Travel Time Estimation and Urban Key Routes Analysis Based on Call Detail Records Data: A Case Study of Guangzhou City

Weimin Mai[1,2], Shaohang Xie[1,2], and Xiang Chen[1,2(✉)]

[1] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
chenxiang@mail.sysu.edu.cn

[2] Key Lab of EDA, Research Institute of Tsinghua University in Shenzhen (RITS), Shenzhen 518075, China

Abstract. Nowadays, the study of urban traffic characteristics has become a major part of city management. With the popularization of the mobile communication network, the call detail records(CDR) have become important resources for the study of urban traffic, containing abundant temporal and spatial information of urban population. To excavate the traffic characteristics of Guangzhou city, this paper focuses on two aspects: travel time estimation and the analysis of key routes in urban area. First, we propose a method of estimating urban travel time based on traffic zones division using the traffic semantic attributes. According to the features of users flow extracted from CDR, we determine the traffic semantic attributes of the areas covered by base stations. With these semantic attributes, we cluster the cell areas into several traffic zones using a K-means method with a weighting dissimilarity measure. Then travel time between different positions in Guangzhou is estimated using the key locations of traffic zones, with an accuracy of 67%. Furthermore, we depict the key routes of Guangzhou city utilizing a DBSCAN method with the users' trajectories extracted from the CDR data. The results of obtained routes are validated by actual traffic conditions and provide some extra discoveries. Our works illustrate the effectiveness of CDR data in urban traffic and provide ideas for further research.

Keywords: Travel time estimation · Urban key routes · Call detail records

1 Introduction

The management of transportation has become a fundamental part in building a smart city. Traditional measurement of traffic conditions relies on on-road sensors. However, for most of cities, such real-time sensor systems are not sufficient

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 659–676, 2019.
https://doi.org/10.1007/978-3-030-32388-2_55

due to their limited coverage and expensive cost. The quantity of mobile users is increasing that even in a small city there are huge numbers of mobile phone users. By 2013, mobile phone penetration rates have reached 128% and 89% in developed and developing countries, respectively [23]. CDR data in mobile cellular network contains abundant spatial and temporal information of urban populations since users in CDR data can be regarded as samples among the whole population of the city and the locations in CDR can be regarded as the samples of complete and continuous trajectories. The cost of collecting and storing CDR data is much lower than building an on-road sensor system, which motivates us to explore the generalizable methods of extracting traffic features based on CDR data.

Studies in many traffic aspects have been conducted using CDR data such as traffic flow estimation [4,12,28], mobility pattern analysis [5,8], daily trajectories analysis [18,19]. In this paper, we will mainly focus on utilizing CDR data to excavate traffic characteristics of urban area from two aspects: travel time estimation and the analysis of urban key routes. Both the knowledge of travel time between different locations of the city and the understanding of key routes bring us effective ways to find and settle the problems of urban congestion, providing strategies on planning urban roads. The awareness of commuting time can also guide urban residents on some daily services such as navigation and route selection. In addition, advertising and commercial site selection can be more targeted along these hot routes.

In this paper, Our contributions consist of two aspects:

- First, we propose a method of estimating urban travel time based on traffic zones division using the traffic semantic attributes. We associate the traffic time estimation with traffic zones, taking advantages of the inter-zone and intra-zone traffic characteristics. However, instead of dividing traffic zones with geographic information only, we introduce traffic semantic attributes obtained from features of users flow extracted from CDR into the process of traffic zone division, which makes the traffic zones more consistent with traffic ground true. Thus the estimation of travel time can be more convincing.
- Secondly, we apply a DBSCAN method to discover the key routes of Guangzhou city with the users' trajectories extracted from the CDR data. The obtained routes are validated by actual traffic conditions.

The rest of the paper is organized as follows. In Sect. 2, we take an overview of the study area and the data we use, and describe the process of data preprocessing. In Sect. 3, we determine the peak/off-peak hours of the city, which serve as the target periods of the following work. In Sect. 4, we explain how to estimate the travel time with traffic zones dividing method based on traffic semantic attributes and then evaluate the accuracies of the results. In Sect. 5, the key routes of the city are generated and analyzed according to actual traffic conditions. In Sect. 6 we will make a conclusion to the whole paper.

2 Data Description and Preprocessing

2.1 CDR Data

In this study, CDR of 3 weeks in May 2018 in Guangzhou are used, including Connection Management(CM) and Mobility Management(MM) records. The privacy-related information such as IMSI and cell-phone number have been encrypted and mapped into some certain strings corresponding to the original records.

Data cleaning is conducted because records in the original data may miss information in some fields and not all fields are useful for traffic study. The most important information of traffic analysis mainly includes three aspects: person, time and place. Therefore, we omit some fields other than IMSI, time and base station location and delete those records with missing fields. Then all remaining records are reformatted as shown in Table 1.

Table 1. Reformatted CDR

User ID	*Time*	*Cell ID*	*Longitude, Latitude*
001	2018-05-17 23:14:50	14	(113.2896, 23.0303)
002	2018-05-18 00:16:59	20	(113.2880, 23.0199)
...	...	...	...

The field *Time* is the middle instant of the duration of a record, representing when this record occurs. Note that we cannot know the exact location of a user when he generates a record, but can only estimate his location roughly through the location of the base station he accesses. The coverage radius of a base station in dense areas ranges from tens of meters to hundreds of meters, while the coverage radius in suburban areas can be several kilometers.

2.2 Study Area

The study was conducted in Guangzhou, a megacity in China with area more than 7,000 square kilometers and a population over 10 million.

If the whole city is considered into the study, some traffic characteristics tend to be covered by the average effect of the surrounding counties. Besides, traffic flow of a mega city may differ a lot between weekdays and weekends. Therefore, we only consider records of weekdays and limit the study area to downtown area of Guangzhou which contains 2225 base stations.

2.3 Data Preprocessing

The total number of valid reformatted records is about 1.57 billion. Since our purpose is to study the traffic characteristics such as travel time and key routes,

we should first extract the OD (origin-destination) records that reflecting the users' location changing and the corresponding time from the CDR data. We group the reformatted records by $UserID$ and sort them in ascending order according to $Time$. An OD record is generated from two adjacent ordered CDRs with the same user but different locations, which is shown in Table 2.

Table 2. OD records

User ID	*Origin location*	*Destination location*	*Occurrence time*	*Duration* (*min*)
033	(113.5054, 23.0655)	(113.5251, 23.0674)	2018-05-21 17:57:31	31
...	...	...	...	...

In Table 2, $OccurrenceTime$ is the middle instant of the $Time$ of two adjacent ordered CDRs, which is used to describe which period the OD records belong to. *Duration* means the difference of the $Time$ of adjacent ordered CDRs.

Not all the OD records obtained above can reflect traffic characteristics because of the noise. On the one hand, duration or travel speed of some OD records deviate from the actual urban traffic conditions seriously. For example, *Duration* in an OD record can be several hours but travel time between the furthest two positions in a city cannot be more than 130 min. Such a record can only indicate that the user has moved but can provide little valuable information of the actual travel time. On the other hand, although a user did not take a travel, his positions may switch repeatedly among several adjacent base stations when the signal received by the user were at a low level. OD records generated from these conditions are useless to the study of urban traffic, so they are filtered out by the steps listed below:

- Delete the OD records whose *Duration*s are shorter than 5 min or longer than 130 min in consideration of the size of Guangzhou.
- Delete the OD records whose speeds are less than 5 km/h or more than 90 km/h in consideration of the traffic speed defined by the ratio of the linear distance between two positions to the duration.
- Delete the OD records generated by a user who switching repeatedly among several base stations.

About 280 million valid OD records are obtained after the above steps, which will become the basis of the following sections.

3 Determine Peak/Off-Peak Hours

Traffic characteristics vary at different slots in a day. The study on travel time and key routes is more valuable when we focus on specific periods instead of the whole day. Thus, in this section, we first determine the peak and off-peak hours of the city based on OD records.

An index proposed in [10] is adopted. A day is divided into 24 hour-slots, for each $slot_i$, a ratio is calculated as follows:

$$ratio_i = \frac{count\,(min)}{t_{ci}}. \tag{1}$$

t_{ci} is the total number of ODs in $slot_i$, while min_i corresponds to the ODs whose *Duration* are almost the shortest among the ODs with the same *OriginLocation* and *DestinationLocation*. The ratio should be smaller in peak hours than in off-peak hours. The results of the each slot arranged in ascending order are shown in Table 3.

Table 3. OD conditions in each hour-slot

slot	$count\,(min)$	t_{ci}	$ratio_i$
8	2123878	17826920	11.914%
18	2297994	18782738	12.235%
7	1664880	13433255	12.394%
17	2276326	18251610	12.472%
19	1845927	14795896	12.476%
12	1910879	15294000	12.494%
13	1822975	14498632	12.573%
9	2136199	16787798	12.725%
11	2013125	15644336	12.868%
20	1701205	13209582	12.879%
16	2127293	16494872	12.897%
6	790046	6076080	13.003%
14	1988096	15247447	13.039%
10	2055462	15763727	13.039%
15	2038210	15525905	13.128%
21	1626446	12379002	13.139%
22	1392017	10354612	13.443%
1	535763	3855155	13.897%
0	704587	5060986	13.922%
23	1009759	723423	13.958%
5	451215	3157483	14.290%
2	444051	3079807	14.418%
3	409281	2822401	14.501%
4	405517	2709823	14.965%

From the ratios of all hour-slots, we can conclude three peak periods of the city, which are consistent with our daily knowledge. Morning peak is defined as

slot7-slot9, noon peak is defined as slot11-slot13 and evening peak is defined as slot17-slot19. The off-peak hours mainly lie in early morning.

4 Urban Travel Time Estimation

Traffic patterns vary from region to region. Based on the inter-features and intra-features of two different traffic zones, a method of estimating travel time was proposed in [10]. However, the method they used to divide traffic zones focused mainly on geographical proximity but ignored the information on traffic. In this section, we will introduce a method to estimate travel time based on traffic zones division with traffic semantic attributes.

4.1 Traffic Zones Division

To estimate the travel time, we should divide the traffic zones in advance. A traffic zone can be regarded as the combination of a group of cell-areas covered by adjacent base stations. The main idea of dividing the traffic zones is to extract the traffic features of the cell-areas and then partition the cell-areas into different clusters using these features. Dong et al. [7] proposed a method to cluster cell-areas according to the semantic attributes of the base stations, which is adopted in our work.

First, we idealize the coverage of the base stations, considering that they do not overlap with each other. Then, the coverage area of a base station can be represented by its corresponding Voronoi polygon [9], and a traffic zone is composed of several adjacent polygons. Part of the Voronoi plot is shown in the Fig. 1. The polygons of solid line are the areas covered by the base stations and the blue dots in the polygon represent the position of the base stations.

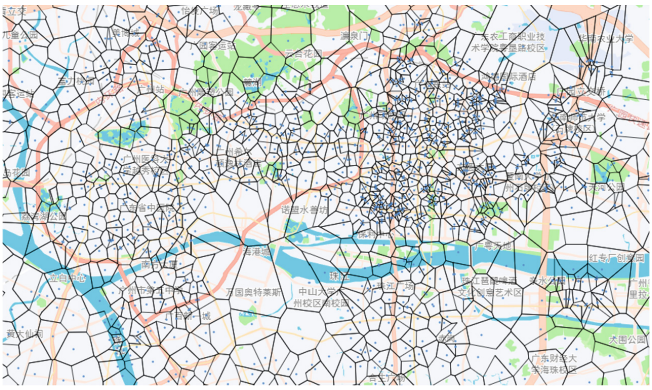


Fig. 1. Voronoi polygons division (Color figure online)

As mentioned above, we need to extract the traffic semantic attributes of cell-areas as one of the characteristics before dividing the traffic zones. This

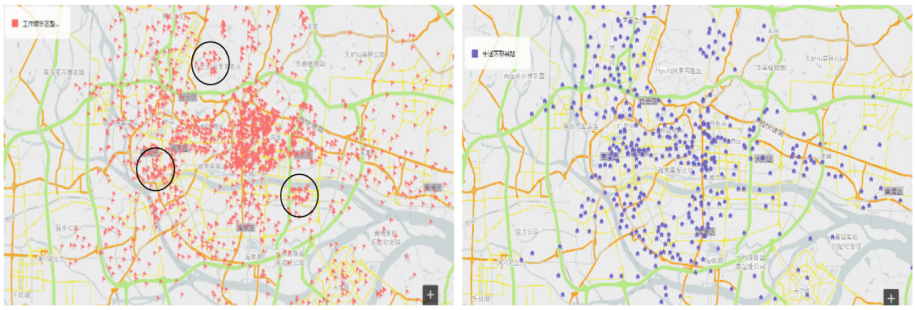
attribute is used to describe the main function of a cell-area in urban traffic networks. Traffic features such as real-time volume(V_t), hourly inflow(V_i), hourly outflow(V_o), hourly increment flow(V_{inc}), peak and valley values are considered when we use K-means methods to cluster all cell-areas into two categories of traffic semantics: working-entertainment areas and residential areas [7]. A 104-dimension feature vector $x_i = \left(x_i^1, x_i^2, \cdots, x_i^{104}\right)$ is constructed for every cell-area, the meaning of each dimension of the vector is described in the Table 4.

Table 4. Vector to determine traffic semantic attributes

Demension	Description
$\left(x_i^1, x_i^2, \cdots, x_i^{24}\right)$	Slot V_t, generated according to the total number of different users in a cell-area of each slot
$\left(x_i^{25}, x_i^{26}, \cdots, x_i^{48}\right)$	Slot V_i, generated according to the number of incoming users in a cell-area of each slot
$\left(x_i^{49}, x_i^{50}, \cdots, x_i^{72}\right)$	Slot V_o, generated according to the number of outgoing users in a cell-area of each slot
$\left(x_i^{73}, x_i^{74}, \cdots, x_i^{96}\right)$	Slot V_{inc}, generated according to the increment number of users in a cell-area of each slot, calculated from $V_{inc} = V_i - V_o$, a negative result is valid
$\left(x_i^{97}, x_i^{98}, \cdots, x_i^{104}\right)$	Peak values and valley values of the four kinds of features above

Results of two kinds of attributes are shown in Fig. 2. The base stations of the first attribute occupy a high proportion in Central Business District and other typical working-entertainment areas like Pazhou Exhibition Area, Baiyun Mountain Tourist Area and Beijing Road Business Area. Therefore, the base stations in the first cluster are tagged with working-entertainment zone attribute. By contrast, the distribution of base stations in the second cluster is not as concentrated as that of the first cluster. However, their proportion in typical residential areas like Liwan District and Baiyun District is relatively high so the base stations in the second cluster are tagged with residential attribute. These semantic attributes of cell-areas will be used as a dimension in feature vector when merging the cell-areas into traffic zones.

We believe that a traffic zone can be formed by connecting the adjacent cell-areas. Therefore, for traffic zones division, the most important factors that influence the results are geographical latitude and longitude, that is, the closer the cell-areas are, the more likely they are divided into the same zone. However, it is obviously unreasonable to only rely on the geographical latitude and longitude information, since this may ignore some information that matches with the actual traffic conditions. If a group of cell-areas is mostly of residential attribute while another group is mostly of entertainment attribute, it is more appropriate to divide them into different traffic zones. When the traffic attributes are taken into



(a) Cells of working-entertainment areas (b) Cells of residential areas

Fig. 2. Traffic semantic attributes

consideration, the boundary of traffic zones can be more diversified to match the more complex actual traffic conditions.

Using geographical longitude and latitude, traffic semantic attributes and the difference of average V_{inc} between morning peak and evening peak, we construct a 4-dimension feature vector for each cell area, where the traffic semantic attribute is either 0 or 1 (0 means the working-entertainment area and 1 represents the residential area). Then a K-means clustering method with a weighting dissimilarity measure [7] was applied to acquire traffic zones.

The K-means method needs to define the number of traffic zones to be divided in advance. Different values of K will lead to different division results. As for a medium or large city, it is generally divided into 50–100 traffic zones [26]. Since we only study on the downtown area of Guangzhou, the number of traffic zones is reduced to 20–60. However, method in [7] does not explain how to evaluate the clustering results to find the appropriate parameters. When evaluating whether a result is effective to estimate the travel time, the primary consideration is whether both the numbers of OD records within and between traffic zones are appropriate. A ratio described in [10] is chosen to determine the proper number K of traffic zones, which is explained by Eq. (2). The smaller the ratio is, the more appropriate the division result is.

$$Ratio = \frac{count(intraTravel) + count\,(invalidInterTravel)}{count\,(totalTravel)}. \tag{2}$$

As shown in Fig. 3, the base stations are regarded as points on a two-dimensional plane. Suppose $Z1$ and $Z2$ are two traffic zones, lines with red tags are $interTravel$s while lines with blue tags are $intraTravel$s. If the number of the $interTravel$s between two zones is less than a threshold value, then all these $interTravel$s are called $invalidInterTravel$s.

The ratio results are shown in Fig. 4. Finally, K = 37 is chosen as the parameter of K-means algorithm and the city is divided into 37 traffic zones as shown in Fig. 5. For each traffic zone, a major junction within it is selected as the key

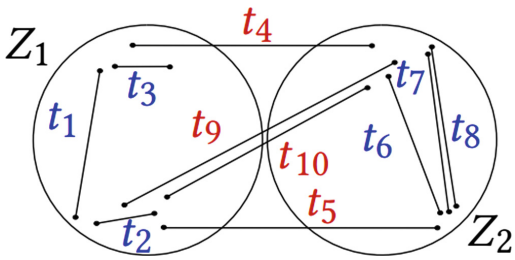


Fig. 3. *InterTravel* and *intraTravel* [10] (Color figure online)

location of this zone. The blue markers represent the original clustering centroids and the red car-shape markers are the selected key location.

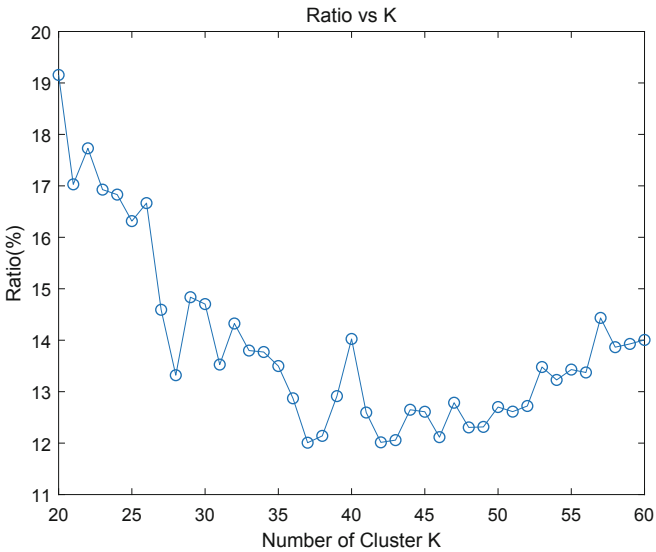


Fig. 4. Ratio of different K

4.2 Travel Time Estimation

By the steps above, we have divided the Guangzhou City into several traffic zones with traffic key locations. Now we apply the travel time estimation algorithm using key locations [10] to estimate the travel time between two positions in the city. The algorithm is explained in Algorithm 1.

Tables 5 and 6 show the $interTime_{avg}$ in morning peak and evening peak. Due to space limitation, we select only the first 13 zones to display. Some cells

(a) Study area (b) Traffic zones (c) Selected key locations

Fig. 5. Results of traffic zones division (Color figure online)

Algorithm 1. Travel time estimation using key locations

1: Calculating $time(C_i, C_j)$ for each $interTravel$ with origin $P_i \in Z_i$ and destination $P_j \in Z_j$, where C_i and C_j are the key locations of Z_i and Z_j respectively:
$time(C_i, C_j) = time(P_i, P_j) \times \frac{Dist(C_i,C_j)}{partialDist(P_i,P_j)+shortestDist(P_i,C_iC_j)+shortestDist(P_j,C_iC_j)}$
2: For each pair of distinct combination of C_i and C_j, averaging all corresponding $time(C_i, C_j)$ obtained above to get $interTime_{avg}(Z_i, Z_j)$;
3: Calculate $intraSpeed_{avg}(Z_i)$ of each zone Z_i using $intraTravel$ corresponding to that zone;
4: Estimating travel time between any two positions $P_i \in Z_i$ and $P_j \in Z_j$:
$time_{est}(P_i, P_j) = \frac{shortestDist(P_i,C_iC_j)}{intraSpeed_{avg}(Z_i)} + interTime_{avg}(Z_i, Z_j) \times \frac{partialDist(P_i,P_j)}{dist(C_i,C_j)} + \frac{shortestDist(P_i,C_iC_j)}{intraSpeed_{avg}(Z_i)}$

are marked as "/", which denotes there are no sufficient OD records to estimate the travel time between the corresponding zones.

As can be seen from Tables 5 and 6, most $interTime_{avg}$ between two certain zones differ little in different peaks. However, for some specific zones, there are obvious differences between morning and evening peak. For example, it takes only 23 min in the morning peak hours while about 57.6 min in the evening to travel from Zone3 (lies in the center of the city) to Zone19 (lies in the northwest of the city). This result is consistent with the actual traffic conditions in Guangzhou that during the evening peak hours, the roads from the center of the city to Baiyun District are seriously congested due to the people returning home from work.

Table 7 shows the $intraSpeed_{avg}$ of the traffic zones. We can see that $intraSpeed_{avg}$ is relatively stable in each zone during different peaks.

To evaluate the effectiveness of this method in our study, we estimate the travel time of 30000 location pairs generated randomly in our study area, and then compare the results with the estimation in AMAP API. When the time deviation is less than 20%, the estimation of travel time is regarded to be accurate. The final accuracy of our method is about 67%. Inaccurate estimations occur mainly in the conditions that the two positions of the pair are very close to each other(travel time is less than 7 min) or they lie in two remote zones (travel time is more than 90 min).

Table 5. *interTime* of morning Peak. Unit: minute

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	/	26.4	23.7	25.2	21	47.9	/	/	36.7	17	/	41.6	/
1	19.1	/	25.4	26.6	16.7	30.7	/	34.9	21.7	18.2	25.2	22.9	19.1
2	/	27.4	/	31.4	24.8	43.3	29.3	31.3	26.1	23.9	25.1	26.8	/
3	60.5	33.5	34.9	/	36.1	30.6	89.8	33.3	32.3	27.6	34.1	36.8	60.5
4	31.8	16.3	22.8	28.3	/	32	/	26.5	25	19	26.9	18.6	31.8
5	49.5	37.1	37.8	31.4	36.5	/	37.4	56.2	23.6	21.9	48.2	30.7	49.5
6	/	31.4	33.1	52	/	/	/	19.6	/	/	35.1	/	/
7	/	30.5	29.1	28.4	26.9	52.3	20.5	/	24.4	/	25.1	53.8	/
8	14.9	26.2	30.1	24.5	28.3	22.5	/	24	/	18	46.1	27.9	14.9
9	18.2	20.2	23.6	22.9	16.8	39.1	/	25.7	27.5	/	/	26	18.2
10	/	24.4	21.8	29.4	24.8	36.5	30.8	44.9	32.3	30.8	/	30.7	/
11	36.7	23.7	23.5	34.4	21	32.8	/	55.5	26.8	19.5	30.1	/	36.7
12	/	26.4	23.7	25.2	21	47.9	/	/	36.7	17	/	41.6	/

Table 6. *interTime* of evening Peak. Unit: minute

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	/	52.5	/	25.2	20.3	45.4	/	/	35.8	16.8	/	46	/
1	14.7	/	26.1	35.7	16.7	32.7	/	32.9	22	19.7	25.7	23.7	31.7
2	23	27.1	/	34.6	23.7	45.1	31	31.9	26.1	24.7	24.3	26.7	14.9
3	55.1	38.4	36.9	/	42.6	29.7	90.6	38.1	36.4	23.7	35.2	46.5	37.3
4	33.1	16.1	22.8	34.8	/	33.1	/	50.9	24.7	23.4	27.3	19.2	26.1
5	/	32.6	41.6	33.2	33.3	/	/	60.6	21.9	20.1	46.7	30.3	30.8
6	/	30.3	30.8	50.2	/	/	/	22.7	/	/	32.1	/	15.1
7	52.1	35.7	29.2	29.9	26.1	49.8	18.8	/	27	27.7	27.6	50.9	27.3
8	29.1	24.8	30.9	30.2	24.3	24.1	/	24.2	/	20.2	41.7	27.8	20.8
9	18.9	19.2	19.5	30.3	19.1	31.3	/	/	26.9	/	/	15.9	22.4
10	37.6	24.8	25	27.3	27.2	42.3	38	56.3	36.5	37.1	/	32.9	25.9
11	35	22.8	23.6	40.6	19.8	33.5	/	71.5	27.3	22.7	30	/	33.6
12	/	27.6	12.9	34.1	21.7	43.1	/	27.9	19.7	21.4	25.8	38.9	/

Table 7. *interSpeed* of morning Peak and evening Peak. Unit: m/minute

Zones ID	0	1	2	3	4	5	6	7	8	9	10	11	12
Morn	348	327	362	435	353	375	437	407	354	399	461	388	322
Eve	334	327	364	411	359	369	423	408	358	404	450	390	316

Thus, it can be seen that OD information extracted from CDR data can reflect the information of travel time between traffic zones to some extent. Dividing traffic zones based on traffic semantics is a relatively helpful method for travel time estimation.

5 Urban Key Routes Analysis

In the previous section, we study the travel time estimation of Guangzhou city, revealing the characteristics in aspect of traffic zones. In this section, we focus on characteristics of another aspect: the key routes of the city.

With OD records, we can generate a trajectory for each user. The general idea of finding the key routes is to group the mobile phone users' trajectories into several clusters according to their similarity, and then from each cluster we can obtain a common route. Users' trajectories extracted from the CDR data are composed of a series of base stations' locations where the user stayed, which means these trajectories are of coarse granularity. Since the trajectories differ in length and time interval, it is difficult to cluster the trajectories according to Euclidean distance or cosine similarity.

To settle these problems, we use a DBSCAN method with length of common sub-track [16] as distance metric, which is more robust to trajectory noise and able to deal with clusters of any shape. The similarity of two trajectories Tr_i, Tr_j is defined as the ratio of their longest common subsequence [6] Tr_{ij}^{LCS} to the length of the shorter trajectory. See Algorithm 2 for details.

Algorithm 2. Obtaining urban key routes with DBSCAN method

Step 1: Calculating the distance matrix D of all trajectories.

$$sim(Tr_i, Tr_j) = \begin{cases} 0, & min(L(Tr_i), L(Tr_j)) < \delta \\ \frac{L(Tr_{ij}^{LCS})}{min(L(Tr_i), L(Tr_j))}, & min(L(Tr_i), L(Tr_j)) \geq \delta \end{cases}$$

$D_{i,j} = 1 - sim(Tr_i, Tr_j)$

D is stored by sparsed matrix.

Step 2: Trajectories clustering.

Using distance matrix D, cluster all trajectories into a set of clusters C with DBSCAN method.

$C = \{C_1, C_2, \cdots, C_m, \cdots\}$ where $C_m = \{Tr_1, Tr_2, \cdots, Tr_n, \cdots\}$

Step 3: Trajectories merging.

for $C_m \in C$ **do**

    $Tr_{repre}^m = Union(Tr_{ij}^{LCS}), Tr_i, Tr_j \in C_m$

end for

The representative trajectories of all clusters obtained above indicate the key routes of the city. Taking the evening peak hours defined in Sect. 3 as an example, we use OD records of one day, grouping them by users and ordering them chronologically to from users' trajectories.

Among all trajectories in this period, some contain more than 50 points while more than 90% trajectories contain no more than 11 points. In order to simplify the following task, we sample 11 points randomly on those trajectories with length more than 11. After the process mentioned above, we obtain 803,909 trajectories in total. The density of non-one values of the distance matrix is about 0.1%.

The parameters of DBSCAN algorithm are MinPts = 1000 and eps = 0.5, which lead to the results of 11 clusters. Clustering results of all trajectories clusters are shown in Table 8.

Table 8. Trajectories number of the clusters

Cluster ID	Number of trajectories
1	133740
2	3578
3	2176
4	1816
5	3335
6	3002
7	1356
8	1846
9	3631
10	2710
11	1356
Noise trajectory	645363
Full trajectory	803909

To generate smooth representative routes, we do not merge all Tr^{LCS} of a cluster into a trajectory. Instead, for each cluster, we count the number of each Tr^{LCS} and arrange them in ascending order. Only the Tr^{LCS} whose number is greater than 90% of the largest number of the cluster will be merge. The key routes in the downtown area of Guangzhou are shown in Fig. 6.

As can be seen from Table 8, in all trajectories clusters, the largest one consists of more than 0.13 million trajectories while the other ten clusters are relatively small. The key route of the first cluster is merged by many paths from Yuexiu District to Liwan District.

We can see that most of the key routes (Route 3, 5, 6, 8, 9, 10, 11) pass through the CBD area(west part of Tianhe District and east part of Yuexiu district) of Guangzhou. These routes correspond to North Guangzhou Avenue, Tiyuxi Road, Tianhe Road and several nearby roads in actual traffic networks. In the actual traffic network, each of the roads mentioned above has a high traffic flow. These roads locate near the most prosperous business areas, linking the regions of office buildings and the biggest shopping mall of Guangzhou city.

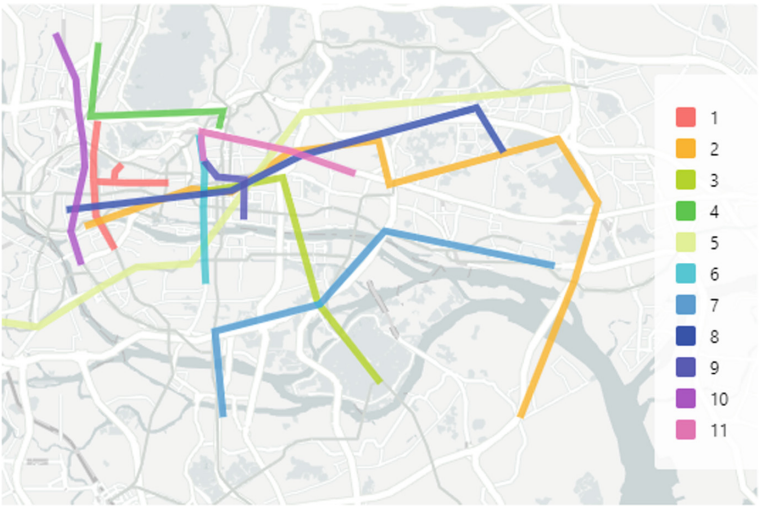


Fig. 6. Key routes of Guangzhou in evening peak

The traffic of the key routes mainly flows from the center of the city to the periphery of the city such as Baiyun District(in the northwest of the city) and Huangpu District(in the east of the city). This is a typical phenomenon of a metropolis. Most of the citizens in Guangzhou come from other cities and many of them tend to choose their house in the areas like Baiyun District, Panyu District and Huangpu District because of the price. In the evening peak, the car flows of driving home cause the roads scaterring from the center to suburb busy.

Three key routes (Route 1, Route 2, Route 8) locate closely in Liwan District in the west of the city, showing that these regions are crowded. When it comes to the traffic ground truth, these regions are the old town of Guangzhou, where there are several relatively narrow roads owning to the limitations of previous urban planning. It is worth mentioning that one of the key routes (Route 3) flows towards the city center from a periphery area in the southeast of the city. We find that the starting point of this route lies in Guangzhou Higher Education Mega Center, where most of the people working there live outside this area since it is mainly used for education. At evening peaks, a great quantity of students go to the city center of entertainment, which also contributes to this distinctive key route.

The discoveries mentioned above validate the key routes obtained from CDR data and can serve as a reference for urban roads planning. The results also demonstrate the applicability of mobile CDR data in the field of urban traffic.

6 Related Work

Researches Based on CDR: Ubiquitous mobile phones and the massive records they generate present new opportunities to excavate the spatial and temporal information of urban areas, which motivates many researchers to conduct studies on multiple aspects about cities with call detail records. For instance, Blumenstock et al. [2], Smith-Clarke et al. [20] and Frias-Martinez et al. [21] study the strategies for predicting economic levels of urban regions respectively. Moreover, Blumenstock et al. [3] also analysis the socioeconomic status of individuals combining CDRs with other survey ground truth. A bunch of works aim to localize the residential areas of the CDR users [9] with the or explore the periodic patterns on different scales [13,22,24]. Human location and mobility have become the major research fields with a lot of works such as [1,15]. Recently, owing to the limitations of data granularity, more and more studies take advantages of mobile traffic data gradually. Different from the research mentioned above, however, we mainly focus on estimating the traffic characteristics.

Traffic Time Estimation: Traffic time estimation and its application are traditional issues in the research field of transportation. Most of these studies are based on data of road networks. [14] designs a travel time model that separates trip travel times into link travel times and intersection delays. A fixed point approach is proposed in [17] to estimate travel time with simultaneous path inference. [25] introduces a good method to estimate time combining the temporal and historical contexts learned from different trajectories. [10] shows a good example to analyze travel time with OD information extracted from CDRs, associating the travels with traffic regions. Our work draw lessons from the idea of analyzing time according to traffic zones and take advantages of CDR data.

Urban Region Division: Many studies have worked on divide the urban areas into different regions to discover what functions they provide to the whole cities, which benefit the urban planning a lot. [29] proposes a topic-model-based method to divide urban regions, setting a precedent for combining POI data and human mobility in this field. Some other works such as Dong et al. [7] and Xu et al. [27] employ clustering-based algorithms to classify regions into several groups automatically. The regional semantics of these above-mentioned works may differ since they focus on different aspects of urban features.

Urban Trajectories Analysis: As for the analysis towards urban trajectories, Schlaich et al. [18] develop a method to generate trajectories of mobile phone users using the location-updating sequences and compare them with actual road networks. Hoteit et al. [11] put forward an idea of utilizing interpolation methods to estimate travel trajectories and find the positions of hotspots.

7 Conclusion

In this paper, based on the CDR data, we introduce traffic semantic attributes into the process of traffic zones division before estimating the travel time between

any two positions of the city, making the inter-zone and intra-zone charateristics more consistent with actual traffic conditions. The accuracy of our proposing travel time estimating method can achieve 67%. This method provides ideas for estimating the travel time in urban area in a low-cost way. Furthermore, making use of the individual trajectories extracted from the CDR data, we analyze the key routes in Guangzhou city using a DBSCAN trajectories clustering method. These routes can be validated by the actual conditions of the city, providing some new discoveries and visual references for the planning of urban roads and further study on urban traffic. Both of these two aspects of urban traffic study highlight the value of CDR data in urban traffic and provide ideas for further research.

Acknowledgement. The work is supported in part by NSFC (No. U1711263), Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20170816151823313), States Key Project of Research and Development Plan (No. 2017YFE0121300-6) and MOE-CMCC Joint Research Fund of China (MCM20160101).

References

1. Becker, R., et al.: Human mobility characterization from cellular network data. Commun. ACM **56**(1), 74–82 (2013)
2. Blumenstock, J., Cadamuro, G., On, R.: Predicting poverty and wealth from mobile phone metadata. Science **350**(6264), 1073–1076 (2015)
3. Blumenstock, J., Eagle, N.: Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In: Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, p. 6. ACM (2010)
4. Caceres, N., Romero, L.M., Benitez, F.G., Castillo, J.M.D.: Traffic flow estimation models using cellular phone data. IEEE Trans. Intell. Transp. Syst. **13**(3), 1430–1441 (2012)
5. Calabrese, F., Diao, M., Lorenzo, G.D., Ferreira, J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. Part C **26**(1), 301–313 (2013)
6. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3nd edn, pp. 390–396. The MIT Press, Cambridge (2009)
7. Dong, H., et al.: Traffic zone division based on big data from mobile phone base stations. Transp. Res. Part C **58**, 278–291 (2015)
8. Frias-Martinez, V., Soguero, C., Frias-Martinez, E.: Estimation of urban commuting patterns using cellphone network data. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp 2012, pp. 9–16. ACM, New York (2012). https://doi.org/10.1145/2346496.2346499. http://doi.acm.org/10.1145/2346496.2346499
9. Frias-Martinez, V., Virseda, J., Rubio, A., Frias-Martinez, E.: Towards large scale technology impact analyses: automatic residential localization from mobile phone-call data. In: Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, p. 11. ACM (2010)
10. Hasan, M.M., Ali, M.E.: Estimating travel time of Dhaka city from mobile phone call detail records. In: International Conference on Information and Communication Technologies and Development (2017)

11. Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., Pujolle, G.: Estimating human trajectories and hotspots through mobile phone data. Comput. Netw. **64**, 296–307 (2014)
12. Järv, O., Ahas, R., Saluveer, E., Derudder, B., Witlox, F.: Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. PLoS ONE **7**(11), e49171–e49171 (2012)
13. Järv, O., Ahas, R., Witlox, F.: Understanding monthly variability in human activity spaces: a twelve-month study using mobile phone call detail records. Transp. Res. Part C: Emerg. Technol. **38**, 122–135 (2014)
14. Jenelius, E., Koutsopoulos, H.N.: Travel time estimation for urban road networks using low frequency probe vehicle data. Transp. Res. Part B: Methodol. **53**, 64–81 (2013)
15. Palchykov, V., Mitrović, M., Jo, H.H., Saramäki, J., Pan, R.K.: Inferring human mobility using communication patterns. Sci. Rep. **4**, 6174 (2014)
16. Qin, S., Zuo, Y., Wang, Y., Xuan, S., Dong, H.: Travel trajectories analysis based on call detail record data. In: Control and Decision Conference (2017)
17. Rahmani, M., Koutsopoulos, H.N., Jenelius, E.: Travel time estimation from sparse floating car data with consistent path inference: a fixed point approach. Transp. Res. Part C: Emerg. Technol. **85**, 628–643 (2017)
18. Schlaich, J., Otterstätter, T., Friedrich, M., et al.: Generating trajectories from mobile phone data. In: Proceedings of the 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies (2010)
19. Sevtsuk, A., Ratti, C.: Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. J. Urban Technol. **17**(1), 41–60 (2010)
20. Smith-Clarke, C., Mashhadi, A., Capra, L.: Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 511–520. ACM (2014)
21. Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Prediction of socioeconomic levels using cell phone records. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 377–388. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22362-4_35
22. Trasarti, R., et al.: Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. Telecommun. Policy **39**(3–4), 347–362 (2015)
23. ITU: International Telecommunication Union (2013). https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf. Accessed 4 Mar 2019
24. Vieira, M.R., Frías-Martínez, E., Bakalov, P., Frías-Martínez, V., Tsotras, V.J.: Querying spatio-temporal patterns in mobile phone-call databases. In: 2010 Eleventh International Conference on Mobile Data Management, pp. 239–248. IEEE (2010)
25. Wang, Y., Zheng, Y., Xue, Y.: Travel time estimation of a path using sparse trajectories. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 25–34. ACM (2014)
26. Xiao-Dan, L.I., Yang, X.G.: Study on traffic zone division based on spatial clustering analysis. Comput. Eng. Appl. **45**(5), 19–22 (2009)
27. Xu, F., Li, Y., Wang, H., Zhang, P., Jin, D.: Understanding mobile traffic patterns of large scale cellular towers in urban environment. IEEE/ACM Trans. Netw. (TON) **25**(2), 1147–1161 (2017)

28. Yi, H., Edara, P., Sun, C.: Traffic flow forecasting for urban work zones. IEEE Trans. Intell. Transp. Syst. **16**(4), 1761–1770 (2015)
29. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 186–194. ACM (2012)