Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Validating the Origin−Destination estimation algorithm for bus systems using smart card data

Mona Mosallanejad[a*], Sekhar Somenahalli[b], Akshay Vij[c], David Mills[d]

[a] *PhD Candidate, University of South Australia,* Adelaide 5095, South Australia
[b] *Senior Lecturer, University of South Australia,* Adelaide 5095, South Australia
[c] *Senior Research Fellow, University of South Australia,* Adelaide 5095, South Australia
[d] *Project Officer, Department of Planning, Transport & Infrastructure,* Adelaide 5095, South Australia

Abstract

Using smart card information provides transit planners with access to a large source of spatial-temporal data. Based on Adelaide's metroCARD data, this study used a new heuristic algorithm to estimate an accurate public transport Origin−Destination (OD) matrix, using SQL software and the trip chain model. Unlike other cities, Adelaide passengers do not swipe metroCARD when alighting and passenger's destination information is not stored in their metroCARD. Hence, this study used a number of assumptions to accurately derive their destination information. In case of transfer journeys, the methodology derived the passengers' alighting points using the Euclidian distance between the alighting point and the next boarding stop. It did this by making appropriate assumptions including the minimum walking distance.

This paper developed an appropriate validation technique to verify the accuracy of the estimated Origin−Destination (OD) matrix. A new survey method was developed in which volunteers provided their metroCARD information, then based on the trip chain model, the data was analysed. By interviewing the volunteers and comparing the two results, the model's accuracy has been validated. Results elaborate that the method used is 98% accurate and can be utilised elsewhere. Accurately estimating public transport Origin−Destination (OD) matrix will facilitate transit planners in route rationalisations; which will lead to higher public transport patronage.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Origin−Destination matrix; Public Transport; Trip Chain Model; Smart card; Validation

* Corresponding author. Tel.: +61449230448.
E-mail address: mona.mosallanejad@mymail.unisa.edu.au

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

1. INTRODUCTION

Transport planners design transit facilities to accommodate and influence people's travel behaviour regarding the use of public transport instead of private vehicles. An increasing number of vehicles in metropolitan cities causes problems due to traffic congestion, air pollution and fuel consumption (Ma et al. 2013). The alternative public transport system needs to be more reliable so that its patronage is increased. The demand for public transport depends on various factors such as travel time, weather and service breakdown (Morency et al. 2007) and these can be estimated from the Origin−Destination (OD) matrix. An Origin−Destination (OD) matrix identifies the boarding and alighting stations of individual passengers and then aggregates this data within a population. It can then identify travel patterns for a specific route over a specific period of time.

The demand can be more accurately calculated with the development of Information and Telecommunication Technology, which is changing the amount, type, and quality of data available to planners and managers. Public transport agencies increasingly adopt the use of Automatic Data Collection Systems (ADCS). A large amount of boarding data are being collected by the transport agencies on an ongoing basis. In comparison with traditional transport surveys, which are usually time-consuming, expensive and only suited to identify a snapshot, smart card data can be used to examine the whole network on a regular basis which is ideal for transport planners. These data sets can be employed to infer accurate estimates of passenger Origin− Destination (OD) patterns. Understanding public transport passenger Origin-Destination (OD) flows are crucial to improving the planning and operation of transit systems.

In this research, a one-month (May 2017) dataset was used. This data was provided by the Department of Planning Transport and Infrastructure (DPTI) in Adelaide, South Australia. A new methodology has been developed, using SQL software based on the trip chain model, to create an Origin−Destination (OD) matrix for Adelaide's bus system users and as a result, estimate the demand for the system. Adelaide was chosen for this study because unlike other cities, commuters only scan their smart card upon boarding, not alighting. This means that the algorithm must be generic and therefore applicable elsewhere. For this reason, it is important to develop appropriate validation techniques so that the estimated Origin− Destination (OD) matrices can be accurately verified. In this paper, a new survey method was developed and conducted by using a sample from the main data set. Comparing the results from the trip chain model and interviewing the volunteers served to validate the algorithm for Origin−Destination (OD) matrix estimation.

Developing approaches for estimating accurate Origin−Destination (OD) matrices from smart card data is critical for transit planners (Alsger et al. 2015). Having knowledge of travel demand will facilitate the design of appropriate public transport routes, which leads to the optimisation of schedules. In turn, this will enhance public transport patronage with the potential of improving the public transport system's performance.

2. ORIGIN−DESTINATION ESTIMATION METHODS

The primary function of the smart card is to collect a fare, but it can also be utilised for finding the travel pattern. Usually, smart card data does not directly provide the information which is required for planners (Kurauchi & Schmöcker 2016). Many procedures have examined the estimation of Origin−Destination (OD) matrices based on smart card data for public transport since the 19th century. These methodologies may vary depending on: firstly, the availability of passengers' trip data; and secondly, the duration which can vary according to previous studies, from 1 week to 1 year. Before the evolution of new technologies for collecting data, most studies for inferring travel patterns were based on household and on-board survey data. Survey data was generated using alternative methods to estimate an Origin−Destination (OD) matrix. These methods are:

- Non-iterative algorithm (Tsygalnitsky 1977)
- Fluid mechanics: for example, estimating an Origin−Destination (OD) matrix for bus routes (Tsygalnitsky 1977)
- Passenger on-off counts and checker records at each stop (Simon & Furth 1985)
- Constrained least squares and Fratar model: this growth factor model estimated the bus boarding matrix by counting the number of passengers (Gur & Ben-Shabat 1997)
- Fuzzy theory (Markus et al. 2000)

The introduction of the Automatic Fare Collection system made it possible to develop further methods for estimating an Origin−Destination (OD) matrix. Initially, a new methodology was proposed to compare Origin−Destination (OD) trips versus the number of passengers (Barry et al. 2002). Since then, researchers have explored the potential of smart card data to infer trip rates, turnover rates and travel behaviour to improve planning aims (Bagchi & White 2005; Utsunomiya, Attanucci & Wilson 2006). The following methods have been developed based on Automatic Data Collection System for Origin−Destination matrix (OD) estimation:

- Furness model (Lianfu et al. 2007)
- Fusion approaches (Kusakabe & Asakura 2011)
- Multiple linear regression (Kalaanidhi & Gunasekaran 2013)
- Iterative proportional fitting (Cui 2006; Gordon et al. 2013; Horváth, Horváth & Gaál 2014; Li, Yuwei 2007)
- Maximum likelihood estimation (Cui 2006; Ickowicz & Sparks 2015; Li, Yuwei 2007)
- Inferring the alighting station via the straightforward algorithm and iterative method (Chapleau et al. 2008; Seaborn et al. 2009; Zhao 2004; Zhao et al. 2007)
- Trip chain model (Ali, Kim & Lee 2015; Alsger et al. 2018; Munizaga, MA & Palma 2012; Nassir et al. 2011; Wang 2010)

Time-dependent Origin−Destination (OD) matrix using smart card can also improve public transport planning. This matrix was estimated from the passenger counts system both at boarding and alighting stations, and based on the forecasting method linking boarding and alighting data (Horváth 2012). This method also included transfer time and was validated based on an application in the Hungarian capital city. To verify the impacts of mode choice, a new methodology was developed for forecasting individual passengers' travel behaviours. The method was based on an accessibility index and evaluated transit amenities by utilising the composite impedance gravity method, in order to estimate the impact of travel time for both in-vehicle and off-vehicle. The estimation was based on multiple linear regression (Kalaanidhi & Gunasekaran 2013). Yang and Jun (2018), developed a new methodology to visualise the travel patterns of transit commuters in Seoul, South Korea, by calculating the trajectory and using Carto to create a map.

The moth-flamed optimisation (MFO) algorithm is a new population-based metaheuristic algorithm that investigates the celestial navigation of moths to estimate the Origin−Destination (OD) matrix (Heidari, Moayedi & Abbaspour 2017). Li, Tian et al. (2018), compared different studies of using smart card information to examine transit passengers' travel behaviours and provided a comprehensive review of them.

The trip chain model is a recently devised method for determining travel patterns and travel behaviours. Trip chaining model was utilised by Barry et al. (2002) for estimating destinations for the first time (Li, Tian et al. 2018). Although there is not an exact definition for the trip chain, a basic definition is that each chain consists of one or more stops to the next destination and a trip chain is specified based on the number of stops to the next destination. The algorithm which we will use here to estimate the alighting stop is based on the trip chain model (Alsger et al. 2016; Langlois et al. 2016; Li, Tian et al. 2018), which is a recent method for determining behavioural attributes of passengers' trips.

Smart card data can also be employed to infer travel behaviour (Langlois, et al. 2016). Initially, some researchers believed that smart card data could not provide all information such as trip purposes (M Bagchi & White 2003). More recently researchers have used such data to generate the required information. Kusakabe and Asakura (2011) used the fusion approach to estimate the purpose for a trip based on two criteria: 1) arrival time at the station; and 2) duration of stay between alighting and next boarding. Another methodology based on the trip chain model was developed a few years go to estimate trip purposes (Lee & Hickman 2014).

The literature review identified following gaps for estimating public transport OD matrices from smart card data.

- The accuracy of OD matrix estimation by using a trip chain model is still debatable and require further investigation.
- Distinguishing transfer from the activity is an issue for most researchers, so new assumptions should be considered and need to validate them.
- The validation of assumptions for the trip chaining model needs a detailed investigation.

3. DATA STRUCTURE

When using a smart card, the requirement is to tap, swipe or wave the card at the station, stop or vehicle. Regarding the flat fare policy and some zonal fare policies, commuters should tap once after boarding, and it records only a single transaction. However, in some cities, the exit reader is available as well, and the fare policy is based on distance or zone. For each trip, two records are available both for boarding and alighting (Kurauchi & Schmöcker 2016).

The data used in this paper is based on the "metroCARD" database which is used in Adelaide and is collected by the Department of Planning Transport and Infrastructure (DPTI) for a specific period - May 2017. Each metroCARD data contains spatial and temporal information. In Adelaide where the flat fare policy operates, commuters validate their cards when they board a public transport vehicle but not for alighting. Three modes of public transport are available, these being bus, train and tram. The information for each smart card transaction contains card identification, time, date, transport mode used, fare type, stop code, stop label, route code and validation type (refer to Table 1). Seven types of tickets are available in Adelaide: 2-Section fare, Daytrip (1-day pass), SV (metroCARD), 28 Day pass, Visit, Other (Miscellaneous metroCARD-based validation products) and Tickets (magnetic ticket products). In Adelaide when passengers swipe their card and pay an initial transaction, the fare is valid for 2 hours. Passengers can also utilise any other public transport transfer services modes within this two hour timeframe without incurring any further cost.

Table 1. Individual metroCARD information

Media code	Fare type	Transport mode	Date & time	Stop code	Latitude	Longitude	Route code	Direction
807***CB	SV	4	2017-05-01 09:49:35	8089	-34.979759	138.525912	Tram	1
6AD***07	SV	1	2017-05-01 10:02:20	2658	-34.890404	138.585119	235	1
94E***FB	TICKETS	1	2017-05-01 10:39:15	3351	-34.924343	138.598468	271	1
94E***FB	TICKETS	1	2017-05-01 10:43:01	3335	-34.924022	138.604979	271	1
94E***FB	TICKETS	1	2017-05-03 10:43:05	3335	-34.924022	138.604979	271	1
11C***89	28DAY	1	2017-05-05 10:46:32	3285	-34.920343	138.607313	271	1
707***27	OTHER	1	2017-05-01 11:04:05	2072	-34.870071	138.638452	271	1
5AE***CC	SV	5	2017-05-03 18:16:46	40001	-34.831641	138.695056	GAW	0

Note: Transport mode: 1 = Bus, 4 = Tram, 5 = Train

There are some deviations from the one-swipe rule: railway stations in Adelaide operate under a closed system, and swiping is required for both boarding and alighting, and various systemic and user issues mean that transfers between the train and other modes cannot be estimated directly from the metroCARD. In addition, there is a free tram zone in Adelaide where passengers do not need to swipe their cards; this means that the tram boarding point is not available. Given these limitations, this study focuses on bus users.

4. METHODOLOGY FOR ESTIMATING THE ORIGIN–DESTINATION MATRIX

One of the most common methods for estimating the destination is the trip chain model. As mentioned previously, each smart card can provide the boarding location and boarding time of each trip. The problem is differentiating the destination from the alighting point, because the alighting stop is not always the destination and commuters may alight to transfer to another bus or another mode of public transport. The trip chain model assumes the destination of one trip is located in the vicinity of the next boarding within an acceptable walking distance. Some assumptions that were considered in this algorithm are:

- The initial boarding location of a trip leg is the 'origin'.
- A passenger's alighting point is assumed to be within walking distance of the next boarding stop

- Passengers return to the place where they first boarded that day, or to some other nearby station.
- Commuters take the first available service after arriving at a boarding place.
- Each smart card is used by a single commuter and cannot be used by multiple passengers.
- Commuters who use the public transport system do not use any other mode of transport on that same day.

The description of some of the terms used in this paper are listed below.

- Media code: the unique identifier for each metroCARD in Adelaide
- Time threshold: the waiting time between two consecutive transactions.
- Trip leg: the trip for an individual commuter between boarding and alighting stops.
- Walking distance: the maximum distance between two consecutive trip legs that commuters walk to transfer to another public transport service.
- Trip ID: identifies an ID for each trip, which is unique for every service.
- Route ID: identifies a unique ID for each route.
- Stop ID: identifies a unique ID for an individual stop or station entrance; a multiple route ID may use the same stop.
- Service ID: contains a unique ID of the available service for one or more routes.
- Block ID: identifies the block to which a specific trip belongs. A block can consist of a single trip or more for the same vehicle.

4.1. Estimating the alighting stop

A new heuristic algorithm is used to estimate stop-level origins and destinations, based on the boarding transactions in the metroCARD datasets. The algorithm used to estimate the alighting stop is shown in Fig. 1. This flowchart was used for finding the alighting stop and not the destination because not all alighting stops are the destination of a trip leg. For OD estimation, some terms like trip ID and service ID were extracted from the Google Transit Feed Specification (GTFS) dataset. In the database provided by DPTI, the stop ID for each metroCARD is different from the stop code in GTFS data, and these need to be matched. Once that was done, the data based on the transaction time was sorted, and a metroCARD ID was selected. Based on the trip chain model, the subsequent transaction in each trip leg is a key point for inferring the alighting stop. By considering the following transaction of a metroCARD (the next boarding), the alighting stop was estimated by calculating the minimum Euclidian distance. Based on the algorithm, for each transaction, the trip ID, service ID and block ID from ‘stop_times.txt’ in GTFS data were selected. These criteria are unique for each service for various modes of public transport: for example, a bus which departs at a specific time from its origin has its own trip ID, service ID and block ID, which may be different from the subsequent bus. By matching the time of each transaction in metroCARD data with the arrival and departure time in GTFS data, and by considering the day that the commuter swiped the card, a trip ID is chosen. If there is no trip ID relevant to the metroCARD data, an interval of five minutes was considered for selecting the trip ID. If in this period no trip ID was selected, then the next available trip ID was chosen for the algorithm by considering a delay.

In Adelaide, some buses change their route ID in the middle of the route for some specific hours, especially before entering the central business district (CBD). This is known as a thru-linking route. For finding the thru-linking route for these services, the first stage is to infer the stop at which the route ID changed to another one: in other words, by identifying the last stop for the current route ID, the changing location can be inferred. To find the last stop, the data were sorted based on arrival time. Then, based on the trip ID which was selected for the transaction and the existing route ID, the last stop and its arrival time were chosen. By entering the chosen stop and relevant time in the timetable database, the available route could be selected. Routes with the same service ID and block ID could be chosen and labelled as thru-link routes.

In the next step, the Euclidian distance was calculated between all stops along the current route and the following transaction (next boarding). By using the stop code and route ID, subsequent stops based on distance could be identified. The latitude and longitude of these stops were labelled X0, Y0, and the latitude and longitude of the successive transaction (next boarding) were labelled X1, Y1. Based on the equation 1 the Euclidean distance needs to be calculated.

$$\text{Euclidean distance} = \sqrt{(X1 - X0)^2 + (Y1 - Y0)^2} \quad (1)$$

For the next stage, the stop ID with minimum Euclidean distance is selected. The distance was compared with the maximum acceptable walking distance of 1000 m, derived for Adelaide through sensitivity analysis; this distance will vary from city to city. If the distance to the selected stop is less than the walking distance, then it was labelled 'alighting stop'; otherwise, the alighting stop was labelled 'cannot be inferred'.

In some cases, the alighting stop could not be inferred if the distance to the next boarding was higher than the acceptable walking distance. Manual analysis showed that the GPS system incorrectly selected stops in certain situations due to their proximity to a stop on the other side of the road. If the alighting stop could not be inferred, then the opposite stop was considered in the algorithm to check whether the alighting stop could be estimated or not. By adding this stage to the algorithm, an additional 5% of the alighting stops could be estimated.
Sometimes a commuter uses different modes of transit during the day; for example if returning with colleagues or friends with private cars. In this instance, the alighting stop could not be inferred due to lack of information, and this made it impossible to track some trips during the full day. If the alighting stop could not be inferred in previous sections, then the travel pattern was investigated at this point, and multi-day transactions for individual media codes were used to estimate the destination. It was possible to check if the commuter on subsequent days uses public transport from the same stop or zone; and if on other days of the week the alighting stop could be inferred and the travel pattern was consistent, the estimated stop was taken as the alighting stop for the day which could not be inferred before, and the alighting stop was replaced.

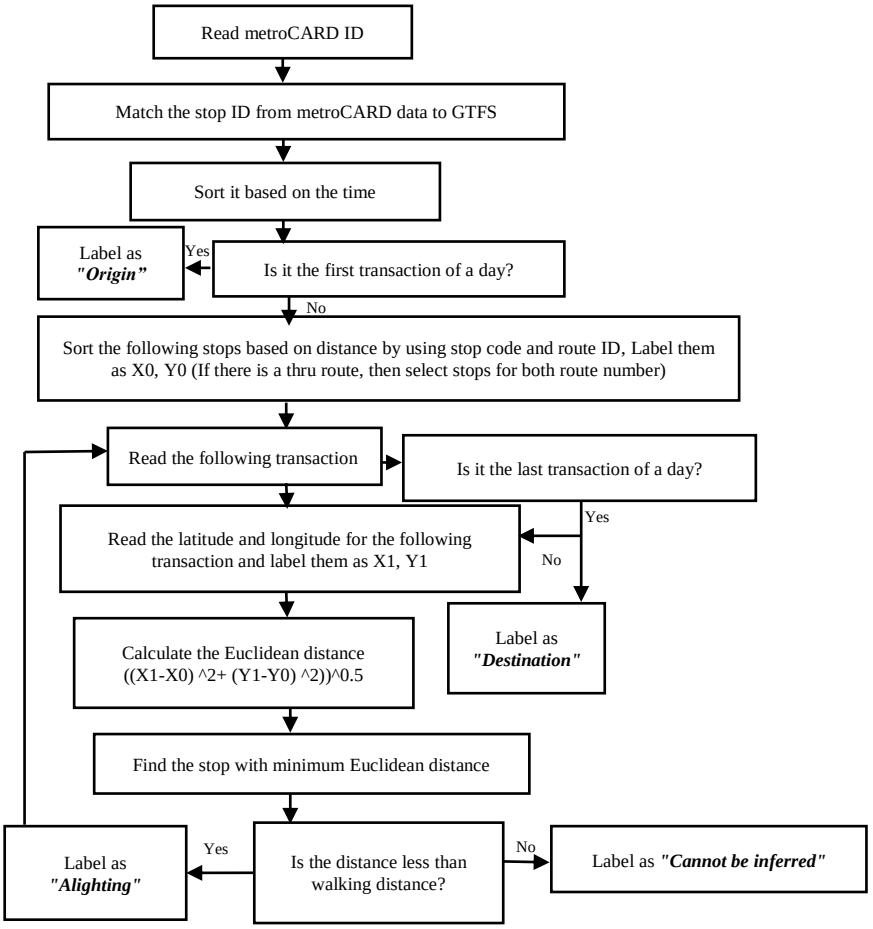

Fig. 1. Estimation of alighting stop

Fig. 2 depicts an example of a trip chain model for inferring a passenger's alighting stop. If a commuter starts the

trip at stop i on route 1 and the next transaction is at stop j on route 3; then the alighting point can be estimated. As mentioned earlier, some routes in Adelaide change their route ID, but passengers are not required to revalidate their cards. For example, if route 1 changes to route 2 as shown in Fig. 2 (a thru-linking route), the Euclidian distance is used to find the alighting stop; all distances from stops in route 1 and route 2 to stop j, ED1, ED2, ED3 and ED4, should be calculated (see Fig. 2) and the stop with the minimum Euclidian distance selected as the alighting stop: this should be less than the acceptable walking distance. For instance, if the first boarding is at stop i and the second boarding at stop j, then the commuter alighted at stop m in route 2 (the thru-linking route for route 1). Also, stop i is the origin of the first trip leg because it is the first transaction of a day. If the next transaction will be k, this is the last transaction of a day and based on the assumptions the destination should be near the origin of a day i. By using the minimum Euclidian distance from stop k to i by route 4, the alighting stop will be i which is the last destination of a day, and there is no other transaction afterwards.

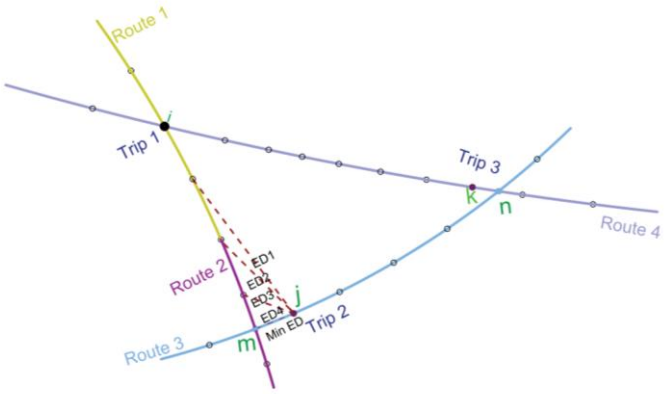

Fig. 2. An example of a trip chain for inferring the alighting stop

4.2. Estimating the alighting time

For estimating the alighting time, a trip ID which is related to each service needs to be determined. Each service from the origin has a unique trip identifier, so by selecting the trip ID for the boarding transaction, it is possible to estimate the alighting time as well. The trip ID is selected, using route ID and stop ID for boarding transaction based on the boarding time and date. If the trip ID is in 5-minute intervals, then alighting time is selected based on alighting stop and trip ID. If this is later than the boarding time, then it is labelled as alighting time.

4.3. Destination estimation

Based on the below-mentioned algorithm, after finding the alighting stop, four categories should be considered for inferring the destination: First, it should be checked if the data is related to the last transaction of a day or not; if yes it means the alighting stop which was inferred should be labelled as a destination. If the alighting stop for the last transaction of a day could not be inferred in the previous step, the destination cannot be estimated. Next, it should be checked if a commuter has used the same route twice or used the parallel route to reach the destination. In this instance, it can be concluded this is an activity since there is no necessity for anyone to alight from a direct route and take the same or parallel route again. It means the alighting stop is the destination point.

Time threshold between two subsequent transactions can be considered another criterion for finding the destination of each trip leg. If the time threshold is less than 20 minutes, then the commuter transfer to another bus and the inferred

stop is just the alighting point. For transactions with a time threshold more than 20 minutes and less than an hour, the label of short activity should be considered, and the alighting stop is the destination. For all the transactions with a time threshold, i.e. more than one hour which is related to long activity, the label of destination should be assigned.

The last criterion for investigating the destination stop is the distance between the boarding stop and the subsequent alighting stop. If this value is less than 400 meters, then one should label the alighting stop as the destination (Nassir, Hickman & Ma 2015).

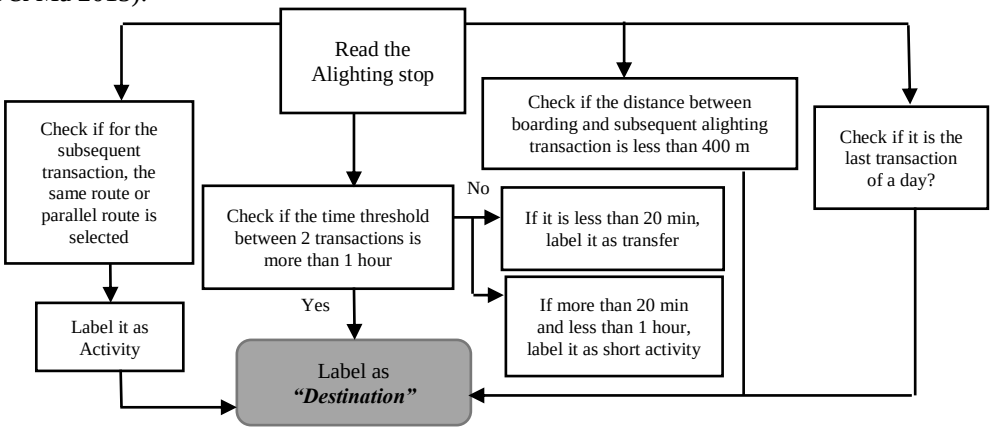

Fig. 3. Estimation of the destination stop

4.4. Sensitivity analysis

A sensitivity analysis is a method to show the uncertainty on the output of a mathematical model due to the uncertainty in its input. The Geoffrey E. Havers (GEH) statistic is a formula which can be used to evaluate the accuracy level of a model and it is widely used for traffic engineering and traffic modelling. The GEH is applied to every single pair of the estimated OD matrix and if GEH is less than five, indicating a good fit (Hollander & Liu 2008). The GEH formula is calculated based on the below equation:

$$\mathrm{GEH} = \sqrt{\frac{2\,(x_i - y_i)^2}{x_i + y_i}} \qquad (2)$$

Where x_i is sample measurement and y_i is reference measurement

- Walking distance

In this section, sensitivity analysis is done for different walking distance, and the relevant errors for each pair are calculated. For this purpose, 5 scenarios for different walking distances were introduced and the survey database was analysed based on trip chain model. The generated OD matrix based on survey database is considered as a reference and the result from the interview is considered as a sample measurement. The relationship between the errors in each estimated matrix and its related walking distance is demonstrated in Fig 4. (Chu et al. 2015).

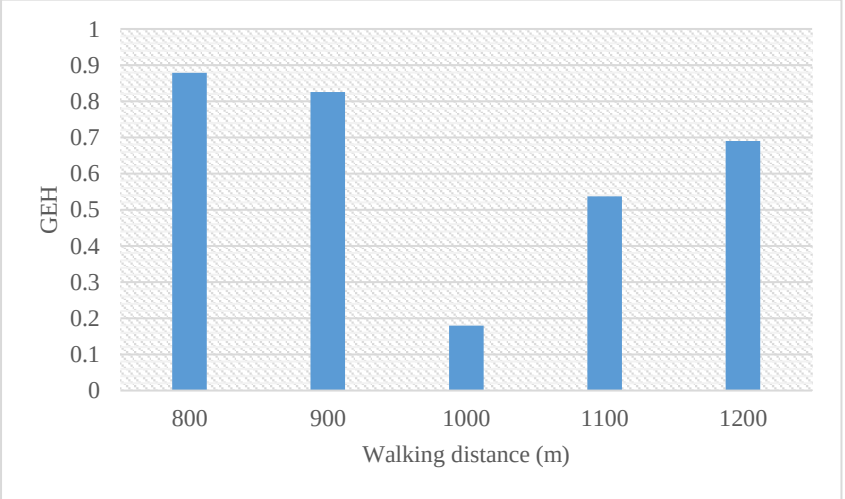

Fig. 4. Relative errors in OD estimation for different walking distance

Though GEH less than one indicates a good fit, walking distance threshold of 1000 metres shows significantly lower GEH and hence this distance was chosen as a threshold distance.

- Transfer time

Commuters tend to reach a destination without transferring to another bus or another mode of the transit system, but in situations where there are no cross-suburban routes, they must alight and use another bus or mode. Smart cards are unable to collect transfer information and some other factors which can affect passenger's travel behavior. So some assumptions need to be considered to investigate the transfer locations. As mentioned, it is assumed that if the time threshold between alighting stop and next boarding is less than 20 minutes, then transfer occurred in between. In this section, sensitivity analysis for confirming this assumption elaborated.
For undertaking sensitivity analyses, various scenarios were tested using different assumptions and algorithm was executed.
It means that for each time threshold, the data from the survey database analysed and the result compared with the interview. The generated OD matrix based on survey database is considered as a reference, and the result from the interview is considered as a sample measurement. The relationship between the errors in each estimated matrix and its related time threshold is demonstrated in Fig. 5.
Though all time thresholds are below the acceptable errors, 20-minute threshold was adopted for further analysis due to its low GEH value.

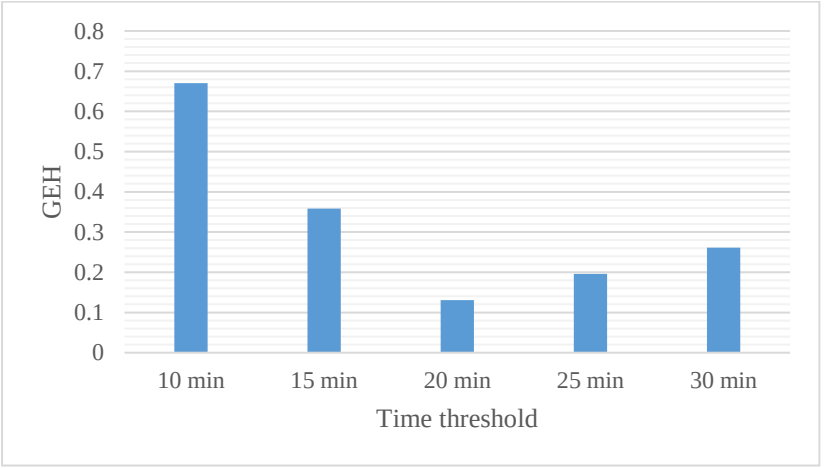

Fig. 5. Relative GEH for the different time threshold

Also Fig. 6 shows the number of transfers from both trip chain model and interview for different time thresholds. This analysis also confirmed that the time threshold adopted in this study are appropriate.

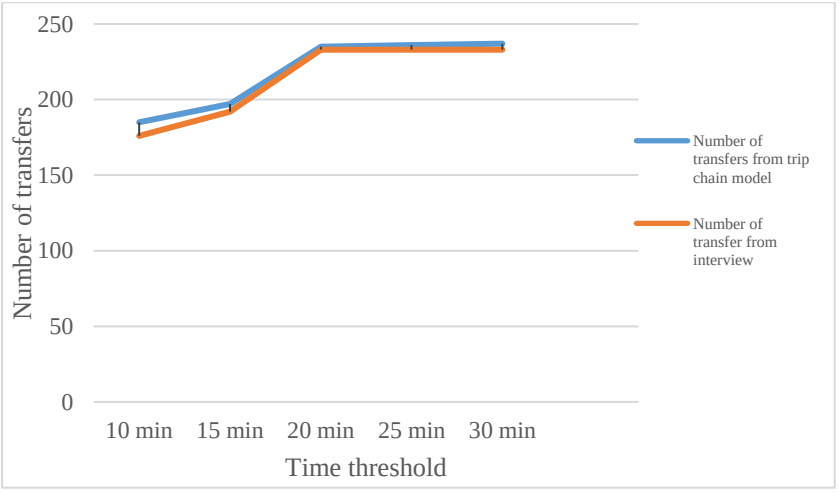

Fig. 6. Number of transfers from trip chain model and interview for the various time thresholds

5. VALIDATION METHOD

The best way to examine the model's accuracy is to validate the results. This was done through a survey in which fifteen volunteers were recruited.

Earlier studies validated their results by undertaking a household survey or utilising data from a closed system where both boarding and alighting statistics are available. For example, Barry, Freimer and Slavin (2009) successfully validated these assumptions by surveying passenger counts at the exit and entrance of the subway station. Later, Devillaine et al. (2012) validated their findings from the smart card by undertaking a travel survey, where the users' smart card IDs were recorded. Munizaga, M et al. (2014) validated the assumptions which were used in the trip chain

model by using travel survey of a small group of volunteers. Their validation has correctly identified the trips for 90% of the cases. In Brisbane where the 'tap on, tap off' system is used, the data includes both boarding and alighting information. Therefore, the trip chain model assumptions are validated by utilising the go card dataset (Alsger 2016; He et al. 2015).

5.1. Estimating the sample size for a survey

For undertaking a survey the most effective criterion is considering the reasonable sample size. Otherwise, the result will be biased. While increasing the sample size will lead to fewer errors, bigger sample size is more costly. Estimating the sample size is critical for obtaining accurate results, and it is necessary to investigate how much an increase in the sample size will lead to justifiable results with fewer errors. In the context of survey objectives, two rationales can be considered. The first one is estimating the certain population parameters and the second is to test the statistical hypothesis. In this paper, the objective of the survey is the first rationale which is related to population parameters. For estimating the required sample size for population parameters, some factors should be taken into account (Richardson et al. 1995): firstly, the variability of parameters over the population; secondly, the required degree of precision; and thirdly, population size.

Some approaches that considered estimating the sample size, for instance, Ceder (2016), employed a procedure involving a survey for Origin−Destination (OD) matrix. They did this by taking into account the percentage of passengers who travel between specific origins and destinations, the population of each suburb and the accuracy of each cell in Origin−Destination (OD) matrices. Previous studies' sample sizes vary as follows: 37 volunteers (Ebadi & Kang 2016), 53 (Munizaga, M et al. 2014), 306 (Lee & Hickman 2014) and 8000 households (Seaborn et al.2009).

Other approaches which are available for the discrete variable to estimate sample size and will be used in this paper were based on a random sample method. In this dataset which includes discrete variables, the standard error for estimating a proportion p is given in equation 3 (Richardson et al. 1995).

$$s.e.(p) = \sqrt{!\frac{N-n}{n}!\,!\frac{p(1-p)}{n}!} \qquad (3)$$

n: the sample size
N: population

By assuming 95% of confidence with the result, the sample size should be estimated based on the population of the whole dataset. In the present study, only the number of commuters who utilised buses as their mode of transport is considered, the number of transactions per day for these passengers can be considered as "N" which is 139187 transactions. Importantly, it should be noted that for calculating the population of the whole dataset, the first week of May 2017 is considered and a day with more transactions (Wednesday) is selected to obtain a more accurate result. The number of transactions which is required for the confidence of 95 per cent is 105 and in this paper 407 transactions are used to validate the result.

5.2. Survey

In this research, after obtaining ethics approval, a survey was conducted by randomly recruiting fifteen volunteers who usually utilise bus services. These participants represented various socio-economic sections and spread evenly across the Adelaide metropolitan area. The objective of this survey was to compare the accuracy of estimated OD results with the actual travel pattern of the metroCARD users. The Department of Planning Transport and Infrastructure (DPTI) was willing to release the selected survey participant's unique details and their travel pattern provided they receive written consent from the participants. Accordingly, survey participant's smart card details were collected after undertaking their written consent. Fig. 7 shows all the boarding locations for the selected sample, and it clearly demonstrates that the sample covers most part of Adelaide.

The next step for doing a survey is matching the unique ID from each metroCARD and the database. The Department of Planning Transport and Infrastructure (DPTI) provided the media code (unique identifier in the dataset) for the selected metroCARD numbers, and two sources of data can be matched by using the relevant ID. Also because the dataset was for May 2017 and most volunteers may forget their travel pattern during that month, and usually surveys which rely on respondents' memories are inaccurate (Hwang et al. 2017), a new database was provided by DPTI for January to May 2018, specifically for selected media codes. For fifteen participants over a five-month period, 1683 transactions were collected; it should be stated here that because the volunteers were selected randomly, they may use all modes of public transport. For validating the results, only bus transactions were taken into account.
The new dataset was analysed based on the trip chain model and its assumptions; then by interviewing the volunteers, the results were validated. Each participant from the group was asked about his/her travel patterns during the specific time, and then results were compared, which was achieved through the metroCARD analysis. Among 1177 transactions only 944 were further analysed due to data integrity issues.

By interviewing the volunteers about the exact origins and destinations and also transfer stops the result can be validated. All participants were asked about their travel patterns during the relevant time by showing them the result from the analysis, and their responses were compared with the smart card analysis. From this 944 inferred OD pairs, 926 transactions estimated accurately based on the interview. There were no discernible differences between the travel patterns derived from the trip chain model and the actual travel patterns of the volunteers, and the results were 98% accurate (See table 2).

Table 2. Survey data information

Number of volunteers	15
Number of transactions (5 months)	1686
Number of transactions for the bus system	1177
Number of inferred OD pairs	944
Number of accurate OD based on an interview	926
Accuracy level	98.09

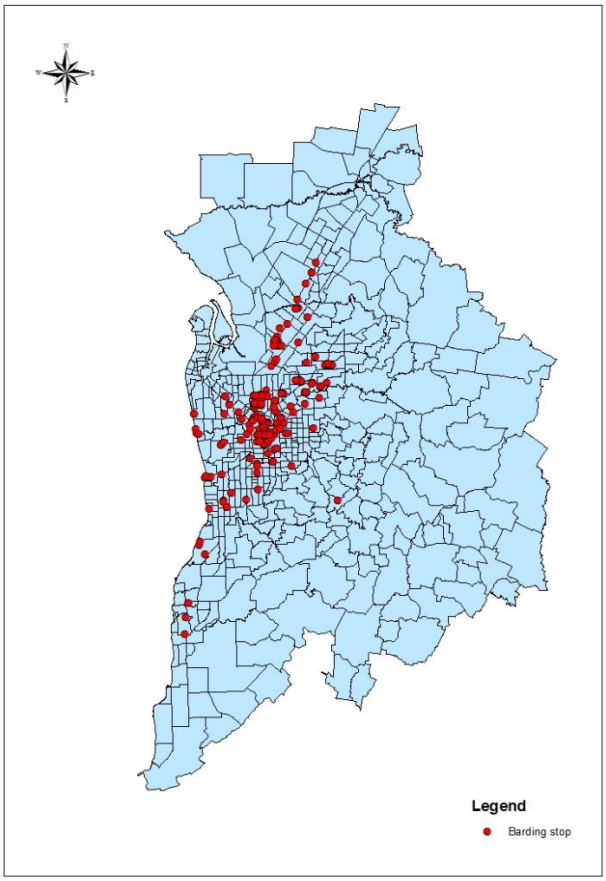

Fig. 7. An example of a trip chain for inferring the alighting stop

5.3. Limitations

Some of the issues and limitations obtained from the survey analysis are mentioned below:

1. If a volunteer uses another mode of transport such as a personal car, bicycle, walking more than the assumed walking distance during the day, the algorithm cannot infer the destination. Based on the trip chain model assumptions, it emerged that commuters should only use the public transport system. This is a limitation of the analysis due to lack of information.

2. Some of the bus routes run different services, and their stop locations may vary during different days. For example, route M44 runs from Stop F2 Grenfell St - Northside, while on some days it runs from Stop D3 Currie St - Northside. The problem which occurs here is that the nearest bus stop to the next boarding will be inferred based on the algorithm. Although the alighting stop is the same for commuters during different days, due to changes in boarding stops the destination may vary.

3. If a commuter starts the first transaction of a day far from home and walks more than the assumed walking distance, the last transaction of a day can be inferred in the vicinity of the first boarding. For example, in some cases, a volunteer starts the first transaction of a day with a train and returns via bus, and because the bus stop is situated near her/his home, some differences may occur in inferring the destination. However, in this paper, the focus is on the bus system, and the analysis was done on those passengers who use bus services. Those transactions which were related to other modes were omitted from the sample, and therefore, this problem can be ignored.

6. CONCLUSIONS

As public transport agencies increasingly adopt the use of Automatic Data Collection System (ADCS), a large amount of boarding data are being collected by the transport agencies on an ongoing basis. Advances in Information and Telecommunication Technology are changing the amount, type and quality of data available to planners and managers. In comparison with traditional surveys, which are usually time-consuming, expensive and only suited to identify a 'snapshot' for a specific context, smart card data could be used to examine the whole network on a regular basis which is ideal for transport planners. These data sets offer opportunities to infer good estimates of passenger Origin-Destination (OD) patterns. This approach used various improvements over traditional methods for improving the estimated OD pair accuracy. These include (i) minimising the GPS errors by using the stops on the opposite side of the road (ii) increasing the OD estimation accuracy by observing commuter travel pattern over a week period and (iii) improving the estimated OD accuracy by using the parallel routes.

Understanding the public transport passenger Origin−Destination (OD) flows is crucial to improving the planning and operation of public transport systems. Consequently, the validity of the estimation method should be investigated. In this paper, a new survey method was conducted to validate the algorithm for Origin−Destination (OD) estimation. The results elaborated that the method used in this paper is 98% accurate and can be employed elsewhere. An accurate estimation of public transport Origin−Destination (OD) will greatly assist government agencies in route rationalisation; it will lead to higher public transport patronage. For further studies, the results of census data can be used to validate the algorithm, and sensitivity analysis can be considered for other assumptions. Finally, for further studies the purpose of trips can be estimated based on smart card information.

Acknowledgements

The authors are grateful to the Department of Planning, Transport and Infrastructure (DPTI), for providing data for this research.

References

Alfred Chu, K. & Chapleau, R., 2008. Enriching archived smart card transaction data for transit demand modeling, Transportation Research Record: Journal of the Transportation Research Board, no. 2063, pp. 63-72.

Ali, A., Kim, J. & Lee, S., 2015. Travel behavior analysis using smart card data, KSCE Journal of Civil Engineering,

pp. 1-8.

Alsger, A., Mesbah, M., Ferreira, L. & Safi, H., 2015. Public Transport Origin-destination Estimation Using Smart Card Fare Data, Transportation Research Board 94th Annual Meeting.

Alsger, A., Assemi, B., Mesbah, M. & Ferreira, L., 2016. Validating and improving public transport origin–destination estimation algorithm using smart card fare data, Transportation Research Part C: Emerging Technologies, vol. 68, pp. 490-506.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L. & Hickman, M., 2018. Public transport trip purpose inference using smart card fare data', Transportation Research Part C: Emerging Technologies, vol. 87, pp. 123-137.

Alsger, A., 2016. Estimation of transit origin destination matrices using smart card fare data, School of Civil Engineering, The University of Queensland.

Bagchi, M. SDG & White, P., 2003. Use of public transport smartcard data for understanding travel behaviour, Proceeding of the European Transport Conference (ETC) HELD, Strasbourg, France.

Bagchi, M. & White, P., 2005. The potential of public transport smart card data, Transport Policy, vol. 12, no. 5, pp. 464-474.

Barry, J., Newhouser, R., Rahbee, A. & Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data, Transportation Research Record: Journal of the Transportation Research Board, no. 1817, pp. 183-187.

Barry, J., Freimer, R. & Slavin, H., 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City, Transportation Research Record: Journal of the Transportation Research Board, no. 2112, pp. 53-61.

Ceder, A., 2016. Public transit planning and operation: Modeling, practice and behavior, CRC press.

Chapleau, R., Trépanier, M. & Chu, K., 2008. The ultimate survey for transit planning: complete information with smart card data and GIS, Proceedings of the 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability, pp. 25-31.

Cui, A., 2006. Bus passenger origin-destination matrix estimation using automated data collection systems, Massachusetts Institute of Technology.

Devillaine, F., Munizaga, M. & Trépanier, M., 2012. Detection of activities of public transport users by analyzing smart card data, Transportation Research Record: Journal of the Transportation Research Board, no. 2276, pp. 48-55.

Ebadi, N. & Kang, J.E., 2016. Constructing Activity-Mobility Patterns of University at Buffalo Students Based on UB Card Transactions.

Friedrich, M. Mott, P. & Noekel, K., 2000. Keeping passenger surveys up to date: A fuzzy approach, Transportation Research Record: Journal of the Transportation Research Board, no. 1735, pp. 35-42.

Gordon, J., Koutsopoulos, H., Wilson, N. & Attanucci, J., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data, Transportation Research Record: Journal of the Transportation Research Board, no. 2343, pp. 17-24.

Gur, Y. & Ben-Shabat, E., 1997. Estimating bus boarding matrix using boarding counts in individual vehicles, Transportation Research Record: Journal of the Transportation Research Board, no. 1607, pp. 81-86.

He, L., Nassir, N., Trépanier, M. & Hickman, M., 2015 Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems, CIRRELT.

Heidari, A., Moayedi, A. & Abbaspour, R.A., 2017. Estimating Origin-Destination matrices using an efiicient moth flame- based spatial clustering approach, International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, vol. 42.

Hollander, Y. & Liu, R., 2008. The principles of calibrating traffic microsimulation models, Transportation, vol. 35, no. 3, pp. 347-362.

Horváth, B., 2012. A simple method to forecast travel demand in urban public transport, Acta Polytechnica Hungarica, vol. 9, no. 4, pp. 165-176.

Horváth, B., Horváth, R. & Gaál, B., 2014. A new iterative method to estimate origin-destination matrix in urban public transport, Transport Research Arena Europe, pp. 14-17.

Hwang, JH., Kim, H., Cho, S., Bellemans, T., Do Lee, W., Choi, K., Cheon, SH. & Joh., C-H 2017. An examination of the accuracy of an activity-based travel simulation against smartcard and navigation device data, Travel Behaviour and Society, vol. 7, pp. 34-42.

Ickowicz, A. & Sparks, R., 2015. Estimation of an origin/destination matrix: application to a ferry transport data, Public Transport, vol. 7, no. 2, pp. 235-258.

Kalaanidhi, S. & Gunasekaran, K., 2013. Estimation of Bus Transport Ridership Accounting Accessibility, Procedia-Social and Behavioral Sciences, vol. 104, pp. 885-893.

Kurauchi, F. & Schmöcker, J-D., 2016. Public Transport Planning with Smart Card Data, CRC Press.

Kusakabe, T. & Asakura, Y., 2011. Behavioural data mining for railway travellers with smart card data, Behavioural Data Mining for Railway Travellers with Smart Card Data.

Langlois, GG., Koutsopoulos, HN. & Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users, Transportation Research Part C: Emerging Technologies, vol. 64, pp. 1-16.

Lee, S. & Hickman, M., 2014. Trip purpose inference using automated fare collection data, Public Transport, vol. 6, no. 1-2, pp. 1-20.

Li, Tian, Sun, D., Jing, P. & Yang, K., 2018. Smart Card Data Mining of Public Transport Destination: A Literature Review, Information, vol. 9, no. 1, p. 18.

Li, Yuwei & Cassidy, MJ., 2007. A generalized and efficient algorithm for estimating transit route ODs from passenger counts, Transportation Research Part B: Methodological, vol. 41, no. 1, pp. 114-125.

Lianfu, Z., Shuzhi, Z., Yonggang, Z. & Ziyin, Z., 2007. Study on the method of constructing bus stops OD matrix based on IC card data, Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on, IEEE, pp. 3147-3150.

Ma, X., Wu, Y-J., Wang, Y., Chen, F. & Liu, J. 2013. Mining smart card data for transit riders' travel patterns, Transportation Research Part C: Emerging Technologies, vol. 36, pp. 1-12.

Morency, C., Trepanier, M. & Agard, B., 2007. Measuring transit use variability with smart-card data, Transport

Policy, vol. 14, no. 3, pp. 193-203.

Munizaga, M., Devillaine, F., Navarrete, C. & Silva, D., 2014. Validating travel behavior estimated from smartcard data, Transportation Research Part C: Emerging Technologies, vol. 44, pp. 70-79.

Munizaga, M.A. & Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile, Transportation Research Part C: Emerging Technologies, vol. 24, pp. 9-18.

Nassir, N., Khani, A., Lee, S., Noh, H. & Hickman, M. 2011. Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system, Transportation Research Record: Journal of the Transportation Research Board, no. 2263, pp. 140-150.

Nassir, N., Hickman, M. & Ma, Z.L., 2015. Activity detection and transfer identification for public transit fare card data, Transportation, vol. 42, no. 4, pp. 683-705.

Richardson, A.J., Ampt, E.S. & Meyburg, A.H., 1995. Survey methods for transport planning, Eucalyptus Press Melbourne.

Seaborn, C., Attanucci, J. & Wilson, N., 2009. Analyzing multimodal public transport journeys in London with smart card fare payment data, Transportation Research Record: Journal of the Transportation Research Board, no. 2121, pp. 55-62.

Simon, J. & Furth, P.G., 1985. Generating a bus route OD matrix from on-off data, Journal of Transportation Engineering, vol. 111, no. 6, pp. 583-593.

Tsygalnitsky, S., 1977. Simplified methods for transportation planning, Master's thesis, Massachusetts Institute of Technology Cambridge, MA.

Utsunomiya, M., Attanucci, J. & Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning, Transportation Research Record: Journal of the Transportation Research Board, no. 1971, pp. 119-126.

Wang, W., 2010. Bus passenger origin-destination estimation and travel behavior using automated data collection systems in London, UK, Massachusetts Institute of Technology.

Yang, H. & Jun, C., 2018. Visualization of Public Bus Passenger Travel for Travel Pattern Analysis, Adjunct Proceedings of the 14th International Conference on Location Based Services, pp. 121-126.

Zhao, J., 2004. The planning and analysis implications of automated data collection systems: rail transit OD matrix inference and path choice modeling examples, Massachusetts Institute of Technology.

Zhao, J., Rahbee, A. & Wilson, N.H., 2007. Estimating a Rail Passenger Trip Origin‐Destination Matrix Using Automatic Data Collection Systems, Computer‐Aided Civil and Infrastructure Engineering, vol. 22, no. 5, pp. 376-387.