Transportation Research Part C 68 (2016) 490–506

Contents lists available at ScienceDirect

Transportation Research Part C

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

Validating and improving public transport origin–destination estimation algorithm using smart card fare data ☆

Azalden Alsger *, Behrang Assemi, Mahmoud Mesbah, Luis Ferreira

School of Civil Engineering, The University of Queensland, Australia

ARTICLE INFO

Article history:
Received 22 August 2015
Received in revised form 18 March 2016
Accepted 6 May 2016
Available online 20 May 2016

Keywords:
O–D estimation
Validation
Public transport smart card fare data
Trip-chaining method

ABSTRACT

Smart card data are increasingly used for transit network planning, passengers' behaviour analysis and network demand forecasting. Public transport origin–destination (O–D) estimation is a significant product of processing smart card data. In recent years, various O–D estimation methods using the trip-chaining approach have attracted much attention from both researchers and practitioners. However, the validity of these estimation methods has not been extensively investigated. This is mainly because these datasets usually lack data about passengers' alighting, as passengers are often required to tap their smart cards only when boarding a public transport service. Thus, this paper has two main objectives. First, the paper reports on the implementation and validation of the existing O–D estimation method using the unique smart card dataset of the South-East Queensland public transport network which includes data on both boarding stops and alighting stops. Second, the paper improves the O–D estimation algorithm and empirically examines these improvements, relying on this unique dataset. The evaluation of the last destination assumption of the trip-chaining method shows a significant negative impact on the matching results of the differences between actual boarding/alighting times and the public transport schedules. The proposed changes to the algorithm improve the average distance between the actual and estimated alighting stops, as this distance is reduced from 806 m using the original algorithm to 530 m after applying the suggested improvements.

1. Introduction

The use of smart card fare data has been rapidly increased in the field of transit network planning, behaviour analysis and demand forecasting (Pelletier et al., 2011). These data have become a valuable source of information for public transport origin–destination (O–D) estimation, allowing a better understanding of individuals' travel patterns and analysing the variability of transit users' behaviour (Morency et al., 2007; Kusakabe and Asakura, 2014; Kieu et al., 2015; Langlois et al., 2016).

Recently, a number of studies have used different methodologies to infer the O–D matrices for public transport trips using smart card fare data (Barry et al., 2009; Alfred Chu and Chapleau, 2008; Munizaga et al., 2010; Wang, 2010; Nassir et al., 2011; Gordon et al., 2013). Most automated fare collection systems record passengers' boarding information but not their alighting information. The lack of details on alighting stops is therefore the result of the currently used automated fare collection systems in which passengers are not required to tap off their cards upon alighting. Given this limitation of the data

☆ This article belongs to the Virtual Special Issue on Data-driven smart-city-enabled traffic system modeling, analysis and optimization.

* Corresponding author.
E-mail addresses: a.alsger@uq.edu.au (A. Alsger), b.assemi@uq.edu.au (B. Assemi), mahmoud.mesbah@uq.edu.au (M. Mesbah), l.ferreira@uq.edu.au (L. Ferreira).

http://dx.doi.org/10.1016/j.trc.2016.05.004
0968-090X/

produced by the majority of the existing systems, the O–D estimation results based on the smart card fare data needs to be re-evaluated, before its use in the analysis of individuals' travel behaviour (Bagchi and White, 2004; Munizaga et al., 2014).

The trip-chaining method, described later in this paper, is normally used to construct a passenger's travel sequence by connecting trip-legs recorded by his/her smart card usage. A few studies have attempted to evaluate this method and its assumptions. Farzin (2008) validated the 2006 estimated O–D results obtained from 5% of all transit trips in São Paulo, Brazil with the 1997 O–D household survey results. The results were not convincing due to the data limitations and the 10-year time-lapse between the datasets used for the comparison. Barry et al. (2009) validated the results obtained from the analysis of smart card data, with the data collected by passenger counting at the exit and entrance of subway stations as well as boarding and alighting at bus stops. Their study is based on two major assumptions: a very high percentage of passengers return to their previous alighting station to start their next trip, and a high percentage of passengers finally return to the first station they started their first trip of the day. Gordon et al. (2013) used the assumption of last destination as the closest to the first origin in their methodology. However, these assumptions need to be validated with a more reliable data source which includes accurate boarding and alighting details of public transport passengers.

The accessibility and quality of the additional data required for further evaluation of the trip-chaining method have been usually a big challenge. Devillaine et al. (2012) proposed a method for evaluating smart card analysis results with travel surveys, where the users' smart card IDs are recorded as a part of the survey. Chow (2014) evaluated an online approach to conduct customer surveys at a public transit agency by linking prior trip history into the survey. Munizaga et al. (2014) applied exogenous validation (information from travel surveys and personal interviews of a small sample of volunteers), in addition to endogenous validation (information from the same dataset), to validate the assumptions of the trip-chaining method, given the lack of alighting information in the main dataset.

Alsger et al. (2015) evaluated the common trip-chaining method assumptions using a unique smart card fare dataset obtained from TransLink, the public transport authority of South-East Queensland (SEQ), Australia. The important advantage of this dataset for the evaluation of the trip-chaining method assumptions is that it includes both boarding and alighting times and locations for each passenger of the public transport services that comprise buses, trains and ferries. The study focused on the individual assumptions (allowable transfer time, allowable walking distance and last destination of a given day being the same as the first origin of that day) of the trip-chaining method, in a situation where actual boarding and alighting information were known. Table 1 summarises the findings and gaps of the existing literature:

However, none of the above-mentioned studies has implemented and validated the whole estimation algorithm with a reliable dataset. Hence, the objective of this paper is to validate and improve the accuracy of the existing trip-chaining method through an in-depth evaluation of the public transport O–D matrices based on passengers' actual boarding and alighting data. The results highlight the impact of the method's assumptions on the accuracy of O–D estimation. Furthermore, a revised algorithm is proposed and empirically evaluated to improve the accuracy of the trip-chaining method.

The remaining sections of this paper are organised as follows. The next section explains the data description and preparation procedure. The research methodology is then described, which comprises the implementation of the existing O–D estimation method, the validating procedure and the improvement of the method by suggesting a revised algorithm. The results of the implementation and evaluation of the existing trip-chaining method are provided next. These results are then

Table 1
Summary of the findings and gaps of the existing literature.

Component		Studies	Findings	Gaps
Estimation assumptions	Walking distances (buffer zones)	Cui (2006), Wang (2010), Nassir et al. (2011), and Munizaga and Palma (2012)	Different walking distances were chosen to infer alighting stops (e.g., 400, 800, 1000 and 1100 m)	Different values were used for the assumptions of the O–D estimation. None of these studies have implemented and validated the whole estimation algorithm with a reliable dataset
	Transfer times	Bagchi and White (2004), Nassir et al. (2011), Kieu et al. (2013), Ma et al. (2013), and Hofmann and O'Mahony (2005)	Different transfer times were chosen to connect trip-legs to infer O–D trips (e.g., 30, 60 and 90 min)	
	Last destination assumptions	Barry et al. (2002), Nassir et al. (2011), Munizaga and Palma (2012), and Gordon et al. (2013)	Some studies assumed the last destination as the first origin, where others assumed it as the closest stop to the first origin	
Validation attempts	Additional data requirement for validation	Farzin (2008), Barry et al. (2009), Devillaine et al. (2012), Munizaga et al. (2014) and Chow (2014)	Additional data (e.g., travel survey, personal interviews, and passenger counting) were used for validation	The accessibility and quality of the additional data required for further evaluation of the trip-chaining method are usually a concern

used to discuss, propose and evaluate the revised algorithm which improves the accuracy of the existing trip-chaining method. Finally, some conclusions and suggestions for future work are presented.

2. Data description and quality improvement

The smart card (known as GoCard) fare data analysed in this study were obtained from TransLink, the public transport authority of South-East Queensland (SEQ), Australia. Data for one weekday (Wednesday 20 March 2013) were analysed over the SEQ bus, train and ferry network: this day was selected as it was in the middle of the week, was not a public holiday and had normal weather condition. In the SEQ network, a transaction record is generated each time a passenger boards and alights. Each transaction contains information comprising: the operation date, run, route, direction, ticket number, smartcard ID, boarding time, alighting time, boarding stop and alighting stop.

An important aspect of this system is that it includes both boarding and alighting times and locations, that is, where the passenger gets on or off a public transport vehicle. Transferring activities are not directly obtained. The data were filtered with some transactions excluded, such as duplicate transactions and when no boarding or alighting stops were recorded. Some transactions were also excluded owing to missing data, such as the bus route number. Smart card IDs with only a single transaction were also excluded, as the trip-chaining procedure requires at least two transactions per smart card ID. If any transaction of a selected card ID holder was excluded, the rest of the card ID transactions for that card ID holder were also excluded, as the transactions have to be in sequence to be chained. The one-day data initially included 184,074 transactions. After excluding 22,628 records (i.e., 12.3%) in the data cleaning process, the final dataset included 161,446 transactions (corresponding to 63,597 smart card ID holders).

To improve the quality of the dataset and obtain a more accurate evaluation of the O–D estimation method, a closer look at the transactions reveals an issue that occurs due to the nature of the existing smart card fare management system and is applicable to all similar systems (Robinson et al., 2014). When passengers forget to tap off their smart cards when alighting, the system should estimate the alighting information. The existing SEQ smart card fare management system has used the last stop of the route on which a passenger has forgotten to tap off his/her smart card as the 'actual' alighting stop. Fig. 1 shows an example of the trips, where this issue has occurred.

As shown in Fig. 1(a), the first alighting stop for a passenger is the last stop of the route. However, the second boarding stop is at a distance of 10.23 km from the first alighting stop, whilst the time difference between the first alighting and the second boarding is 4.33 min. Hence, the first alighting stop is very likely to have been selected by the smart card fare management system, as the passenger has forgotten to tap off his/her smart card when alighting. The first alighting stop is more likely to be located where shown in Fig. 1(b), given the following simple rules to estimate it:

Fig. 1. Example of smart card transactions with a wrong alighting stop. (a) Wrong transactions and (b) potentially actual transactions.

- Find the stops with the same code as the first boarding stop in the schedule.
- Choose the stop for which, the scheduled boarding time is the closest to the actual boarding time.
- Extract the sequence of stops from the schedule starting at the selected stop.
- Exclude the stops with a scheduled arrival time later than the next boarding time.
- Choose the closest stop to the next boarding stop.

A conservative approach was used in this study to detect the potentially wrong alighting stops in the dataset. First, for each individual transaction in the dataset, the alighting stop is estimated using the estimation method discussed above. Then, a list of transactions is extracted from the dataset for which, the alighting stop is the last stop of the corresponding public transport route and the distance between the actual stop and the estimated alighting stop is more than 800 m. Next, a subset is chosen from this list for which, the direct distance between the alighting stop and the next boarding stop (the first boarding stop, if it was the last transaction of the day) is more than 1000 m. Although the number of such transactions is very limited in the dataset (i.e., 674), these transactions were excluded from the validation process, as they might adversely impact on the validation results.

3. Methodology

To evaluate and improve the existing trip-chaining estimation algorithm, the algorithm is implemented and validated by using the unique smart card fare data described in the previous section. Although the boarding and alighting data exist in the dataset, the alighting information is assumed to be "unknown" for validation purposes. Based on this assumption, the public transport O–D matrix is generated based on different transfer time and distance criteria as the main assumptions of the method. The accuracy of the resulting matrices is evaluated based on the actual O–D matrix generated from the complete boarding and alighting data. Some methodological improvements are also proposed, based on the validation results.

3.1. Implementation of existing O–D estimation algorithm

This section explains the trip-chaining method and its assumptions, as proposed by the current literature (Wang, 2010; Nassir et al., 2011; Gordon et al., 2013). The trip-chaining method basically develops a list of public transport passengers' trips, by connecting the corresponding trip-legs for each smart card holder, when some certain criteria for a transfer between the trip-legs are met. These criteria are often based on the assumption that passengers choose a public transport stop to board, which is the closest stop to their previous alighting stop (virtually their current location). This method creates a buffer zone around every boarding stop based on an assumed walking distance to infer the previous alighting stop. The general assumptions of the trip chaining method are:

- *Allowable transfer time:* Different values of allowable transfer time have been used by previous research. The allowable transfer time threshold ranges from 30 min (Bagchi and White, 2004; Nassir et al., 2011); to 60 min (Kieu et al., 2013; Ma et al., 2013); and even 90 min (Hofmann and O'Mahony, 2005).
- *Allowable walking distance (buffer zone):* Similar to allowable transfer time, different values of allowable walking distance for transfers have been used by previous research. Table 2 provides a summary of the applied allowable walking distances by previous research.
- *Last destination:* Two different assumptions for the last destination have been used by previous studies. The most common assumption is to choose the first origin in a given day as the last destination of the day (e.g., Nassir et al., 2011; Munizaga and Palma, 2012). The alternative, more realistic assumption is to choose a stop on the last trip-leg's route which is the closest stop to the first origin in a given day as the last destination of the day (Gordon et al., 2013).

Fig. 2 shows how these assumptions of the trip-chaining method work. As shown, a passenger boards at (B_1) and alights at (A_1) in his/her first trip-leg. Then the passenger walks to the next boarding stop (i.e., B_2) to start his/her second trip-leg, which ends at (A_2). As the alighting information is not usually recorded (which is not the case in the applied dataset in this study), a buffer zone is considered around (B_2) to estimate (A_1). Thus, an alighting stop (A_1') is used instead of actual, unknown (A_1) in the trip-chaining method, where (A_1') is inside the buffer zone and satisfies the trip-chaining method assumptions. It should be noted that the time between (B_2) and (A_1) could be short, which is enough for a transfer only, or long, which could involve an activity. If this time is less than the allowable transfer time, the first and the second trip-legs are chained to create an O–D trip (in which B_1 is the origin and A_2 is the destination). In this paper, the O–D trip is defined as the movement of a passenger from an origin to a destination using public transport services. An O–D trip may

Table 2
Values of allowable walking distances used by the literature.

Allowable walking distance	400 m	800 m	1000 m	1100 m
Relevant studies	Wang (2010) and Zhao et al. (2007)	Nassir et al. (2011)	Munizaga and Palma (2012)	Cui (2006)

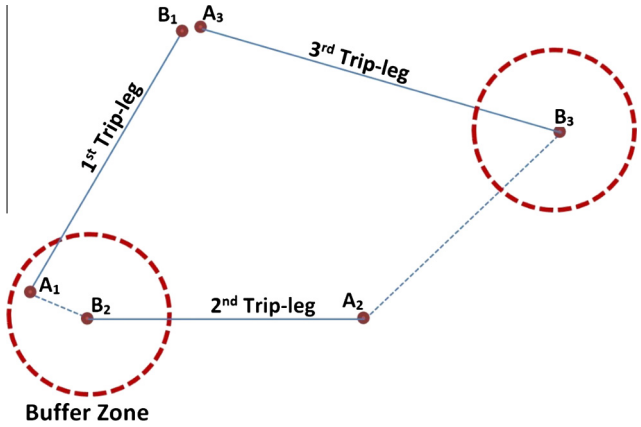

Fig. 2. Demonstration of trip-chaining method. B_i and A_i are respectively boarding and alighting stops. The time between each B_i and the consecutive A_i is the in-vehicle time.

have one or multiple trip-legs including transfers between trip-legs. However, it is only the component within the public transport system. Access to public transport is not part of the O–D trip.

The existing O–D estimation algorithm is presented in Fig. 3, as implemented in this study. This figure presents the O–D estimation algorithm where the alighting stops are not known (as is the case in the majority of smart card fare management systems). Fig. 3 also presents this study's proposed improvements to the existing algorithm, as explained later in the paper. As explained, the algorithm's adjustable parameters are allowable transfer time, allowable walking distance, and the last destination assumptions.

The algorithm basically estimates the alighting stops and chains the trip-legs wherever there is a transfer among them to generate public transport passengers' O–D trips. To accelerate the search process through all General Transit Feed Specification (GTFS) files (including public transit bus, train and ferry records), a search list for each route is created.

For each smart card ID's trip-leg, the search is performed through the schedule of the corresponding route to estimate the best-fitting alighting stop, as follows:

1. Choose a smart card ID from the database and then select the first trip-leg of the corresponding passenger for a given day.
2. Search through the database to find the closest stop to the location of the passenger's next trip-leg on the same service with a sequence number greater than the boarding stop.
3a. If the distance between the inferred alighting stop and the next boarding stop is greater than the specified allowable walking distance, the alighting stop is labelled as the trip's destination.
3b. If the distance between the inferred alighting stop and the next boarding stop is less than the specified allowable walking distance, check the inferred alighting time. The inferred alighting time has to be less than the next boarding time; otherwise, the inferred alighting stop is labelled as the trip's destination.
3c. If the time difference between the inferred alighting stop and the next boarding stop (i.e., transfer time) is less than the specified allowable transfer time, label the trip-leg as a transfer; otherwise the inferred alighting stop is labelled as the trip's destination.
If the currently processed trip-leg is the last trip-leg of the smart card holder, check this assumption in two different ways (the last alighting is the same as the first boarding of that day; and the last alighting is the closest to the first boarding of that day).
4. The search process continues for all card ID holders for the given day.

3.2. Validation procedure

Given the availability of both boarding and alighting data in this study's dataset, the actual O–D trips are generated to validate the existing trip-chaining method's assumptions (as shown in Fig. 4). In this process, trip-legs are chained together to obtain the passengers' actual O–D trips, based on different allowable transfer times (T). If the transfer time is less than the allowable transfer time between two consecutive trip-legs, the first trip-leg is labelled as a transfer; otherwise, the first alighting stop is labelled as the O–D trip's destination.

The generated actual O–D trips are compared in this study with the estimated O–D trips to validate the existing trip-chaining method. For each smart card ID, the validation procedure compares the estimated origin stop/zone with the actual origin stop/zone. The next step compares the estimated destination stop/zone with the actual destination stop/zone for each O–D trip matched in the first step. This comparison gives the matching percentage of the estimated and actual O–D trips that have the same origin and destination stops/zones.

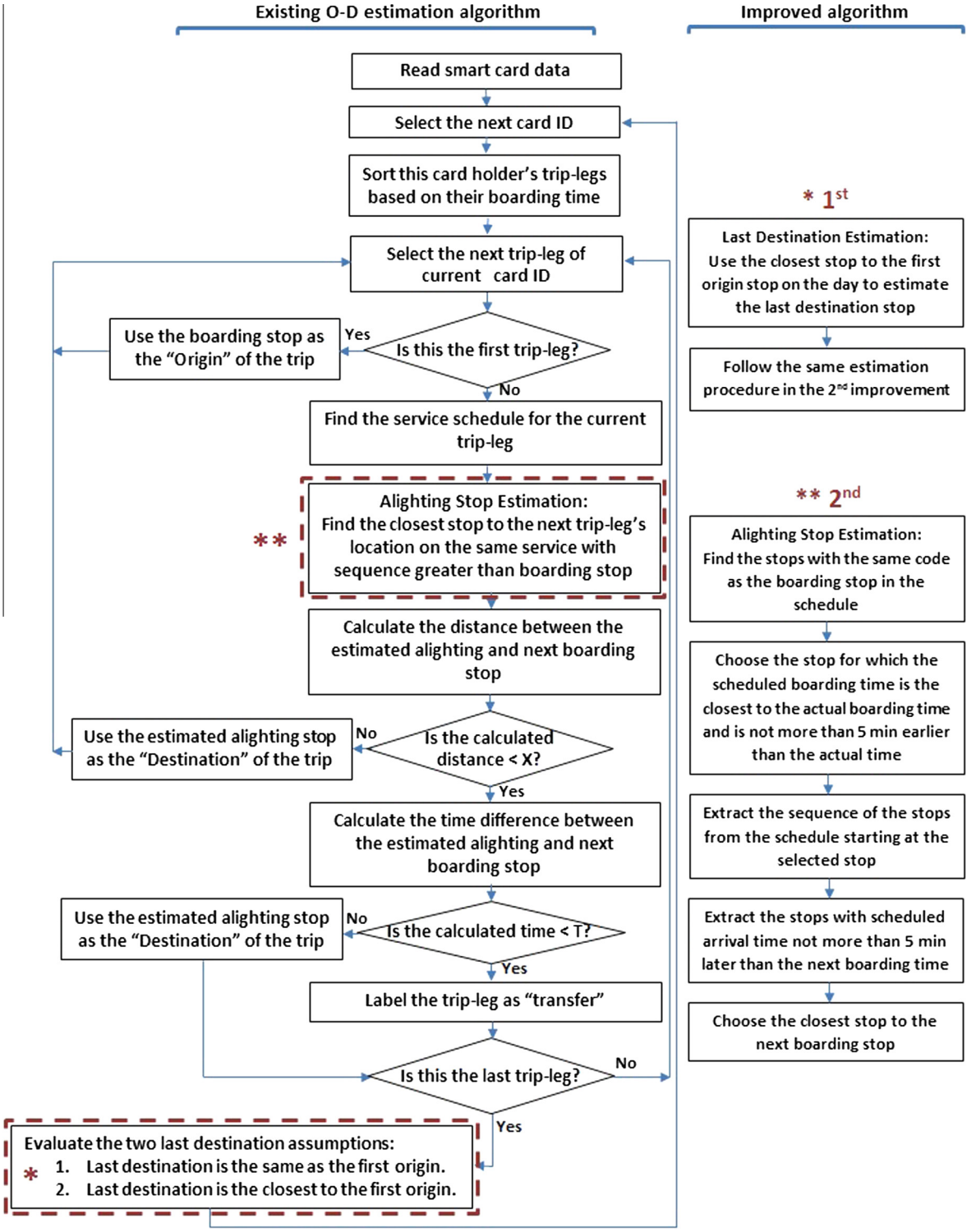

Fig. 3. Existing and improved O–D estimation algorithm. *T* is allowable transfer time, and *X* is allowable walking distance. The left flowchart shows the existing algorithm in which, the potential points for improvement are highlighted with the dotted boxes. The right flowchart shows the proposed improvements to the existing algorithm.

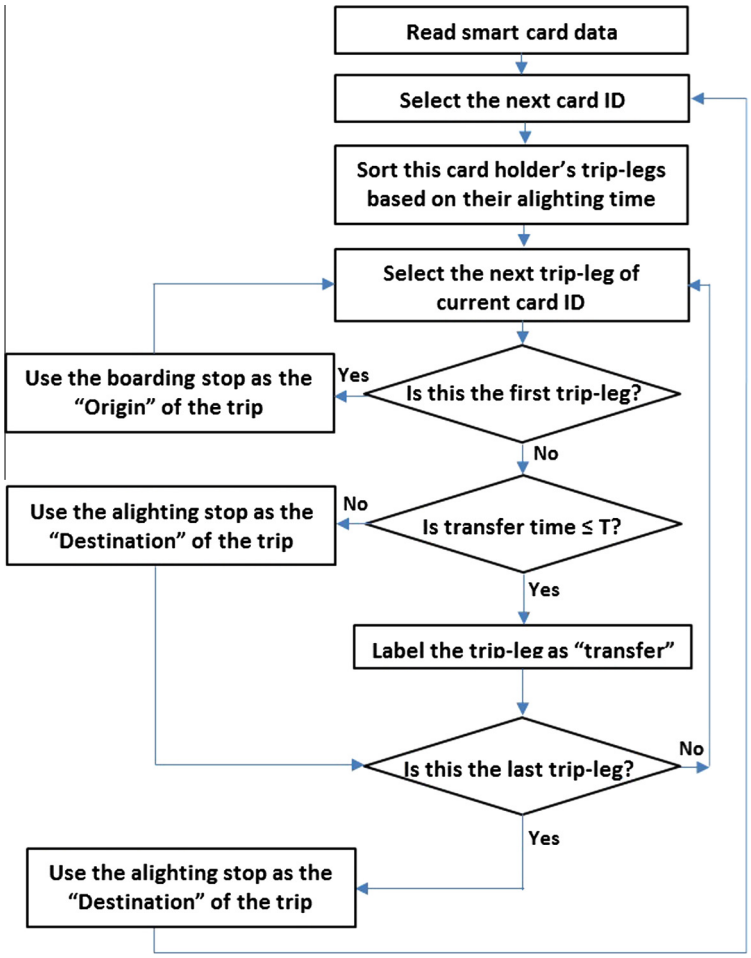

Fig. 4. Generating actual O–D trips using trip-chaining method and actual alighting data. *T* is allowable transfer time (30, 60 and 90 min).

3.3. Improvement procedure

With the availability of both boarding and alighting stops' data in the smart card fare dataset used in this study, the estimation errors are evaluated and some improvements are proposed to the existing trip-chaining algorithm. The results of the validation process are used to find the deficiencies and inaccuracies of the current algorithm. Some improvements to the algorithm are then proposed following an exploratory approach, as shown in Fig. 3. Finally, the improved algorithm is implemented and the results are evaluated to validate the proposed solution. The empirical results of the study are presented next.

4. O–D estimation and validation results

This section discusses the results of the evaluation of the existing trip-chaining method (illustrated in Fig. 3), with different values for the allowable transfer time and walking distances (i.e., the main O–D estimation method's assumptions). The values used in the evaluations of this study are 30, 60 and 90 min for the allowable transfer time, and the allowable walking distances as provided in Table 1. The last destination assumption is also investigated in two different ways as explained previously. To generate the actual O–D matrices, trip-legs of each smart card holder are chained based on the allowable transfer time, as explained in Fig. 4. To generate the estimated O–D matrices, the alighting stop for each trip-leg is estimated based on the allowable walking distance. Then, the trip-legs are chained based on the allowable transfer time, as explained in Fig. 3. The following subsections present the results of the evaluation of the trip-chaining method assumptions.

4.1. Allowable transfer time and walking distance assumptions

Fig. 5 shows the generated number of actual and estimated O–D trips at different allowable transfer times and walking distances. As shown in this figure, for all transfer times (i.e., 30, 60 and 90 min) and distances (as in Table 1), the number of

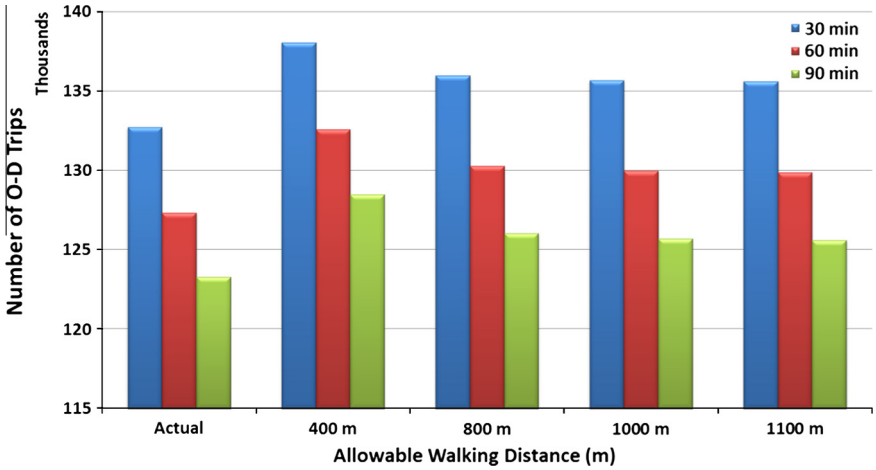

Fig. 5. Number of actual and estimated O–D trips at different walking distances and transfer times.

estimated O–D trips is greater than the number of actual O–D trips. The increment in the estimated number of O–D trips, compared to the actual number, is due to the method assumptions where trip-legs are chained based on different allowable transfer times and walking distances. However, the highest number of estimated O–D trips is at the allowable walking distance of 400 m, and this number drops as the walking distance increases from 800 m to 1100 m. Moreover, there is no significant difference between the total number of estimated O–D trips when the allowable walking distance increases beyond 800 m. As shown in Fig. 5, the total number of estimated and actual O–D trips is higher at the 30 min transfer time compared to the 60 and 90 min transfer times. As transfer trips (O–D trips that have at least one transfer), 18% of the O–D trips with 30 min allowable transfer time are transfer trips compared to 21% and 23% with 60 and 90 min allowable transfer time, respectively. As a 60 min allowable transfer time, 91% of the transfer trips have just one transfer compared to 8% with two transfers. About 80% of the transfer trips have an average transfer walking distance of 400 m at all allowable transfer times.

4.2. Last destination assumption

The assumed buffer walking distance cannot be applied to the last trip-leg of a passenger in a given day. To deal with this issue, two different assumptions have been made in previous studies. The first assumption is the last destination is the same as the first origin in a given day. The second assumption is the last destination is the closest to the first origin in a given day. In this section, the two assumptions regarding the last destination are validated as explained in Fig. 3.

To have a closer look at the results, the estimated matrix with 60 min allowable transfer time and 800 m allowable walking distance is chosen for further evaluation of this assumption. An in-depth investigation of the first assumption shows that 11.6% (72.6% of erroneously estimated O–D trips) of all matched O–D trips, for which the destination is estimated at more than 800 m from the actual destination, are the final transactions of the day for the corresponding passengers. The results show that the average distance is 5059 m and the maximum distance is 36,527 m for these trips. The main reason for this discrepancy is that the trip-chaining method assumes the last destination to be the first origin of the day for the corresponding passenger. Fig. 6(a) illustrates such trips where the last destination of a passenger is estimated to be the same as the first origin (i.e., point B) of that day, although the distance between the actual last destination (i.e., point C) and the estimated last destination is short (1.5 km).

When the second assumption is applied, a significant improvement is obtained in the matching between the actual and estimated last destinations, as shown in Fig. 6(b). The applied procedure for the last trip-leg of the day is as follows:

- Use the last trip's route ID to find the last alighting stop for each smart card holder.
- Find the public transport stop on this route which is the closest to the first boarding stop of the day for the same passenger.
- Choose this stop as the last destination, as shown in Fig. 3.

The applied procedure for the second assumption was proven to work even for long distances between the actual and estimated last destination. Fig. 7(a) presents an example of such trips where the distance between the actual last destination (i.e., point C) and the estimated last destination (i.e., point B) based on the method's assumptions is more than 10 km. After applying the second assumption procedure, the last destination stop is estimated to be the same as the actual last destination as shown in Fig. 7(b).

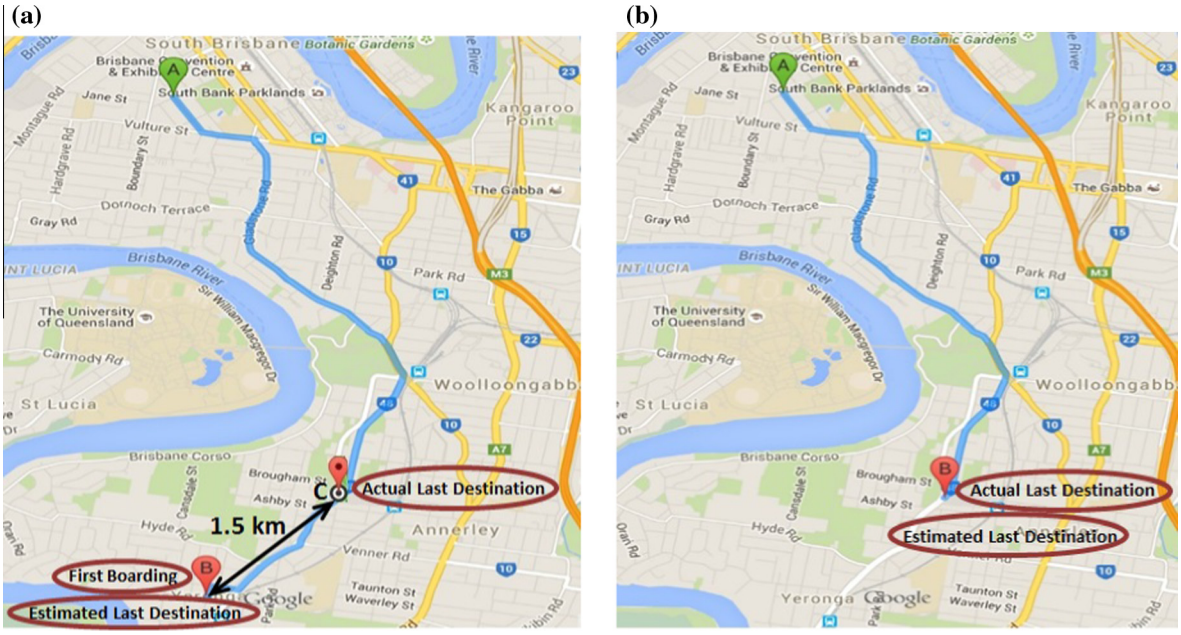

Fig. 6. Example of the last destination assumptions. (a) First assumption and (b) second assumption.

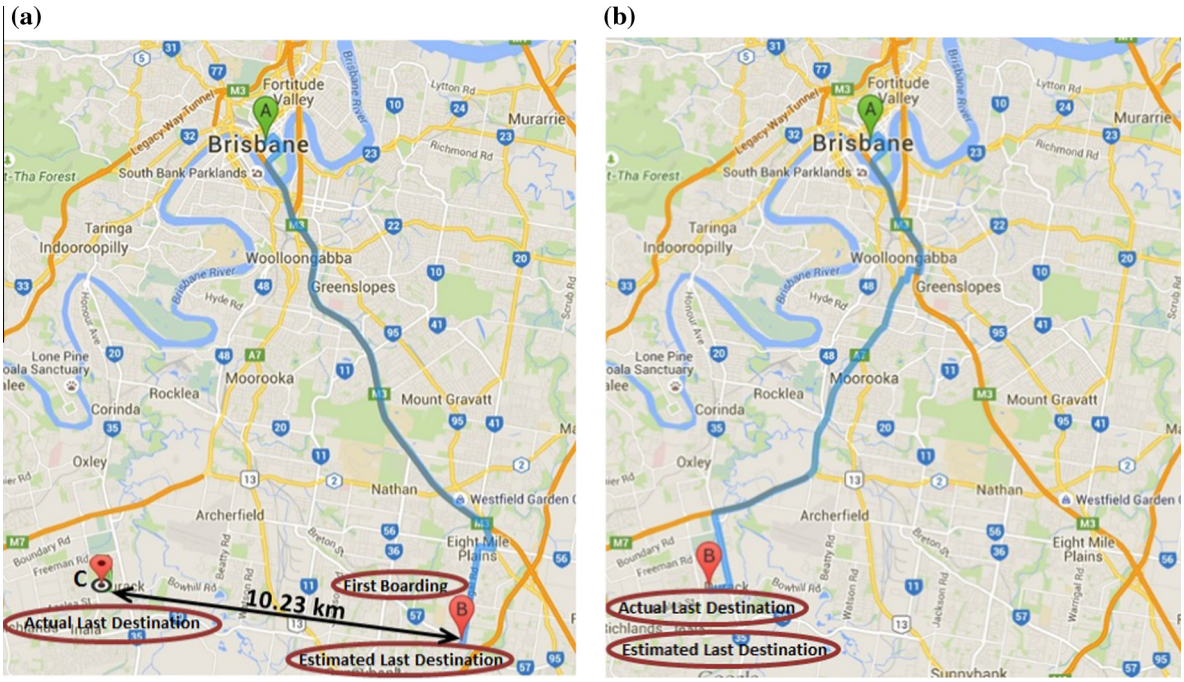

Fig. 7. Example of a long distance between actual and estimated last destination. (a) Original estimation and (b) improved estimation.

4.3. Validation results

To evaluate the accuracy of the existing trip-chaining method and its assumptions, the estimated O–D matrices are compared with the actual O–D matrix for each specific allowable transfer time, with this process previously explained in Section 3. The validation process is performed at both the aggregated level (zone level) and the stop level. The aggregated level validation results are mainly of interest of zonal demand studies. However, the stop level validation provides more details regarding the distances between the actual and estimated trip destinations, which is especially useful in investigating

the trip-chaining method errors. The aggregated level is based on the Brisbane Strategic Transport Model (BSTM) which consists of 1515 zones (BSTM User's Guide, 2009).

4.3.1. Aggregated level (zone level)

Table 3 shows the validation results at different allowable transfer times and different allowable walking distances at the aggregated level (zone level). An estimated O–D trip is considered matching with an actual O–D trip, when the origin and the destination zones match for the two trips. The first origin is considered as the last destination in the applied trip-chaining method for this analysis.

As shown, the matching percentages range between 65.2% and 67.2%. There is very little improvement in the matching results as the allowable walking distance increases (especially when the walking distance increases beyond 800 m). Moreover, the matching results for estimated O–D trips with 30 min allowable transfer time are slightly better compared to O–D trips with 60 and 90 min allowable transfer times. Overall, it can be concluded that the matching percentage for the trip-chaining method at the aggregated level is 67.2% maximum, given different values used for the allowable transfer time and allowable walking distance.

4.3.2. Stop level

At a more detailed level of validation, the direct distance between each estimated destination and its corresponding actual destination is calculated for 30, 60 and 90 min allowable transfer times and 800 m walking distance. To calculate the direct distances (the shortest distance over the earth's surface) between each two points in the algorithm, the Haversine formula was used, as follows:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 * \cos\varphi_2 * \sin^2(\Delta\lambda/2)$$

$$c = 2 * a\tan 2(\sqrt{a}, \sqrt{1-a})$$

$$d = R * c$$

where φ is latitude, λ is longitude and R is the earth's radius (i.e., 6371 km) (Sinnott, 1984).

The results of the stop level validation (matching percentages at different distances) are presented in Fig. 8 (the distance is binned at 50 m intervals). The cumulative matching percentages are also demonstrated in Fig. 8 to better demonstrate the overall accuracy of the exiting trip-chaining method. The figure shows a reasonable degree of matching at short distance intervals (i.e., 57% for 0–50 m and 8% for 50–100 m). Overall, 76% and 84% of matching can be achieved within distance differences of 400 m and 800 m between the estimated and the actual destinations, respectively. Although the distance difference between an estimated destination and its corresponding actual destination may be very short, this does not necessarily mean that they both fall in the same zone.

4.4. Distribution and analysis of errors

A closer look at the results of the trip-chaining estimation with 60 min allowable transfer time and 800 m allowable walking distance (not much improvement is obtained beyond these thresholds, as explained above) reveals the distribution of estimation errors, and thus, some potential points for improvement. The results of the trip-chaining estimation indicate that 16% of the destinations were estimated with a distance more than 800 m from the actual destination. Of 130,274 O–D trips estimated by the trip-chaining method, 127,367 can be matched with the actual O–D trips extracted from the dataset. Of these 127,367 O–D trips, 20,310 (i.e., 16%) have their destinations estimated within a distance of more than 800 m from

Table 3
Estimation matching at different allowable transfer times and distances at the aggregated level.

Allowable walking distance	Actual[a]	Estimated			
	/	400 m	800 m	1000 m	1100 m
Allowable transfer time = 30 min					
Extracted O–D trips	132,735	138,069	135,982	135,700	135,611
Matching	N/A	87,601 (66.0%)	88,799 (66.9%)	88,932 (67.0%)	89,132 (67.2%)
Allowable transfer time = 60 min					
Extracted O–D trips	127,332	132,581	130,274	129,969	129,877
Matching	N/A	83,278 (65.4%)	84,562 (66.4%)	84,725 (66.5%)	84,787 (66.6%)
Allowable transfer time = 90 min					
Extracted O–D trips	123,281	128,487	126,022	125,697	125,600
Matching	N/A	80,332 (65.2%)	81,712 (66.3%)	81,890 (66.4%)	81,902 (66.4%)

[a] For the actual matrices extracted from the original dataset, the distance threshold is not applied.

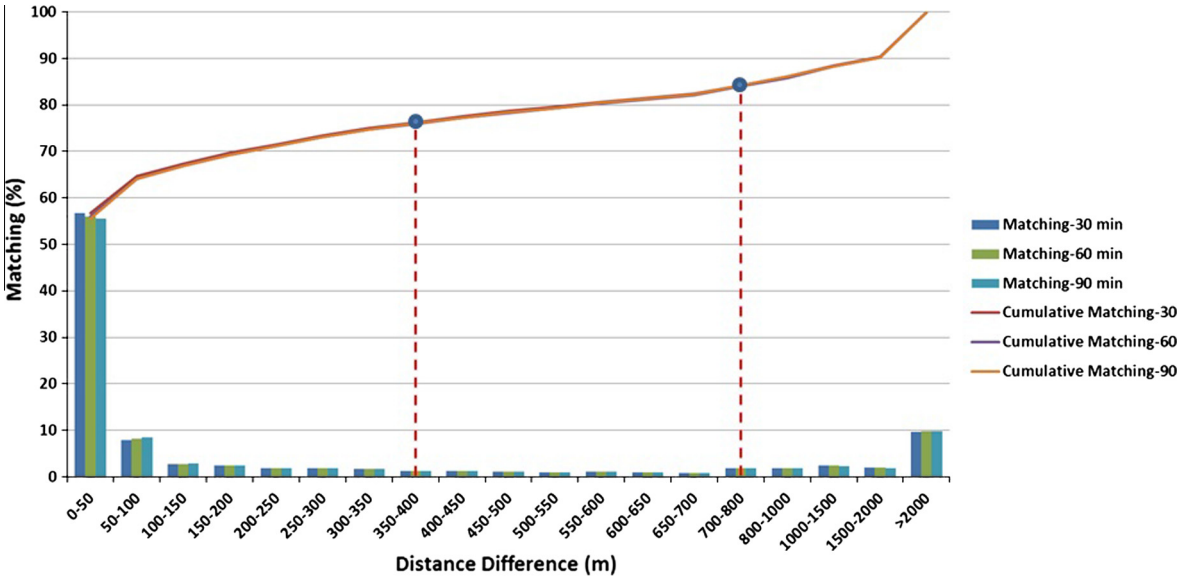

Fig. 8. Distribution of distance between the actual and estimated destination stops given different allowable transfer times.

the actual destinations. Fig. 9 illustrates the distribution of estimation errors, for distances above 800 m between estimated and actual destinations.

Overall, the average and the variance of the distance between the estimated and actual destinations are 806 m and 2327 m respectively. The maximum distance between an estimated destination and its corresponding actual destination, however, is 36,527 m. Given this outcome, it is necessary to have a closer look at the results to find the potential deficiencies of the trip-chaining method, and to develop improvements to the algorithm.

A heuristic investigation of the results and the trip-chaining method and its assumptions reveals two major problems of the method. The first issue is due to the assumption that the last destination is the same as the first origin in a given day. The second issue is due to the time difference between service schedules and actual boarding/alighting times of the public transport services. These problems are explained in more details next.

An in-depth investigation of the results shows that 11.6% (72.6% of erroneously estimated O–D trips) of all matched O–D trips, which their destination estimated more than 800 m far from the actual destinations, are the final transactions of the day for the corresponding passengers. The results show that the average distance is 5059 m and the maximum distance is 36,527 m for these trips. The main reason for this discrepancy is that the trip-chaining method in its original form assumes the last destination to be the first origin of the day for the corresponding passenger. As discussed earlier in this study, some previous studies have addressed this issue by using more realistic assumptions (e.g., Gordon et al., 2013).

Of the remaining 4.4% (27.4% of erroneously estimated O–D trips) of all matched O–D trips, the average distance between the estimated destination and the actual destination is 2990 m. Figs. 10 and 11 show the distribution of estimation errors (in terms of the difference between the sequence number of estimated and actual destination stops) based on different number of sequence order a destination stop may have in the schedule. These figures indicate that when there are more than 15 dif-

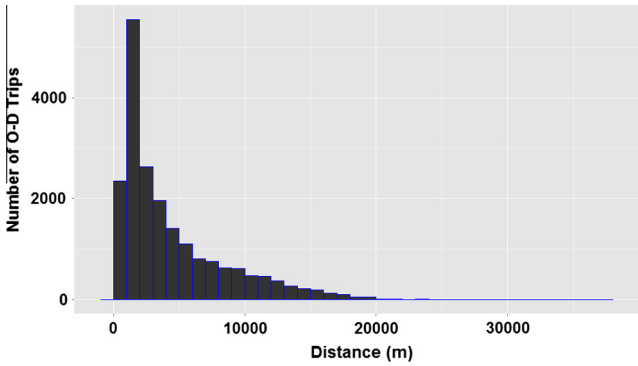

Fig. 9. Distribution of errors for distances above 800 m.

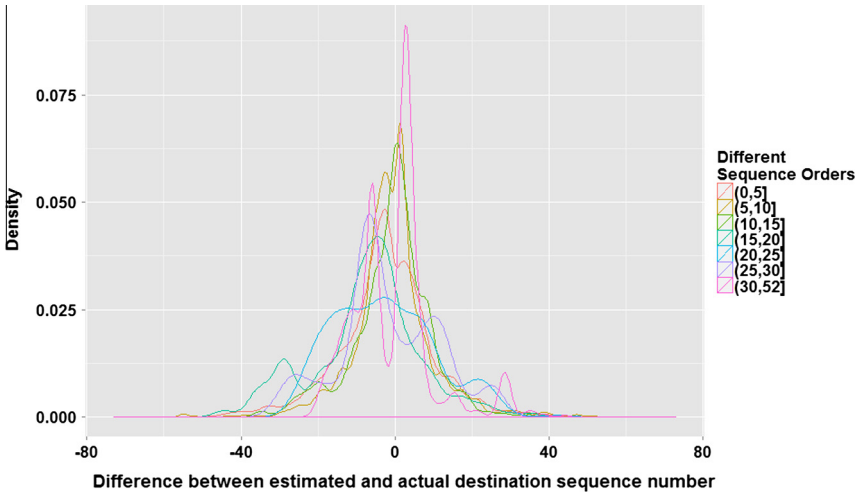

Fig. 10. Distribution of errors based on different number of sequence order of destinations.

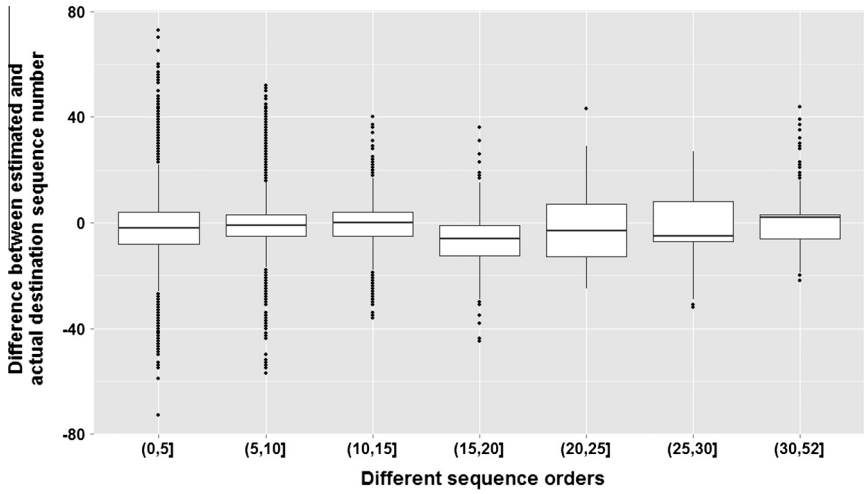

Fig. 11. Boxplot of errors based on different number of sequence order of destinations.

ferent variations in the sequence number of a destination stop at different times of a day, it is more likely to have bigger estimation errors.

A closer look at the erroneously estimated trips reveals at least two issues due to differences between the service schedules and the actual boarding/alighting times of the public transport services:

1. The first issue especially occurs when the transfer time between two consecutive transactions is very close to 60 min which is explained in the following section.
2. The second issue is that the evaluation of erroneously estimated O–D trips shows that the above-mentioned differences between scheduled and actual boarding times can cause the trip-chaining method to use the previous or the next sequence of stops from the schedule instead of the sequence that corresponds to actual transactions. This problem occurs when the time difference between a transaction's actual boarding time and that in the schedule is very different.

Whilst addressing these issues can improve the accuracy of the trip-chaining method (as shown in the next section), there are other issues related to the estimation of the trip destinations that cannot be simply addressed. After applying the proposed improvements, 17,052 O–D trips have their destination estimated with more than 800 m of distance from the corresponding actual destinations. These trips belong to 15,833 unique smart card IDs. Of the respective public transport passengers, 14,644 have only one O–D trip, 1160 have two O–D trips, 28 have three O–D trips and only one has four O–D trips on the day of the study. The lack of at least two O–D trips on a day adversely affects the possibility of improving the accuracy of the trip-chaining estimation.

The distance between an alighting stop and its consecutive boarding stop has also a significant negative impact on the O–D estimation accuracy, as suggested by the literature (e.g., Gordon et al., 2013) too. Fig. 12 shows the distribution of the estimation error for each alighting stop, based on the distance between the actual alighting stop and its consecutive boarding stop. As shown, there is a significant correlation between the errors and the distance between actual alighting stops and the subsequent boarding stops. These results suggest that the algorithm's accuracy significantly declines, when there is a long distance between an alighting stop and its consecutive boarding stop. This is mainly attributed to the use of other modes of transport (especially car) between the two stops.

Figs. 13 and 14 show two examples of the estimated and actual destinations, where the distance among them is more than 800 m. The estimation algorithm estimates the destination as the closest to the next boarding, Figs. 13(a) and 14(a). However, the actual destination is not the same as the estimated destination, Figs. 13(b) and 14(b), and this is due to the passengers' travel behaviour. The actual destination in Fig. 13(b) is at Sunnybank shopping centre where it is most likely that the passenger alighted at this stop to do some shopping or other personal activities.

Fig. 14(b) supports the above finding that some estimation errors are due to the travel behaviour of passengers. Although the distance between the estimated destination and next boarding stop is shorter than the distance between actual destination and next boarding stop Fig. 14(a), the passenger chooses to alight in a stop that it is more convenient for walking (considering the highway and the extra walking) than estimated destination in Fig. 14(b).

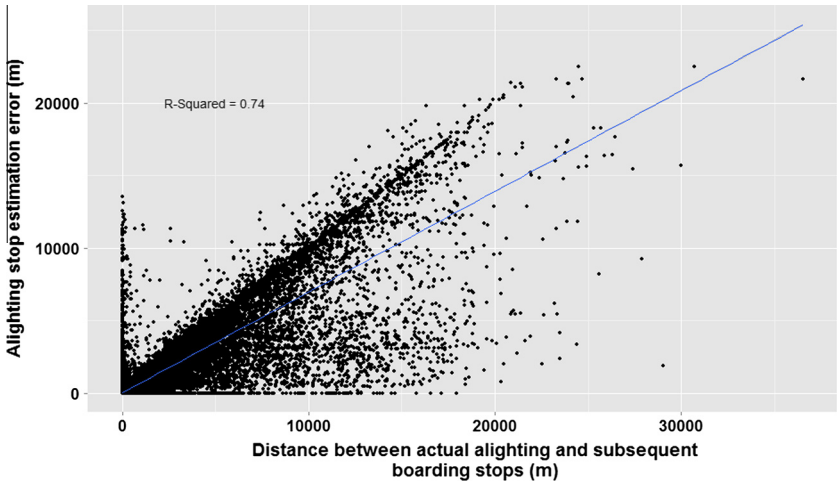

Fig. 12. Distribution of errors based on distance between actual alighting and consecutive boarding stops.

Fig. 13. Example of wrong estimation due to passengers' travel behaviour (i.e., shopping). (a) Estimated destination and (b) actual destination.

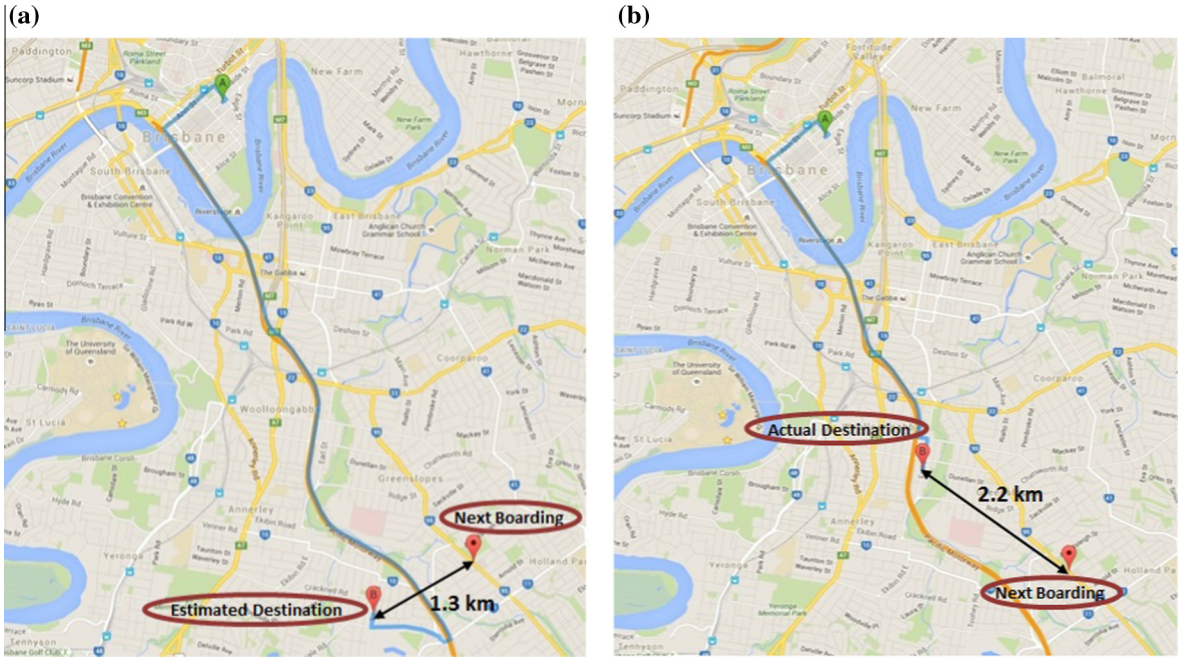

Fig. 14. Example of wrong estimation due to walking obstacles. (a) Estimated destination and (b) actual destination.

5. Improving current trip-chaining algorithm

The last destination assumption is updated in the proposed trip-chaining method by applying the following rules for the last trip-leg of the day, where the improved algorithm corrects the estimation and finds the exact actual last destination, as shown in Figs. 6 and 7:

- Use the last trip's route ID to find the final alighting stop for each smart card holder.
- Find the public transport stop on this route which is the closest to the first boarding stop of the day for the same passenger.
- Choose this stop as the final destination, as shown in Fig. 3.

As discussed in the previous section, the differences between the scheduled and actual boarding times can cause problems in the estimation of O–D trips. Fig. 15 presents an example of two consecutive trip-legs that are chained by the trip-chaining method, as the transfer time between them is less than 60 min based on the service schedules. Fig. 15 shows the boarding and alighting time for each trip-leg and the time (in minutes) between two consecutive trip-legs from the original dataset. However, the actual data indicate that the time lapse between the two trip-legs is a little more than 60 min and

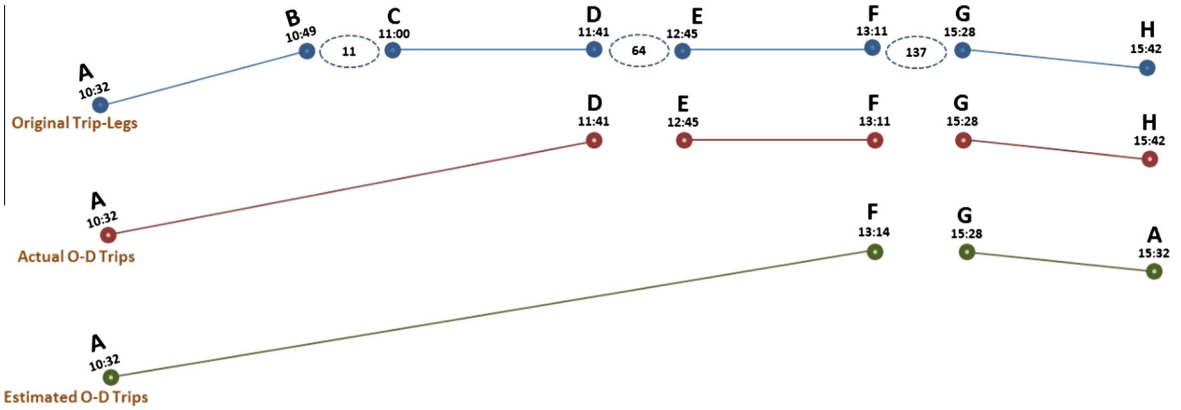

Fig. 15. Example of a transfer between two trip-legs with time difference close to 60 min.

Table 4
Improved algorithm results.

Algorithm[a]	Original trip-chaining method	Improved trip-chaining method: phase I	Improved trip-chaining method: phase II
Zoning level matching (%)	66.4	72.4	72.6
Number of O–D trips with destination at more than 800 m distance from the actual destination	20,310 (16%)	17,957 (14.1%)	17,052 (13.4%)
Average distance between the estimated destination and the actual destination	806 m	621 m	530 m
Maximum distance between the estimated destination and the actual destination	36,527 m	26,499 m	23,575 m

[a] With 800 m allowable transfer distance and 60 min allowable transfer time.

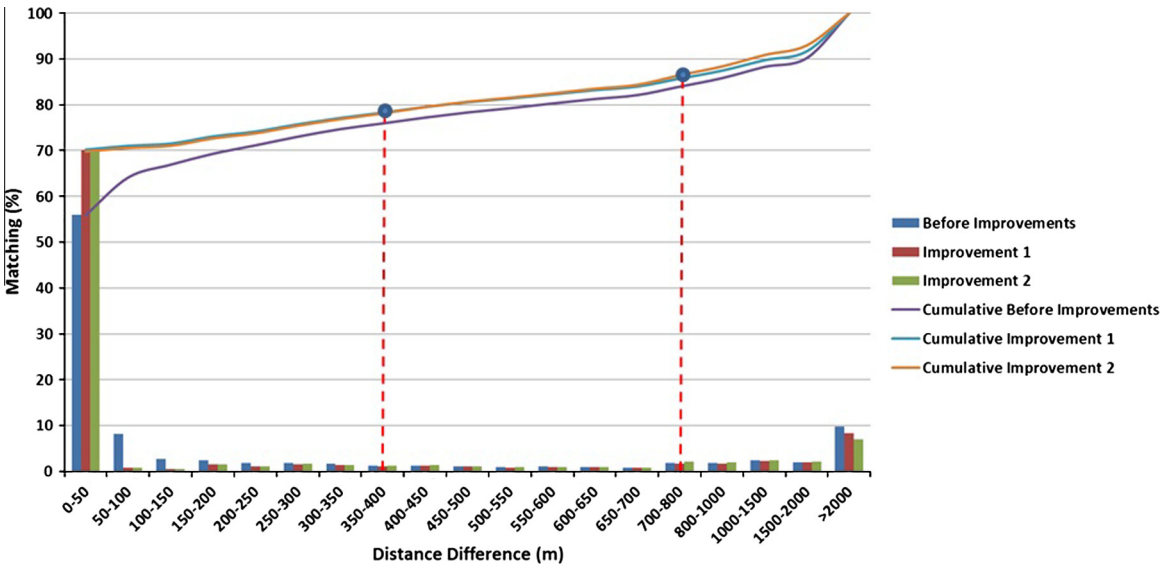

Fig. 16. Distance between actual and estimated destination stops after applying improvements.

thus, they are not chained, when using the actual data for O–D extraction. For the estimated O–D trips, on the other hand, the trip-legs were chained as the service schedule indicates that the time between the two trip-legs is less than 60 min.

To address the issues caused by the inconsistencies between the scheduled and actual boarding times, two improvements are applied to the trip-chaining method. First, the time offsets between the sequence of stops extracted from the schedule for each specific route are applied instead of the scheduled boarding times. Second, the algorithm is updated to estimate the alighting stop for each trip-leg, as shown in Fig. 3.

Table 4 shows the results of the applied improvements to the trip-chaining method compared to the results of the original algorithm. As the first improvement (the final destination of the day is the closest to the first origin of the day) is applied, a significant improvement (72.4%) in the matching percentage is achieved. After applying the second improvement for inconsistencies between the scheduled and actual boarding times, a further very slight improvement in the matching percentage at the zoning level is seen (72.6%). In addition, the average distance between the estimated destination stop and its corresponding actual destination stop decreases from 806 m to 530 m after applying the improvements.

Fig. 16 shows the results of the algorithm improvements, in terms of the distance between the actual destinations and the corresponding estimated destinations. A significant improvement in the matching percentage (70.2%) is achieved at the first 50 m after applying the improvements. However, the cumulative matching is also calculated to show improvements in the accepted level of accuracy compared to that before applying the improvements. Before applying the improvements, 76% matching is achieved within 400 m in comparison to 78.3% matching achieved after applying the improvements. The same can be seen at 800 m where the matching percentages are 84% and 86.5% before and after applying the improvements, respectively. It can be concluded also from Fig. 16 that a significant improvement in the cumulative matching at short distances is achieved compared to a slight improvement at long distances.

6. Conclusions

Recently, smart card fare data have been widely used to generate public transport O–D matrices. Although these data produce more extensive O–D matrices compared to the traditional data sources of public transport trips, the validation of the

estimation method and its assumptions need to be confirmed. Most smart card fare data, however, lack the passengers' alighting details, which hinders a thorough validation of the existing O–D estimation algorithm.

The unique data used in this study enabled us to more accurately evaluate the trip-chaining method and its assumptions in a large, real-world public transport network. The data include both boarding and alighting details of all public transport passengers in the wide region of South East Queensland (SEQ), Australia.

To validate the exiting O–D estimation algorithm and the corresponding trip-chaining method, this study implemented the algorithm and applied it to the SEQ smart card fare data. To resemble the usual smart card fare data settings for which the algorithm has been developed, this study initially excluded the alighting details from the O–D estimation process. Then, the results of the estimation were compared with the actual O–D matrices developed using the complete data (i.e., including the alighting details).

Different allowable transfer times and walking distances were applied to test their impact on the estimated matrices. At the zone level, the match between the estimated O–D matrices and their corresponding actual O–D matrices varied between 65% and 67%, given different values for the algorithm assumptions. The results showed that increasing the allowable walking distance beyond 800 m has no significant impact on the matching percentages. Moreover, the allowable transfer time has a small impact on the matching percentage, as the matrices with 30 min allowable transfer time are slightly more accurate compared to those with 60 and 90 min allowable transfer times. Obviously, the level of matching between the estimated and actual O–D matrices (aggregated accuracy) depends on the accepted level of distance between a given actual stop and the corresponding estimated stop at the destination (i.e., expected level of accuracy).

An in-depth investigation of the existing trip-chaining method revealed some important issues that need to be addressed. Firstly, it is necessary to clean the data and fix the issues caused by the method applied by the smart card data management system to handle special cases. One special case that was investigated by this study is the transactions for which, the corresponding passengers have forgotten to tap off their smart cards at alighting. The current system assumes that passengers have alighted at the last stop of the respective route, in such cases. As such an assumption adversely impacts on the results of the algorithm validation, this study proposed a method to detect and exclude the transactions with a high likelihood of inaccurate alighting details caused by missing data.

Secondly, this study showed that the O–D estimation algorithm's assumption used to infer the last destination causes extensive inaccuracies in the estimation results. The algorithm basically assumes that the final destination of each passenger in a given travel day is the same as the passenger's first origin in the same day. This assumption was found to be true in 66.4% of the cases. As suggested by some studies, this study modified the original algorithm by choosing the closest stop on the route to the first boarding stop of the day as the last destination to evaluate its impact on the estimation accuracy. Accordingly, the match between the estimated and actual destinations was improved to 72.4%, confirming the validity of the revised assumption.

Thirdly, this study found that the differences between actual boarding times and the corresponding schedule boarding times cause some inaccuracies in the estimation results. Hence, the study proposed an improvement to the existing algorithm to use the time offsets between the stops on a given route extracted from the route's schedule, instead of using the scheduled boarding times for the estimation process in the trip-chaining method. The evaluation of the proposed change showed that the average distance between the estimated and the corresponding actual destinations was improved, as the distance was reduced from 806 m to 530 m.

This study used the smart card fare data of South East Queensland, Australia for validating the existing O–D estimation algorithm. The proposed algorithm validation approach is, however, applicable in other regions, subject to data availability. A recent evaluation of the O–D estimation algorithm showed that the algorithm achieves rather similar level of accuracy based on the data used in this study, as well as the smart card fare data collected in Quebec, Canada (He et al., 2015). The lessons learnt in the current validation exercise are applicable to other regions with a smart card management system, even when the alighting details are not recorded. However, the application of the findings to other regions with different public transport network characteristics should be done with caution, as further evaluation of the algorithm is necessary before any generalisation. In future research, it is useful to further evaluate the algorithm with multiple days of smart card fare data, as well as data from other regions.

Acknowledgements

The authors would like to acknowledge the support of the Australian Research Council (grant DE130100205). The authors are grateful to TransLink (the public transport authority of South-East Queensland, Australia) for providing the data for this research.

References

Alfred Chu, K., Chapleau, R., 2008. Enriching archived smart card transaction data for transit demand modeling. Transport. Res. Rec.: J. Transport. Res. Board 2063, 63–72.

Alsger, A., Mesbah, M., Ferreira, L., Safi, H., 2015. Use of smart card fare data to estimate public transport origin–destination matrix. Transport. Res. Rec.: J. Transport. Res. Board 2535, 88–96.

Bagchi, M., White, P., 2004. What role for smart-card data from bus systems? Munic. Eng. 157, 39–46.

Barry, J., Freiner, R., Slavin, H., 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transport. Res. Rec.: J. Transport. Res. Board 2112, 53–61.
Barry, J., Newhouser, R., Rahbee, A., Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. Transport. Res. Rec.: J. Transport. Res. Board 1817, 183–187.
BSTM User's Guide, 2009. BSTM Multi-Modal Model Development, Network and Zoning, Brisbane Strategic Transport Model (BSTM), Queensland Transport & Main Roads, pp. 5–9.
Chow, W., 2014. Evaluating Online Surveys for Public Transit Agencies using a Prompted Recall Approach Master's Dissertation. Massachusetts Institute of Technology.
Cui, A., 2006. Bus Passenger Origin–Destination Matrix Estimation using Automated Data Collection Systems Master's Dissertation. Massachusetts Institute of Technology.
Devillaine, F., Munizaga, M.A., Trépanier, M., 2012. Detection activities of public transport users by analyzing smart card data. Transport. Res. Rec.: J. Transport. Res. Board 2276, 48–55.
Farzin, J.M., 2008. Constructing an automated bus origin–destination matrix using farecard and global positioning system data in São Paulo, Brazil. Transport. Res. Rec.: J. Transport. Res. Board 2072, 30–37.
Gordon, J., Koutsopoulos, H., Wilson, N., Attanucci, J., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transport. Res. Rec.: J. Transport. Res. Board 2343, 17–24.
He, L., Nassir, N., Trépanier, M., Hickman, M., 2015. Validating and Calibrating a Destination Estimation Algorithm for Public Transport Smart Card Fare Collection Systems. No. CIRRELT-2015-52.
Hofmann, M., O'Mahony, M., 2005. Transfer journey identification and analyses from electronic fare collection data. In: Intelligent Transportation Systems, Proceedings IEEE, pp. 34–39.
Kieu, L., Bhaskar, A., Chung, E., 2015. A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. Transport. Res. Part C: Emerg. Technol. 58, 193–207.
Kieu, L., Bhaskar, A., Chung, E., 2013. Mining temporal and spatial travel regularity for transit planning. Australasian Transport Research Forum (ATRF), 36th, Brisbane, Queensland, Australia.
Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: a data fusion approach. Transport. Res. Part C: Emerg. Technol. 46, 179–191.
Langlois, G., Koutsopoulos, H., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transport. Res. Part C: Emerg. Technol. 64, 1–16.
Ma, X., Wu, Y.J., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transport. Res. Part C: Emerg. Technol. 36, 1–12.
Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14, 193–203.
Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smart card data from Santiago, Chile. Transport. Res. Part C: Emerg. Technol. 24, 9–18.
Munizaga, M.A., Palma, C., Mora, P., 2010. Public transport O–D matrix estimation from smart card payment system data. In: 12th World Conference on Transport Research, Lisbon, Paper No. 2988.
Munizaga, M.A., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behaviour estimated from smartcard data. Transport. Res. Part C: Emerg. Technol. 44, 70–79.
Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M., 2011. Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. Transport. Res. Rec.: J. Transport. Res. Board 2263, 140–150.
Pelletier, M., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transport. Res. Part C: Emerg. Technol. 19, 557–568.
Robinson, S., Narayanan, B., Toh, N., Pereira, F., 2014. Methods for pre-processing smartcard data to improve data quality. Transport. Res. Part C: Emerg. Technol. 49, 43–58.
Sinnott, R.W., 1984. Virtues of the haversine. Sky Telescope 68, 159.
Wang, W., 2010. Bus Passenger Origin–Destination Estimation and Travel Behaviour using Automated Data Collection Systems in London, UK PhD Dissertation. Massachusetts Institute of Technology.
Zhao, J., Rahbee, A., Wilson, N., 2007. Estimating a rail passenger trip origin–destination matrix using automatic data collection systems. Comput. Aided Civ. Infrastruct. Eng. 22, 376–387.