Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System

Lijun Sun
National University of Singapore
Future Cities Laboratory
Singapore-ETH Centre
sunlijun@nus.edu.sg

Der-Horng Lee
National University of Singapore
dhl@nus.edu.sg

Alex Erath
Future Cities Laboratory
Singapore-ETH Centre
erath@ivt.baug.ethz.ch

Xianfeng Huang
Future Cities Laboratory
Singapore-ETH Centre
huang@arch.ethz.ch

ABSTRACT

Rapid tranit systems are the most important public transportation service modes in many large cities around the world. Hence, its service reliability is of high importance for government and transit agencies. Despite taking all the necessary precautions, disruptions cannot be entirely prevented but what transit agencies can do is to prepare to respond to failure in a timely and effective manner. To this end, information about daily travel demand patterns are crucial to develop efficient failure response strategies. To the extent of urban computing, smart card data offers us the opportunity to investigate and understand the demand pattern of passengers and service level from transit operators.

In this present study, we present a methodology to analyze smart card data collected in Singapore, to describe dynamic demand characteristics of one case mass rapid transit (MRT) service. The smart card reader registers passengers when they enter and leave an MRT station. Between tapping in and out of MRT stations, passengers are either walking to and fro the platform as they alight and board on the trains or they are traveling in the train. To reveal the effective position of the passengers, a regression model based on the observations from the fastest passengers for each origin destination pair has been developed. By applying this model to all other observations, the model allows us to divide passengers in the MRT system into two groups, passengers on the trains and passengers waiting in the stations. The estimation model provides the spatio-temporal density of passengers. From the density plots, trains' trajectories can be identified and passengers can be assigned to single trains according to the estimated location.

Thus, with this model, the location of a certain train and the number of onboard passengers can be estimated, which can further enable transit agencies to improve their response to service disruptions. Since the respective final destination can also be derived from the data set, one can develop effective failure response scenarios such as the planning of contingency buses that bring passengers directly to their final destinations and thus relieves the bridging buses that are typically made available in such situations.

ACM UrbComp'12, August 12, 2012. Beijing, China

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining, Sptial database and GIS*

General Terms

Experimentation

1. INTRODUCTION

Rapid transit systems are increasingly becoming the most important mode of public transportation in many large cities around the world owing to its faster velocity, higher reliability, and larger capacity, as compared with other transport modes. Understanding the demand characteristics of such systems is central to the public transport agencies and operators, so as to manage and improve their services. During a typical weekday, the demand of the bus and rapid transit systems have distinct spatio-temporal characteristics, which have been captured using smart card data, as shown in [7] and [8] respectively. Park et al. [8] studied the demand characteristics of different public transport modes, in particular the rapid transit system, based on the smart card data records in Seoul, South Korea.

The implementation of an automated fare collection system allows public transport agencies to collect large quantities of data, recording passengers activities with detailed time and space information. It has been recognized that there are large potential benefits of using this data to improve public transport planning and operation [9]. As a result, an increasing number of researchers have been using

such data to analyze public transport systems characteristics and passenger behaviors. Bagchi and White [3] have demonstrated the feasibility of obtaining turnover rates, trip rates and the proportion of linked trips from smart card data, which can be further used to adjust such services. For some entry-only smart card systems, trip destination information is not recorded but needs to be imputed. Different methodologies haven been proposed to estimate the origin-destination pairs and alighting time [7, 4]. Jang [5] has studied the travel time and transfer activities in Seoul, South Korea using smart card data, which provides a comprehensive travel time map and basic understanding of transit services. By analyzing smart card data collected in Outaouis, Canada, Agard et al. [1] have identified different trip habits based on the predefined user types and variabilities of trips against time. Utsunomiya et al. [12] pointed out that demand pattern varies with day in week, therefore, different operation schedules should be provided for each day. Some researchers also focus on data processing methods and aim to get more meaningful information from smart card data. In [2], different types of analyses are conducted to support further planning purposes. Potential usage and challenges have also been highlighted.

Lee et al. [6] used smart card data from Singapore which contains detailed boarding/alighting activities to conduct an analysis on bus service reliability, including trajectories, occupancy of buses and in particular the headway distribution along the route since bus bunching occurs at times. Based on this approach, different operating strategies can be applied and tested in a simulation environment with passenger demand as inputs.

To data, information dedicated to identify passenger locations within a MRT system based on smart card data remains scant. Identifying trajectories and occupancies of trains is significant to transit agencies in order to improve the service level by designing timetable, adjusting velocity and increasing/decreasing dwell time at stations, however, these information is difficult to obtain from the operators' point of view. This is different from data generated from bus systems, as rapid transit data records do not feature any time information regarding when passengers board or alight from a train, which leads to difficulties in describing the trajectories and occupancies of trains. Fortunately, smart card data provides us the opportunity to extract this information. In this present study, smart card data is used to extract the spatio-temporal demand variation of the MRT(Mass Rapid Transit) system.

In the light of smart card records of passengers' tapping in and tapping out of the system, a model has been proposed to detect different travel time elements. This model can be regressed based on the assumption that the observations with the least duration between each origin and destination pair record over a given day travel through the system has no waiting time. The regressed parameters can then be employed to indicate the most probable location of every passenger, which further results in a realistic description of passengers' spatio-temporal density and trains' trajectories.

The present paper is organized as follows. Firstly, information regarding the featured smart card data and the MRT service which would be used as a case study, are introduced in the following section. In Section 3, a travel time regression model is proposed and estimated with data records from the passengers with minimum travel time. Section 4 presents the methodology to extract the spatio-temporal density of passengers and trajectories of trains based on the proposed travel time model. Finally, conclusions and an outlook on further research and applications are discussed in the concluding section.

2. RELATED WORK

Spatio-temporal distribution of traffic demand provides potential benefits to both transit operators and passengers. To investigate the spatio-temporal demand of urban road network, a number of efforts have been made with different kinds of urban dataset regarding urban road network. For example, taxies play an important role in urban road networks as probe vehicles, in particular, GPS-equipped taxies can generate large quantity of trajectory data. In most cases, urban taxi drivers are quite familiar with the road network they drive on everyday, especially the spatio-temporal distribution of traffic demand for each time slot, in other words, what they have is human sensed real-time traffic information, therefore, they have more intuitive intelligence and experience in finding fastest route for given origin, destination at certain time. Given this assumption, Yuan et al. [13] has designed a two-stage routing algorithm which leads to a smart driving system based on historical GPS trajectory data. With the help of this system, a fastest route can be generated given the departure time, origin and destination of passengers.

The fastest route generating problem is further studied with GPS trajectory data from urban taxies in [14]. Compared with [13], a Cloud-based computing system is developed to generate real-time fastest driving route in this paper. Regarding to estimating real time traffic information on urban road network, the methodology for estimating the travel time and spatio-temporal traffic density on road surfaces was introduced by using GPS trajectory data [13, 14].

The research conducted by Zheng et al. [15] is also based on GPS trajectories data from urban taxis, whereas it focused on the aspect of urban computing and application. In this study, the flaws in the existing urban planning of Beijing was detected with the mentioned dataset, the result of which provided comprehensive view on urban planning problems. This will help the city planners in decision making in conceive future plans to a larger extent.

In this paper, we focus on investigating the passengers' spatio-temporal density of subway system. To meet this objective, this study is conducted with the help of smart card data which record urban transit activities. Furthermore, the density data also provides us the opportunity to extract trains' trajectories, which are quite important in understanding the service level.

3. DATA PREPARATION

The smart card data used in this study was collected by a fare collection system, kindly provided by the Singapore Land Transport Authority (LTA). The smart card is Singapore's single largest contactless stored value smart card system and is mainly used for payments on public buses and Mass Rapid Transit (MRT) trains since April 2002. For this study, only records with both boarding and alighting stops being on the East West MRT line are selected since it is the most busiest rapid transit service in Singapore, as shown in Figure 1.

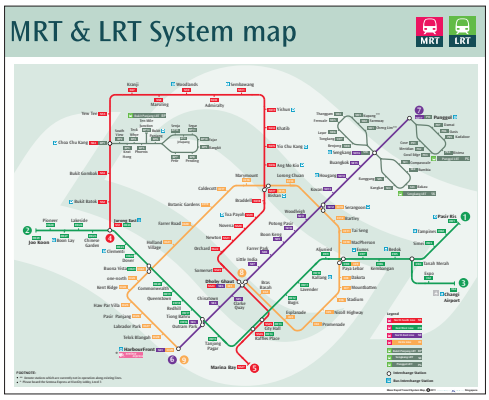


Figure 1: MRT and LRT system map in Singapore [10]

3.1 Smart Card Data

The smart card system was first introduced in Singapore with the aim of providing a convenient, automated fare collection public transportation system. Compared with other smart card data sources stated in [7] and [11], the most significant advantage of the smart card dataset is that it contains precise timing and location information for both boarding and alighting. Hence, transfer information can also be derived. This serves as a basis to generate information on load profile, spatio-temporal variation, and the waiting time of passengers.

In the smart card based fare collection system, the fare charge is calculated based on travel distance, trip mode and different passenger types, so any other information describing these three characteristics can be obtained from the data set.

This present study is conducted based on smart card records of one entire week in April, 2011 provided by the Land Transport Authority (LTA) of Singapore. To test the presented methodology of identifying spatio-temporal density and train trajectories, a one day sample is used.

3.2 Case Study - EW MRT Services

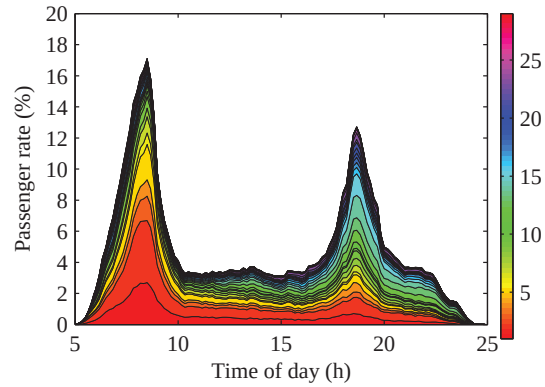
In this study, the East-West (EW) MRT service, which is known as the green line, is chosen to investigate the demand characteristics and test the proposed strategies in order to identify passengers' spatio-temporal density and trains' trajectories. This service has 29 stations moving in both directions. Figure 1 shows the general map of MRT and LRT (Light Rapid Transit) systems in Singapore. The case study examined is service that is on the green line, but the two stations on the extension line leading to Changi Airport are not included [10].

For this study, records of the time taken for passengers to tap in and tap out are used, along with boarding and alighting stations, and passenger types. Other information such as the locations of stations along the routes and characteristics of stations are obtained from supplementary information provided by LTA.

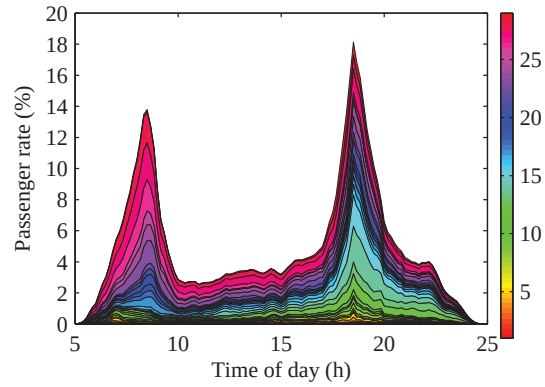
4. DEMAND PATTERN

In this section, travel demand patterns based on the data extracted from smart card data are described. With the help of the smart card data, it is possible to estimate how many passengers are in the MRT system at a given time t, for each station. To this end, records with tapping in time, known as $t_{in} < t$ and tapping out time $t_{out} \geq t$ are identified as passengers in the MRT system.

Figure 2 shows the number of passengers at each station during the course of the day, for train services in both directions, on a Monday in April 2011. It is observed that the demand for each direction has its own characteristics and both have significant morning and evening peaks.



(a) EW1-EW29



(b) EW29-EW1

Figure 2: Demand characteristics on EW line

Figure 2a and Figure 2b show a distinct morning peak at 8:30 am for both directions. Likewise, the evening peak can be observed at 6:30 pm. The different shapes of the two graphs indicate significant commuting in both directions with the morning commute direction from EW1 to EW29 being somewhat more distinctive.

It can be seen as well that in the morning peak, most of the demand originates from the first and last few stops along the line, while in the evening peak, most of the demand departs from the middle section of the line/service. In fact, this pattern maps effectively with land usage in Singapore. The predominant residential locations are located along the outskirts of Singapore and the work locations are centralized at the middle part of the city. During a typical weekday, most of the trips generated in the morning and evening peaks

are commuters who travel to their work locations and back home respectively.

Such demand characteristics provide a basic understanding of an MRT service and the travel demand patterns of commuters over a typical weekday. The characteristics can be helpful to fine tune demand responsive train schedules or to define a more reliable strategy regarding the operation of MRT services.

5. TRAVEL TIME AND LOCATION OF PASSENGERS

5.1 Travel Time

Unlike the bus system, the MRT boarding and alighting times of individual passengers cannot be extracted directly. The tapping in and tapping out of their smart card takes place at the ticket gantry of the MRT station which is typically located on another floor of the station, typically one level below the entrance of station. Therefore, we cannot assign passengers to single trains directly. To make this information available to transit operators, a model describing passenger's movement between tapping in and tapping out would be required. In this study, such a model is proposed.

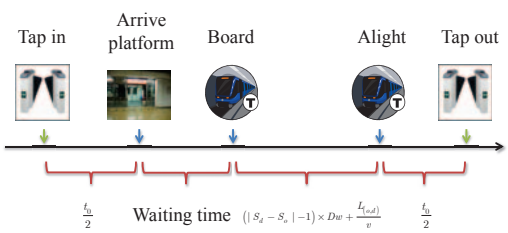


Figure 3: Activity chain of a typical subway trip

Figure 3 shows the typical activities for an MRT train ride. The trip begins with passenger tapping in at the ticket gantry. The passenger then makes his/her way to the platform, boards a train to travel on his/her journey, alights, and ends the journey by tapping out at another ticket gantry at his/her final destination. The waiting time can be calculated as the interval between passenger's arrival at the platform and upon boarding the train. In the smart card dataset, the exact time of tapping in and tapping out are recorded. The interval between these two activities is the total time which a passenger spends in the MRT system. However, as stated, the boarding time and alighting time can not be obtained directly because of the uncertainty in the length of waiting time. This however needs to be imputed.

From all the passengers having the same origin and destination pair, the passenger with the minimum travel time can be located. In this study, the travel velocity of trains is assumed to be constant, therefore, the passenger with the minimum travel time also has the minimum waiting time. Due to the large quantity of data used in this study, the waiting time of these fastest passengers are assumed to be zero, which means that the passengers can board a train immediately upon arriving at the platform.

The time interval between boarding and alighting is assumed to comprise two parts. First, the total running time between every two adjacent stations, and secondly the total dwell time at internal stations. From this, a general travel time model can be formulated as follows:

$$T - T_w = t_0 + (|S_d - S_o| - 1) \times Dw + \frac{L_{(o,d)}}{v} \qquad (1)$$

where T_w is the waiting time while t_0 comprises two parts, the time spent tapping into the station to the time when a passenger arrives at the platform, and the time spent alighting from the train to tapping out of the station. S_o and S_d are the index of the stations, thus $|S_d - S_o 1| - 1$ is the number of stations a passenger has passed, excluding the origin and destination stations. Dw is the average dwell time at each station, which is assumed to be a constant value for all stations without considering the boarding and alighting demand. $L_{(o,d)} = |D(d) - D(o)|$ is the distance from the origin station to the destination station and v is the velocity of the trains. Thus, in this proposed model, only T_w, Dw and v are unknown.

Based on the minimum travel times for each origin-destination pair, this travel time model can be estimated with the fastest passengers who generally have $T_w = 0$. In this regression analysis, the minimum travel time records with an origin same as destination are removed so that the size of the regression data for both directions is $N^2 - N = 841$. The results of this is a travel time model are shown in Figure 4 and Table 1.

Table 1: Regression result of travel time model

Parameters	Value	t stat	p value
$t_0(s)$	109.75	48.5787	0.0000
$Dw(s)$	65.76	61.4119	0.0000
$v(m/s)$	21.63	61.8186	0.0000
R^2		0.9981	

The regression results indicate that dwell time at stations is about $65s$ and that the travel velocity of trains is about $22m/s$, which are in accordance with effective values. Figure 4 shows the observed travel time for the fastest passengers from smart card dataset and the predicted travel time based on the proposed model.

5.2 Determining Location

Given the variability of the platform waiting time and availability of records of both the tapping in and tapping out of the smart card, it would be wise to use the latter for determining the passengers' location. Based on the previous travel time model, a passenger's location L at certain time t for two directions can be described by following Equation 2,

$$T_a - t = \begin{cases} \frac{D(d)-L}{v} + (|S_d - S_n| - 1) \times Dw + \frac{t_0}{2} & \text{if } D(d) \geq L \\ \frac{L-D(d)}{v} + (|S_d - S_n| - 1) \times Dw + \frac{t_0}{2} & \text{if } D(d) < L \end{cases} \qquad (2)$$

where T_a is the time when a passenger taps out of the station, and n, S_n are the number of stations which the passenger has journeyed through and the location of that station respectively. In Equation (2), t_0 is likewise divided equally into two parts, so only $\frac{t_0}{2}$ is considered for determining the

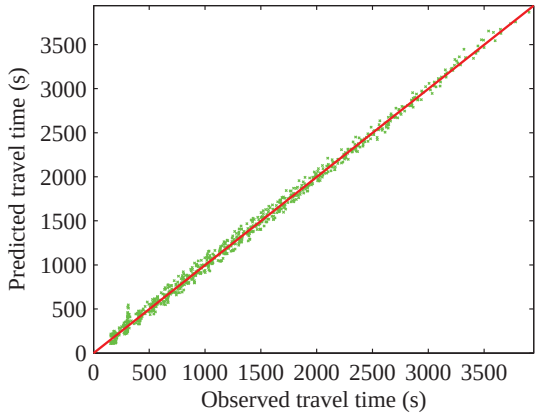


Figure 4: Predicted versus observed travel time for the fastest passengers

passenger's location based on the tapping out smart card record.

The temporary location of any passengers boarding at station k traveling in direction 1 (fulfilling $D(d) > D(o)$) can then be described by Equation 3.

$$L(k) = \left(t - T_a + (|S_d - k| - 1) \times Dw + \frac{t_0}{2}\right) \times v + D(d) \tag{3}$$

To distinguish between passengers waiting on a platform an travelling on a train, Equation (4) is proposed. For all the possible stations in $o, \cdots, k, \cdots, d$, if the first station k^* can be found which satisfies Equation (4), the permanent estimated location of the passenger is $L(k^*)$, where $P(k^*)$ is the location of station k^*.

$$L(k^*) - P(k^*) \geq 0 \tag{4}$$

For the opposite direction, the same method can be applied assuming that the passenger has just passed station k, then the temporary estimated location of this passenger is

$$L(k) = \left(T_a - t - (|S_d - k| - 1) \times Dw - \frac{t_0}{2}\right) \times v + D(d) \tag{5}$$

Then, for all the possible stations in $o, \cdots, k, \cdots, d$, if the first station k^* can be found which satisfies Equ (6), the estimated location of the passenger is $L(k^*)$, where $P(k^*)$ is the location of station k^*.

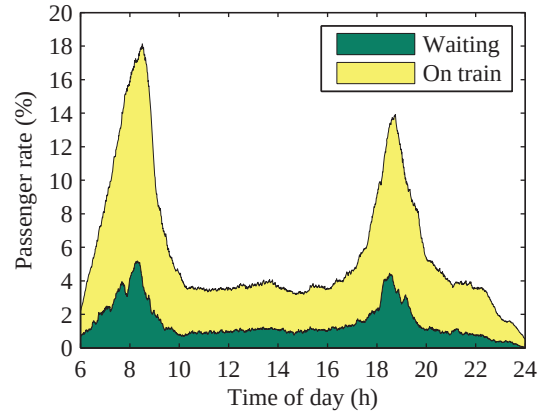
$$P(k^*) - L(k^*) \geq 0 \tag{6}$$

5.3 Waiting Passengers

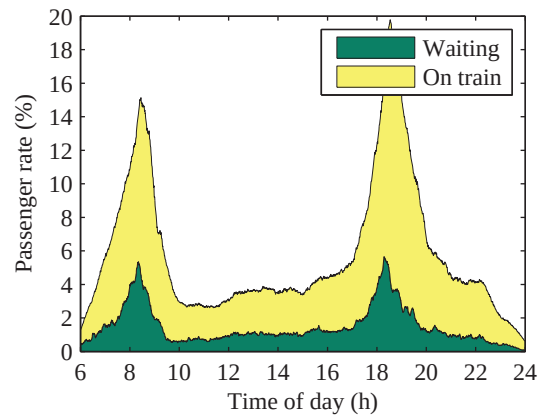
Based on the location model in Section 5.2, if for all the possible stations $o, \cdots, k, \cdots, d$, no station k^* satisfies Equation 4, it must be assumed that the passenger is in the MRT system but not on a train which, according to the travel time model, means that the passengers is either on the way to the platform or waiting there.

In other words, based on location estimation procedure the demand in the subway system can be contiously categorized into two groups: passengers who are on board the trains and passengers who are waiting for their trains.

Figure 5 shows the number of waiting passengers and on board the trains for both directions.



(a) EW1-EW29



(b) EW29-EW1

Figure 5: Demand of waiting and onboard passengers on EW line

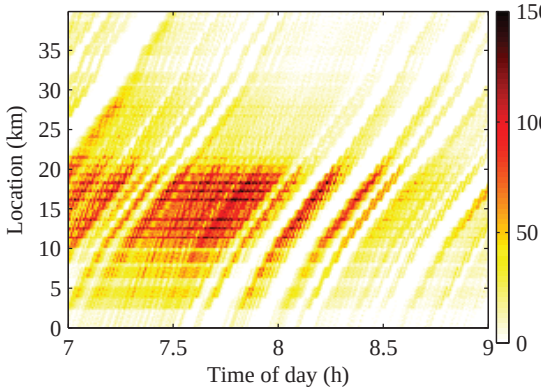
Compared with Figure 2, Figure 5 provides time-volume relationship for both trains and platforms. This serves as a basis for the spatio-temporal density model presented in the next section. Furthermore, for any point in time and any station, the number of passengers located at the respective station can be derived, which is crucial in the event of train breakdowns or evacuations, in order to determine an effective response strategy.

6. SPATIO-TEMPORAL DENSITY AND TRAJECTORIES

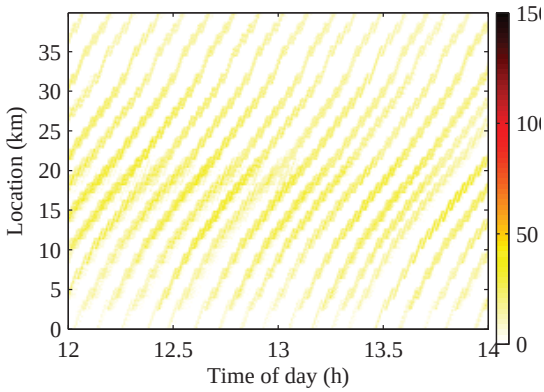
Section 5 describes the data processing to determine passengers's locations based on the proposed travel time model. In this section, the results of applying the described method using the smart card data records of a Monday in April 2011. The travel time model has been regressed with the travel time of the fastest passengers with the same data set. These two models make it possible to extract the spatio-temporal density of passengers onboard a train. Furthermore, the trajectories of trains can also be identified based on the spatio-temporal density figure.

Based on the location estimation model, for all the pas-

sengers onboard as shown in Figure 5, their locations at any time t can be determined. As a next step, the spatio-temporal density relationship can be constructed using the estimated number of passengers within a certain length interval. Figure 6 shows the spatio-temporal density of passengers, from $7am$ to $9am$ in the morning and $12pm$ to $2pm$ in the afternoon respectively for one direction, in intervals o f $100meters$ and $30seconds$. The colors indicate the passenger density who are onboard a train at a certain time and location. Intuitively, the location estimating model will work better for passengers with less travel time for each origin-destination pair, because there would be more variations for longer travel times for certain origin-destination pair, such as the cumulative difference in dwell time and velocity.



(a) EW1-EW29,7a.m.-9a.m.



(b) EW1-EW29,12p.m.-2p.m.

Figure 6: Spatio-temporal density of passengers on EW line($pax/100m$)

Despite some decentralization in the density figure due to non-observed variability, distinct spatio-temporal relationships can be detected as well. This applies especially to graph (a) which plots the density distribution for midday. However, for peak hours, as depicted in graph (b), the assignment to single trains does not appear to be so straightforward. Here, additional information such as effective train operations on a given day or at least the trains schedule would help to consolidate decentralized density observation to individual train trajectory. After such a procedure, passenger loadings could be determined for every train along the entire Ease-West line. Such information is ideally suited to serve as a basis for developing failure response strategies. Since origin and destination pairs for every observation are known, one could use such data in the event of service disruption for the route planning of contingency buses which would act as a substitute for the disrupted train service. Currently, such buses typically run along the interrupted track section, and serve as bridge services. However, depending on the demand patterns and spatial distribution of the final destination, other strategies such as direct buses to highly frequented destinations might provide better service for affected passengers. Because of the very limited time for replacement service planning after an incident, failure response plans need to be prepared in advance and be readily available in the event of an incident. Compared to a system based on real-time information, the retrospective nature of this study is therefore advantageous. However, given the changing demand patterns over a day, a series of different service dispatch plans would need to be prepared to suit the prevailing demand conditions at a given point in time optimally.

7. CONCLUSION

In this article, the demand characteristics of the case study of one MRT service was investigated using smart card data collected in Singapore, with the objective of identifying effective commuter loadings for every train service.

A travel time model has been proposed by reconstructing a typical MRT trip into segments. The model was regressed using the data collected from the fastest passengers for each origin-destination pair. Based on the regression results, a location estimation model was developed to distinguish between passengers travelling on trains and waiting on platforms.

The location estimation model was then applied to all MRT train passengers. Based on the resulting spatio-temporal density plot, it appears feasible to group observations together to individual train trajectories. Such information in turn, has great potential to improve current disruption response plans. Optimizing demand responsive failure response plans based on origin destination demand data, however, is a complex and extensive problem, especially since a multitude of such plans would need to be prepared given the demand fluctuations over a day. Further research would therefore need to focus on developing heuristics that allow one to generate failure response efficiently.

The proposed model can be improved by accounting for station specific access and egress times t_0 given the different layouts of MRT stations. In terms of applying this to real world scenarios, the scope of the analysis needs to be extended from a single line to the whole MRT network which would require consideration of transfers. We will conduct a further study in the future.

8. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Singapore Land Transport Authority for supporting this research and providing all the required data. Furthermore, the financial support from the National Research Foundation of Singapore, the funding authority of the Future Cities Laboratory, is greatly appreciated.

References

[1] 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*, Eaint-Etienne, France, May, 17-19 2006.

[2] K. Alfred Chu and R. Chapleau. Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (2063):63–72, 2008.

[3] M. Bagchi and P. White. The potential of public transport smart card data. *Transport Policy*, 12(5):464–474, 2005.

[4] J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, (1817):183–187, 2002.

[5] W. Jang. Travel time and transfer analysis using transit smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, (2144): 142–149, 2010.

[6] D. Lee, L. Sun, and A. Erath. Study of bus service reliability in singapore using fare card data. In *12th Asia-Pacific Intelligent Transpotation Forum*, Kuala Lumpur, Malaysia, April 2012.

[7] M. Munizaga and C. Palma. Estimation of a disaggregate multimodal public transport originĺcdestination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18, 2012.

[8] J. Park, D. Kim, and Y. Lim. Use of smart card data to define public transit use in seoul, south korea. *Transportation Research Record: Journal of the Transportation Research Board*, (2063):3–9, 2008.

[9] M. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19 (4):557–568, 2011.

[10] SMRT. MRT & LRT system map, 2012. `http://www.smrt.com.sg/trains/images/All_CCL_CLE_System_Map_LRT_lines_big.jpg` [Online; accessed May 18, 2012].

[11] 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*, 11(1):1–14, 2007.

[12] M. Utsunomiya, J. Attanucci, and N. Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board*, (1971):119–126, 2006.

[13] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 99–108, San Jose, California, 2010.

[14] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 316–324, San Diego, California, USA, 2011.

[15] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 89–98, Beijing, China, 2011.