Estimating Crowd Flow and Crowd Density from Cellular Data for Mass Rapid Transit

Guanyao Li, Chun-Jie Chen, Wen-Chih Peng and Chih-Wei Yi
Department of Computer Science
National Chiao Tung University, Hsinchu, Taiwan
{gli,cjchen10167,wcpeng,yi}@cs.nctu.edu.tw

ABSTRACT

Mass rapid transit(MRT) is playing an increasingly important role in many cities due to its large carrying capacity, high speed and punctuality. Understanding the crowd flow and crowd density for MRT is crucial for smart city and urban planning. The traditional way to this task is by using smart card data. However, we can only know the number of passengers entering or exiting the station from smart card data. When and where the passengers change their MRT lines still remain unknown. Nowadays, each user has his/her own mobile phones and from the cellular data of mobile phone service providers, it is possible to know the users' transportation mode and the fine-grained crowd flows. As such, given a set of cellular data, we aim to estimate the crowd flow of MRT passengers and crowd density of stations as well as routes. To achieve these goals, we firstly propose an efficient and scalable approach to detect MRT trips with a pre-defined reference system. Then based on the detection result, we estimate the crowd flow and crowd density by grouping and counting the MRT trips. Extensive experiments are conducted to evaluate the detection and estimation approaches on a real dataset from Chunghwa Telecom, which is the largest telecommunication company in Taiwan. The results confirmed that our approaches are suitable for MRT trips detection, crowd flow and crowd density estimation. Finally, we provide case studies to present some applications and demonstrate the usefulness of our approaches.

KEYWORDS

crowd flow and crowd density estimation, urban computing, cellular data

ACM Reference format:
Guanyao Li, Chun-Jie Chen, Wen-Chih Peng and Chih-Wei Yi. 2017. Estimating Crowd Flow and Crowd Density from Cellular Data for Mass Rapid Transit. In *Proceedings of , Halifax, Nova Scotia, Canada, August 14, 2017 (UrbComp'17),* 9 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Mass rapid transit(MRT) is playing an increasingly important role in many cities due to its large carrying capacity, high speed and punctuality. Understanding the crowd flow and crowd density for MRT is essential for urban planning, public transport network planning, and public transport timetable arrangement. It is also helpful for passengers to plan their trips. What's more, it helps the telecommunication company to determine where to deploy additional cell tower in order to provide better service. In this paper, we focus on estimating crowd flow and crowd density for MRT.

In the previous studies, users' smart card data, which record when and where users enter or exit an MRT station, have been used to estimate the MRT passenger flow[5]. However, the smart card only records the origin and destination stations of a user. Whether a user changes lines during the trip, and when and where the change occurs are usually unknown. We cannot estimate the exact passenger number for those transfer stations, since the number of transfer passengers is unknown. For example, there are several MRT lines in Figure 1 and A, B, C, D and E are MRT stations. If a passenger takes the MRT from station A to station D, she can choose the route $\langle A, B, C, D\rangle$ in which she changes from the blue line to the red line at station C or she can choose the route $\langle A, B, E, D\rangle$ in which she changes from the blue line to the green line at station B. In both cases, the smart card system only records the original station A and destination station D but it cannot identify which transfer station the passenger chooses.

To deal with the limitation of using smart card data, we use cellular data instead. Nowadays, each user has his/her own mobile phones. When users use their mobile phones to call, send messages or access the internet, the phones are connected to a nearby cell tower. Even if the user does not use the phone, it will still connect to a nearby cell tower every hour. Using the location of cell tower to denote a user's approximate location, it is possible to know the users' transportation mode and the fine-grained crowd flows.

However, there are several issues to estimate crowd flow and crowd density from cellular data for MRT. First, the coverage of the cell tower is large and there is intersection of coverage of some cell towers. Thus, we can not infer the exact location, the moving direction and the speed of a user from the cellular data directly. Second, the cellular data sampling rate is not static. It depends on the strength of the signal and whether the user uses her mobile phone during her trip. Thus, there may be no data occurring when use past some MRT stations. Third, some MRT routes are similar to routes of other transportation modes, which leads to the difficulty in identifying MRT trips and other trips.

To deal with the above issues, we firstly propose an efficient and scalable approach for detecting both indoor and outdoor MRT trips. Some cell towers are selected as reference towers to build a reference system. Based on the reference system, a matching-based approach is proposed to detect individual MRT trips. Our approach

UrbComp'17, August 14, 2017, Halifax, Nova Scotia, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Figure 1: An example of an MRT network

takes both spatial and temporal features into consideration. External data, including MRT route network data and MRT travel time data are introduced to our approach. Then, based on the results of our detection approach, we estimate the crowd density of stations and routes as well as the crowd flow of origin-destination stations in a route.

In summary, the major contributions of our paper are outlined as follows:

- We study on a novel and fundamental problem in urban computing, i.e., estimating crowd flow and crowd density from cellular data for Mass Rapid Transit.
- We propose an efficient and scalable approach to detect both indoor and outdoor MRT trips with cellular data considering the cell tower property, spatial and temporal factors.
- We conduct comprehensive experiments over the data from the largest telecommunication company in Taiwan, the results demonstrate the efficiency and scalability of our approaches.
- We provide case studies to show the result of crowd flow and crowd density estimation for MRT.

The remainder of this paper is organized as follows: Section 2 discusses related works. Section 3 shows the dataset and data preprocess. Section 4 presents our MRT trip detection approach. The estimation approaches are discussed in Section 5. The evaluation results of our approach are presented in Section 6. Section 7 shows case studies. Section 8 concludes this paper.

2 RELATED WORKS

Our work is about MRT trip detection as well as crowd density and crowd flow estimation. In this section, we discuss some related works.

There have been some works about transportation mode detection. The prior works [10] [16] [17] proposed methods to detect user transportation mode from user GPS data. Compared to the GPS data, there are two challenges of using mobile phone data: it is inaccurate in determining the position of a user, and the data sampling rate is irregular[2]. To deal with these issues, [2] presented methods for mapping trajectories of cell tower latitude-longitudes to transport networks. It defined stay region, extracted trajectories between stay regions as trips and then mapped the trips to the transport network. Extracting users' stationary stay locations from cellular data as the origin and destination of trips was prevalent in the prior research, it has also been discussed in [1] and [6]. Compared to [1] [2] [6], our proposed approach does not have to detect the stay region for each user. Instead, we mapped the cell tower to the MRT station directly utilizing a pre-defined reference system. The paper [5] proposed a method to extract trips from user call detail record(CDR) data and utilized the data from the public transport smart card system to distinguish transportation mode. A probabilistic method consisting of a Hidden Markov Model and two sub-models was proposed in [15] to identify transportation modes(driving, biking and walking). Labeled data were necessary for training in [15]. In [12], users from the same origin to the same destination were clustered into three subgroups(driving, walking and public transit) according to their travel time to infer transport mode share. Other works [7] [9] utilized the signal strength to detect the transportation mode. However, the signal strength data are not available in our work. An algorithm for MRT trip detection was proposed in [4] which utilized the property that the indoor MRT stations in Singapore are served exclusively by indoor cell towers, and cell phones outside the MRT network cannot access those towers. However, the limitations of this work are obvious. The algorithm is limited to detecting indoor MRT trips. And for other cities without the exclusive property, the algorithm in [4] is not suitable.

For crowd flow and crowd density estimation, [3] built a hard- and software system to estimate the number of passengers in a vehicle. [8] presented approaches to estimate crowd density and pedestrian flows using Wi-Fi and bluetooth data. In [13], a bluetooth scan based method was proposed to detect the crowd density. In our work, we focus on using mobile phone data which can be obtained more easily as no extra sensors or devices are needed.

3 DATASET AND DATA PRE-PROCESS

We use users' mobile phone cellular data from Chunghwa Telecom, which is the largest telecommunication company with a market share of 38% in Taiwan. The cellular data we used are the records of cell towers the mobile phone is connected to. A mobile phone is connected to a cell tower in two cases: active network events or passive network events[4]. Active network events include calling, sending messages or accessing the internet. Passive network events include switches between network zone or after one hour of inactivity.

An example of the user mobile phone data is shown in Table 1. The data consist of a user ID, the longitude and latitude of the connected cell tower, as well as the time stamp. The user ID was anonymized by the hashing process. Thus, the personal information of the user is unknown to the authors during the study.

Because there is intersection of the coverage of some cell towers, one major problem of using cellular data for mobility modeling is the oscillation problem[14]. Oscillation occurs when the mobile phone switches between cell towers very quickly instead of being connected to one cell tower. We denote two cell towers as t_i and t_j. Assume that the cell towers connected by a mobile phone are

Table 1: Overview of the Dataset

User ID	time stamp	longitude	latitude
-87556096	00:59:19	121.587	25.048
-87556096	00:59:20	121.59	25.04
-87556096	00:59:21	121.587	25.048
-87556096	01:59:23	121.587	25.048
...	...	...	...
-87556096	16:02:01	121.5	25.041
-87556096	16:02:06	121.5	25.041

Table 2: Example of processed data

User ID	start time	end time	longitude	latitude
-87556096	00:59:19	01:59:23	121.587	25.048
...	...	...	...	...
-87556096	16:02:01	16:02:06	121.5	25.041

$\langle t_i, t_i, t_j, t_i, t_i \rangle$. If the connection of the mobile phone switches between $\langle t_i, t_j \rangle$ and $\langle t_j, t_i \rangle$ very quickly, the connection to t_j can be regarded as resulting from the oscillation problem. To reduce the effect of the oscillation problem, we remove the records which resulted from the oscillation problem. In this example, the connection to t_j will be removed in our data pre-process. After that, if two consecutive records are the same cell tower, the two records will be merged. An example of the pre-process result of Table 1 is presented in Table 2. The second record in Table 1 is removed as oscillation data and three records of the cell tower $\langle 121.587, 25.048 \rangle$ and two records of the cell tower $\langle 121.5, 25.041 \rangle$ in Table 1 are merged as one record respectively in Table 2.

After data pre-process, we define each record as $\ell = \langle u, t_c, t_s, t_e \rangle$, where u is the user ID, t_c is a cell tower, t_s is the start time and t_e is the end time. Then the user data can be denoted as a sequence of data records: $L = \langle \ell_1, \ell_2, ..., \ell_i, ...\ell_n \rangle$.

4 MRT TRIP DETECTION

The MRT trip detection algorithm plays the key role in our work. The intuitive idea for detecting MRT trips from a user's cellular data is to compare the similarity of the user's trajectory and the MRT route. However, it is not efficient since we have to calculate the distance between each MRT station and each cell tower in the user data.

The signal coverage of a cell tower is limited; thus the mobile phone is usually connected to a nearby cell tower. An observation is shown in the Figure 3. From the observation, we know that the mobile phone of a MRT passenger is usually connected to a cell tower near the MRT station. Utilizing this property, we propose an efficient approach in this work.

The overview of the approach is presented in Figure 2. Given the cellular data of the user as the input, the detection approach will output the detail MRT routes of the user. The detection approach consists of three sub-approaches. To deal with the issue of large coverage and intersection of coverage of cell towers, we propose a tower-station matching approach with the help of a reference system, considering the cell tower property and the spatial factors.

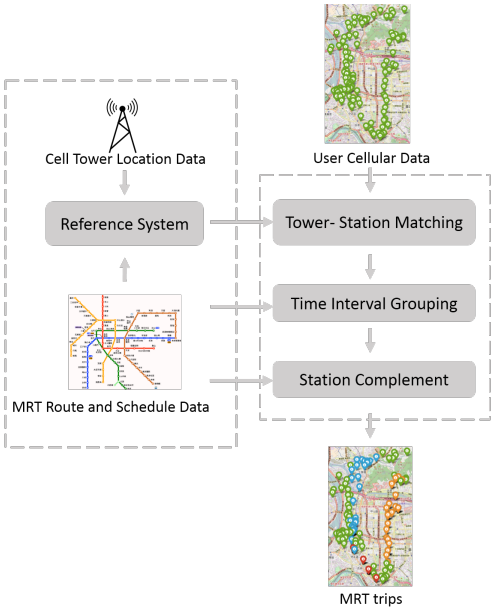


Figure 2: The overview of the MRT trip Detection approach

Figure 3: An observation for the MRT passenger

The time interval grouping approach that utilizes the external data and temporal factors is proposed to overcome the problem caused by the issue that different transportation modes share similar routes. To deal with the irregularity of cellular data sampling rate, we propose a station complement approach to infer the detail route of users.

4.1 Reference System Building

To detect MRT passengers, our idea is to detect whether the user's mobile phone connects to a cell tower near MRT stations. We use the reference system consisting of reference towers of each station to achieve this goal efficiently. The idea of the reference system is inspired by the work[11]. The reference towers are cell

towers near MRT stations and to which an MRT passenger's mobile phone will be connected with high possibility. In the example in Figure 3, cell towers t_1, t_2 and t_3 can be used as reference towers for MRT stations s_1, s_2 and s_3 respectively.

To build the reference system, we select some cell towers as reference towers for each MRT station. Most MRT stations are large and there are several gates; we therefore selected reference towers for each gate when building the reference system. We define two types of reference systems in our work: the K Nearest Tower reference system(KNT) and D meters Coverage Tower reference system(DCT).

4.1.1 KNT Reference System. A mobile phone is usually connected to the nearest cell tower with the highest probability. But if two cell towers are close to each other, there is intersection of their coverage. In this case, even in the same position, different mobile phones may be connected to different cell towers. Thus, for the KNT reference system, we select the k nearest cell towers as the reference towers for each gate of the station. If a cell tower is the K nearest cell tower for more than one MRT station, we regard it as the reference tower for the station to which it is closest so that each reference tower in the reference system only serves one MRT station.

4.1.2 DCT Reference System. The radius of the cell tower's coverage is limited. It ranges from several hundred in urban areas to several thousand meters in the suburbs[2]. The closer to the cell tower, the higher the possibility that the cell phone will be connected to it. In the DCT reference system, we consider the radius of the coverage. If a station is in the D meters coverage of a cell tower, the cell tower will be selected as the reference tower for the station. To make sure that each reference tower only serves one station, if there is more than one station within its D meters coverage, we define the tower as the reference tower of the closest station.

Figure 4 shows the overview of all the cell towers in Taipei(the green point), the gates of MRT stations(the red point) and the reference towers of each station with different reference systems(the blue point). Compared to the non-reference towers, the reference towers are very close to the corresponding MRT stations. It illustrates that the reference towers can be used as the distinction to identify the MRT trips and non-MRT trips.

4.2 Tower-Station Matching

After the reference system is built, we can use these cell towers as a distinction between MRT trips and non-MRT trips. In the user's data, if a connected cell tower $\ell_i.t_c$ is a reference tower of MRT station s_i, the data record ℓ_i will be matched to the station s_i.

For example, given the reference system shown in Table 3 and the user data $\langle \ell_0, \ell_1, \ell_2, ...\ell_9 \rangle$, if the cell tower connected record $\langle \ell_0.t_c, \ell_1.t_c, \ell_2.t_c, ...\ell_9.t_c \rangle$ is $\langle t_0, t_1, t_6, t_2, t_3, t_7, t_8, t_4, t_8, t_9 \rangle$, the raw data are matched to be $\langle s_1, s_2, s_3, s_4 \rangle$ since t_1, t_2, t_3, t_4 are reference towers for s_1, s_2, s_3, s_4 respectively.

4.3 Time Interval Grouping

In the tower-station matching step, we detect the MRT trip from the spatial dimension and identify some potential MRT stations. But only considering the spatial dimension is not enough. If the user was driving a car past the MRT stations or taking a bus with a similar route, her mobile phone may also be connected to some reference towers. However, the travel time from one station to another is distinct for different transport modes. And the speed of MRT is higher than the speed of other public transportation modes and cars. In this step, we take the temporal factors into consideration and use the travel time as a distinction to detect trips.

Table 3: Example of a reference system

Station	Reference Tower	Station	Reference Tower
s_1	t_1	s_4	t_4
s_2	t_2	s_4	t_4
s_3	t_3		

Given the station sequence obtained from the last step, we group the station to a trip based on the time interval and the real travel time between every two consecutive stations in the station sequence. For two consecutive stations s_i and s_{i+1}, we denote the time interval of the two stations as $TI(s_i, s_{i+1})$ and the real travel time as $TT(s_i, s_{i+1})$. If $TI(s_i, s_{i+1}) < TT(s_i, s_{i+1}) + \theta$, $\langle s_i, s_{i+1} \rangle$ is regarded as an MRT trip. θ is set to be 60s in our work. Otherwise, s_i is regarded as the destination station of the last MRT trip while s_{i+1} is regarded as the origin station of the new MRT trip. The real travel time from one station to another is available at $DataTaipei$[1], which is an open data platform of the Taipei government. The time interval of two consecutive stations in the station sequence is calculated as follows: if $\ell_i.t_c$ and $\ell_{i+1}.t_c$ is the reference tower of s_i and s_{i+1} respectively, then, $TI(s_i, s_{i+1}) = l_{i+1}.t_s - l_i.t_e$.

We also consider the connection duration for grouping. If the duration of staying in a station s is more than β, the station s is regarded as the destination station of the last MRT trip and also the origin station of the new MRT trip. β is set to be 30min in the experiment. The duration of staying in a station is the duration of being connected to the reference tower of the station.

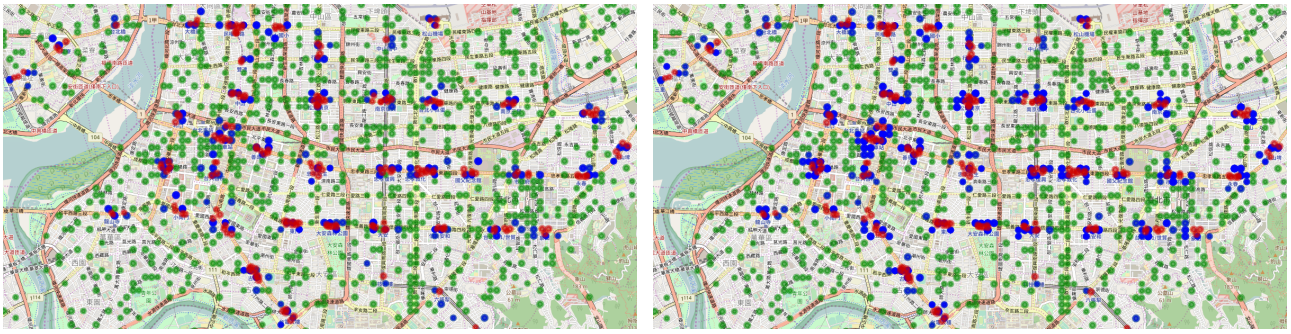
We continue the example in Section 4.2. If $TI(s_1, s_2) < TT(s_1, s_2) + \theta$, $TI(s_2, s_3) < TT(s_2, s_3) + \theta$ while $TI(s_3, s_4) > TT(s_3, s_4) + \theta$, then $\langle s_1, s_2, s_3 \rangle$ and $\langle s_4 \rangle$ are detected as two candidate MRT trips.

4.4 Station Complement

To better capture the origin station and destination station of the MRT trip, given the candidate MRT trip $\langle s_1, s_2 ... s_n \rangle$, we check the connected cell tower preceding the first station t_p and the connected cell tower behind the last station t_b in the raw data. If the distance from t_p to an MRT station s is smaller than a distance threshold Δd and the time interval is smaller than $TT(s, s_1) + \Delta t'$, we extend the MRT trip and regard s as the origin station instead. For the destination station complement, we adopt the same strategy. Δd is set to be 500m and $\Delta t'$ is set to be 10min.

In the above example, given the candidate trip $\langle s_1, s_2, s_3 \rangle$ and the raw data $\langle t_0, t_1, t_6, t_2, t_3, t_7, t_8, t_4, t_8, t_9 \rangle$, the connected cell tower preceding the reference tower of s_1 is t_0. If the distance from t_0 to the station s_5 is smaller than Δd and the difference between the time interval and travel time satisfies the time threshold, the origin station of the MRT trip is updated to be s_5. The connected tower

[1] http://data.taipei/



(a) With KNT reference system(K = 2)

(b) With DCT reference system(D = 200)

Figure 4: The overview of all cell towers, reference towers and MRT stations

after the reference tower of s_3 is t_7. If the distance from t_7 to the station s_6 is smaller than Δd and the time interval satisfies the time constraint, the destination station of the MRT trip is updated to be s_6. Assume that the trip $\langle s_4 \rangle$ does not change after complement. In our work, we define that each trip consists of at least two stations. Thus, the exact MRT trip is $\langle s_5, s_1, s_2, s_3, s_6 \rangle$.

After that, we complement the MRT trip according to the MRT network. If two consecutive stations s_i and s_j obtained in the last step are not consecutive stations in the MRT network, we insert the stations between s_i and s_j in the MRT network into the trip. If there are multiple routes from s_i to s_j in the MRT network, we choose the route whose travel time is closest to the time interval. For those stations complemented in this step, we estimate the time stamp of the station according to the real travel time from the station with a time stamp.

5 CROWD FLOW AND CROWD DENSITY ESTIMATION

Based on the results of our proposed MRT trip detection approach, we define three kinds of passengers:

- (In-Passenger)The passenger who enters and departs from the same station.
- (Out-Passenger)The passenger who arrives at the station by MRT and goes out of the station.
- (Transfer Passenger) The passenger who arrives at the station by MRT, changes the MRT line and departs from the station as the transfer station.

The smart card system only records the origin and destination stations, so we can only estimate the number of in-passengers and out-passengers from the smart card system data. However, since we can obtain the exact MRT routes of a user with our proposed detection approach, we can estimate the number of the three kinds of passengers based on the detection result. In this paper, we focus on estimating the crowd density of a station, the crowd density of different lines at a transfer station and the crowd flow of origin-destination stations in a route.

The crowd density of a station is the sum of the number of the three kinds of passengers. We obtain the number of the trips whose origin, destination or transfer station is the target station from the detection result.

The crowd density of different lines at a transfer station is the passenger number of different lines at the transfer station. Passengers at a transfer station may choose different lines. Estimating the crowd densities of different lines at the transfer stations will help the MRT company to arrange the schedule more reasonably. We count the trips whose origin, destination or transfer station is the target station and the line is the target line.

The crowd flow of origin-destination stations in a route is the number of passengers from the given origin station to the given destination station by the route. Understanding the crowd flow between any origin-destination stations is essential. It is helpful for route recommendation, route arrangement and schedule arrangement. To estimate the crowd flow, we count the trips that from the origin station to the destination station by the route among all trips in the detection result.

Some case studies about the crowd density and crowd flow estimation will be presented and discussed in the Section 7.

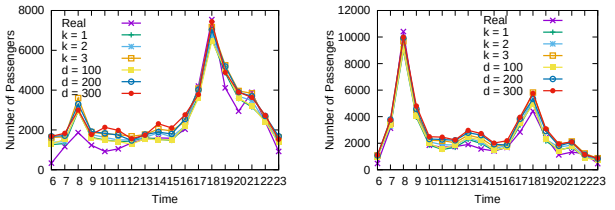
6 EVALUATION

To evaluate the performance of our approach, we conducted our experiments in three parts. In the first part, we evaluated the accuracy of our detection and estimation approaches. In the second part, we discussed the precision and recall of our detection approach. Finally, we tested the efficiency and scalability of our detection approach.

6.1 Performance of the estimation approach

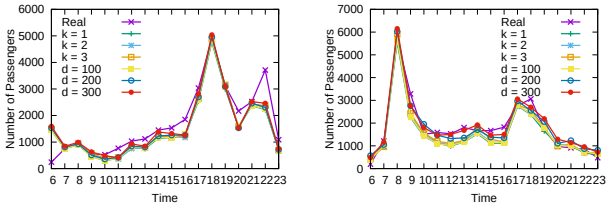
To evaluate the accuracy of our detection and estimation approach, we estimated the in-passenger and out-passengers number of two stations based on the trip detection result and compared the estimation results with the ground truth released by the Taipei Metro Company. The Pearson's correlation is used as metric for evaluation.

We applied our approach to the data of 10% of Chunghwa Telecom users in Taipei from 2017/01/06 to 2017/01/12. The number of users is around 310, 000. In the KNT reference system, K was set to be 1, 2 and 3; and D was set to be $100m$, $200m$ and $300m$ in the DTC reference system. The MRT stations we selected for evaluation



(a) Comparison of in-passengers (b) Comparison of out-passengers

Figure 5: Comparison of passengers at Nanjing Fuxing Station on 2017/01/06



(a) Comparison of in-passengers (b) Comparison of out-passengers

Figure 6: Comparison of passengers at Taipei 101 Station on 2017/01/06

are the Nanjing Fuxing Station and the Taipei 101 Station, which are two of the most important stations in Taipei. To estimate the number of in-passengers and out-passengers, we firstly detected the MRT trips of all users. Then we counted the trips according to the origin and destination at hourly intervals. The market share of Chunghwa Telecom is 38% and the data we used were 10% of total users. After obtaining the inference result γ, we calculated the approximate result α as : $\alpha = \gamma/(\text{sampling size} * \text{market share})$.

The results for Nanjing Fuxing Station on 2017/01/06 are shown in Figure 5. The results for Taipei 101 Station on 2017/01/06 are shown in Figure 6. Figures 5(a) and 6(a) present the comparison between the estimation results of in-passengers and the ground truth. From the comparison, we can learn that our estimation results have a very similar trend to the ground truth. The good correspondence confirms that our proposed approach is suitable for MRT trip detection and crowd density estimation. The comparison of number of out-passengers in Figures 5(b) and 6(b) leads to the same conclusion. The comparison of the data with other days is also similar to the results for 2017/01/06. Because of the page limitation, we only present the results for 2017/01/06.

From the figures, we can also learn that when K and D become larger, the estimation result increases. That is because the larger the K and D, the more cell towers are selected as reference towers and so more MRT trips will be detected. If the K and D are too large, the result will be overestimated. And if they are too small, the estimation result will be underestimated.

We also used the Pearson's correlation to further estimate the correlation. The result in presented in Table 4. The Pearson's correlation between the ground truth and all our approaches are larger than 0.9, which confirms the good correspondence and shows the good performance of our approaches.

Table 4: Correlation between the estimation result and ground truth

Type	Correlation	Type	Correlation
K = 1	0.956394	D = 100	0.954600
K = 2	0.956323	D = 200	0.956437
K = 3	0.956169	D = 300	0.960197

Table 5: Evaluation of the detection result

Type	Precision	Recall	F-1
K = 1	0.9144	0.7565	0.8280
K = 2	0.7858	0.7678	0.7767
K = 3	0.7680	0.7927	0.7802
D = 100	0.8315	0.7379	0.7819
D = 200	0.7629	0.7990	0.7805
D = 300	0.6402	0.8101	0.7152

6.2 Performance of the detection approach

In this part of the experiment, we used a small dataset in which the truth of the MRT trips had been labelled by the users for evaluation. The dataset consists of 7 days data of 10 users in Taipei. To validate the effectiveness of the approach, the precision, recall and F-1 score are investigated. We firstly define the precision and recall for one trip.

Definition 6.1. (Precision of a detected trip) Given a detected trip T_i with $|T_i|$ stations, if the trip exists and the real trip T'_j consists of $|T'_j|$ stations, then the precision of the trip is $p_i = (|T_i \cap T'_j|)/|T_i|$. Otherwise, $p_i = 0$.

Definition 6.2. (Recall of a real trip) Given a real trip T'_j with $|T'_j|$ stations, if the trip is detected and the detection trip has $|T_i|$ stations, the recall is $r_j = (|T_i \cap T'_j|)/|T'_j|$. Otherwise, $r_i = 0$.

Then given the set of detected trips $\langle T_1 ... T_i ... T_n \rangle$ and a set of real trips $\langle T'_1 ... T'_j ... T'_m \rangle$, the precision is $Precision = (\sum_{i=1}^{n} p_i)/n$, the recall is $Recall = (\sum_{j=1}^{n} r_j)/m$ and the F-1 score is $F_1 = 2 * (precision * recall)/(precision + recall)$.

The precision, recall and F-1 score of our approach on the dataset is shown in 5. All of them confirm the good performance of our approach. In Table 5, the precision decreases when K or D becomes larger while the recall increases. That is because when more cell towers are selected as the reference towers, more trips will be detected as MRT trips. But at the same time, some of the non-MRT trips are detected as MRT trips by mistake. When considering the F-1 score, the approach with $K = 1$ has a better performance than other approaches.

6.3 Efficiency and Scalability

To evaluate the efficiency and scalability of our detection approach, we sampled the cellular data on 2017/01/06 with different

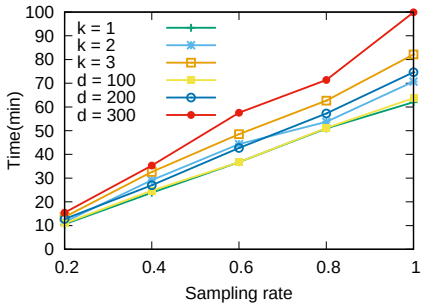


Figure 7: Running time of the detection approach with various reference systems

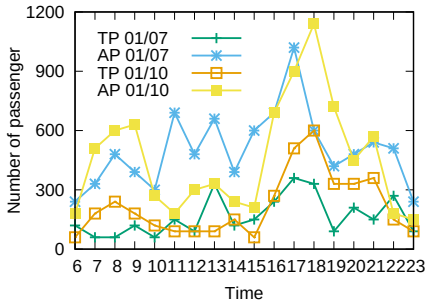


Figure 8: Number of passengers from Minquan W.Rd. Station to Taipei Main Station

sampling rate: 0.2, 0.4, 0.6, 0.8 and 1. We ran our proposed approach with various reference systems on the datasets of different size. The running time is shown in Figure 7. The running time of our approach increases when the size of the dataset increases. The increment of running time for different size of datasets is nearly linear, which illustrates that the detection approach is scalable. In addition, given the data with the same size, the running time of our approach with the KNT reference systems increases when K increases. The reason is that, the larger the K is, the more trips are detected as MRT trips, which results in more matching, grouping and complement. The running time of the approach with DCT reference systems also increases when D becomes larger resulted from the same reason.

7 CASE STUDIES

In this section, based on the result of our MRT trip detection approach, we present the overview of the crowdedness of MRT stations and MRT routes at different times in Taipei. Then we present case studies about crowd flow and crowd density estimation. We utilize the KNT reference system with $K = 1$ to obtain the MRT trip result.

7.1 Overview of the crowdedness

It is crucial to know the change in the crowd density in a day. After detecting the MRT trips, we visualize the trips with the heat map at different times of the day to show the variety of the crowd density of the routes and stations. We use the data of different times(morning, noon, afternoon and evening) on 2017/01/12. The result is presented in Figure 9. The results showed that the crowdedness in the morning(Figure 9(a)) is very similar to the crowdedness in the afternoon(Figure 9(c)). This is because the route by which people go to work in the morning and go back home in the afternoon is very similar. In addition, the hot regions in Figures 9(a), 9(c) and 9(d) are mainly transfer stations. But, there are more hot regions in Figures 9(b) and not all of the hot regions are transfer stations. The comparison shows that the transfer stations play different roles at different time in a day.

7.2 Crowd density of a station

We aim to show the crowd density of a station for each hour. Figure 10 shows the results for Taipei Main station, which is one of the most important and crowded MRT stations in Taipei. We use the data from 2017/01/07 to 2017/01/10. From the figures, we can learn about the estimation number of the three kinds of passengers for each hour. The estimation number of transfer-passengers helps us have a more comprehensive understanding of the crowd density of the transfer stations. In Figure 10(a), the trend of number of passengers on a workday(2017/01/09 and 2017/01/10) is different from the trend on the weekend(2017/01/07 and 2017/01/08). There are two peaks in workday's trend, one is around 8 : 00 which is the time people go to work and one is around 18 : 00 which is the time people go home after work. But for the weekend, there is only one peak at around 17 : 00. The results in Figures 10(b) and 10(c) lead to a similar conclusion.

7.3 Crowd density of different lines

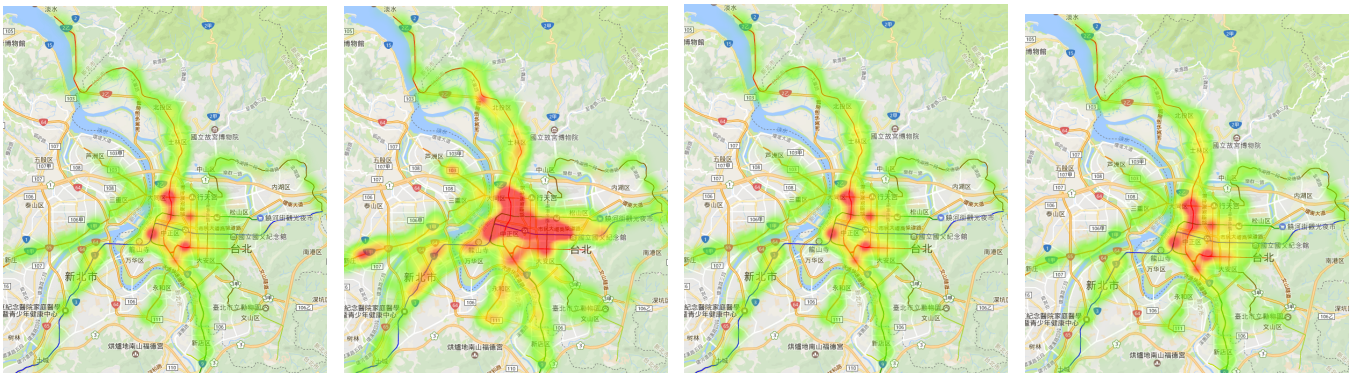
We take the Taipei Main Station for example in the case study. It is the intersection of the MRT blue line and MRT red line. We estimate the crowd densities of the two lines for each hour. Figure 11(a) presents the number of passengers departing from Taipei Main Station by different lines each hour on 2017/01/07(weekend) and 2017/01/10(workday). Similarly, Figure 11(b) presents the number of passengers arriving at Taipei Main Station by different lines for each hour. From the result, we learn that the Blue line is busier than the Red line at the Taipei Main Station.

7.4 Crowd flow of Origin-Destination Stations

Given the origin station and destination station, we estimate the crowd flow from the origin station to the destination at hourly intervals. Figure 8 shows the number of all passengers(AP) from Minquan W.Rd. Station to Taipei Main Station as well as the transfer passengers(TP) who are from Minquan W.Rd. Station and then change their MRT line at Taipei Main Station. The figure reveals that the trends are different on the weekend(2017/01/07) and on workdays(2017/01/10). Both the peak time as well as the number of transfer passengers and all passengers differ between the weekend and workdays.

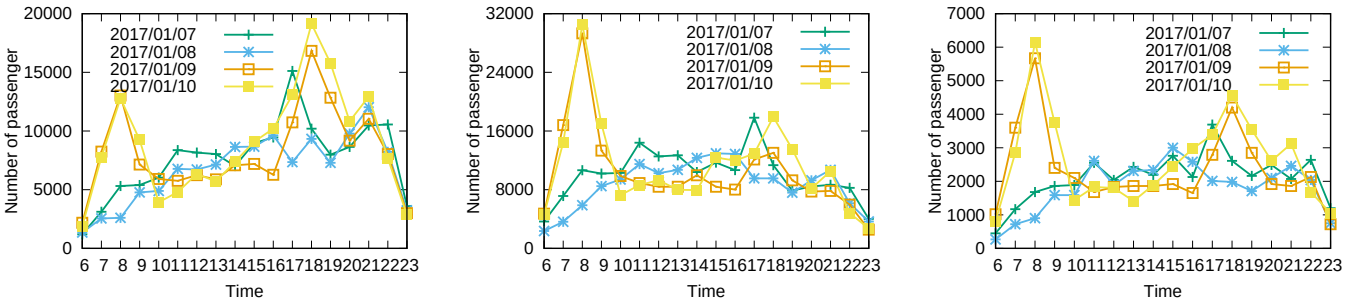
8 CONCLUSION

In this paper, we aimed to estimate the crowd flow and crowd density for MRT stations and routes. We firstly proposed an efficient and scalable approach to detect MRT trips from cellular data. In



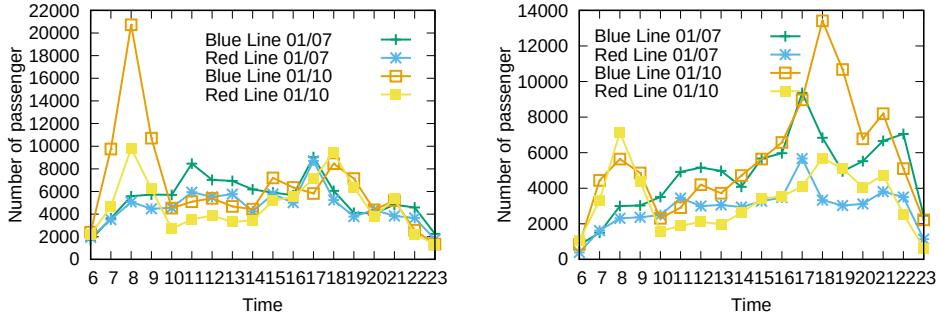
(a) 07 : 00 ~ 09 : 00 (b) 12 : 00 ~ 14 : 00 (c) 17 : 00 ~ 19 : 00 (d) 21 : 00 ~ 23 : 00

Figure 9: Overview of the crowdedness in Taipei



(a) Number of in-passengers (b) Number of out-passengers (c) Number of transfer-passengers

Figure 10: Crowd density of Taipei Main Station



(a) Number of passengers departing from Taipei Main Station (b) Number of passengers arriving at Taipei Main Station

Figure 11: Crowd density of different lines at Taipei Main Station

our proposed detection approach, both spatial and temporal factors were considered. Then we estimated the crowd flow and crowd density of stations and routes based on the MRT trip detection result. We conducted extensive experiments to show the effectiveness, efficiency and scalability of our approaches on the data from Chunghwa Telecom, which is the largest telecommunication company in Taiwan. We also provided several case studies. The case study demonstrated that we could have a more comprehensive understanding of the crowd flow and crowd density for MRT utilizing our proposed detection and estimation approaches.

REFERENCES

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies* 58 (2015), 240–250.

[2] Manoranjan Dash, Kee Kiat Koo, Thomas Holleczek, Ghim-Eng Yap, Shonali Priyadarsini Krishnaswamy, and Amy Shi-Nash. 2015. From Mobile Phone Data to Transport Network–Gaining Insight about Human Mobility. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, Vol. 1. IEEE, 243–250.

[3] Marcus Handte, Muhammad Umer Iqbal, Stephan Wagner, Wolfgang Apolinarski, Pedro José Marrón, Eva Maria Muñoz Navarro, Santiago Martinez, Sara Izquierdo Barthelemy, and Mario González Fernández. 2014. Crowd Density Estimation for Public Transport Vehicles.. In *EDBT/ICDT Workshops*. 315–322.

[4] Thomas Holleczek, Shanyang Yin, Yunye Jin, Spiros Antonatos, Han Leong Goh, Samantha Low, Amy Shi-Nash, et al. 2015. Traffic measurement and route recommendation system for mass rapid transit (mrt). In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1859–1868.

[5] Thomas Holleczek, Liang Yu, Joseph Kang Lee, Oliver Senn, Carlo Ratti, and Patrick Jaillet. 2014. Detecting weak public transport connections from cellphone and public transport data. In *Proceedings of the 2014 International Conference on Big Data Science and Computing*. ACM, 9.

[6] Shan Jiang, Joseph Ferreira, and Marta C Gonzales. 2016. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* (2016).

[7] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*. ACM, 2.

[8] Lorenz Schauer, Martin Werner, and Philipp Marcus. 2014. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 171–177.

[9] Timothy Sohn, Alex Varshavsky, Anthony LaMarca, Mike Y Chen, Tanzeem Choudhury, Ian Smith, Sunny Consolvo, Jeffrey Hightower, William G Griswold, and Eyal De Lara. 2006. Mobility detection using everyday gsm traces. In *International Conference on Ubiquitous Computing*. Springer, 212–224.

[10] Leon Stenneth, Ouri Wolfson, Philip S Yu, and Bo Xu. 2011. Transportation mode detection using mobile phones and GIS information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 54–63.

[11] Han Su, Kai Zheng, Haozhou Wang, Jiamin Huang, and Xiaofang Zhou. 2013. Calibrating trajectory data for similarity-based analysis. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 833–844.

[12] Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 318–323.

[13] Jens Weppner and Paul Lukowicz. 2013. Bluetooth based collaborative crowd density estimation with mobile phones. In *Pervasive computing and communications (PerCom), 2013 IEEE international conference on*. IEEE, 193–200.

[14] Wei Wu, Yue Wang, Joao Bartolo Gomes, Dang The Anh, Spiros Antonatos, Mingqiang Xue, Peng Yang, Ghim Eng Yap, Xiaoli Li, Shonali Krishnaswamy, et al. 2014. Oscillation resolution for mobile phone cellular tower data to enable mobility modelling. In *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, Vol. 1. IEEE, 321–328.

[15] Dafeng Xu, Guojie Song, Peng Gao, Rongzeng Cao, Xinwei Nie, and Kunqing Xie. 2011. Transportation modes identification from mobile phone data using probabilistic models. *Advanced Data Mining and Applications* (2011), 359–371.

[16] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 312–321.

[17] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 247–256.