Journal of Transport Geography xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Smart card data-centric replication of the multi-modal public transport system in Singapore

Xiaodong Liu, Yuan Zhou, Andreas Rau*

TUM CREATE, 1 CREATE Way, #10-02, CREATE Tower, 138602, Singapore

ARTICLE INFO

Keywords:
Public transport
Smart card data
Direct assignment
Supply and demand

ABSTRACT

This paper proposes an innovative method of replicating the multi-modal public transport system in Singapore with high precision using smart card database. It replicates the operation of public transport system with known exogenous passenger demand and provides many operational details, including passenger inter-modal trip chains, operational timetable, and detailed transfer behaviour. The paper elaborates on the methodology of the replication including data cleaning, filtering, processing and converting the collected data to meaningful information such as bus journey trajectories and metro system timetable. Thereafter, actualised passenger trip chains are directly assigned to the replicated public transport supply. The resulting replication covers almost 96% of trips made in public transport in Singapore. It provides solid quantitative information on several aspects to support decision making, including precise temporal and spatial travel demand analysis, transfer pattern analysis, traffic condition investigation and bus utilisation analysis.

1. Introduction

With the progressive growth of travel demand in cities, public transport systems have become life lines for daily commute. Unfortunately, public transport infrastructures consume large financial investments and hence, impact daily operations. In this respect, public transport modelling is crucial in providing a comprehensive understanding of certain scenarios to decision makers so that they could respond accordingly to improve the service quality. Conventional ways of public transport modelling include the four-step, activity-based, and the emerging agent-based approach. Most of them depend on data from surveys and help decision makers understand macroscopic situation in the current system or future scenarios so that the public transport system could be optimised accordingly.

The widely implemented smart card systems collect precise information on passenger trips in public transport systems. With high penetration rate, such smart card database, as well as other big data in transportation systems provide rich information on passenger patterns and operational details such as exact boarding and alighting time and location, passenger inter-modal trip chains. It brings in a new opportunity to replicate the current situation so that many details of actual operation could be revealed. With the collected detailed information, replications of public transport systems could be created. Such replications, compared to conventional models, provide better, detailed and comprehensive operational information at microscopic level which could be directly used for improving the existing operation of public transport systems. However, such approach lacks the capability of forecasting future scenarios. Therefore, the application of such approach is to describe current scenario and should be complemented with conventional approaches to conduct prospective analyses to support decision making. This research aims to replicate the public transport system in Singapore using the smart card database. The resulting replication is expected to provide operational details at microscopic level and be utilised by decision makers for different aspects, including understanding passenger public transport travel behaviour, identifying operational bottlenecks.

In this paper, conventional public transport modelling approaches and recent research using smart card data are reviewed in Section 2. The data sources, development environment and methodology of the replication process are introduced in Section 3. Results and applications, as well as a comparison between the proposed replication and conventional models are presented in Section 4. The last section draws a conclusion from the study while providing the scope for future research.

2. Literature review

Throughout the history of transport modelling, the most prominent approach has been the four-step model, which remained mostly unaltered from the 1960s (Ortúzar and Willumsen, 2011). It provides a

* Corresponding author.
E-mail address: andreas.rau@tum-create.edu.sg (A. Rau).

https://doi.org/10.1016/j.jtrangeo.2018.02.004
Received 21 December 2016; Received in revised form 14 December 2017; Accepted 19 February 2018

systematic framework for modelling both private and public transport. As its name suggests, this classic model is presented as a sequence of four sub-models: trip generation, distribution, mode choice, and assignment. Some advanced approaches handle the four steps simultaneously instead of following the certain sequence (Lohse et al., 1997; Vrtic et al., 2007). Because travel is derived demand due to change of activities undertaken at different locations by individuals or groups, a lot of research has been done to review the relationships between travel and activities as well as their interactions with individuals or households. The relationship between certain activities and their constraints on time and location was proposed by Hägerstraand (1970). Recker et al. (1986a,b) examined the travel pattern within households and developed a model which enumerates feasible activity-travel patterns and selects the ones most likely to be chosen by other household members. The dynamics of activity patterns and their influence to travel was investigated by van der Hoorn (1979, 1983) using multi-day activity diary. The activity-based models have generally improved the quality of modelling with the inclusion of relationships between different trips by a sequence of different activities and interactions between individual and other household members. A further extension of such approach is the agent-based approach (Helbing and Balietti, 2012) which involves more details compared to the others including individual characteristics and interactions between different agents. The agent-based models appear to be ideal for studying the interdependencies between individual activities.

All three models, namely four-step, activity-based and agent-based models as mentioned above are widely used for modelling both general transport systems and public transport systems. However, they have certain limitations. First of all, large amount of surveyed data is normally required. The intensive requirement on manpower, time and financial investment limits the coverage of the survey and hence influences the data quality. For example, the Household Interview Travel Survey (HITS), a national-wide survey in Singapore, is conducted every 4 or 5 years and covers around 10,000 households per survey. The latest HITS 2012 lasted for 1 year from June 2012 to May 2013 (Land Transport Authority Singapore, 2013). Also, the mathematical models used by the conventional approaches provide only estimated results and require careful calibration. For example, the gravity model commonly applied in the distribution sub-model in the four-step approach and the Monte Carlo process used by the activity-based models require careful calibration (Ortúzar and Willumsen, 2011). Improper model calibration may lead to biased results. Thus, the produced results, with much estimation involved, provide a macroscopic analysis on actual operations of transport systems wherein many details including demand analysis during specific times of day cannot be revealed.

As introduced in Section 1, the implementation of smart cards in the public transport system provides large amount of precise information and enables many new methods for research in public transport (Bagchi and White, 2005; Pelletier et al., 2011). With the information-rich database collected by smart cards, actualised passenger trips and real travel ODs can be extracted (Munizaga and Palma, 2012) and passenger travel behaviour can be better understood (Agard et al., 2006; Du et al., 2017). Furthermore, with large amount of data collected, many operational details of public transport systems can be revealed. For example, passengers' spatial-temporal distribution within public transport system can be extracted (Sun et al., 2012). Such information can be further used to identify operation bottlenecks and help to improve the quality of service. Besides the applications in public transport, the collected smart card database could also be used to evaluate general operation of transportation systems (Liu et al., 2016). However, studies (Bueno et al., 2017; Hamre and Buehler, 2014) show that to conduct prospective analyses, the application of smart card data still needs to be complemented with traditional approaches. Recent studies on both conventional and big data-oriented research show a convergence and the joint effort of both approaches result in mutual benefits (Chen et al., 2016).

This research aims to develop methodologies to fully utilise the collected smart card data for replicating the actualised public transport system with several operational details. The resulting replication is expected to directly contribute towards a concrete understanding of the current condition of the public transport system.

3. Replication of the public transport system in Singapore

A smart card database was provided by the Land Transport Authority (LTA) of Singapore. The database contains substantial information on all trips made with smart cards from 1 August 2013 to 31 October 2013, altogether for 92 days. A total of 517,203,124 trips were recorded for both bus and metro systems, on an average more than 5.5 million trips per day. The smart card system in Singapore is very advanced. It has a very high penetration rate of 97% while the remaining 3% trips paid by cash on-board (Prakasam, 2008). The system requires passengers to check in and out of the system. This requirement is far from being the usual case worldwide and can be found in some cities (e.g. Singapore and Beijing). The check-outs are of great importance to the replication because they complete passenger information including alighting time and location. It also enables the identification of passenger trip chains by considering alighting time and boarding time to the next trip. With high penetration rate and comprehensive information collected, the database could sufficiently represent the travel patterns of public transport users. Based on this condition, the exact data processing and replication methodology were developed according to the smart card database and other complementary data.

To reproduce the operational conditions of Singapore's public transport system, this research uses a direct assignment method to assign all collected complete passenger trip chains to the public transport network. PTV Visum was used for the direct assignment. With direct assignment, passenger trip chains are not assigned to routes estimated by behaviour models, rather, each segment of the trip chain is directly assigned onto bus or metro line route with known check-in and check-out information. Different from other methods requiring demand of each origin-destination (OD) pair, direct assignment requires complete information on individual trip chains, including starting time and stop, ending stop, line routes taken of each segment of trip chains. As a result, the original database needs to be further processed to provide information on passenger trip chains and exact operational trajectories of buses and trains.

3.1. Data source

The smart card system in Singapore requires passengers to check in for boarding and check out for alighting in both bus and metro systems. Each trip is recorded in the database based on successful check-in and check-out. Such mechanism provides complete and accurate information on stop and time where passengers board and alight. Table 1 lists all data types collected by the smart card system. They can be grouped into three categories, namely identifications, ride information and service information. The identifications include journey ID and card ID. The journey ID is unique for each passenger trip chain in the public transport system. Multiple trips within one trip chain share the same journey ID. The smart card system considers consecutive trips with transfers less than 45 min as trips made for a same journey. Card ID and passenger type (student, adult or elderly people) are also collected as passenger information. The ride information records specific data on each trip. Mode of service (metro or bus), boarding and alighting stop, check-in time, ride time and distance are all recorded. Additionally, for all trips made with buses, the bus service line and direction, with bus plate number and departure sequence number are also collected. The combination of this data could be used to identify unique bus journeys of certain days.

The smart card database on its own does not provide complete information for replicating an entire public transport system. For

Table 1
Collected trip information by the smart card system in Singapore.

Category	Data type	Remarks
Identifications	Journey ID	
	Passenger information	Card ID and passenger type
Ride information	Mode	Metro or bus
	Origin	Stop ID
	Destination	Stop ID
	Check-in time	Day and time of day
	Ride time	Time difference between check-out and check-in
	Distance	Absolute travel distance along this trip
Service information (for buses only)	Line information	Service number and direction
	Bus information	Plate number and bus departure sequence number

Table 2
Data sources for replicating the public transport system.

Category	Type of data	Source and collection time
Network geometry	Road network	• Provided GIS shape files[a], 2012
	Stop coordinates	• Provided GIS shape files, 2012 • OpenStreetMap, 2014
Public transport supply	Stop sequences	• Provided data[b], 2014 • Data mining from the smart card database, 2013
	Metro and bus time profiles • Departure headways • Travel times between stops • Dwelling times	• Field survey, 2015 • Data mining from the smart card database, 2013
	Transfer times • Stop-to-stop walking time (bus system) • Gate-to-platform walking time (metro system)	• Field survey, 2015 • Google Maps, 2015
Public transport demand	• Complete personal trip chains • Stop-to-stop OD matrices	• Data mining from the smart card database, 2013

[a] The GIS shape file is provided by the Singapore Land Authority.

[b] The data is provided via LTA Singapore's DataMall (Land Transport Authority Singapore, 2016).

example, road network geometry is needed as a base for public transport network. Operational details such as walking time from fare gates to platforms within metro stations are also required. The other required information is either provided (e.g. geography information system (GIS) shape file, line route stop sequences) or surveyed (e.g. travel time between stops of the metro system, walking time between fare gate and platforms) (Zhou et al., 2015). Table 2 lists all data involved in the replication work and their sources. The collected data sets were adjusted to the same time horizon (August to October 2013) to fit the provided smart card database.

3.2. *Development environment*

The replication work was conducted mainly on a server running Windows Server 2012 R2. The server is equipped with two Intel Xeon CPUs (E5-2640 v3 @ 2.6GHz) and 128GB RAM. Multiple software programs were used. PTV Visum is the main tool used for the replication and visualisation. Direct assignment is also conducted using PTV Visum. Microsoft SQL Server 2014 is deployed for database management. Python programming language is used for data processing and batch-process work in PTV Visum.

3.3. *Data cleaning and filtering*

A small amount of erroneous data was detected in the database. Two main types of errors were discovered. First, a few thousand lines of data were found containing negative travel times. Due to the small share, such data was removed. Second, some other data sets were found to have recorded non-existing boarding or alighting stops of the recorded bus line route. The erroneous data was corrected using valid information on correct boarding or alighting stops, service route and travel distance.

The provided database covers travel patterns for three consecutive months which includes weekdays, weekends, school holidays, public holidays, and days with special public transport operation. A preliminary analysis of the database showed that the travel behaviour on Fridays, weekends, holidays and days with special public transport operation deviates significantly from regular weekdays (Monday to Thursday). Hence, aiming at the most crucial situation during normal weekdays, only data records from 41 days (Monday to Thursday without special events) were retained and further utilised for the subsequent steps.

3.4. *Replication of public transport supply*

The fundamental principle to replicate the public transport supply is to convert the individual trip data records into bus and metro operational data records.

While boarding on or alighting from buses, passengers are required to check in and check out inside the vehicles. The collected timestamps could be used to estimate arrivals and departures of buses. When passengers enter or leave the metro system, they check in and check out at fare gates, but not inside trains. The collected timestamps compared to the train arrivals or departures, contain walking time between fare gates and platforms and waiting time at platforms. Hence, the timestamps collected by metro system cannot be used to estimate train arrivals or departures. As a result, the replication of bus and metro systems were conducted in different ways. For bus system, the operational information was directly retrieved from the check-in and check-out data. For metro system, smart card data was complemented with surveyed data (e.g. walking time between platform and the fare gate) to speculate the operational information.

3.4.1. *Bus system*

As introduced previously, the combination of bus service number, direction, plate number and departure sequence number could be used to identify unique bus journeys of the day. Thus, passenger check-in and check-out information could be categorised per bus journey. Since the check-in and check-out in buses are accomplished on-board, the timestamps of all passengers boarding or alighting certain buses at particular stops could be used to identify the arrival and departure times and then extended to compile timetables (Michalski et al., 2016). In some case, check-in and check-out data is not sufficient or not available due to lack of boarding or alighting passengers. As a result, some arrival and departure times of buses at certain stops were not directly available. Such missing information was interpolated by linear regression using valid trajectories from other bus journeys covering the same sections during similar period of day. Fig. 1 shows an example of bus journey trajectories of a bus line route in the morning. There were five bus journeys presented. Along the five trajectories, there are dots presented by triangles or squares. Those dots represent the arrival and departure times of buses at certain stops. The ones with squares are calculated from the smart card database while the triangles are estimated. The

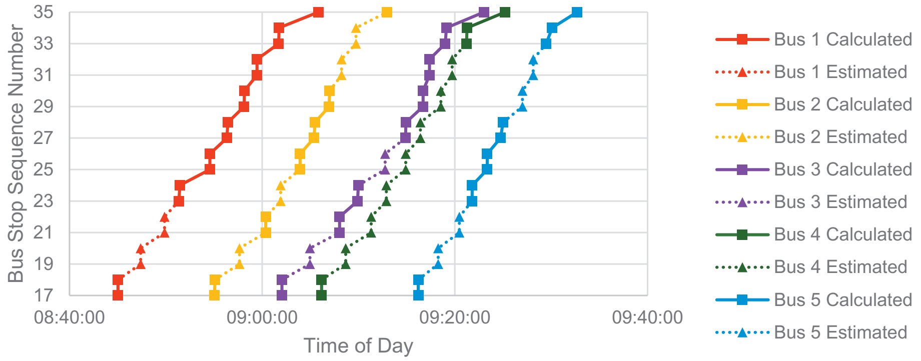


Fig. 1. Example of bus journey trajectories.

dots in the figure are linked with either solid lines or dotted lines. The solid lines represent calculated trajectories and the dotted lines are interpolated.

The interpolation of travel times was done using Eqs. (1) and (2). For example, the arrival time of stop 2 for bus journey number 1 was originally unknown. The average travelling time between stop pair 1 and 2, stop pair 2 and 3 can be calculated by given data of other journeys. The travelling time between stop 2 and stop 3 is about 1.6 times of that between stop 1 and stop 2. Thus, the travelling time between stop 1 and stop 3, which is 6 min and 59 s, was divided into 2 min and 39 s and 4 min and 20 s. Unknown dwelling times were estimated using the average valid dwelling times from the same line route with similar period of day (15-minute interval).

$$\hat{T}^{j}_{(i,i+1)} + \hat{T}^{j}_{(i-1,i)} = T^{j}_{(i-1,i+1)} \tag{1}$$

$$\frac{\hat{T}^{j}_{(i,i+1)}}{\hat{T}^{j}_{(i-1,i)}} = \frac{\sum_{h=1}^{N} \frac{T^{h}_{(i,i+1)}}{T^{h}_{(i-1,i)}}}{N} \tag{2}$$

where:

$(i-1,i)$ are two consecutive stops along bus line route;

$\hat{T}^{j}_{(i,i+1)}$ is the interpolated travel time of bus j from stop i to stop $i+1$;

$T^{j}_{(i,i+1)}$ is the calculated travel time of bus j from stop i to stop $i+1$;

N is the total number of data records with calculated travel times from stop $i-1$ to i and from i to $i+1$ within similar period of day (15-minute interval) of the travel time to be estimated.

The operational trajectories were calculated for all bus journeys within the 41 chosen weekdays. These disaggregated results exhibit a comprehensive view on bus operations. But they also raised the complexity of the public transport network for direct assignment. The disaggregated trajectories were first used for direct assignment. However, due to the complexity of route searching of direct assignment, the assignment could not be finished. To reduce the complexity, all bus journeys were grouped into 15-minute intervals for evaluating the operation conditions based on the travel times over the entire journeys. These intervals were again aggregated into six different periods (early morning from 3:00 to 6:15, morning peak from 6:15 to 8:45 am, interpeak from 8:45 am to 16:15, afternoon peak from 16:15 to 19:30, early evening from 19:30 to 22:00 and late evening 22:00 to 3:00 of the next day) based on the evaluation to further reduce complexity (Michalski et al., 2016). For each of the six periods of day, averages of the travel times between stops and the dwelling times at each stop of each line route were used to define the bus services' time profile. The departure headways were calculated based on the 15-minute aggregation of originally generated bus trajectories. Thus, all bus services were replicated from the smart card database and implemented on top of the existing road network.

3.4.2. Metro system

Different from buses, check-in and check-out of metro system take place at fare gates. Thus, arrival and departure times of trains cannot be directly calculated using the check-in and check-out times in the database. Data mining was applied to replicate the timetables for metro system which consisted of 9 service lines in 2013. Because alighting passengers always check out in a relatively shorter period, a peak is observed in the plot of alighting passenger volume vs. time. By investigating the check-out passenger volume, the arrival of trains could be estimated. However, checking-out passengers at a metro station could alight from different line routes in the metro system. To identify the direction of arriving trains which the checking-out passengers were taking, corresponding boarding information of each trip was used to identify the directions of arriving trains so that the check-out information could be used to estimate train arrivals. Fig. 2 illustrates an example of resulting passenger check-out peaks at a station on a certain metro line route.

Parallel to the data mining work, a survey (Zhou et al., 2015) on metro operations was conducted to collect detailed information on departure headways, travel times between stops, and train dwelling times at stops. Walking times between different platforms as well as between platforms and fare gates were also measured for representation of the transfers inside the metro system. The survey was conducted for three different periods of day (morning peak, off peak and afternoon peak) over the entire metro system. Actual train arrival times were calculated by offsetting the peaks by the surveyed walking time from the platform to the fare gate.

There are some limitations in the estimation of train arrivals at stations. If there were only a few arriving passengers, the arrival would not be detected. The arrivals could also be wrongly detected if there were passengers who did not exit immediately. With the estimated arrivals from the succeeding and preceding stops and surveyed travel times, the mistaken arrival estimations were either eliminated or interpolated. Table 3 shows an example of the replicated train arrival timetable. The times with asterisks were interpolated with similar approach for bus journey interpolation (as shown in Eq. (1)). Wrongly detected arrivals were removed by investigating travel times from previous station or to next station. The wrongly detected arrivals are additional arrivals at a certain station. Using them to calculate the travel times normally result in shorter travel times compared to other correctly detected arrivals. By detecting outlying travel times, the wrongly detected train arrivals were removed.

Complete arrival time tables were created for all metro services. However, this approach provided only the arrival time estimations. Other missing information on metro services such as departure times at terminals and train dwelling times were calculated based on the arrival timetable and the surveyed information. For example, the train departure times at each station were calculated as estimated arrival times plus surveyed dwelling times. And departure times at terminals were estimated by deducting the travel times from terminals to the second

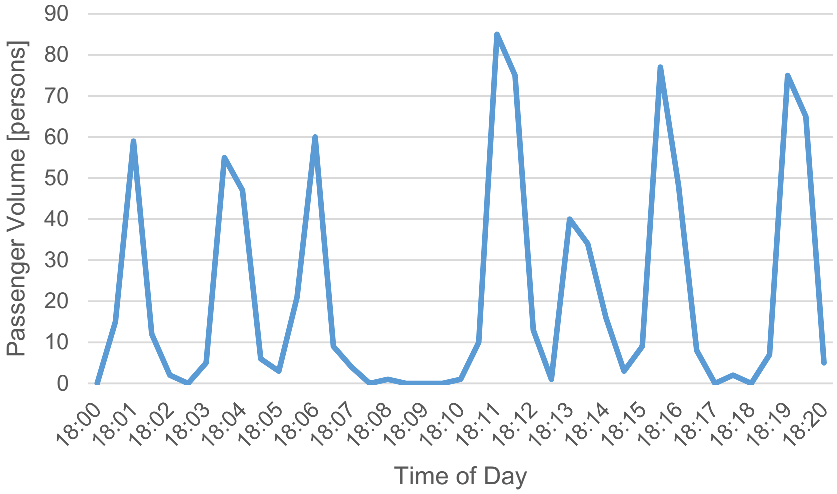


Fig. 2. Check-out passenger volumes.

Table 3
Estimated train arrival timetable.

Train no.	Stop 1	Stop 2	Stop 3	Stop 4	Stop 5	Stop 6
Train 1	06:20:06	06:22:45[a]	06:27:05	06:30:40	06:33:31	06:35:54
Train 2	06:24:08	06:26:47	06:30:51	06:34:29	06:37:38	06:40:07
Train 3	06:28:57	06:31:36	06:35:56	06:39:20[a]	06:42:34	06:45:04
Train 4	06:37:24	06:39:44	06:43:54	06:47:27	06:50:31[a]	06:53:13[a]
Train 5	06:39:08	06:41:47	06:46:28	06:49:54	06:53:05	06:55:25

[a] Estimated arrival times by interpolation.

stations along the service routes.

3.4.3. Transfers

The smart card database collects information on passengers' boarding and alighting from a particular bus line route, as well as entering and leaving the metro stations. No detailed information within the metro system was collected by the smart card system. Therefore, the transfers were classified as internal transfers within the metro system and external transfers.

As introduced previously, a survey was conducted to obtain detailed information including transfer times within the metro system. Since there was no change in station infrastructure between 2013 and 2015, the surveyed transfer times in 2015 were found reliable to represent the situation in 2013. Additionally, the fare gates were also modelled as dummy platforms which serve as access point of the metro system so that the transfers from bus stops to metro stations can be considered as an external transfer from bus stops to fare gates followed by an internal transfer from fare gates to platforms.

The external transfer times were calculated based on the collected smart card database. The smart card database collects information on individual trips as separate data records. No transfer information was directly stored in the database. But as introduced previously, identical journal IDs are shared among different trips within same journeys. By sorting different trips with the same journey ID by the boarding times, the chronological order of the trips was clearly obtained. The alighting stops of a certain trip and the boarding stop of its succeeding trip indicate a transfer between the two stops (transfer stop pair). The time difference between the corresponding alighting and boarding was also calculated to estimate the transfer time. Thus, the actual transfer links between stops could be identified and the corresponding transfer times were estimated.

It is worth mentioning that the calculated transfer times do not contain only transfer walking time. According to the definition of transfer by the smart card system in Singapore, a maximum of 45 min of transfer time is allowed. If the transfer time is longer than 45 min, the next check-in will be regarded as start of a new trip chain. The 45-minute duration is rather long and many activities such as shopping and dining could be finished during the period. As a result, the calculated transfer times could contain such activities. Additionally, if passengers transfer to buses, the calculated transfer times contain waiting times at stops as well. But if passengers transfer to metro stations, the boarding times of the succeeding trip clearly indicate the time when passengers arrived at the fare gates. Considering the different situation, the calculations of external transfer times were made in different ways.

The check-in times to the metro system indicate the times when passengers pass the fare gates, no waiting times were included in the calculated transfer times. The transfer times from buses to metro could be directly estimated based on the calculated times. Due to the large amount of available data, many transfer times were retrieved for each transfer stop pair. As shown in Table 3, the retrieved transfer times vary substantially and may contain certain errors. An outlier filter was applied to remove outlying calculated transfer times with both static and dynamic limits. According to the smart card system, 45 min was first applied as a static upper limit. Another dynamic outlier filter was applied using the interquartile range method (Tukey, 1977). The interquartile range has dynamic upper and lower limits for the dataset as shown in Eq. (3).

$$\begin{aligned} L_{up} &= Q_3 + 2 \times |Q_3 - Q_1| \\ L_{down} &= Q_1 - 2 \times |Q_3 - Q_1| \end{aligned} \tag{3}$$

where:

L_{up} is the upper limit for data filtering;
L_{down} is the lower limit for data filtering;
Q_i is the ith quartile of the dataset.

Data-falls outside the limits are normally considered outliers and are filtered. Medians of the filtered datasets were used as representative transfer walking times (Fig. 3).

Different from the transfers to metro stations, calculated transfer times to bus stops contain certain portion of waiting times. With no supporting information on bus operations, the calculated transfer times could not be used to estimate actual transfer walking times. As a result, the transfer walking times from metro to buses were replaced by the calculated ones in the opposite direction (from buses to metro). And the transfer times between bus stops were calculated as distance measured by a third-party map divided by the average walking speed of 74 m/min in Singapore (Tanaboriboon et al., 1986).

With all bus and train trajectories replicated and passenger transfer network created, Singapore's public transport system supply was developed comprehensively. The operational details including the

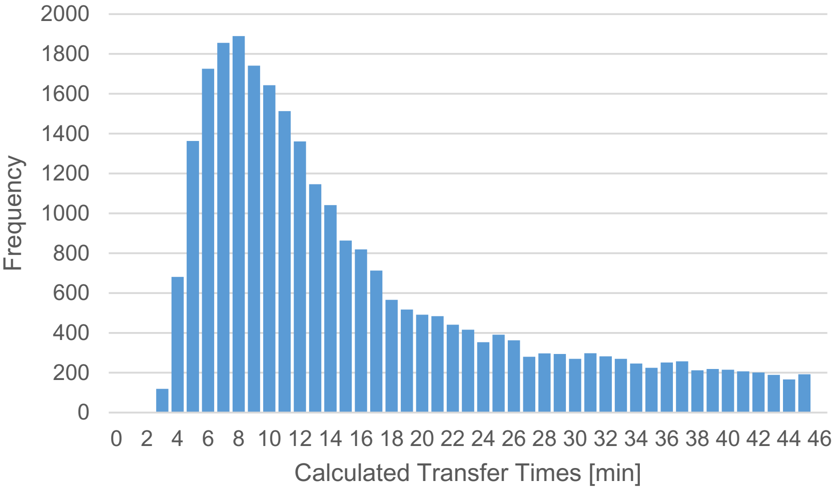


Fig. 3. Retrieved transfer times of one transfer stop pair.

transfers and operational timetables were generated from the collected smart card database and stored.

3.5. Replication of public transport demand

The conventional public transport models rely on defined zones. A zone in the network can be a residential block, a business building or a school. Passengers start their trips from each zone and then finish the first mile to access the public transport network. After the trips in public transport systems, passengers finish the last mile to a certain destination zone. The travel demand OD matrices are created based on the zones in the network and their characteristics.

With smart card system, no information on first or last mile is collected. But the system provides precise information on passengers starting stop, ending stop and exact paths in the public transport system. Stop-to-stop OD matrices were created from the smart card database.

As mentioned before, the complexity of public transport network raise challenges in terms of computational power required by the direct assignment method. The large number of to-be-assigned datasets have the same issues. As a result, the 41 chosen days were further selected for direct assignment. Days with the most stable travel patterns were chosen to represent weekdays in the three-month period. They were chosen based on the mean error of the passenger travel demand of each OD pair compared to the average demand of 41 days (Eq. (4)). The calculated mean errors are divided by the average number of trips of the 41 days to represent the relative difference between daily travel demand and the average travel demand (Table 4).

$$Mean\ \ Error_j = \frac{\sum_{i=1}^{N} \left| OD_{i,j} - \overline{OD_i} \right|}{N} \tag{4}$$

where:

j is the jth day of the selected 41 days;

$OD_{i,\ j}$ is the passenger travel demand of the ith OD pair in the jth day;

$\overline{OD_i}$ is the average passenger travel demand of the ith OD pair;

N is the total number of OD pair created from the smart card database.

Ten days out of the 41 days (19 to 22 August, 26 to 29 August, 2 to 3 September 2013) with the least mean error were chosen for the direct assignment. Those days were consecutive ten normal weekdays (Monday to Thursday) from the second week after a four-day long holiday (two public holidays and a weekend, from 8 to 11 August 2013). The selected ten days contain totally 59,896,509 data records.

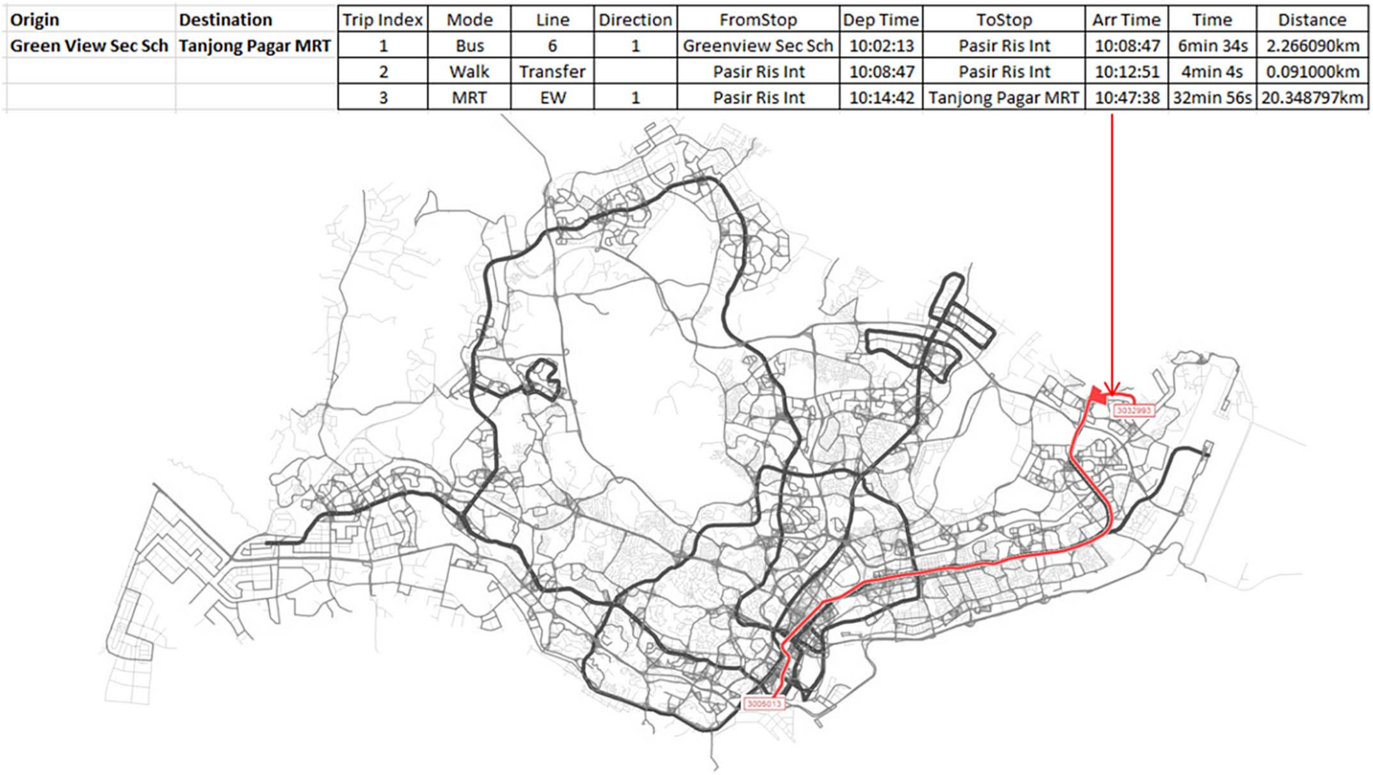
Table 4
Relative difference between daily travel demand and the average travel demand.

Dates	Difference [%]	Dates	Difference [%]
8/22/2013	6.56%	8/12/2013	7.14%
8/27/2013	6.64%	10/9/2013	7.14%
8/28/2013	6.65%	10/29/2013	7.15%
8/20/2013	6.67%	10/24/2013	7.18%
8/19/2013	6.74%	9/24/2013	7.22%
9/3/2013	6.76%	8/15/2013	7.25%
8/26/2013	6.80%	9/25/2013	7.26%
8/29/2013	6.80%	9/26/2013	7.29%
8/21/2013	6.80%	10/17/2013	7.29%
9/2/2013	6.81%	10/21/2013	7.30%
8/14/2013	6.83%	10/30/2013	7.47%
10/2/2013	6.84%	10/23/2013	7.49%
10/8/2013	6.89%	10/28/2013	7.50%
9/30/2013	6.93%	10/16/2013	7.52%
8/13/2013	6.93%	8/5/2013	7.74%
10/7/2013	6.96%	10/31/2013	8.02%
10/10/2013	7.01%	8/1/2013	8.03%
9/4/2013	7.01%	8/6/2013	8.35%
10/1/2013	7.04%	10/14/2013	8.68%
10/22/2013	7.08%	9/5/2013	8.91%
10/3/2013	7.12%		

The smart card database collects each trip segment as individual data records and identifies passenger trip chains with a unique journey ID. Hence, complete passenger trip chains can be obtained by ordering the start times of all trips with same journey ID. Additionally, by investigating the alighting stop and time of a certain trip segment and the boarding time and stop of its next segment, passenger transfer trips are also generated.

As shown in Fig. 4, a passenger took bus line 6 in the morning from stop “Green View Sec Sch” to “Pasir Ris Int”. The bus trip took him/her 6 min and 43 s. After alighting, he/she walked for 4 min and 4 s to the metro station and took a train. The later metro trip to “Tanjong Pagar MRT” station took him/her 32 min and 56 s. With all three segments generated from smart card database, the complete trip chain was incorporated into PTV Visum for direct assignment. All trip chains within the selected 10 days were reconstructed in such way for direct assignment. The 59.9 million trip segments records were combined into 41.9 million passenger trip chains.

Table 5 presents the direct assignment result. 41,598,426 out of 41,934,463 of collected passenger trip chains from the selected 10 days were successfully assigned. There exist 336,037 trip chains that could not be assigned. Those trips failed because they are trips made with irregular or night services. Because the operational replication focuses



Origin	Destination	Trip Index	Mode	Line	Direction	FromStop	Dep Time	ToStop	Arr Time	Time	Distance
Green View Sec Sch	Tanjong Pagar MRT	1	Bus	6	1	Greenview Sec Sch	10:02:13	Pasir Ris Int	10:08:47	6min 34s	2.266090km
		2	Walk	Transfer		Pasir Ris Int	10:08:47	Pasir Ris Int	10:12:51	4min 4s	0.091000km
		3	MRT	EW	1	Pasir Ris Int	10:14:42	Tanjong Pagar MRT	10:47:38	32min 56s	20.348797km

Fig. 4. Passenger trip chain example for direct assignment.

Table 5
Direct assignment result.

Type	No. of trips [trips]
Successfully assigned	41,598,426
Unsuccessfully assigned with departure time specified	0
Unsuccessfully assigned with no available services	336,037
Unsuccessfully assigned with required walk links missing	0
Total	41,934,463

on normal operation on weekdays, the irregular or night services are not implemented. So, all trips made with such services were not successfully assigned. Therefore, excluding trips made by cash (3%), mistaken data which could not be corrected (0.3%) and unsuccessfully assigned trips made with irregular or night services (0.7%), the remaining approximately 96% travel demand has been successfully accounted for.

As introduced previously, the public transport supply implemented in the replication has a certain level of aggregation. To match the aggregated public transport supply and precisely collected passenger trip chains, a 15-minute tolerant interval was implemented to match passengers' departure times and public transport's departure times at stops. This results in certain level of errors (up to 15 min) to the direct assignment results, compared to actual scenario in terms of temporal travel demand distribution. But due to the nature of direct assignment, results are precise in terms of spatial travel demand distribution.

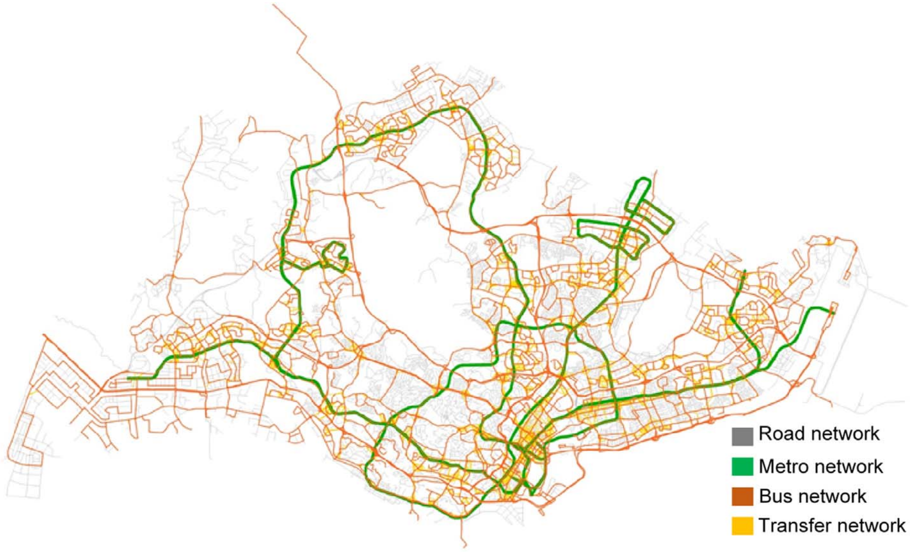


Fig. 5. Replicated public transport network in Singapore.

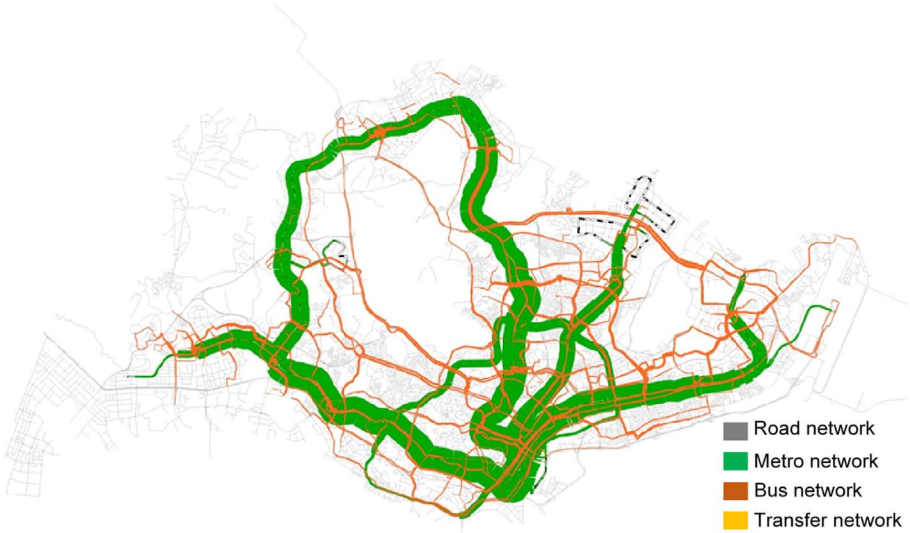


Fig. 6. Public transport passenger demand over the entire network.

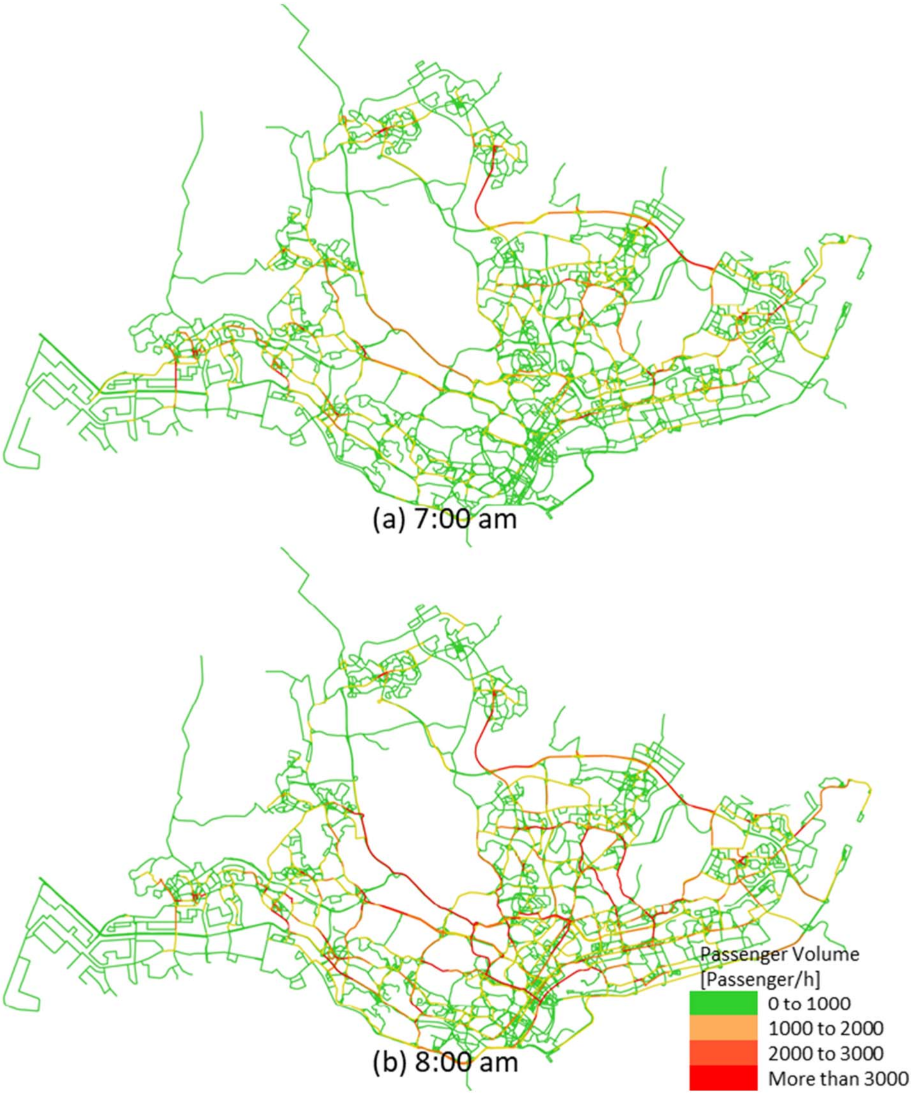


Fig. 7. Hourly passenger distribution in bus network during morning peak hours.

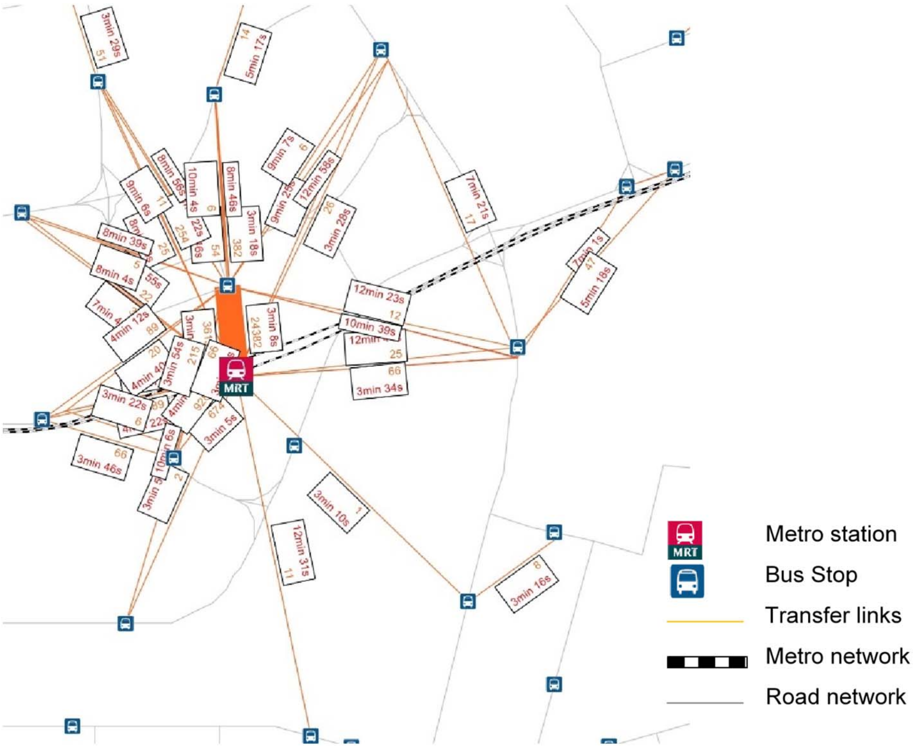


Fig. 8. Passenger demand from 8:00 to 9:00.

Table 6
Bus occupancies along journeys on an example bus line in Singapore.

Stop no.	On-board passenger	Design capacity	Occupancies
1	51	88	57.95%
2	53	88	60.23%
3	53	88	60.23%
4	49	88	55.68%
5	43	88	48.86%
6	41	88	46.59%
7	58	88	65.91%
8	55	88	62.50%

4. Application of the replication

Compared to the conventional models, the disaggregated replication with precise assigned passenger trips, contains accurate temporal and spatial information on both supply and demand sides of the public transport system. Many operational details, including corridor-specific demand, bus operation bottlenecks, passenger transfer patterns and route choice preference could be analysed. These operational details could directly provide quantitative results to the decision makers to understand the actual operational situation and improve the service quality.

As shown in Fig. 5, the operational details including the transfers and operational timetables were generated from the collected smart card database and implemented in the replication.

With the direct assignment method, passenger demand over the selected 10 days were all directly assigned to public transport supply with exact path information. As shown in Fig. 6, travel demand along every segment of the public transport network is presented. Due to the large gap of travel demand between bus and metro system, bus lines with relatively lower demand are not well visualised. It is still obvious that the metro system is carrying the majority of trips while buses are covering the rest of demand in the entire city. As introduced before, the supply implemented has a certain level of aggregation. Therefore, each assigned passenger trip has minor offsets in time compared to the original database.

Furthermore, by filtering the travel demand by time of day, temporal distribution of public transport passenger can be revealed. Fig. 7 illustrates the hourly travel demand on the bus network in Singapore during the morning peak hours. Compared to conventional approaches which build models based on conditions during the survey periods, the disaggregated replication contains trip information with precision of a few minutes. The demand changing dynamics provides a clear view of the trend of public transport ridership over the entire system. Highly demanded road sections during specific times of day can be identified to support public transport planning and fleet operation.

Additionally, operation details of the system such as transfer patterns could also be retrieved from the replication. As shown in Fig. 8, the realised transfer network with expected transfer walking time and transferring passenger volumes on particular links were also obtained from the direct assignment. In the figure, the numbers in the rectangle indicate the transferring time and passenger volumes. The red number indicates the walking time and the yellow number is the transferring passenger volume. Such information could directly help decision makers to understand the most in-demand transfer locations and specific links in order to prioritise improvement in infrastructure for transferring.

Other than describing the travel demand, characteristics of the traffic condition could also be identified. As shown in Fig. 1, bus trajectories were generated from the smart card database. Many aspects were observed from the figure. For example, time and location of the buses slowing down (with lower slop in the trajectories) as well as bunching (intersection or converging of trajectories) were identified. This could help to identify the bus bunching problem (Wang, 2016) as well as to investigate the general traffic conditions over the network and reveal the times and locations of traffic congestions.

Moreover, the bus occupancies over the operations could also be retrieved (Table 6). With known plate numbers from the database and third-party data resource (sgWiki, 2015), the design capacity of each bus in Singapore was found. The bus vehicle utilisation can be indicated by the occupancy of the vehicles or the number of passengers compared

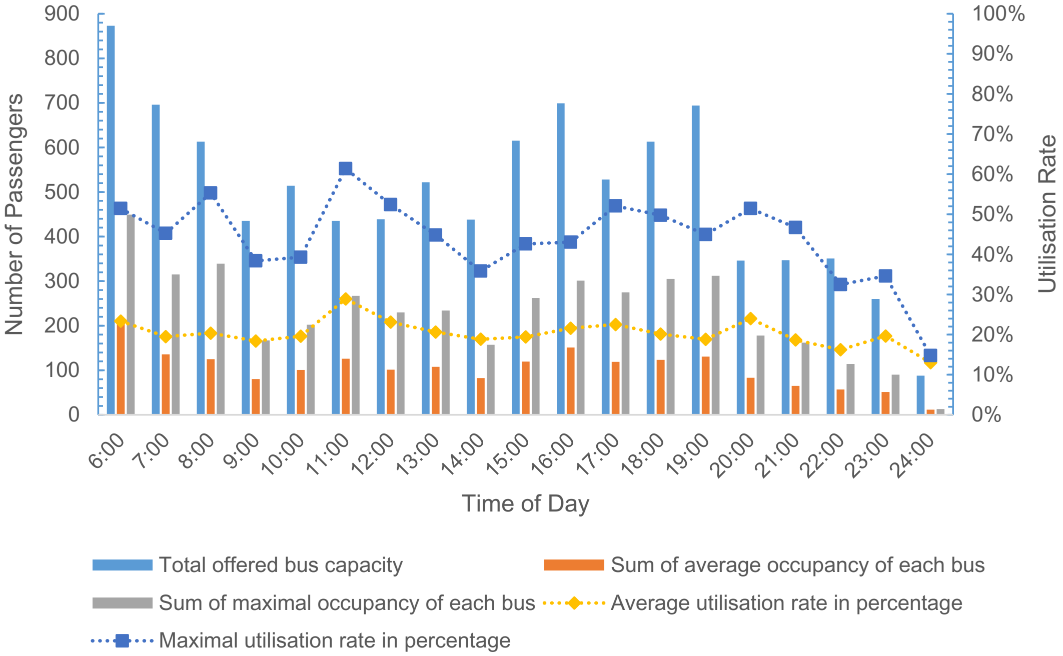


Fig. 9. Bus occupancies against offered capacity over a day.

Table 7
Comparison of public transport modelling approaches to the replication.

Approach	Four-step model	Activity-based model	Agent-based model	Smart card data-centric replication
Survey requirement	Yes	Yes	Yes	Yes, but minor
Data requirement	High	High	High	Very High
Computational power requirement	Low	Low	High	High
Mathematical method involved	Yes	Yes	Yes	No
Calibration requirement	Yes	Yes	Yes	No
Result detail level	Low	Low	High, but estimated	Very high, only actualised
Capability to hypothetical scenarios	Yes	Yes	Yes	No

to its design capacity. On-board passenger numbers were attained by the calculation of numbers of passengers boarding and alighting at each stop along the specific bus journeys. The results can benefit the operators and agencies with dynamic occupancy information of each bus along each section of service routes, so that the route planning and fleet operation could be optimised accordingly.

Based on the study of bus occupancies, the general utilisation of bus fleet over the day could also be estimated. Fig. 9 shows an example of average and maximum occupancies of an example bus line against the offered capacity over the day. Average and maximum utilisation rates are also shown. It can be observed that this bus line provides sufficient supply during morning and evening peak hours. Meanwhile, it also manages to maintain the utilisation rate of buses within a stable range during day time. This study helps to identify insufficient or excess supply of bus lines over the day and leads to a re-organization of bus fleet for operators.

Many of the above-mentioned outputs from this replication work have been reported to and implemented by Singapore's public agencies. For example, the trip chain information is being used for trip planning and the transfer demand information is used to identify key spots for infrastructure planning.

Additionally, as introduced, the replication is capable of representing current situations. But it lacks proper mathematical models for forecasting future scenarios. However, the direct assignment results can be further aggregated to compute O-D demand. Thus, with both demand per OD pair and exact passenger distribution, the replication could help to calibrate passenger behaviour model (e.g. route choice models).

Compared with the conventional approaches, the smart card-centric replication of public transport system has certain advantages and disadvantages (Table 7). It requires less survey data, but large amount of high quality smart card data. The approach does not require passenger behaviour estimation, but needs very high computational power for data processing and direct assignment. Lastly, the replication provides a very high level of actualised operational information of real situations, but it has no capability of forecasting future scenarios. As a result, the major application of this approach is to support the decision makers to fully understand the current situation to improve the service quality.

5. Conclusion and future work

This study successfully replicated the public transport system in Singapore with data from smart cards. But compared with the conventional approaches, the replication provides several results in terms of operational details which can be directly used by decision makers to improve the quality of service. However, the successful replication highly relies on the quality and quantity of data input. Compared to the smart card systems worldwide, Singapore's system is advanced with a high penetration rate and provides precise information on both check-in and check-out. The comprehensiveness and accuracy of collected database guaranteed the success of the replication.

The replication contains most of the trips made with public transport during the selected 10 days. Trips made by cash (3%), mistaken data which could not be corrected (0.3%) and unsuccessfully assigned trips made with irregular or night services (0.7%) account for a sum of 4% missing trips. As a result, the replication covers approximately 96% of travel demand on public transport system. There exist certain errors (up to 15 min) in terms of passenger departure times due to the aggregated supply implemented. But it is precise in terms of spatial demand distribution because of the nature of direct assignment.

Therefore, the details presented by the replication are reliable in terms of coverage and could be used to represent the entire population's usage of public transport.

During the replication work, bus journeys were aggregated to reduce complexity and the real passenger inter-modal trip chains were directly assigned to the replicated supply. The resulting replication contains a certain level of difference compared to the original database. A fully disaggregated replication could be done if there are fewer constraints on the computational power.

Many applications can be derived from the replication. For example, the transfer demand can be used for infrastructural planning, bus occupancies can be used to optimise fleet operation and planning. There is still a large scope for future work in this study. Particularly, the current work focuses on daily commute during weekdays; the analysis can be further extended to analysis for different types of days (weekdays, weekends, and public holidays) or different types of passengers (students, adults and elder people). Such studies could provide better insights on the public transport system operation.

Due to the nature of replication, the applications are limited to understanding current scenarios. As introduced previously, another major future work is to use the aggregated OD demand and the actualised trip distribution provided by the direct assignment to validate passenger route choice behaviour. This work could overcome the limitations of this data-centric approach and provide capabilities for forecasting future scenarios and conducting attitudinal or prospective analyses.

Acknowledgment

This work was financially supported by the Singapore National Research Foundation (NRF) under its Campus for Research Excellence And Technological Enterprise (CREATE) programme.

References

Agard, B., Morency, C., Trépanier, M., 2006. Mining public transport user behaviour from smart card data. IFAC Proc. Vol. 39 (3), 399–404.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12 (5), 464–474.

Bueno, P.C., Gomez, J., Peters, J.R., Vassallo, J.M., 2017. Understanding the effects of transit benefits on employees' travel behavior: evidence from the New York-New Jersey region. Transp. Res. A Policy Pract. 99, 1–13.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299.

Du, B., Zhou, Y., Rau, A., 2017. Imputation of passengers' trip activities based on smart card data. In: mobil.TUM 2017 International Scientific Conference on Mobility and Transport, Munich, 4–5, July.

Hägerstraand, T., 1970. What about people in regional science? Pap. Reg. Sci. 24 (1), 7–24.

Hamre, A., Buehler, R., 2014. Commuter mode choice and free car parking, public transportation benefits, showers/lockers, and bike parking at work: evidence from the Washington, DC region. JPT 17 (2), 67–91.

Helbing, D., Balietti, S., 2012. Agent-based modeling. In: Helbing, D. (Ed.), Social Self-Organization. Agent-Based Simulations and Experiments to Study Emergent Social Behavior. Springer, Heidelberg, pp. 25–70.

Land Transport Authority Singapore, 2013. Household Interview Travel Survey 2012: Public Transport Mode Share Rises to 63%. https://www.lta.gov.sg/apps/news/page.aspx?c=2&id=1b6b1e1e-f727-43bb-8688-f589056ad1c4, Accessed date: 6 December 2016.

Land Transport Authority Singapore, 2016. DataMall. https://www.mytransport.sg/content/mytransport/home/dataMall.html, Accessed date: 6 December 2016.

Liu, X., Rau, A., Busch, F., 2016. Intersections' level of service evaluation using smart card data in Singapore. In: 23rd ITS World Congress, Melbourne, Australia, 10–14, October.

Lohse, D., Teichert, H., Dugge, B., Bachner, G., 1997. Ermittlung von Verkehrsströmen mit n-linearen Gleichungssystemen unter Beachtung von Nebenbedingungen einschließlich Parameterschätzung (Verkehrsnachfragemodellierung: Erzeugung, Verteilung, Aufteilung). Schriftenreihe des Instituts für Verkehrsplanung und Straßenverkehr H. 5/1997. Fakultät Verkehrswissenschaften, "Friedrich List". Technische Universität, Dresden.

Michalski, G.-M., Zhou, Y., Liu, X., Rau, A., 2016. Bus system modelling using smart card data in Singapore. In: 23rd ITS World Congress, Melbourne, Australia. 10–14, October.

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. C Emerg. Technol. 24, 9–18.

Ortúzar, J.d.D., Willumsen, L.G., 2011. Modelling Transport. John Wiley & Sons, Chichester, West Sussex, United Kingdom.

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. C Emerg. Technol. 19 (4), 557–568.

Prakasam, S., 2008. The Evolution of e-payments in Public Transport - Singapore's Experience. Land Transport Authority, Singapore.

Recker, W.W., McNally, M.G., Root, G.S., 1986a. A model of complex travel behavior: part II—an operational model. Transp. Res. A Gen. 20 (4), 319–330.

Recker, W.W., McNally, M.G., Root, G.S., 1986b. A model of complex travel behavior: part I—theoretical development. Transp. Res. A Gen. 20 (4), 307–318.

sgWiki, 2015. sgWiki Buses. http://www.sgwiki.com/wiki/Buses, Accessed date: 1 November 2015.

Sun, L., Lee, D.-H., Erath, A., Huang, X., 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. In: The ACM SIGKDD International Workshop, Beijing, China. 12/8/2012 - 12/8/2012. ACM Press, New York, NY, USA, pp. 142.

Tanaboriboon, Y., Hwa, S.S., Chor, C.H., 1986. Pedestrian characteristics study in Singapore. J. Transp. Eng. 112 (3), 229–235.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison Wesley.

van der Hoorn, T., 1979. Travel behaviour and the total activity pattern. Transportation 8 (4), 309–328.

van der Hoorn, T., 1983. Experiments with an activity-based travel model. Transportation 12 (1), 61–77.

Vrtic, M., Fröhlich, P., Schüssler, N., Axhausen, K.W., Lohse, D., Schiller, C., Teichert, H., 2007. Two-dimensionally constrained disaggregate trip generation, distribution and mode choice model: theory and application for a Swiss national model. Transp. Res. A Policy Pract. 41 (9), 857–873.

Wang, J., 2016. Analysis of Singapore's Bus Bunching Problems Based on Smart Card Data. (Master's thesis: Munich, Germany).

Zhou, Y., Michalski, G.-M., Wang, J., Aslam, Z., 2015. Field survey on MRT/LRT in Singapore. TUM CREATE, Singapore.