Methods for pre-processing smartcard data to improve data quality

Steve Robinson a,⇑, Baskaran Narayanan b, Nelson Toh b, Francisco Pereira a

a Singapore-MIT Alliance for Research and Technology, Singapore 138602, Singapore
b Public Transport Quality, Land Transport Authority, Singapore 575701, Singapore

A R T I C L E   I N F O

Article history:
Received 22 February 2014
Received in revised form 11 October 2014
Accepted 14 October 2014
Available online 7 November 2014

Keywords:
Smartcard
Automatic Fare Collection system
Automatic Vehicle Location system
Public transport travel patterns
Data quality
Trip identification

A B S T R A C T

In recent years smartcards have been implemented in many transit systems around the world as a means by which passengers pay for travel. In addition to allowing speedier boardings there are many secondary benefits of smartcard systems including better understanding of travel patterns and behaviour of travellers. Such research is dependent on the smartcard correctly recording the boarding stop, and where available the alighting stop. It is also dependent on the smartcard system correctly aggregating individual rides into trips.

This paper identifies causes for why smartcard systems may not correctly record such information. The first contribution of the paper is to propose a set of rules to aggregate individual rides into a single trip. This is critical in the research of activity based modelling as well as for correctly charging the passenger. The second contribution of the paper is to provide an approach to identify erroneous tap-out data, either caused by system problems or by the user. An approach to detecting this phenomenon is provided. The output from this analysis is then used to identify faulty vehicles or data supply using the “comparison against peers approach”. This third contribution of the paper identifies where transit agencies and operators should target resources to improve performance of their Automatic Vehicle Location systems. This method could also be used to identify users who appear to be tapping out too early.

The approaches are tested using smartcard data from the Singapore public transport network from one week in April 2011. The results suggest that approximately 7.7% of all smartcard rides recorded the passenger as alighting one stop before the bus stop that they most probably alighted at. A further 0.7% of smartcard rides recorded the passenger as alighting more than one stop before the bus stop that they most probably alighted at. There was no evidence that smartcards overestimated the distance travelled by the passenger.

1. Introduction

In recent years smartcards have been implemented in many transit systems around the world as a means by which passengers pay for travel. The primary benefits of smartcards are speedier boarding and reduced costs by the operator in handling cash. However there are many secondary benefits of smartcard systems including better understanding of travel patterns and behaviour of travellers. Such research is dependent on the smartcard correctly recording the boarding stop, and where possible the alighting stop. Such research is also dependent on the smartcard system correctly aggregating...
individual rides into trips. Much of the research assumes that this is the case. However there are many reasons why such basic information may not be correctly recorded by the smartcard system.

This paper identifies causes for why smartcard systems may not correctly record such information. The first contribution of the paper is to propose a set of rules to aggregate individual rides into a single trip. This is critical in the research of activity based modelling (Bowman and Ben-Akiva, 2001), as well as for correctly charging the passenger. The second contribution of the paper identifies where transit agencies and operators should target resources to improve performance of their Automatic Vehicle Location systems. This method could also be used to identify users who appear to be tapping out too early. These three areas of research are then brought together in a unified methodology for ensuring the quality of smartcard data. It is recommended that all users of smartcard data should implement this methodology before further analysing travel patterns and travel behaviour using smartcard data. The methods suggested in this paper could also be used by transit agencies to reduce certain revenue loss.

The paper is organised as follows. Section 2 provides a background to smartcard technology and a short literature review into research on the usage of smartcard data. Section 3 then lists various reasons why smartcard systems may output erroneous data. Section 4 describes some of the data that will be used later on in the paper. Section 5 suggests some rules for aggregating rides into trips. Section 6 then proposes a method for identifying faulty tap-out data. Section 7 then outlines a method for identifying the cause of erroneous smartcard data. Section 8 proposes a methodology for pre-processing smartcard data. Finally Section 9 presents conclusions and further work.

1.1. Clarification of terminology used in this paper

In this area of study there is often more than one word for the same meaning. This paper will use the definitions based upon Transmodel as given in reference (European Committee for Standardisation, 2006). The main ones used are defined below:

- **Ride**: This describes the movement of a passenger on a single vehicle, typically a bus or train. The ride begins at the stopping point where the passenger boarded the vehicle, and ends at the stopping point where the passenger alighted the vehicle.
- **Stopping Point**: This describes a location where a passenger may board or alight a public transport vehicle. A stopping point is typically a bus stop or a train platform.
- **Trip (shortened from PT Trip)**: This describes the movement of the traveller from an origin location to a destination location. The origin location is assumed to be the first stopping point that the passenger entered the public transport network. Likewise the destination location is assumed to be the last stopping point from which the passenger exited the public transport network.
- **Vehicle Journey**: This describes the movement of a public transport vehicle through a defined sequence of stopping points.
- **Transfer (similar to TransXChange: Connection Link)**: This describes the movement of a passenger from one public transport vehicle to another public transport vehicle. This may involve the passenger having to walk to a different stopping point. It should be noted that the term “transfer” has been used to generalise the fact that a passenger may not have to change stopping points, but only vehicles.

It should also be noted that there is inconsistency in the literature as to whether to use “smartcard” or “smart card”. This paper uses the former.

2. Background and literature review

Smartcards have been used in the transport industry since the UPass was launched in Seoul in 1996 (Lee, 2010). This was followed by the launch of the Octopus card in Hong Kong in September 1997 (World Bank Group, 1999) and the Oyster card in London in June 2003 (BBC, 2013). This section firstly provides details on how smartcards work and the benefits in reduction of boarding time they provide. The section will then discuss how data from smartcards can be used by transport planners and academics to study OD and travel time estimation. The reader is also recommended to read the comprehensive literature review on smartcard data written by Pelletier et al. (2011). Surprisingly very little literature was found which investigates the quality of smartcard data.

2.1. How smartcards work

Smartcards are cards, typically shaped like a bank card, which have an embedded integrated circuit and are commonly called Proximity Integrated Circuit Cards (PICC) or proximity cards for short. They communicate with a reader. The reader generates an alternating magnetic current at a frequency of 13.56 MHz (Atmel, 2013). This induces an electric current in the antennas of the smartcard which allows two way communication to take place. Communication is undertaken between
smartcard and reader using modulation at 857.5 kHz. Communications are based upon the ISO 14443 protocol. Typically the maximum distance of communication between card and reader is approximately 10 cm (Atmel, 2013).

When a passenger places their smartcard in the proximity of the reader, the reader will deduce the appropriate fare, or validate the presence of a season ticket, and update information on the card such as the location where the tap-in occurred and the current card balance (Cubic, 2013). The Electronic Ticket Machine (ETM) is able to do this since it is typically interfaced with the bus Automatic Vehicle Location (AVL) system. The reader will also send this information to a central computer on the bus. This information may be immediately transmitted back to a central server, or uploaded once the bus returns to the depot.

2.2. Benefits of smartcards in reducing boarding time

A significant benefit of such smartcard systems is that they significantly reduce transaction times and hence reduce dwell times. The Highway Capacity Manual (TRB, 2000) suggests that the boarding time per passenger is 1.2 s using a smartcard as opposed to 1.8–2 s for coin payment without change. Work by Cundill and Watts (1973) suggests that this time for cash payment might be underestimated and give boarding times of 2.3–5.0 s per passenger for cash payments where no change is offered. In a more recent study, Tirachini (2013) found that boarding time with exact cash took around 5.66 s, and where change was given 12.73 s. He found magnetic payments took between 4.61 for devices on the right hand side, and 5.56 s for those on the left hand side. Milkovits (2008) also found that the typical smartcard transaction time is 1.5–2 times quicker than magnetic card payments.

2.3. Use of smartcards for OD estimation

There are other benefits to using smartcards beyond a reduction in boarding time. Smartcard data can potentially provide much valuable information to transport planners and academics. One of the applications of smartcard data is to estimate origin–destination patterns of travellers. One major challenge to doing this is caused by fixed fares, regardless of distance travelled. This means that the passenger only has to tap-in on their smartcard when they board the bus and not when they alight. This results in the boarding stop and time being known but the alighting stop typically being unknown. There is a significant amount of research in determining the destination from this.

Trépanier et al. (2007), proposed a model for inferring the destination of a passenger for bus systems where passengers only have to tap-in their smartcard on boarding but not on alighting. The basic concept that they utilise is to minimise the distance between the alighting stop on one trip and the boarding stop on the next trip. Several modifications are made to this approach. If there is no other trip in the day, then previous days are searched to find a similar boarding stop. The assumption is that the alighting stop will also be similar. They tested their method using smartcard data from Gatineau, Quebec in Canada.

Chu and Chapleau (2008) built upon this work to determine bus transfers. They noted that the transfer information in smartcard data may often be unreliable. When determining transfers they assumed an average walk speed of 1.2 m/s. They used direct distance between bus stops, but realised that this may not be the case. They therefore added a buffer of 5 min in their estimated transfer time to account for this and other measurement errors. Importantly they noted the importance of detecting “return trips”. They tested for this by checking whether in a second ride the passenger took the same bus line but in the opposite direction.

Munizaga and Palma (2012) developed a similar method to Trépanier et al. (2007) to estimate the alighting stop. However it was modified to take into account that buses may loop back. Thus an upstream bus stop may be further away from the boarding stop of the next ride than a downstream bus stop. However passengers may still alight at this upstream stop if the total trip travel time is reduced by alighting at the earlier stop.

Seaborn et al. (2009) suggested some rules for determining if rides recorded by smartcard systems should be aggregated together. They suggested a maximum allowable bus-to-bus transfer time of 45 min. They verified how good this model was by comparing the predicted number of public transport journeys per day with survey data. The research highlights the difficulty of obtaining truth data by which to measure any attempt to aggregate rides together into trips.

Wang et al. (2011) proposed a way of merging smartcard and AVL data to determine the boarding stop of a passenger and infer the probable alighting stop. They applied this approach to London smartcard data, and were able to validate this method by comparing the estimated OD matrices against OD matrices recorded from manual surveys.

Gordon et al. (2013) built on earlier work of Seaborn et al. (2009) and Wang et al. (2011) in analysing London smartcard data. In particular they included additional rules on correctly inferring transfers. A transfer was defined as a transition between two consecutive rides that do not contain a trip-generating activity. They incorporated a test on walking transfer speed. If this was infeasible an additional possible transfer point was sought. They also recognised the importance of identifying return trips. This was done using a circuitry rule which compared the Euclidean distance of two adjacent rides with the distance between the origin of ride 1 and destination of ride 2. They also defined a minimum distance between trip origin and trip destination for a trip comprised of more than one ride.

Devillaine et al. (2012) also developed rules to propose if separate rides should be aggregated into the same trip. This work was done in the context of inferring whether an activity had taken place. They demonstrated their approach using smartcard data from Santiago and Gatineau.
Utsunomiya et al. (2006) analysed 7 days worth of transaction data for 62,500 smartcards from the Chicago CTA system. Unlike much of the literature they had access to the home addresses of the smartcard users. They undertook analysis such as average distance between home and first transit trip, as well as analysing daily usage patterns. They also highlighted some issues with the smartcard data. This included missing transactions due to faulty equipment or a passenger error, and recording the wrong bus number as the bus driver sometimes failed to manually set up the electronic ticket machine correctly.

The above research all highlights the usefulness of smartcard data in determining OD travel patterns. Such data is used by government authorities and operators to make planning decisions concerning the transport network.

2.4. Use of smartcards for deriving other performance metrics

Trépanier et al. (2009) suggested that there are many other performance metrics which can be measured using smartcard data. They used 21 million smartcard transactions from January 2005 to March 2007 from Gatineau in Quebec to measure metrics such as person-kilometres travelled, average bus occupancy, and number of trips performed by each age category (student, adult, senior). They also used estimated statistics such as the speed of each trip using smartcard data. However information from speed is derived originally from an AVL system so could be obtained without the smartcard system.

Pelletier et al. (2011) suggest that smartcard data can be used at the strategic level (i.e. long term planning), tactical level (i.e. fine tuning a given service), and operational level (i.e. schedule-adherence, person-km for each run, etc.).

Ma et al. (2013) demonstrated how smartcard data can be used to determine the regularity of travel patterns of individual travellers over a period of time. They do mention how there were several smartcard transactions with a date of 1st January 1900 which had to be filtered out.

The applications of smartcard data are all reliant on the raw data being correctly. This is not always the case as the next section will discuss.

3. Problems with data from smartcards

This section provides a comprehensive list of problems that may be encountered with smartcard data. This is something which the literature appears to have overlooked on the whole, other than the issue concerning the lack of alighting stop measurements. Pelletier et al. (2011) did highlight some issues such as implementation costs. It was also noted that smartcards did not collect all the information that researchers and planners wanted, such as user background, trip purpose or assessment of a service.

There are many potential problems with data obtained from smartcards and these are elicited below in Table 1. The problems are grouped into the following categories:

- **Software**: issues that are caused by erroneous software. This also includes the business rules and logic incorporated in the software
- **Data**: Issues that are caused by erroneous data
- **Hardware**: issues that are caused by faulty hardware
- **User**: issues that are caused by the user, both accidental and deliberate

The problems given in Table 1 can affect smartcard data recorded from all modes of transport. Those problems which tend to occur only with bus or tram use are marked with an asterisk in the “Name” column.

This paper will develop some approaches to identify and mitigate some of the errors raised above.

4. Data used to validate methodology

To test the approaches suggested in this paper real world data from Singapore has been used. This data was provided by the Land Transport Authority of Singapore (LTA).

Singapore has an extensive bus and light rail network (called the MRT). As of summer 2013 there are 4 major MRT lines (North-South line, East-West line, North-East line, and Circle line). There are approximately 4700 bus stops in Singapore served by 365 bus lines. However this number is increasing as part of a service improvement program.

Singapore uses a smartcard technology called EZLink card for the public transport system. Fares are distance based. This means that passengers have to both tap-in and then tap-out from the bus. When a passenger taps-into a bus, the fare that would have to be paid if the passenger continued to the last stop is deducted. Then when the passenger taps-out to alight, the actual fare that is payable is calculated, and the difference between what was paid earlier is refunded. In Singapore it is possible to transfer buses within the same trip. The rules for valid transfers will be given later.

EZLink data used for this research came from the time period of Monday 11th April to Sunday 17th April 2011 and 28th April 2011. Data was provided for each ride and included:

- Anonymised card id – NB. It is not possible to identify the owner of the card
- Trip id – i.e. the unique identifier of the trip
Summary of problems with data from smartcards.

Table 1
Summary of problems with data from smartcards.

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Erroneous trip definition</td>
<td>Software</td>
<td>The smartcard system may not correctly aggregate the sequence of rides together when defining a trip. This will be discussed more in Section 5</td>
</tr>
<tr>
<td>1.2</td>
<td>Bugs in the system</td>
<td>Software</td>
<td>Smartcard systems contain significant amounts of software, from the reader to the back-end systems. If this software is erroneous then there can be significant consequences</td>
</tr>
<tr>
<td>1.3</td>
<td>Hacking</td>
<td>Software</td>
<td>Data such as the balance are stored in the smartcard and updated by the reader. Research by Nohl and Plötz (2007) suggest that the cryptography used on these cards could be hacked</td>
</tr>
<tr>
<td>1.4</td>
<td>Lack of destination</td>
<td>Software</td>
<td>Many smartcard systems do not require the user to tap-out. This means that the alighting stop is not measured and must be inferred. This was discussed in Section 2.2</td>
</tr>
<tr>
<td>1.5</td>
<td>Missing requirements</td>
<td>Software</td>
<td>It is important to ensure that the smartcard system retains all the required data that is likely to be needed in future analysis. Ma et al. (2013) identify one such missing requirement in the smartcard system implemented in Beijing where in some cases, the boarding stop is not recorded</td>
</tr>
<tr>
<td>1.6</td>
<td>Time synchronisation</td>
<td>Software</td>
<td>It is important that the time on all readers is synchronised across the system. An easy way to do this is to use GPS time. However even with GPS time it is important to ensure all readers have included the correction for leap seconds</td>
</tr>
<tr>
<td>2.1</td>
<td>Erroneous input data</td>
<td>Data</td>
<td>Smartcard systems also require a large amount of data. For example transit network smartcards will have to have knowledge of the network and fare tables. If this data is wrong there can be significant consequences. As an example the London Oyster smartcard system crashed on Saturday 12th July 2008 due to erroneous data resulting in over 40,000 Oyster cards having to be replaced (SCN, 2008)</td>
</tr>
<tr>
<td>2.2</td>
<td>Faulty location from AVL system*</td>
<td>Hardware</td>
<td>A smartcard reader typically receives the location of the bus from the Automatic Vehicle Location (AVL) system. Therefore if the AVL location is erroneous, then the smartcard boarding or alighting stop will also likely be erroneous. The AVL system can be erroneous due to a myriad of reasons including faulty GPS, faulty odometer, or faulty base data. A list of typical AVL faults can be found in Robinson and Manela (2012)</td>
</tr>
<tr>
<td>3.1</td>
<td>Readers broken</td>
<td>Hardware</td>
<td>Smartcard readers can fail resulting in missing tap-in or tap-out data</td>
</tr>
<tr>
<td>3.2</td>
<td>Smartcards broken</td>
<td>Hardware</td>
<td>Smartcards can break like most other pieces of hardware. Under a Freedom of Information Act request Transport for London reported that in 2008 a total of 28,000 Oyster cards were returned with a technical failure and 67,000 Oyster cards were returned which had been damaged (bent, cut, etc.). NB. As of February 2009 a total of 21.2 m cards had been issued (TfL, 2009)</td>
</tr>
<tr>
<td>3.3</td>
<td>Comms. broken*</td>
<td>Hardware</td>
<td>Communications may break so that it is not possible to transmit smartcard data from the bus to the central server. Although smartcard data is typically backed up inside the bus it may be some time before it is possible to manually transfer the data from bus to backend system</td>
</tr>
<tr>
<td>4.1</td>
<td>Failure to tap in/tap out*</td>
<td>User</td>
<td>A user may fail to tap in or tap out. This may or may not be deliberate. Sometimes the user may not realise that the card has not been read, or sometimes the reader is broken and the user fails to attempt to validate on another reader. This issue is discussed in more detail in Section 6</td>
</tr>
<tr>
<td>4.2</td>
<td>Tap out too early*</td>
<td>User</td>
<td>On some smartcard systems the user has to both tap into the bus and then tap out as they exit. The fare will be determined by the distance travelled. Therefore there is a financial advantage for a user to tap out as soon as possible. This issue is discussed in more detail in Section 6</td>
</tr>
<tr>
<td>4.3</td>
<td>Card stolen or lost</td>
<td>User</td>
<td>Smartcards can be lost or stolen. This may mean that users may change card id over time. According to Transport for London, approximately 308,000 Oyster cards were lost and 16,000 stolen in 2008 (TfL, 2009)</td>
</tr>
<tr>
<td>4.4</td>
<td>Invalid card used</td>
<td>User</td>
<td>Although invalid cards such as stolen or lost cards can be “blacklisted”, there is often a limit to this. Cards may then be removed from this blacklist, so it becomes possible to reuse such cards</td>
</tr>
<tr>
<td>4.5</td>
<td>Multiple cards</td>
<td>User</td>
<td>A user may have more than one smartcard in their wallet. It is feasible that different rides on the same trip may get charged to different cards</td>
</tr>
<tr>
<td>4.6</td>
<td>Card shared</td>
<td>User</td>
<td>Smartcards may be shared between people so a single card id may be related to several people</td>
</tr>
</tbody>
</table>

- Ride number in trip
- Timestamp of start of ride
- Duration of ride – i.e. the time between boarding and alighting
- Boarding stop – i.e. the stop that the smartcard recorded the passenger to board
- Alighting stop – i.e. the stop that the smartcard recorded the passenger to alight
- Line
- Line direction
- Mode – i.e. Bus or MRT

This level of data is typical of that recorded by other transit agencies. For example Transport for London (TfL) make a sample of smartcard data available to researchers on their syndicated website. However the data recorded in Singapore provides one major advantage to that recorded in other cities such as London in that the alighting stop and time are also recorded for all rides. In other cities the alighting stop is typically not recorded as passengers do not have to tap out.

For most of the analysis undertaken in this research, only trips which included a bus-to-bus transfer were included. This is because bus-to-bus transfers allow a check on the consistency of alighting stop of the first ride and the boarding stop of the second ride. The number of bus-to-bus transfers in the 7 days is given in Table 2.
5. Issue 1: Identifying trips from a sequence of rides

The first issue that will be analysed concerns how to group the sequence of rides taken by an individual to form trips. The problem will be described, and a methodology proposed to group rides together.

5.1. Motivation

For transport planners and academics studying origin–destination patterns of travellers, it is important to determine where the trip begins and ends. This may sound like a trivial matter, but in many public transport trips, a passenger will transfer from one service to another service. Key questions to ask when a passenger is seen to take two public transport rides in a relatively short time period are:

- Should these two rides be grouped together as part of the same trip?
- Should an activity be inferred as having taken place between the two rides?

Seaborn et al. (2009) suggest that rides which fall between an activity which was the main purpose for travelling, should be split into two separate trips. Rides either side of an activity which was incidental in the decision to travel, such as buying a newspaper, should be aggregated together into the same trip. Aggregating rides into the correct trips is critical in the research of activity based modelling (Bowman and Ben-Akiva, 2001). This section uses this definition to suggest some rules which can be used to help group rides into trips.

5.2. Suggested solution

5.2.1. Rules for identifying trips from a sequence of rides

A series of business rules are suggested which allow a sequence of rides to be split into a new trip. These rules are listed in Table 3. The first five of these rules are based on those used by the LTA. The last one is a new rule which ensures that two trips forming a “return journey” are always counted as two trips rather than a single trip. This rule uses a similar underlying concept to the “circuity” rule proposed by Gordon et al. (2013), but considers all the rides of a trip rather than only two adjacent trips. De Villaine et al. (2012) also propose a series of rules for inferring transfers, in the context of determining if activities have taken place at a location. They allow a transfer time of up to 30 min. Of course, the list of possible rules given is not exhaustive and more research should be undertaken in this area.

Table 3
Suggested rules for identifying a trip from a sequence of rides.

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Rule description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Maximum number of transfers</td>
<td>There should be a limit to the number of transfers allowed. Further research could be undertaken on this. For example this limit could be dependent on the starting location</td>
</tr>
<tr>
<td>2</td>
<td>Maximum transfer time</td>
<td>There should be a limit to the duration of the transfer. This could be dependent on the transfer that is made. For example the transfer time from bus to train at location A should be less than the transfer time from train to bus at location A, since the transfer to bus will include the time spent waiting for the bus to arrive. Train stations have fixed infrastructure so that the user will already have completed the transfer as they wait for the train</td>
</tr>
<tr>
<td>3</td>
<td>Number of train trips</td>
<td>A passenger may only enter the train network once in the trip</td>
</tr>
<tr>
<td>4</td>
<td>Single journey on a bus line</td>
<td>A passenger may only use the same bus line once in the trip</td>
</tr>
<tr>
<td>5</td>
<td>Maximum trip time</td>
<td>There should be maximum time limit in which a trip has to be completed. Further research could be undertaken to determine a suitable value for this. For example this limit could be dependent on the starting location</td>
</tr>
<tr>
<td>6</td>
<td>Trips should be reasonably direct</td>
<td>The sequence of rides taken to travel from the origin O to desired destination D should be reasonably direct. A metric to determine the directness to a trip is given in Section 5.2.2</td>
</tr>
</tbody>
</table>
5.2.2. Measuring the directness of a trip

The new rule proposed in Table 3 suggests that trips should be reasonably direct. It is first necessary to propose a metric of the directness of a trip. It is based on two measurements:

- **D₁** – Distance from start: This is the geodetic distance (i.e. direct distance) from the start stopping point to the current stopping point.
- **D₂** – Total ride distance: This is the total distance travelled by the user along line of route on all the rides.

The “directness ratio”, \( R \), is then given by the ratio of these two values in Eq. (1):

\[
R = \frac{D_1}{D_2}
\]  

(1)

If the calculated value of \( R \) is less than a critical value \( R_{\text{critical}} \) then the sequence of rides is split into two trips. As an example, if \( R_{\text{critical}} \) is set to 0.1, then if a passenger were to travel 10 km on the bus on more than one ride, then the direct distance from the final exit stopping point to the initial starting stop point would have to be more than 1 km to prevent the system splitting the rides into two separate trips. A value of \( R_{\text{critical}} \) between 0.1 and 0.25 was set to reduce the possibility of any valid trips being found erroneous, although further research should be undertaken to determine what this value should ideally be.

As an example of the usefulness of this rule, Table 4 shows a sequence of four rides taken by a passenger in Singapore on 28th April 2011. It is illustrated in Fig. 1.

From Fig. 1 it can clearly be seen that these four rides are a return journey from Braddell to Ang Mo Kio. However the first 5 rules of Table 3 would not identify this as two trips since; no bus line had been used twice, the train network had only been used once, and no transfer time was unreasonably long.

The OD of the trip would therefore be from Braddell MRT Station bus stop to Braddell MRT station – a distance of only 67 metres! However the use of the directness ratio of rule 6 would identify this sequence of rides as two trips. Given that the total ride distance is 13,500 metres and the direct distance at the end of ride 4 was 67 metres, the critical ratio would only need to be set at 0.005 for this return trip to be identified. In the above example, the group of rides would then be split into two trips; trip 1 containing rides 1, 2, 3, and trip 2 containing ride 4.

Therefore it is strongly recommended that all transport planners and researchers verify how rides are aggregated into trips, and ensure that the “directness of a trip” rule is incorporated. This is an area where further research is encouraged.

6. Issue 2: Identifying erroneous tap-ins and tap-outs

Another important issue to consider when using smartcard data is whether the location stops of the tap-in and tap-outs are correct. To this author’s knowledge the literature makes the assumption that this data is correct. This is a naïve assumption to make and can severely impact analysis done using smartcard data. For example, if an analyst wants to understand transfer behaviour between stopping-points, then it is vital that the exact boarding and alighting stops are known. An error by one stop of measuring either the boarding or alighting stop can significantly impact any estimated walking distance between boarding and alighting transfer stops.

This section will briefly review potential causes of faulty tap-ins and tap outs. It will then suggest an approach to detect erroneous tap-outs.

6.1. Causes of erroneous tap-ins and tap-outs

It is important to review the causes of erroneous tap-ins and tap-outs. Understanding the causes of failure will allow the analyst to design validation checks targeted at these issues. This section elaborates on some of the failures given in Table 1.

The first main cause of erroneous tap-in and tap out data is AVL system failure. The usual functioning of a smartcard system is only to allow passengers to tap-in and tap-out from the bus in the vicinity of a bus stop. Between bus stops the smartcard readers will be temporarily disabled. The ticket machine will obtain its stop-centric location from the AVL system of the bus. Therefore if the AVL location is wrong then the stop location recorded on the card will be wrong. An AVL system

<table>
<thead>
<tr>
<th>Ride</th>
<th>Start time</th>
<th>End time</th>
<th>Start stop</th>
<th>End stop</th>
<th>Duration, secs</th>
<th>Line</th>
<th>Dist ride/m</th>
<th>Total ride/m</th>
<th>Dist from start/m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12:15:48</td>
<td>12:26:01</td>
<td>52161</td>
<td>53229</td>
<td>613</td>
<td>59</td>
<td>2800</td>
<td>2800</td>
<td>918</td>
</tr>
<tr>
<td>2</td>
<td>12:26:01</td>
<td>12:26:01</td>
<td>TRANSFER</td>
<td>TRANSFER</td>
<td>985</td>
<td>88</td>
<td>4400</td>
<td>7200</td>
<td>3861</td>
</tr>
<tr>
<td>3</td>
<td>12:42:26</td>
<td>13:02:03</td>
<td>53239</td>
<td>54571</td>
<td>1177</td>
<td>265</td>
<td>2800</td>
<td>10,000</td>
<td>3404</td>
</tr>
<tr>
<td>4</td>
<td>13:02:03</td>
<td>13:13:41</td>
<td>TRANSFER</td>
<td>TRANSFER</td>
<td>698</td>
<td>505</td>
<td>MRT</td>
<td>13,500</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 4

The directness ration can be used to identify return trips.
can also be running in degraded mode. For example, a common failure of AVL systems is the accuracy of the odometer. If this is detected as being faulty, then the navigation system will use solely GPS sensor input. This will have an impact on the location accuracy when the bus is travelling between bus stops. This is because this is commonly estimated using odometer readings. Therefore if only GPS is available, the exact location of the bus can only be known with certainty when the bus is in the vicinity of a scheduled stop.

To mitigate this, the ticket machine will have been designed to “degrade gracefully”. The ticket machine will believe the bus is always at the last confirmed stop, until it receives positive confirmation that that the bus has arrived at a new stop on its scheduled path. Because of this “fallback mode of operation” it will then be possible for an alighting passenger to tap-out at any point between bus stops. However, until the bus gets within a certain distance of the next scheduled bus stop, the user tap-out will be registered to the previous confirmed bus stop (i.e. the last bus stop visited). This means that smartcard alighting stops recorded are likely to be the previous stop visited rather than the true alighting stop. This is illustrated by Fig. 2. In this example the user can usually only tap out in the vicinity of the stop as shown in the top diagram. However in fallback mode the passenger can tap-out at any point, with the stop only being updated when the bus arrives at the next stop. A way to detect this type of fault is given in Section 6.2. It should be noted that the reason for this fall back behaviour is that the passenger will not be overcharged. Rather the fallback mode is biased towards undercharging the passenger.

![Fig. 1. Example of four rides treated as a single trip rather than two trips.](image)

![Fig. 2. Illustration of how fallback mode caused by AVL failure can result in erroneous smartcard tap-outs.](image)
A second cause of erroneous tap-in and tap-out data are data supply errors. On other AVL systems the location of the bus is determined primarily using an odometer without GPS input. When the bus opens its doors, then the system will correct its location to the nearest stop, or the driver may manually correct the location. Such systems are therefore dependent on the following:

- **Odometer working**: It is necessary for the odometer to be reasonably accurate.
- **Correct distance data**: It is necessary for the on-road distance between adjacent bus stops sent in the data supply to the AVL system to be of certain accuracy.
- **Regular opening of doors**: It is necessary for the doors to open regularly to allow for the automatic location correction to be made.

An example of this type of AVL system failing is illustrated by Fig. 3. In this diagram case 1 shows the normal operation. The passenger is able to tap out just before the bus stop and at the bus stop (i.e. green areas). When the bus moves away from the stop they are not allowed to tap-out (i.e. red areas). In case 2 the data supply is faulty. The distance provided between stops S2 and S3 is too large. As the bus does not open its doors at stop S3 it is not able to correct its location. Thus at point M, when the bus is travelling between stops S3 and S4, it is possible to tap out of the bus and be registered as exiting one stop too early. This failure was caused by erroneous data.

Other causes of erroneous tap-in and tap-out data are described below:

- **Deliberate user error**: The third type of failure is deliberate user error as previously mentioned in Table 1. A passenger is often able to deliberately tap-out early if they desire, and are often unlikely to be caught. The passenger will benefit from this action if the fares are distance based. It should be noted that this error will be evident primarily in the tap-out data. This is because passengers typically tap-in at the door next to the driver. The driver is then able to check that all passengers have tapped-in. This process is made more robust if the driver and operator are incentivised to check that passengers tap-in when boarding.
- **Accidental user error**: The fourth type of failure is accidental user error. This is where the passenger believes they have tapped-in or tapped-out but have failed to do so. This is typically caused by failure of the card reader to read the card correctly. Often the smartcard is stored in a wallet or handbag, which can sometimes impact the signal between card and card-reader. Another scenario where this failed tap-in or tap-out is likely to happen is in crowded buses when there is difficulty getting unobstructed access to the card reader preventing users from visually checking that the card was read correctly.
- **Faulty Card Reader/Card**: The main other type of failure is when the card reader or card itself has faults. An approach to identifying vehicles with this type of hardware failure is given in Section 7.

### 6.2. Approach to detecting erroneous tap-outs

#### 6.2.1. Methodology to detect erroneous tap-outs

From Section 6.1 it was noted that there are two main causes of tap-out error; AVL System error, and deliberate user error. In both these errors the user will have alighted at a different stop on the scheduled trip than the alighting stop recorded by the smartcard. Therefore an approach to detecting erroneous tap-outs is to estimate where a passenger would likely have alighted the bus in order to transfer to another line (bus or MRT). This approach builds upon the literature used to estimate the alighting point in smartcard systems where this is not explicitly recorded – see Section 2.3. The check can be undertaken for all bus-to-bus or bus-to-MRT transfers and follows the following steps. In this explanation a bus transfer is taking place from the alighting stop of ride 1, S1 to the boarding stop of ride 2, S2. Ride 1 is on vehicle journey VJ1.

#### Step 1:

Identify the alighting bus stop of ride 1 and the boarding stop of ride 2 in the trip recorded in the smartcard data. These two stop points define the transfer walk that the passenger was reckoned to have made. Calculate the direct distance between these two stop points. This recorded transfer walking distance is denoted by $W_D$.
Step 2: Identify all stops downstream of the recorded alighting bus stop of ride 1, S1 on vehicle journey VJ1.

Step 3: Calculate the direct distance between the boarding stop S2 and all downstream bus stops on ride 1 identified in step 2. The walking distance between stop s on vehicle journey VJ1 and the boarding stop S2 is given by \( WD_s \).

Step 4: Calculate the distance travelled on the road by the bus between the alighting stop S1 and all downstream bus stops identified in step 2. The on-road distance between the alighting stop S1 and stop s on vehicle journey VJ1 and is given by \( RD_s \).

Step 5: For each downstream stop identified in step 3 calculate the benefit, \( C_s \), if the passenger were to have alighted at a downstream stop, s. It should be noted that the numerator represents the reduction in walking distance if the passenger were to alight at stop s.

\[
C_s = \frac{(WD_s - WD_1)}{RD_s}
\]  

Step 6: Identify the highest value of \( C_s \) and if it is greater than a certain critical value, then assume that the passenger most probably alighted at this stop rather than the recorded stop. A plot of the benefit for over 50,000 bus-to-bus transfers is shown in Fig. 4. The distribution is very bi-modal. A critical value of between 0.2 and 0.7 appeared to clearly identify suspect tap-outs. Analysis of a sample of cases suggested that those above the critical value are all suspect tap-outs, and those below are all likely valid tap-outs.

The key concepts of this methodology are further explained below. Fig. 5 illustrates how this method attempts to find a downstream stop whose direct distance to the next ride boarding stop is less than the direct distance between the smartcard recorded alighting stop and boarding stop. As an example, in Fig. 5 the smartcard records the passenger alighting at stop S11 and then boarded at stop S40. However bus stop S12 is obviously closer to boarding stop S40. However bus stop S12 is obviously closer to boarding stop S40 which suggests that the passenger may have actually alighted at stop S12.

The rationale behind calculating the on-road distance between recorded alighting stop and each downstream stop of step 5 is illustrated by Fig. 6. In this example the passenger is recorded as alighting ride 1 at stop S25 and boarding the next ride 2 at stop S700. This example shows a bus route which loops back. Therefore in this example the downstream stop S42 is closer to the boarding stop than the alighting stop. However it is unlikely that a passenger will actually choose to stay on the bus until stop S42 as the journey time from stop S25 to stop S42 would outweigh the benefit of the shorter transfer walk. A similar concept of reducing overall trip travel time was also used by Munizaga and Palma (2012). Distance along road has been used as a proxy to travel time in the approach described above.

**Fig. 4.** Histogram of the “benefit” of tapping out early. A high benefit suggests that the passenger alighted at a downstream stop to that recorded.
The method described above identifies which tap-outs are likely suspect. However it does not indicate what the cause of this was. Therefore a way of classifying erroneous tap-outs was proposed. It was possible to calculate the number of stops early that a tap-out was. This allows for classification as follows:

- **AVL system error**: If a tap-out was deemed to be one stop too early, then the cause of this was attributed to AVL system error. There are several reasons why this is likely to happen as illustrated in Fig. 2.
- **Deliberate user error**: If a tap-out was calculated to be more than one stop too early, then the cause of the system error was deemed to be deliberate user error. It should be noted that a passenger would derive more financial benefit by tapping out as soon as possible after tapping in.

### 6.2.2. Results

This logic has been applied to the Singapore EZ link data set described in Table 2. Data from all bus-to-bus transfers were used. The results are presented in Table 5. The results show that 7.72% of all tap outs were probably recorded one stop too early due to AVL system issues. Such a result seems reasonable when compared to other AVL systems around the world. By way of comparison, London Buses measures the percentage of scheduled data that is not recorded due to faults somewhere in the AVL system due to either hardware or data supply faults. In early 2009, just after the role out of the new AVL system, this was above 10%. However, after implementation of business processes to actively identify and fix AVL problems, this was brought down to approximately 5% in Autumn 2010, and further decreased to around 2% at the start of 2014 (TfL, 2014).

An approach to identify problematic buses and data supply is provided in Section 7. This approach could also be used to identify users who appear to be tapping out too early.

### 6.2.3. Example: user error

This section provides an example of where the approach has identified a passenger who has most likely tapped out too early. A summary of the relevant rides of the passenger is shown in Table 6 and the various rides shown in Fig. 7.

The passenger boards line 7 but alights two stops later at 19:12:10. At 19:37:09 the passenger boards line 174 at stop 11201. From Fig. 7 it can clearly be seen that line 7 continued to this stop. It is therefore very likely that the passenger tapped out too early.

### 6.2.4. Example: system error

This section provides an example of where the approach has identified a potential system fault. A summary of the relevant rides of the passenger is shown in Table 7 and the various rides shown in Fig. 8.

It can be see that the speed of the transfer walk is very high, 6.8 km was covered in only 22 s! What has most likely happened is that there was a system fault in the first bus. There is only one stop between stops 52071 and stop 72011, since the bus line 5 takes the Pan Island Expressway for 7.7 km between these two stops. The passenger has obviously alighted at stop 72011 but the smartcard system recorded the user as alighting at stop 52071.
7. Issue 3: Identifying faulty vehicles and data supply

Robinson and Manela (2012) proposed the “Comparison against peers” approach to identify vehicles with faulty Automatic Vehicle Location systems at London Buses. This section will implement the same approach using information about potential system errors to determine if there are certain vehicles, or certain bus stops which have a far higher failure rate than their peers. A brief overview of the “comparison against peers approach” will be provided and then the application to faulty vehicle and faulty data supply detection described.

7.1. Comparison against peers approach

The “comparison against peers approach” is a metric which determines how many times more faults a certain individual tends to have compared to other individuals in its peer set. In the context of faults in bus operations, the peer group is all individual vehicles or individual bus stops on the same line. Comparing against “peers” rather than the total population is important. For example consider a line A where there are four separate data supply issues which cause missing data for ALL vehicles on the line. On line B there are no data supply issues. Suppose there are two vehicles, V1 and V2. Vehicle V2 has an intermittent odometer fault which may result in a system error on one stop every trip. If vehicle V1 ran on line A, then in two trips V1 would be recorded as having four faults due to the data supply issues. If vehicle V2 ran on line B, then in two trips V2 would be recorded as having two data supply issues due to hardware failure. However, a quick analysis would suggest that vehicle V1 had double the system errors of vehicle V2 and was hence the faulty vehicle!

The “comparison against peers approach” overcomes this by comparing the average number of faults for each individual bus against the average number of faults for all vehicles operating on the same line. In the context of identifying stops or vehicles which cause the greatest number of system tap-out faults, this ratio is given by Eq. (3). This equation takes into account the fact that a bus might operate on more than one line.

\[
PI_v = \frac{1}{TO_v} \sum_{l \in L} \left( \frac{FO_{v,l} \cdot TL_l}{TL_l} \right)
\]

With \(PI_v\) is the Performance Index for object \(v\), \(FO_{v,l}\) is Number of system tap-out faults for object \(v\) when operating on line \(l\), \(TO_v\) is Total number of tap-outs (both ok and faulty) for object \(v\); \(FL_l\) is Total number of system tap-out faults for all objects operating on line \(l\), \(TL_l\) is Total number of tap-outs (both ok and faulty) for all objects operating on line \(l\), \(L\) is set of lines that the object operated or is related to

This approach has already been implemented successfully by London Buses to identify vehicles with faulty AVL systems. This approach was thus adapted to see whether it could identify faulty vehicles and data supply causing system tap-out errors.

7.2. Data for experiment

Data for this work was obtained by running the approach to detect erroneous tap-outs described in Section 6.2. For every bus-to-bus transfer, the following data was collected:

- Vehicle id
- Line id and direction
- Alighting stop recorded on smartcard
- Status: i.e. valid, system fault, or user fault – see Section 6.2.1.

7.3. Identifying faulty vehicles

The first application of the comparison against peers approach was to identify vehicles which have more system tap-out faults than their peers and hence are likely to have a hardware fault. Thus in Eq. (1), “Object” referred to a vehicle. User faults were excluded from all analysis. In addition all vehicles with less than a total of 100 tap-out records were excluded. In total...
there were 3401 vehicles with more than 100 tap-out records. Table 8 shows the 5 vehicles with the worst PI value, where vehicle ids have been anonymised.

This analysis suggests that vehicle V1 has 9.5 times more system faults than other buses running on the same line. This could be seen to be the case when the number of faults of bus V1 on each of the three lines it operated on was compared against the line average number of system faults, shown in Table 9. This bus should therefore be investigated by mechanics for hardware faults.

At London Buses a similar approach has been used to actively identify and fix vehicles with erroneous AVL systems (Robinson and Manela, 2012). The “% system tap-out faults” metric could therefore be used as a way to identify vehicles with a hardware problem.

7.4. Identifying faulty data supply

The second application of the “Comparison against peers approach” was to identify faulty data supply. As outlined earlier, faulty distance data, or stop location data can result in AVL systems inferring the wrong location, resulting in smartcard records recording the wrong boarding or alighting stop. Since distance between stops may depend on the line, the “Object” in Eq. (1) referred to the combination of stop and line. User faults were excluded from all analysis. In addition all stop-line combinations with less than a total of 50 tap-out records were excluded from the ranking of worst stops.

Table 10 shows a list of the 6 stop-line combinations with the highest Performance Index of system faults. The last column shows the average percentage of system faults over all stops on the line. In the worst case, stop 84381 on line 225, there

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Example of probable user error in tapping out.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ride</td>
<td>Start time</td>
</tr>
<tr>
<td>1</td>
<td>19:10:01</td>
</tr>
<tr>
<td>2</td>
<td>19:12:10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Example of probable system error in tapping out.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ride</td>
<td>Start time</td>
</tr>
<tr>
<td>1</td>
<td>07:02:05</td>
</tr>
<tr>
<td>2</td>
<td>07:37:28</td>
</tr>
</tbody>
</table>

Fig. 7. Example of probable user error in tapping out.
are around 48 times more system faults seen at this stop than seen at other stops on the line. Clearly those stop-line combinations should be analysed further to determine if there are any data supply issues, or other local issues which are causing so many system faults.

### 8. Methodology for ensuring quality of smartcard data

This final section of the paper, brings together the ideas stated in this paper, and suggests a methodology for cleaning smartcard data for further analysis. It is suggested that if the analyst desires to undertake OD analysis, analysis of transfers, or estimates of bus ridership, then the methodology of this section is performed to filter out smartcard records which could distort the findings of such analysis. It should be noted that this method is appropriate primarily for smartcard data where alighting stop information is recorded in addition to boarding data. However the methods proposed in this paper could be adapted to clean smartcard data from systems where alighting information is not available. This would need to use techniques outlined in Section 2.3.

### Table 8

List of 5 vehicles which according to the comparison against peers report have the highest number of system errors.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Vehicle Id</th>
<th>Total tap out records</th>
<th>No system tap out faults</th>
<th>% system faults</th>
<th>PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V1</td>
<td>705</td>
<td>385</td>
<td>54.6</td>
<td>9.52</td>
</tr>
<tr>
<td>2</td>
<td>V2</td>
<td>162</td>
<td>73</td>
<td>45.1</td>
<td>8.51</td>
</tr>
<tr>
<td>3</td>
<td>V3</td>
<td>188</td>
<td>59</td>
<td>31.4</td>
<td>8.16</td>
</tr>
<tr>
<td>4</td>
<td>V4</td>
<td>445</td>
<td>187</td>
<td>42</td>
<td>6.37</td>
</tr>
<tr>
<td>5</td>
<td>V5</td>
<td>627</td>
<td>210</td>
<td>33.5</td>
<td>5.7</td>
</tr>
</tbody>
</table>

### Table 9

Worst vehicle accounts for a significant proportion of system tap out faults on the lines it operates on.

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>Line</th>
<th>Vehicle V1: total records</th>
<th>Column A: vehicle V1: % system tap out faults</th>
<th>Line total records</th>
<th>Column B: line % system tap out faults with vehicle</th>
<th>Col. A/Col. B</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>31</td>
<td>384</td>
<td>52.3</td>
<td>24748</td>
<td>4.5</td>
<td>11.6</td>
</tr>
<tr>
<td>V1</td>
<td>12</td>
<td>297</td>
<td>39.4</td>
<td>15057</td>
<td>7.9</td>
<td>5.0</td>
</tr>
<tr>
<td>V1</td>
<td>196</td>
<td>134</td>
<td>50</td>
<td>15402</td>
<td>7.6</td>
<td>6.6</td>
</tr>
</tbody>
</table>
The suggested approach to pre-processing smartcard data is given in Fig. 9 and described below. It assumes that raw data for each ride is provided.

**Step 1:** The first step will undertake some “basic” checks on the raw ride data. This includes ensuring that data for all compulsory attributes is available. Basic range checks can be undertaken on each attribute. It also ensures that all required look-up tables required to interpret the data are sufficiently populated.

**Step 2:** The second step is to identify tap-in and tap-out data that is likely erroneous as discussed in Section 6. Tap-out data that appears to be erroneous should be corrected automatically before step 3.

**Step 3:** The third step is to aggregate rides into trips as discussed in Section 5.

**Step 4:** An additional step should be to analyse the data from step 2 to identify suspect faulty vehicles and data supply. Further analysis can then be undertaken on these.

### 9. Conclusions and further work

Smartcard systems are increasingly being used for transit fare payment, and the data from such systems are being used to understand demand patterns and traveller behaviour. It is therefore vital to ensure that the raw data being recorded by smartcard systems is correct. The literature has largely overlooked this issue so far, and only basic data quality checks have so far been implemented. This paper has sought to rectify this. It has highlighted why smartcard data may not always be correct. It has suggested three approaches to identifying problematic data. It then used one of these approaches to help identify the cause of the problems – either faulty vehicles or data supply. It is suggested that rigorous data checks are carried out on raw smartcard data before further analysis is undertaken. The process given in Section 8 could be used. The benefits of implementing such a process would be better transport planning data, improved AVL system, and reduced revenue loss.

This research is primarily applicable for smartcard systems where both boarding and alighting data is available. Future work could be to develop techniques which can be used when alighting data is not available. However, it is suggested that all smartcard systems should require passengers to tap-out. This would allow for the collection of useful planning data. A small reduction in fare could be provided to passengers who did tap-out – this is analogous to offering a discounted fare for passengers who do not use cash. This action could be justified by arguing that better planning data can allow for a better allocation of resources.

Further research into identifying other errors outlined in Section 3 is also strongly encouraged. This could include detecting erroneous tap-in data. Once such techniques have been developed, they should be incorporated into a “data cleaning suite” and then implemented before further analysis of the smartcard data is undertaken. Ideally the output of such a “data cleaning suite” should be to an industry standard, so analysis tools developed for one city could be run on smartcard data output from another city.

### Acknowledgements

The authors thank the Land Transport Authority of Singapore for providing the EZLink smartcard transaction data that was used in this research, and in particular Feng Yuan Liu. The authors wish to thank former colleagues at Technical Services...
Group, London Buses, Transport for London, for assistance with obtaining London Buses AVL performance data. The authors also wish to thank two anonymous reviewers for some useful feedback and critique.

This research was supported by the National Research Foundation Singapore through the Singapore MIT Alliance for Research and Technology's Future Urban Mobility IRG research programme.

References


