2014 IEEE 17th International Conference on
Intelligent Transportation Systems (ITSC)
October 8-11, 2014. Qingdao, China

# Analyzing Passenger Density for Public Bus: Inference of Crowdedness and Evaluation of Scheduling Choices

Jun Zhang[†⋆] Xin Yu[†] Chen Tian[†⋆] Fan Zhang[⋆] Lai Tu[†] Chengzhong Xu[⋆§]
[†]Department of Electronics and Information, Huazhong University of Science and Technology, China
[⋆]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[§]Department of Electrical and Computer Engineering, Wayne State University, MI, USA
[†]{zhangjun, yuxin, tianchen, tulai}@hust.edu.cn, [⋆]{zhangfan, cz.xu}@siat.ac.cn

***Abstract*— Bus service is an important public transportation. Besides the major goal of carrying passengers around, providing a comfortable travel experience for passengers is also an important business consideration. The crowdedness inside a bus can directly affect the number of people choosing the bus. Traditional approaches to obtain passenger density rely on field investigations, which are both non-scalable and incomplete. The wide adoptions of smart card fare collection systems and GPS tracing systems in public transportation provide new opportunities. In this paper, we associate these two independent datasets to derive the passenger density, and evaluate the effectiveness of scheduling choices. To our best knowledge, this is the first paper which utilizes smart card data and GPS data to calculate the passenger density of bus service.**

## I. Introduction

Bus service is an important public transportation. It is a critical commute tool for many metropolitan habitants.

For a bus service, besides the major goal of carrying passengers around, providing a comfortable travel experience for passengers is also an important business consideration. Bus comfortability contains many aspects, among which passenger density inside a bus is the most important one. The crowdedness can directly affect the comfortability of people in the bus [1]. If the passenger density of a bus line can be obtained, rescheduling can be adopted based on it for efficiency and users' experience [2].

Traditional approaches to obtain passenger density rely on field investigation. There are two drawbacks. First, it involves expensive labor efforts. Nevertheless, field investigation is not scalable: it is difficult to maintain a record staff all day in every station. Second, it is hard to identify the whole trip of a single passenger. As a consequence, it is hard to estimate the effectiveness of the new schedule.

The wide adoptions of smart card fare collection systems and On Board Units (OBUs) in public transportation provide new opportunities. Each smart card can be identified by a unique serial number. Every time a smart card is taped, details of the transaction are recorded. The OBUs, usually with GPS tracing devices, can record the physical position of the vehicle at different time.

As a result, it is possible to associate the datasets of taping records and position traces to derive the passenger density of bus, and evaluate the effectiveness of scheduling choices. Roughly speaking, if a passenger has a trip record in the database, we want to understand: when and which station, the passenger gets on and off which bus? After deriving all these trips' details, we can get the passenger density of every bus at any time. However, to make this possible, there are three challenges need to be solved.

First, as the Fare Collection Device (FCD) and the OBU on a bus may be equipped and maintained by different venters, their datasets are usually independent. So the FCD ID and OBU ID need be matched to associate the two datasets where FCD ID is the unique identity of the FCD that records smart cards' tappings and the OBU ID is associated with the index of the bus. If we want to identify a passenger's boarding on a particular vehicle, we need to associate the FCD ID with the OBU ID first.

Secondly, the get-on station of each trip needs to be derived. As the original purpose of using a FCD is only for fare collecting, the get-on station is not included in a taping record. Fortunately, the time of each taping is recorded. We can derive the get-on position by querying the GPS trace dataset with the taping time as a key. So synchronizing the time in FCD data and the GPS time is the key issue of this challenge.

Thirdly, the get-off station of each trip also needs to be derived. Unlike metro, usually a passenger who takes the bus just needs to tap the card when get-on. As a result, the get-off transaction of a trip is missing. We need to estimate the place where the passenger most likely to get off.

To our best knowledge, this is the first paper which utilizes smart card data and GPS data to calculate the passenger density of a bus. Specifically, we introduce two indices, the *Passenger Density Index of Bus (PDB)* and *Passenger Density Index at Station (PBS)*, to evaluate the crowdedness of people on a bus and waiting at a station respectively. Definitions are given in Section II. The contributions of this paper include:

- By mining the correlations between smart cards' tapping time and vehicles' trajectories, we match each FCD ID to a vehicle's OBU ID (Section III).
- By analyzing a large dataset of real data from over bus 500 lines and 5 metro lines, we develop a method to get a trip's source/destination stations (Section III).
- With the passengers' origin and destination, we derived the spatial and temporal density information in a case study on Bus Line ♯$B606$ (Section IV).
- According to the analysis of ♯$B606$, we find that the

978-1-4799-6077-4/14/$31.00 ©2014 IEEE

density of this line is nonuniform. We evaluate the effectiveness of three different scheduling approaches to decrease the density (Section V).

## II. Overview

### A. Problem Formulation

Our goal is to estimate the passenger's density of a bus line which contains two aspects. One is the passenger density on a bus and the other is the density of passengers waiting at a station. The input for the estimation are records of passengers' tapping smart card and the GPS trace of buses.

Before we model our system and analyze, we make three assumptions: 1)Each bus is equipped with a fixed OBU and FCD; 2)The passengers using coins to pay their bus fees are much fewer than those using smart cards; 3)The passenger using smart card only holds one unique card. For the first assumption, except for rare situations such as device repairing or maintenance, it is true in the City of Shenzhen. However the OBU and FCD are not originally associated with the bus in database when they were designed. So we have to associate them in our first data preprocess step. For the second assumption, it is not always true but statistically it is. Since currently there is no means to obtain records of passengers using coins, we have to omit this minority, so that we estimate the crowdedness based on counting the tapping records. We also believe this tend to be true for the convenience and wide adoption of smart card. For the third assumption, it is true in most time for most people, which will be the base of deriving the destination of a passenger's trip.

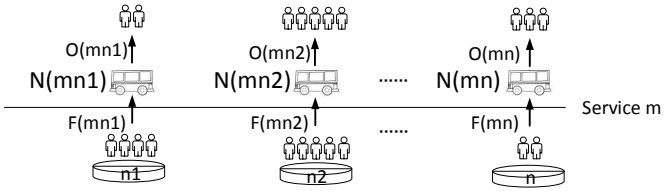


Fig. 1. PDB and PDS

Based on our purpose and the assumptions, the application scenario can be illustrated in Fig. 1 and we can transform the problem of counting passengers into that of counting tapping records, which are what we have in one of our two datasets. To get which station a tapping is made, the second dataset is used. Therefore the crowdedness estimation problem can be formulated into the problem of calculating some metrics with respective with the density. To make this clear, we define two density indices and some corresponding parameters as listed in Table. I.

We define *Passengers Density Index* as the basic measurement for the crowdedness evaluation. To evaluate density for different cases statistically, we can extend the indices over different bus service numbers, stations and time periods. We firstly chose a line to analyze. The bus service number refers to the sequence number of a bus of the line departing from its starting station on a day. Then $\rho_B(m)$ is used to describe the density of vehicle which bus service number is $m$. $\rho_B(p)$ is used to describe the average density of vehicles of the line serving in period $p$. $\rho_B(n)$ is used to describe the density of the vehicle at the Station $n$. With these three indices, we can describe the density of bus in time and space.

TABLE I
DEFINITIONS

| Symbol | Content |
|---|---|
| $\rho_B$ | Passengers Density Index in a Bus (PDB). |
| $\rho_S$ | Passengers Density Index at a Station (PDS). |
| $m$ | Bus Service Number $m$ of a line. |
| $n$ | Bus Station $n$ of a line. |
| $N(m,n)$ | The number of the passengers in the vehicle which Bus Service Number is $m$ between Station $n-1$ and Station $n$. |
| $O(m,n)$ | The number of passengers who get off the vehicle which Bus Service Number is $m$ at the Station $n$. |
| $F(m,n)$ | the number of passengers who get on the vehicle which Bus Service Number is $m$ at the Station $n$. |
| $p$ | different periods in a day. |
| $\mathcal{S}_m$ | the set of all stations that Bus Service Number $m$ pass through. |
| $\|\mathcal{S}_m\|$ | the number of stations in the collection $\mathcal{S}_m$. |
| $B(p)$ | the number of bus services which serve in period $p$ of a day. |
| $B(n)$ | the number of bus services which pass through station $n$ in a day. |
| $L$ | the vehicle load value. |
| $C$ | the station capacity. |

$\rho_B(m)$ focus on the density of different bus service numbers of one line. It is defined in equation 1.

$$\rho_B(m) = \frac{\sum_{n\in\mathcal{S}_m} N(m,n)}{L * \|\mathcal{S}_m\|}. \tag{1}$$

$N(m,n)$ can be calculated by equation 2.

$$N(m,n) = \sum_{i=1}^{n-1} (F(m,i) - O(m,i)) \tag{2}$$

$\rho_B(p)$ focus on the density of different periods of one line. We divided one day into several periods. The density of the $p$ period is defined as followed.

$$\rho_B(p) = \frac{\sum_{m\in p} \rho_B(m)}{B(p)}. \tag{3}$$

$\rho_B(n)$ focus on the density of the vehicle at different stations of one line. It is an average density of the line when it reaches different stations. It is defined as followed.

$$\rho_B(n) = \frac{\sum_m N(m,n)}{L * B(n)}. \tag{4}$$

We use the PBS index $\rho_S(n)$ to measure the density at stations. It is an average density of the station $n$ which belongs to the line in a day. Here, we only consider the passengers who take the line which is analyzed.

$\rho_S(n)$ focus on the density of stations. The average density of station $n$ is defined as followed.

$$\rho_S(n) = \frac{\sum_m F(m,n)}{C * B(n)}. \tag{5}$$

TABLE II
GPS DATASET

| Field | Content | Remarks |
|---|---|---|
| 1 | OBU ID | On Board Unit ID. It is associated with the index of the bus. |
| 2 | Vehicle ID | It is the vehicle registration ID. |
| 3 | Line ID | The line number of the bus |
| 4 | Satellite positioning state | 0 represents positioning, 1 represents unplaced |
| 5 | Longitude | The longitude of the vehicle |
| 6 | Latitude | The latitude of the vehicle |
| 7 | Time | The time of obtaining the location by the GPS device |

TABLE III
SMART CARD DATASET

| Field | Content | Remarks |
|---|---|---|
| 1 | Serial number | It is unique for different records |
| 2 | Card ID | The number of SZT smart card |
| 3 | FCD ID | Fare Collection Device ID. It records smart cards tappings. |
| 4 | Transaction type | 21 represents getting in metro station, 22 represents getting out metro station, 31 represents getting in bus |
| 5 | Time | The time of tapping card |
| 6 | Name | Metro records station name and bus records line name |

### B. The Datasets and System Architecture

Two datasets are involved in the study. The table fields and descriptions are elaborated in Table. II and Table. III.

The two datasets are collections of records in the FCD and OBU of all buses. As is mentioned in the problem formulation, we assume a passenger is uniquely corresponded to a smart card ID, and a FCD ID and an OBU ID are uniquely corresponded to a Vehicle ID. However, only the mapping between OBU ID and Vehicle ID is included in the data table. Matching FCD ID to a OBU or a Vehicle need extra information and some algorithm. Noting that the time of each "event" (either a tapping event or a GPS event) is included in both data table, we can use this information for the matching algorithm. We will discuss the algorthms in Section III.

Therefore, we build a data processing and anaylsis system which is shown in Fig. 2. It contains three layers. In data layer, we collected real probed data from over 500 lines of buses and 5 lines of metro in the database. It includes smart card data and GPS data. Due to the huge amount of data, we use Hadoop based distributed file system to process data. In model layer, we first fused the smart card data and GPS data. It contains matching FCD ID and OBU ID and calculating the passengers' origin. Then, with the passengers' origin, we use our algorithm to predict the passengers' destination. After getting every passenger's OD matrix, we are able to calculate the PDB and PDS indices. In application layer, There are many applications can be designed based on the model. It can be used in analysis on bus scheduling, analysis on bus crowdedness, station construction, bus route planning and so on. In this paper, we made an analysis on bus scheduling. We take three measures to lower PDB and PDS. We also analyzed the effects of different methods.

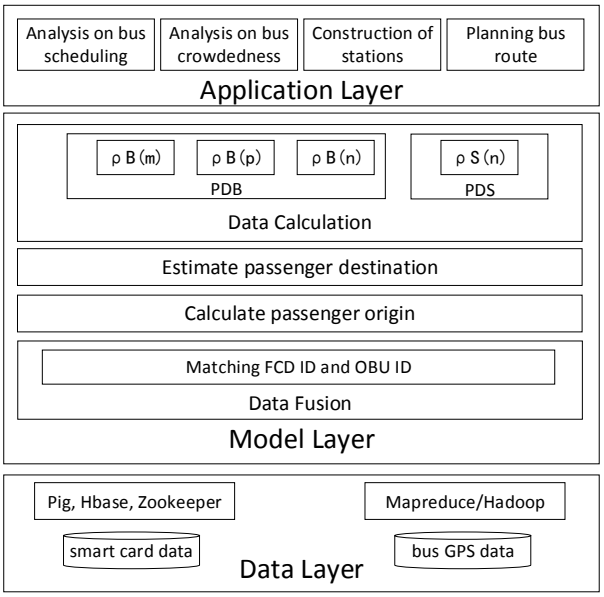


Fig. 2. Approach Chart

## III. DERIVE THE ORIGIN AND DESTINATION OF A PASSENGER'S TRIP

As is mentioned in Section II, to derive the passenger density on a bus and at a station from the data, we need to know where passenger gets on and off the bus. To achieve this, we need first match every FCD to a proper OBU and then estimate the origin and destination.

### A. Match FCD ID and OBU ID

Before we get the origin of a passenger, we need to determine the FCD-OBU pairs. we need to compare the getting station time and the tapping card time. We can choose the minimum time difference as the most likely bus that a passenger gets on. But as GPS time and tapping card time are all equipment time, they are not synchronized. So directly using time as a key to match the pairs may have a high error rate. Here we proposed a new method to match FCD ID and OBU ID.

Matching FCD ID and OBU ID can be described in Fig. 3. Imagine a passenger $A$ takes line $l$ at time $T$. Passenger $A$ may take any bus which belongs to line $l$. According to the time of tapping card and the time of vehicle arriving at a station, we choose the minimum time difference between tapping time and arriving time as the most likely vehicle that the passenger gets on. Without time synchronized, we still can't determine which vehicle the passenger $A$ gets on. But we just mark this pair of FCD ID and OBU ID as a candidate FCD-OBU pair. After analyzing a large amount of data, we choose the pair which have been matched for the

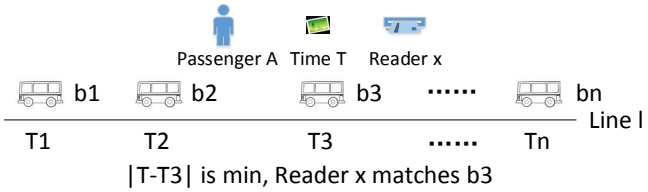


Fig. 3. Match Method

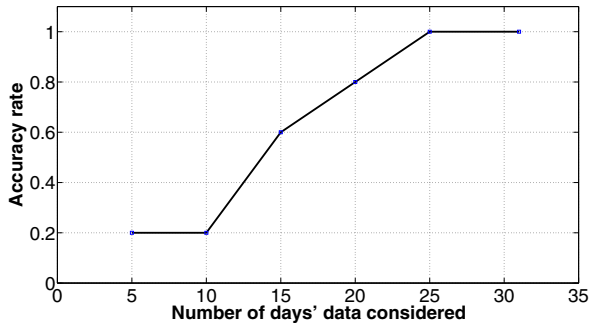


Fig. 4. The accuracy rate of matching result

most times as a FCD-OBU pair. The specific algorithm is shown in Algorithm 1.

**Algorithm 1:** Match FCD ID and OBU ID

1: Find a passenger's record of tapping smart cards from one month's smart card data.
2: Find GPS data of every line that the passenger took.
3: Choose the minimum time difference between tapping card time and arriving time as the most likely vehicle that the passenger gets on.
4: Record the pair of FCD ID and the corresponding OBU ID.
5: Find another passenger and repeat step1 until all the passengers are calculated.
6: Choose the pair which have been matched for the most times as a FCD-OBU pair.

Using this algorithm, we calculated pairs of FCD ID and OBU ID. To evaluate the accuracy of the matching algorithm, we took ten buses and recorded the ten real pairs of FCD ID and OBU ID. We used the ten pairs to check whether our algorithm is effective. The result is shown in Fig. 4. From the figure, we can see that with the amount of data increasing, the accuracy rate of matching goes higher. When we choose 5 days' data, the accuracy rate is only 20%. But When we choose 25 days' data, the accuracy rate can reach 100%. Here,100% means the sample we selected are matched perfectly. The ten vehicles are picked randomly. So we can see in our experiments, it is feasible to use our algorithm with the data of a month's period to match FCD ID and OBU ID.

### *B. Derive a passenger's origin*

With the pairs of FCD ID and OBU ID, we can then calculate the passengers' origin. From the tapping record dataset and the pairs of FCD ID and OBU ID, we can determine which vehicle that the passenger gets on and the time that the passenger taps the card. Here, we can consider when the passenger gets on the bus, she/he will tap the card immediately. Tapping card time is the time she/he gets on the vehicle. With the OBU ID and the getting on time, we can find the location where the vehicle is. After that, we can find a passenger's origin. The algorithm can be described in Algorithm 2

**Algorithm 2:** Find passenger's origin

1: Find a passenger's record of tapping smart cards.
2: Using the real pair of FCD ID and OBU ID, calculate which vehicle that the passenger gets on.
3: Determine her/his origin according to the minimum time difference between tapping card time and arriving time.

### *C. Estimate a Passenger's Destination*

As in ShenZhen one passenger only need to tap the card once to take a bus, so from the records we cannot directly get the destination of a passenger. We need to estimate the passenger's destination from large amount of data. Here we use a method to estimate it, which is based on the trip purpose of a passenger. When a passenger goes out, she/he may tap the card once or more. Here, according to the different times of tapping cards, we can use two different methods to estimate the destination.

*1) Tapping twice or more :* Imagine one passenger takes bus or metro, after she/he gets off at a station, she/he may do two things. One is taking another bus or metro, which we call a connecting trip. The other one is doing nothing. This station is her/his final destination. After a long time, she/he returns. Then we call this trip a round trip. Based on the two different purposes of taking public transportation, we can calculate the destination of a passenger by two methods. If the station is transit station of a connecting trip, the second trip's origin is the first trip's destination. If the second trip is for return, not only the second trip's origin is the first trip's destination, but also the first trip's origin is the second trip's destination. The estimation methods for the different cases can be described in Table IV and Table V. We can determine the second trip's type by calculate the time difference between the first tapping time and the second tapping time. If the time is over 90 minutes, we can determine the second trip is for return. If the time is below 90 minutes, we can determine the second trip is a connecting trip. If one passenger tap the card three times or more, we can confirm the destination one by one.

*2) Tapping once:* If one passenger tap the card once, we take some other measures. Imagine that one passenger takes the bus regular, we know the most common destination of her/him at determined time period, even if one day she/he takes the bus once at the time period, we can also estimate that she/he may go to the most common place. Here workplace and home are the most common destination of a passenger at weekdays. To determine the destination of

TABLE IV
DESTINATION ESTIMATION OF A CONNECTING TRIP

| First Tapping Card Type | Second Tapping Card Type | Estimation Method |
|---|---|---|
| Bus | Bus | The first trip's destination is the second trip's origin. The second trip's destination is unknown. |
| Bus | Metro | The first trip's destination is the second trip's origin. The second trip's destination is known. |
| Metro | Bus | The first trip's destination is known. The second trip's destination is unknown. |

TABLE V
DESTINATION ESTIMATION OF A ROUND TRIP

| First Tap Card | Second Tap Card | Estimation Method |
|---|---|---|
| Bus | Bus | The first trip's destination is the second trip's origin. The second trip's destination is the first trip's origin. |
| Bus | Metro | The first trip's destination is the second trip's origin. The second trip's destination is known. |
| Metro | Bus | The first trip's destination is known. The second trip's destination is the first trip's origin. |

passengers who tap the card once, we analyzed the different purposes of taking public transportation at different time. When a passenger taps the card at morning peak period, she/he may go to work. When a passenger taps the card at evening peak period, she/he may go home. The detailed algorithm is shown in Algorithm 3.

To determine the passengers' destinations who tap the card once, we also need to know the passengers' workplace and home. To determine them, we used a month's data. According to the most frequently origin at different time, we estimate the passengers' workplace and home, which is shown in Algorithm 4.

With these two algorithms, we can determine most destinations of passengers who tap the card once.

**Algorithm 3:** Estimate the destination of passengers who tap the card once

```
if The passenger tap her/his card at weekday then
    Her/his trip purpose may be going to work or home.
    switch (which period the tapping time is in )
    case tapping time is in 6:00-9:00:
        Her/his trip purpose may be going to work.
        Her/his destination is workplace.
    case tapping time is in 11:00-12:30:
        Her/his trip purpose may be going home for lunch. Her/his destination is home.
    case tapping time is in 12:30-14:00:
        Her/his trip purpose may be going to work.
        Her/his destination is workplace.
    case tapping time is in 16:00-20:00:
        Her/his trip purpose may be going home.
        Her/his destination is home.
    default:
        May be difficult to determine
    end switch
else
    May be difficult to determine
end if
```

**Algorithm 4:** Predict passenger's workplace and home

1: Find a passenger's record of taking bus or metro in a month.
2: Take the passenger's most frequently origin in the period of 6:00-9:00 as her/his home. Take the passenger's most frequently origin in the period of 16:00-20:00 as her/his workplace.
3: Find another passenger and repeat STEP1.

## IV. A CASE STUDY: DENSITY OF ♯$B$606 IN SHENZHEN

After calculating the origin and destination of a passenger, we can derive the density of a line. In this paper, we chose the branch line ♯$B$606 in ShenZhen. ♯$B$606 is a branch line but its driving route pass through many residential allotment and working areas. Its first service is at 6:00 and its last service is at 20:00. There are 20 stations one-way. We chose the data of 16th Dec 2013 to analyze. Before we calculate the density of the bus, we first determine how many records we can evaluate its origin and destination. As is shown in Fig 5, we can see there are 3111 records in the dataset of 16th Dec 2013. We can determine 2286 records' origin. This is because if the time difference between tapping time and getting station time is longer than 90 seconds, we will discard it to make sure the origin is accurate. The records decreased about 27 percents. With the origin of the records, we can estimate 2205 records'destination. This is because there are 81 records whose card is tapped once and the tapping time is not in peak periods.

$\rho_B(m)$: The $\rho_B(m)$ index of ♯$B$606 is shown in Fig 6. In Fig 6, the black line stands for the density of upgoing bus. The white line stands for the density of downgoing bus. The horizontal axis represents different bus service numbers. The bus service numbers' order is the time order of its departure. The vertical axis represents the numerical size of $\rho_B(m)$. We can see there are two peak periods in one day's service. They are morning peak and evening peak. In the morning peak the density of downgoing bus is higher and in the evening peak the density of upgoing bus is higher. The density of other periods is very low.

$\rho_B(p)$: The $\rho_B(p)$ index of ♯$B$606 is shown in Fig 7. In Fig 7, we divided one day into five periods. In different periods, passenger's purpose of taking bus may be different. We can see the density of morning peak is the highest. The density of evening peak is higher than other periods. From this figure, we can determine that if we want to ease the density of this line, we should focus on morning peak and

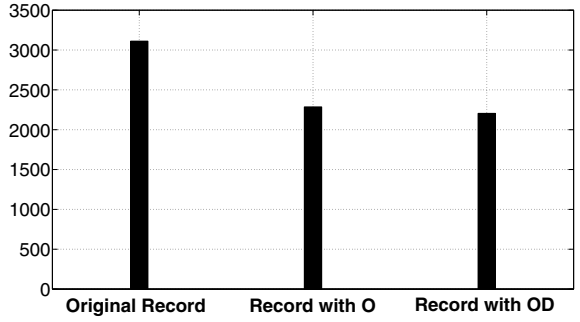


Fig. 5. The number of records which can be calculated

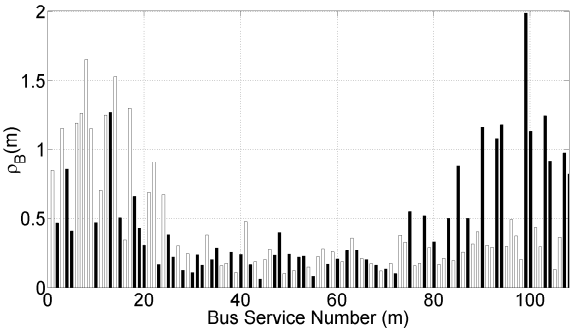


Fig. 6. ♯$B$606's $\rho_B(m)$ on 16th Dec 2013

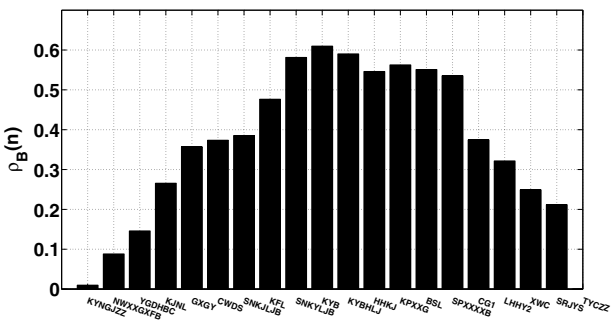


Fig. 8. ♯$B$606's $\rho_B(n)$ of upgoing

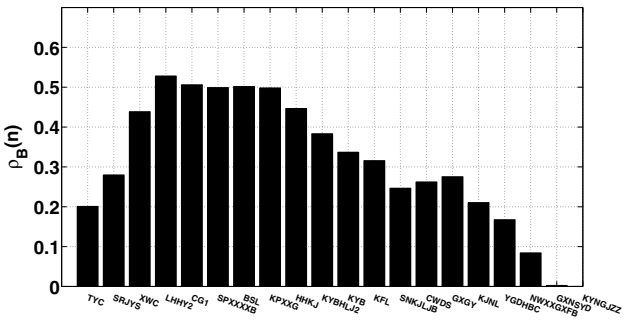


Fig. 9. ♯$B$606's $\rho_B(n)$ of downgoing

evening peak. In the next section, we will have a detailed description.

$\rho_B(n)$: The $\rho_B(n)$ index of ♯$B$606 is shown in Fig 8 and Fig 9. Fig 8 describes the density of upgoing bus. Fig 9 describes the density of downgoing bus. From Fig 8, we can see in the all day service of upgoing, $KYB$, $KYBHLJ$, $SNKYLJB$ are the stations whose vehicle density is most highest. Similarly, from Fig 9, we can see $LHHY2$, $CG1$, $BSL$ are the stations whose vehicle density is most highest in downgoing service. That means if we take ♯$B$606 at these stations, it has a high probability that there are many passengers in the bus.

$\rho_S(n)$: The $\rho_S(n)$ index of ♯$B$606 is shown in Fig 10. It is a three-dimensional figure. The X axis represents the stations of ♯$B$606. We separated the upgoing and downgoing stations. Number 1-20 represents the 20 upgoing stations and number 21-40 represents the 20 downgoing stations. The Y axis represents the bus service number of ♯$B$606. The bus service number order is the time order of its departure. The Z axis represents the $\rho_S(n)$ of♯$B$606's stations. In the figure, we can find in morning peak, the density of stations is the highest. Among them, $TYC$, $CG1$ of downgoing are the most crowd stations. We found these two areas are residential area after investigation. In the morning, there are many people taking bus for work. In evening peak, we found the average density of stations increased. There is not any station whose density is as high as morning peak. After our investigation, we found that is because the working area that ♯$B$606 pass through belongs to the high tech Zones. There are many IT companies in that area. The time of going off work is not fixed.

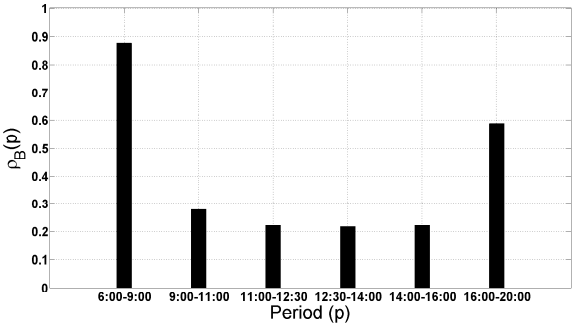


Fig. 7. ♯$B$606's $\rho_B(p)$ on 16th Dec 2013

## V. EVALUATION OF NEW SCHEDULES

From the Section IV, we can see the density of bus and station is not averaged. In morning peak and evening peak, the density increased significantly. This must cause the passengers feel uncomfortable when she/he goes to work with sleepy in the morning or goes home with tired body in the evening. To ease the density of bus and station, we can take two measures.

- Without increasing services and changing vehicles, adjust the departure time to ease the peak periods' density. We can increase the services in peak periods and decrease the services in other periods.

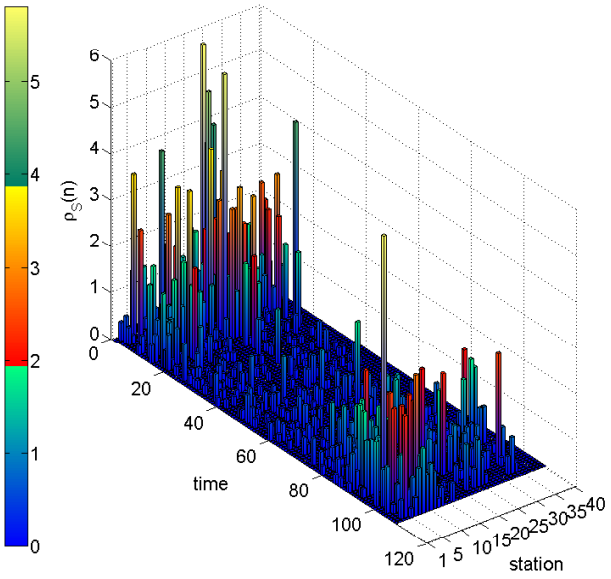


Fig. 10. $\rho_S(n)$ of ♯B606

- Increase the services in peak periods without decreasing the services in other periods.

Using these methods, we simulated the service situation. Before the simulation, we made the following assumptions:

- Passengers' arrival flow obey average distribution. It means that between the two services, passengers get the station evenly.
- All the buses operate normally. It means that if we add a new service between two original services, the new service's running time is half of the sum of the original services' running time.

### A. Change the Departure Time

In this method, we assume that the bus is adequate. That means if we add a new service, it can't happen that there is no bus to drive. We just make sure that the total number of operations is fixed. The scheduling algorithm is increasing the service before the highest density service and decreasing the service whose density is lowest. It can be described as followed.

- **Step1**: Sort the density of morning peak period and evening peak period by descending order.
- **Step2**: Sort the density of other periods by ascending order.
- **Step3**: Increase a new bus number before the highest density service. The new service's departure time is the midpoint of the original highest density service's departure time and its previous service's departure time. The original density becomes half of itself. The density of new service is the same as the original density after changed.

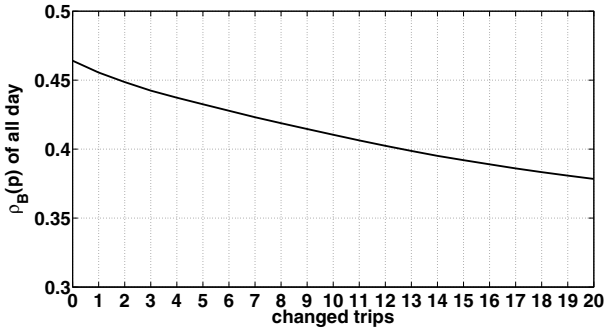


Fig. 11. $\rho_B(p)$ of all day

- **Step4**: Decrease an original service whose density is lowest. After decreased, the density of its next service is the sum of itself and the decreased service's density.
- **Step5**: Repeat Step1.

After the above scheduling method, the average density of the line all day is shown in Fig 11. The horizontal axis represents the number increased in the peak period. The vertical axis represents the $\rho_B(p)$ of all day. From this figure, we can see that by changing the departure time, we can lower the density of the line. But it has a threshold. We can't get a lower density than the threshold.

### B. Increase Service

This method is not good. Because it will rise the operating costs. But here we want to give a trade-off for operators. In this method, we just increase the bus number in morning peak period and evening peak period and we don't decrease the bus number in other period. It can be described as followed.

- **Step1**: Sort the density of morning peak period and evening peak period by descending order.
- **Step2**: Increase a new bus number before the highest density service. The new service's departure time is the midpoint of the original highest density service's departure time and its previous service's departure time. The original density becomes half of itself. The density of new service is the same as the original density after changed.
- **Step3**: Repeat Step1.

With the increasing of service, the density is becoming lower. But it won't maintain the same speed. Fig 12 shows the situation. In the figure, the horizontal axis represents the number increased in the peak period. The vertical axis represents the $\rho_B(p)$ of all day. We can see in morning peak, when the increased service reaches two and nine, the decrease degree of the density begins to decrease. In evening peak, when the increased service reaches one, the decrease degree of the density begins to decrease.

## VI. RELATED WORK

In past, getting information about bus position is based on the on-site recording by hired observers. Larry A. Bowman and Mark A. Turnquist [3] did an observation of seven bus

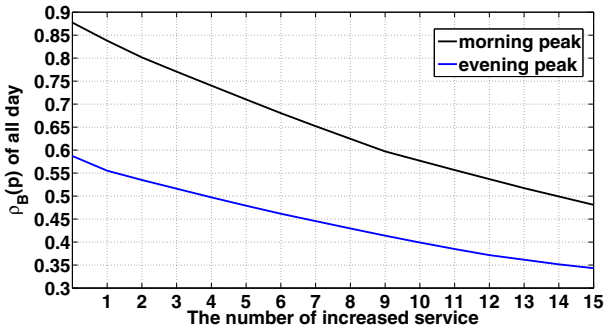


Fig. 12. The change of $\rho_B(p)$ in morning peak and evening peak

stops within 140 hours time, proposed a distribution model of passengers arriving time, and then drew the conclusion that waiting time of passengers is more influential on the reliable behavior of buses. With the Global Positioning System (GPS) rising, the way of getting information changed. James G. Strathman [4] calculated the running time and headways of buses with GPS data, then divided the reliability of buses into peak and off-peak hours indexes. In 2003, Tri-Met, the local transit provider of Oregon, Portland developed a bus dispatch system (BDS) consisting of vehicle location, communications, etc. RL Bertini and A El-Geneidy [5] utilized the archived data to measure the bus performance. At the same time, Smart Card Automated Fare Collection Systems popularity is also increasing in public transport, a large quantity of data is gathered each day in the existing systems. Many researches have been done to collect information from smart card data. Bruno Agard and Catherine Morency [6] presents a common transportation planning / data mining methodology for user behavior analysis from smart card data. Martin and Nicolas [7] presents a model to estimate the destination location for each individual boarding a bus with a smart card. Lijun Sun and Der-Horng Lee [8] present a methodology to analyze smart card data collected in Singapore, to describe dynamic demand characteristics of one case mass rapid transit (MRT) service.

A recent trend of bus analysis is to utilize the Global Positioning System (GPS) and smart cards data to estimate the history of operations of the bus. Many researches have been done to evaluate bus operating conditions. James G.Strathman and Janet R. Hopper [9] evaluated the punctuality of buses of Tri-Met Company in Portland, which was for the operators of Tri-Met. QIN Li-Jun and LV Yan [10] evaluated whether the current bus service is up to the passengers needs using GPS and Smart Card Data. Marie-Pier Pelletier and Martin [11] use smart card data to plan the bus scheduler. In this paper, we owe the priority to the opinions and feelings of public service users, thus evaluating the density of buses and stations,i.e. if such bus lines are comfortable for service users.

## VII. CONCLUSION

In this paper, we presented a series of measurements and their calculation methods for bus crowdedness evaluation. The indices of density on a bus and at a station can be derived from large set of data. We proposed a new method to fuse the GPS data and the smart card data which can reduce the time error in FCD data of and associate the FCD and OBU devices. We further built a model to estimate a passenger's destination. Finally we used the real data to calculate the density of one line. We found that in morning peak and evening peak the density is higher and the density of morning peak is higher than that of evening peak. For this case, we suggested three methods of easing the density and evaluate the effect by simulations.

## ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their valuable comments. This work is supported in part by "National Natural Science Foundation of China (No. 61202303, No. 61100220, No. 61202107)", and by the "Fundamental Research Funds for the Central Universities", and by NSF under grant CCF-1016966.

## REFERENCES

[1] J.-K. Kim, B. Lee, and S. Oh, "Passenger choice models for analysis of impacts of real-time bus information on crowdedness," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2112, no. 1, pp. 119–126, 2009.

[2] A. Wren and N. D. F. Gualda, *Integrated scheduling of buses and drivers*. Springer, 1999.

[3] L. A. Bowman and M. A. Turnquist, "Service frequency, schedule reliability and passenger wait times at transit stops," *Transportation Research Part A: General*, vol. 15, no. 6, pp. 465–471, 1981.

[4] J. G. Strathman, K. J. Dueker, T. Kimpel, R. Gerhart, K. Turner, P. Taylor, S. Callas, D. Griffin, and J. Hopper, "Automated bus dispatching, operations control, and service reliability: Baseline analysis," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1666, no. 1, pp. 28–36, 1999.

[5] R. L. Bertini and A. El-Geneidy, "Generating transit performance measures with archived data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1841, no. 1, pp. 109–119, 2003.

[6] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM*, 2006, pp. 17–19.

[7] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

[8] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. ACM, 2012, pp. 142–148.

[9] J. G. Strathman and J. R. Hopper, "Empirical analysis of bus transit on-time performance," *Transportation Research Part A: Policy and Practice*, vol. 27, no. 2, pp. 93–100, 1993.

[10] Q. Li-jun, L. Yan, Z. Li-Nan, and C. Xu, "Evaluation of the reliability of bus service based on gps and smart card data," in *Quality and Reliability (ICQR), 2011 IEEE International Conference on*. IEEE, 2011, pp. 130–134.

[11] M.-P. Pelletier, M. Trépanier, and C. Morency, *Smart card data in public transit planning: a review*. CIRRELT, 2009.