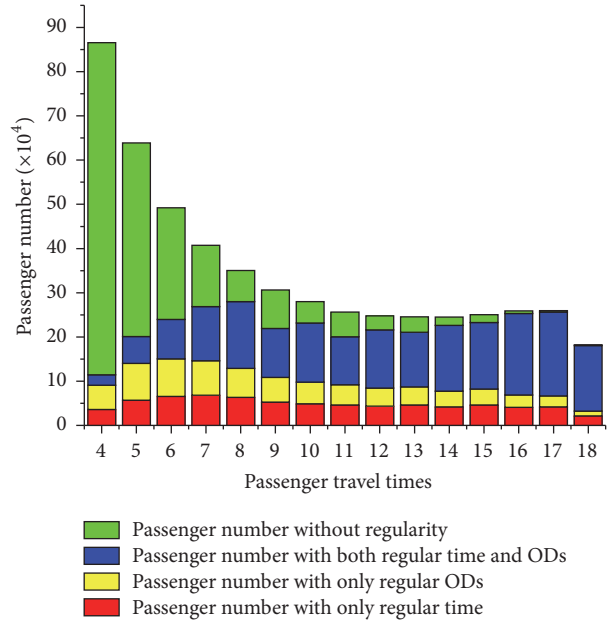

Figure 13: The relationship between four type passenger number and passenger travel times.

passenger number with regular ODs in the top 5 destination stops is far bigger than that in the top 5 origin stops, which means the destination of the passengers is concentrated and the origin of the passengers is dispersed. For Sihui and Liuliqiao North, although they are in the top 5 origin stops and their passenger number with regular destinations is bigger than that with regular origins, which means these two stops are main exchange points for passengers in and outside the 3rd ring. Passenger travel behaviour in the evening is quite opposite to that in the morning. The distribution and detail data of the top five stops in the morning show in Figure 12 and Table 2.

According to the analysis in both temporal and spatial dimension, different types of passengers can be obtained. Some passengers travel only with regular time, some of them travel only with regular ODs, and some of them travel regularly in both dimensions, while others travel without regularity. Passenger number for these four types is shown in Figure 13. We can see that as passenger travel times increased, passenger number without regularity reduced. When the passenger travel times is over 17, only 1% of passengers travelled without regularity.

This paper clusters the bus stations. As a result, more than 2000 areas are identified containing around 7000 bus stops. According to the result of area clustering, this paper analyses the top five origin areas and destination areas with the most passengers and the traffic connection between these areas and other areas in the morning and evening. In the morning, passenger number with regular destinations at the top 5 destination areas is larger than that at the top 5 origin areas, which indicates passengers came from a lot of different areas have several same destinations. The result in the evening is quite opposite to that in the morning. The distribution of the areas shows in Figures 14 and 15, and the detailed data shows in Tables 3 and 4.

Table 3: Detail data for the top five origin areas and the top five destination areas in the morning.

Core Stop	OD Type	Passenger number with regular origins	Passenger number with regular destinations
①Liuliqiao North	OD	21902	35248
②Dabeiyao East	OD	20900	57376
③Qinghe	O	19044	11036
④Yanhuang museum	O	17820	21272
⑤Sihui	O	17744	27048
⑥Zhongguancun South	D	11256	36731
⑦Dongsanqi South	D	12478	28126
⑧Sanyuanqiao	D	17035	27860

Table 4: Detail data for the top five origin areas and the top five destination areas in the evening.

Core Stop	OD Type	Passenger number with regular origins	Passenger number with regular destinations
①Dabeiyao East	OD	52139	19781
②Liuliqiao North	OD	32129	16167
③Zhongguancun South	O	23828	9896
④Sihui	O	23350	13759
⑤Dongzhimen	O	18466	11463
⑥Tongzhoubeiyuan East	D	13350	17211
⑦Qinghe	D	9946	15493
⑧Tongzhoubeiyuan West	D	4732	13790

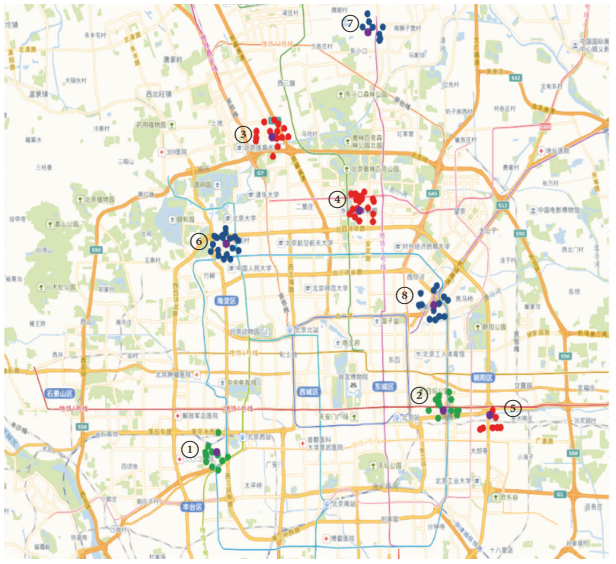

Figure 14: The top five origin areas and the top five destination areas with the most passengers in the morning.

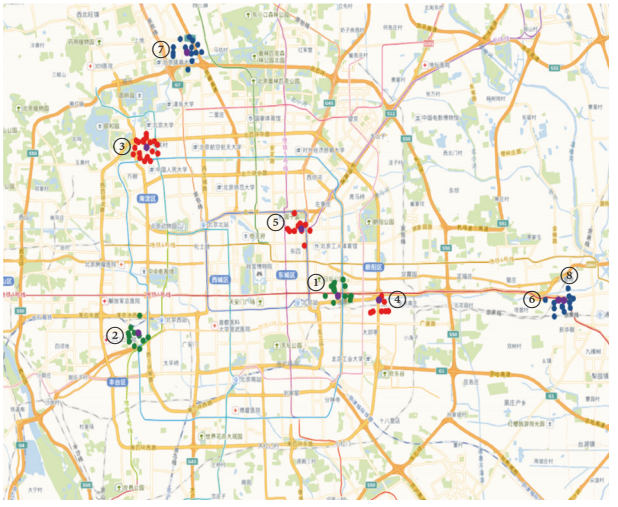

Figure 15: The top five origin area stops and the top five destination area stops with the most passengers in the afternoon.

The paper studies the traffic links between different areas according to the OD data. There are some interesting conclusions based on the study. Among the passengers whose regular destination is CBD area (core stop is Dabeiyao East), 5% and 4.5% of them came from Tongzhoubeiyuan East area and Tongzhoubeiyuan West area, respectively, which are the most closed two areas connecting to CDB. In the evening, 6.6% and 5.4% of the passengers returned to Tongzhoubeiyuan West area and Tongzhoubeiyuan East area, respectively, from CBD area. The distance between CBD and Tongzhoubeiyuan is around 15 km. 24.0% of passengers whose regular origin

Table 5: The top five origin and destination areas in the morning and the most closely connected area to them.

Top5 areas	OD Type	Boarding (alighting) passenger number	Closest connecting Area	Alighting (boarding) passenger number	Dist(m)	Prop(%)
Liuliqiao North	O	21902	Yungang	820	15651	3.7
Dabeiyao East	O	20900	Ritanlu	831	1955	4.0
Qinghe	O	19044	Chengfulu South	1200	4431	6.3
Yanhuang Museum	O	17820	Anzhenqiao West	1158	3268	6.5
Sihui	O	17744	Sihui	4256	0	24.0
Dabeiyao East	D	57376	Tongzhou beiyuan East	2810	14954	4.9
Zhongguancun South	D	36731	Beijing Sport University	1992	4391	5.4
Liuliqiao North	D	35248	Gungang	1857	15651	5.3
Dongsanqi South	D	28126	Tiantong beiyuan	2766	2189	9.8
Sanyuanqiao	D	27860	Beigao	1520	7752	5.5

area is Sihui in the morning went to other stops within itself. 70% of passengers whose regular destination is Dongsanqi South in the morning came from several areas which are located within 4 km around Dongsanqi South area. This phenomenon means Dongsanqi South area is a core traffic area gathering a lot of passengers from other areas to take the subway to the city center. More detailed data shows in Table 5.

4. Conclusions

In this paper, four algorithms are used to analyze the temporal and spatial regularity of passengers traveled by bus based on the large scale data of SCs and the traffic relationship between different traffic areas. At first, this paper proposes a trip reconstruction algorithm gathering SC data by CARDID to improve the calculation efficiency using SPARK platform and analyses the times of passengers traveled by bus, that is, the number of days which have travel records in 18 working days. The proportion of passengers with different travel times comes out based on this study. In the temporal dimension, the proportion of passengers who traveled regularly in temporal dimension is obtained and the relationship between this proportion and the times passengers traveled by bus is also described. In the spatial dimension, this paper proposes a data recognition algorithm based on frequent terms to improve the accuracy of SC data and draws some conclusions similar to that in the temporal dimension. According to the temporal and spatial regularities of passengers, passengers are divided into four types: passengers only with regular travel time, passengers only with regular ODs, passengers with both regular travel time and regular ODs, and passengers without regularity. The number of four type of passengers is also obtained. The paper divides the bus area according to the distance between different bus stops and node importance, mainly analyses the passengers with both regular travel time and regular ODs, and determines the traffic connection between different areas.

Data Availability

The data used in this paper came from Beijing public transportation group. This data only can be used in scientific research with permission. There is no access to a public database or web site.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by National Key Technologies Research & Development program (2017YFC0804900).

References

[1] C. Morency, M. Trépanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transport Policy*, vol. 14, no. 3, pp. 193–203, 2007.

[2] K. K. A. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transportation Research Record*, no. 2063, pp. 63–72, 2008.

[3] K. K. A. Chu and R. Chapleau, "Augmenting transit trip characterization and travel behavior comprehension: multi-day location-stamped smart card transactions," *Transportation Research Record*, vol. 2183, pp. 29–40, 2010.

[4] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," *IEEE Network*, vol. 28, no. 4, pp. 32–39, 2014.

[5] S. Tao, J. Corcoran, I. Mateo-Babiano, and D. Rohde, "Exploring Bus Rapid Transit passenger travel behaviour using big data," *Applied Geography*, vol. 53, pp. 90–104, 2014.

[6] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016.

[7] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1537–1548, 2015a.

[8] M. K. El Mahrsi, E. Côme, L. Oukhellou, and M. Verleysen, "Clustering Smart Card Data for Urban Mobility Analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, 2016.

[9] X. Ma, C. Liu, H. Wen, Y. Wang, and Y. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.

[10] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 274–289, 2017.

[11] C. Seaborn, J. Attanucci, and N. H. M. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," *Transportation Research Record*, no. 2121, pp. 55–62, 2009.

[12] L. Duan, L. Xu, F. Guo, J. Lee, and B. Yan, "A local-density based spatial clustering algorithm with noise," *Information Systems*, vol. 32, no. 7, pp. 978–986, 2007.

[13] X. Xu, J. Jäger, and H. P. Kriegel, "A fast parallel clustering algorithm for large spatial databases," *Data Mining and Knowledge Discovery*, vol. 3, no. 3, pp. 263–290, 1999.