https://dx.doi.org/10.1016/j.trc.2021.103044

Transit OD matrix estimation using smartcard data: Recent developments and future research challenges

Etikaf Hussain[a], Ashish Bhaskar[a]* and Edward Chung[b]

[a]*School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, Australia;* [b]*Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong, China*

*corresponding author: Associate prof. Ashish Bhaskar, School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, Australia; Email: ashish.bhaskar@qut.edu.au

Abstract

In public transport, smartcards are primarily used for automatic fare collection purpose, which in turn generate massive data. During the last two decades, a tremendous amount of research has been done to employ this big data for various transport applications from transit planning to real-time operation and control. One of the smart card data applications is the estimation of the public transit origin-destination matrix (tOD). The primary focus of this article is to critically analyse the current literature on essential steps involved in the tOD estimation process. The steps include processes of data cleansing, estimation of unknowns, transfer detection, validation of developed algorithms, and ultimately estimation of zone level transit OD (ztOD). Estimation of unknowns includes boarding and alighting information estimation of passengers. Transfer detection algorithms distinguish between a transfer or an activity between two consecutive boarding and alighting. The findings reveal many unanswered critical research questions which need to be addressed for ztOD estimation using smartcard data. The research questions are primarily related to the conversion of stop level OD (stOD) to ztOD, transfer detection, and a few miscellaneous problems.

Keywords: smartcard data; public transit OD estimation; transfer detection; origin inference; destination inference; OD scaling

1 Introduction

Advancement in technology has advocated the implementation of the intelligent transport system in the transit sector. Many transit agencies are using Automatic Data Collection (ADC) system, which includes Automatic Fare Collection (AFC) using smartcard, Automatic Vehicle Location (AVL), and Automatic Passenger Counts (APC). AFC data are primarily used for fare collection, which in turn produces huge data that are readily available to transit agencies and planners. Since the last two decades, there is an increasing trend to exploit these big data to generate knowledge for transit applications from long-term planning to real-time operations and control. Few such transit applications are transit Origin Destination matric (tOD) estimation (Alsger et al., 2016, Barry et al., 2002, Munizaga and Palma, 2012, Nassir et al., 2015); travel pattern mining (Cats et al., 2015, Han and Sohn, 2016, Kieu et al., 2015a, Kieu et al., 2015c, Ma et al., 2017, Ma et al., 2013, Morency et al., 2007); trip purpose (Chapleau et al., 2008, Kusakabe and Asakura, 2014, Lee and Hickman, 2014); transit disruption planning (Yap and Cats, 2020, Yap et al., 2020); and identification of Public Transit (PT) improvement areas (Hussain et al., 2020a, Hussain et al., 2020c).

Traffic OD matrix provides mobility demand over a network, which serves as the vital input for large scale transport modelling (Hensher and Button, 2008). When such general OD matrix is disaggregated at modal level, e.g., car and transit OD matrix, it provides more flexibility to plan mobility requirement for each mode, e.g., private vehicles and transit. Car OD along with tOD are the essential input for simulations, the results of which can have a variety of uses such as, adequacy of road infrastructure, signal designing (Cipriani and Fusco, 2004, Pell et al., 2016), congestion evaluation (Choudhury et al., 2011), congestion pricing (Bracher and Bogenberger, 2018), etc. ztOD of an urban area helps in PT planning at strategic, tactical, and operational levels. The potential applications of tOD include but not limited to transit route and network designing; transit resources allocation and distribution across the network (Hussain et al., 2020b); prioritise transit-related projects funding; identification of gaps in demand and supply of transit services (Hussain et al., 2020a); overcrowding (Wang et al., 2015); planning maintenance; and validating large scale assignment models (Tavassoli et al., 2018).

Traffic OD is estimated by employing a range of traditional to advanced datasets. The traditional datasets include sampled household travel surveys (Stopher and Greaves, 2007) and vehicle counts from loops or manual (Zhou et al., 2003). Advanced datasets include Bluetooth data (Behara et al., 2020), vehicle number plate recognition system (Rao et al., 2018), AVL (Guozhen et al., 2011), APC (Cats et al., 2019), mobile phone data (Regt et al., 2017), smartphone location data (Nikolic and Bierlaire, 2017), and social media data (Rashidi et al., 2017). Although the above-stated systems have a higher initial cost, it has low cost throughout their life.

The literature on the use of the smartcard to estimate tOD is still emerging. Earlier, Pelletier et al. (2011), Li et al. (2018), and Faroqi et al. (2018) have presented the review on the use of smartcard data in transportation. Pelletier et al. (2011) provide insights on smartcard technology's evolution, which includes the hardware, data storage and commercialisation. The article also briefly reviews the smartcard data applications for strategic, tactical and operational level planning. The details for individual applications is outside the scope of the paper. Li et al. (2018) reviewed the literature on destination estimation only from entry-only systems and classified the destination estimation models into trip chaining model, probability model, and deep learning model. The study proposed a method to weigh the performance of assorted models. However, the review only considers studies of destination inference. Also, it lacks an in-depth analysis of the studies and only gives details on the models' use. Moreover, Faroqi et al. (2018) briefly summarise studies primarily published after 2010 on smartcard data and group them based on their applications focusing on tOD estimation, mining travel patterns, and trip purpose. The article provides an insight into these applications and does not provide critical and detailed analysis from tOD estimation point of view.

Complementary to the above review, this paper aims to critically review and analyse the current literature on tOD estimation using smartcard data focusing mainly on the origin information estimation, destination inference, transfer detection, zone level tOD estimation, and identify the research needs for future development and enhancements.

To this end, Section 2.1 delineate the AFC data for tOD estimation; Section 2.2 and Section 2.3, respectively provide the details related to data cleansing, and estimation of unknowns, i.e., boarding location estimation, and alighting location estimation. Section 2.4 provides a description and critical assessment of transfer detection rules. Section 2.5 specifies the current literature on the aggregation of tOD from smartcard data, and Section 2.6 elucidate various validation techniques employed for the tOD estimation problem. Section 3 articulates future research directions, and finally, the conclusion is presented in Section 4.

2 tOD estimation problem framework

A typical tOD estimation framework is shown in Figure 1. The framework can be divided into the following four parts; the details of which are explained and critically reviewed in this article.

i. *Data cleansing*

ii. *Transfer detection*

iii. Estimation of unknowns, i.e., boarding and alighting location estimation,

iv. Estimation of the zone to zone transit OD (ztOD) from stop to stop transit OD (stOD), and

v. Validation of proposed methodology

In addition to the above four steps, the entry-only AFC system requires an additional algorithm to infer the alighting stop as it lacks passenger's alighting information (Figure 1). If the data are obtained from an entry-exit system, boarding and/or alighting must only be computed for transactions having missing boarding and/or alighting information. The need to estimate boarding information generally depends on the fields available in the smartcard data. If the smartcard data do not provide boarding location or boarding time or is missing due to system malfunctioning, it would be required to estimate the unknown accordingly.

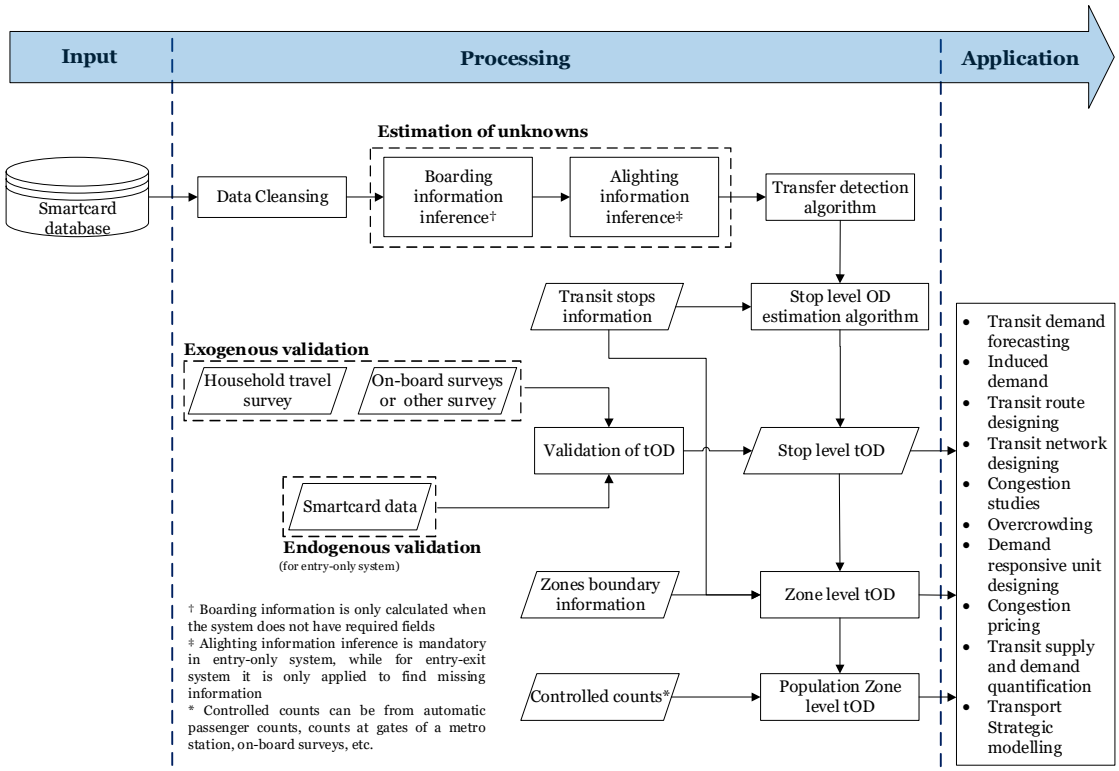

Figure 1 Typical steps of transit OD estimation problem

2.1 AFC data available for tOD estimation

AFC systems can be classified into:

- *Entry-only system:* Here, the passengers are required to tap-in the smartcard only when boarding a transit service (or entering the station in case of rail). Only Tap-in is generally needed in a flat-fare system where the fare is independent of distance travelled. Examples of such systems are Jinlingtong card in Nanjing, China, and Bip! card in Santiago, Chile.
- *Entry-exit system*: Here, the passenger is required to tap-in the smartcard during boarding and tap-out during alighting from the transit vehicle (entering/leaving from the rail station where devices are installed at the station). The entry-exit system is generally preferred in areas having distance-based fare. Examples of such systems include *gocard* in Brisbane, Australia, and *Tmoney* in Seoul, Korea.

In literature, few researchers refer to an entry-only and an exit-entry system as an open system

and closed system, respectively (Lu et al., 2020, Mosallanejad et al., 2019, Nassir et al., 2015, Tang et al., 2020). In general, the closed system should be the one where passengers can typically enter and exit from the gateway, whereas the open system is where passengers have unrestricted entry. To avoid confusion, henceforth this paper follows the above-stated definitions.

Table 1 maps the intermodal datasets used in various studies and provides remark on the usage of the input data, the generated output from the methodology and the aspect of the literature related to tOD estimation. The intermodal consists of bus/trams, subway/metro/rail and ferry. The typical data sets include *vehicle ID*, *boarding stop*, *alighting stop*, *boarding time* and *alighting time*.

Note: in the following sections (section 2.2 to section 2.5) detailed description of those methodologies are provided. Further, the feasibility and reliability in the real-world applications are discussed in section 2.6.

Public transportation agencies have a unique system of ADC in terms of data generated by the system and the installed AFC system (Table 1). Nevertheless, accuracy issues may exist when integrating AFC data with ADC's location data (i.e., AVL data). DoD (2008) has reported 7.8m errors (with 90% confidence interval) in the position estimated by the Global Positioning System (GPS) installed for AVL. The error is expected to increase further when the transit vehicle is near buildings, tunnels, trees, and bridges. Likewise, Kumar et al. (2018) reported 17 m (55 ft) error in GPS based co-ordinates. According to TransLink report, every day, about 14000 trip adjustments, on average, are made to the actual trips made using *gocard* (smartcard) (TCSMS, 2017). Although, there may exist many other reasons for the wrong reading of smartcard data including challenges related to the integration of two big datasets (Lahat et al., 2015), inaccuracy in GPS is one of the prominent source responsible for such errors (Ellison et al., 2017).

Table 1 Available AFC fields for tOD estimation to various researchers

Study	City, Country	Bus					Subway/Metro/Rail					Ferry	Remarks	Output	Work on tOD estimation related to
		ID	BS	AS	BT	AT	ID	BS	AS	BT	AT	All			
Entry-only system OD inference															
(Barry et al., 2002)	New York, US							y		y				Station-to-station OD matrix	Trip chaining
(Hofmann and Mahony, 2005)	N/A*	y**	y		y								*Not available due to security reason ** Route ID is available instead of Bus ID	Transfer detection	Rule-base algorithm
(Cui, 2006)	Chicago, US	y*	y**		y								*Equipment ID of farebox is linked to bus ID. **AVL data are integrated to come up with boarding stop.	Route level population OD matrix estimation	Iterative proportional fitting and maximum likelihood estimation
(Trépanier et al., 2007)	Gatineau, Quebec, Canada		y		y									Destination time and location inference	Trip chaining
(Zhao et al., 2007)	Chicago, US	y**	y*		y			y		y			* for bus, location in coarsely represented by the bus route number, rather than the exact bus location or stop ID ** Bus number and bus route number are taken from AVL data	Rail OD matrix, and transfers within rail and rail to bus	Trip chaining
(Zhang et al., 2007)	Changchun, China	y*			y								* Instead of bus ID, the smartcard data records the route and driver ID. Other data from onboard surveys and surveys from drivers are made available.	Population tOD matrix	Trip chaining, and application of doubly constrained growth factor method to estimate population tOD
(Chu and Chapleau, 2008)	Gatineau & Ottawa, Canada	y	y		y									Transfer detection	Rule-based
(Farzin, 2008)	Sao-Paulo, Brazil	y	y*		y								* Boarding location is not directly estimated, it is determined by integrating AVL and AFC data having Bus ID, and time in common. Stops, AVL and APC data sets are used to infer OD estimation.	Inference of OD matrix	Trip chaining and use of GPS location to infer origin
(Barry et al., 2009)	New York, US				y†			y		y			† the time is rounded to one decimal point of an hour (i.e. 6 minutes)	Multi-modal Zonal OD matrix inference (subway, and bus)	Trip chaining
(Seaborn et al., 2009)	London, UK				y			y	y	y	y		AVL and stops data are integrated with AFC data.	Multi-modal OD matrix, transfers across modes	Trip chaining
(Nassir et al., 2011)	Minneapolis-Saint Pauls, US	y	y		y									Detection of passenger alighting station and transfer detection	Trip chaining
(Li et al., 2011)	Jinan city, China	y	y		y									Inference of alighting stop, zone-to-zone OD	Trip chaining
(Wang et al., 2011)	London, UK		y*		y			y	y	y	y		* Only route number is recorded in AFC data. iBus (AVL) data are employed to estimate the bus stop location.	Origin and destination inference and its validation	Trip chaining
(Munizaga and Palma, 2012)	Santiago, Chile	y	y†		y		y	y		y			† Boarding stop for bus is divided into two parts, 1. transactions made on devices installed in the bus, 2. Transaction made on devices installed at bus station. The latter can give location directly.	Multi-modal large-scale public transport OD matrix	Trip chaining & rule-based transfer detection
(Ma et al., 2012)	Beijing, China	y*				y							* Route number of the bus is reported.	Origin stop location inference	Markov Chain based Bayesian decision tree algorithm

Study	City, Country	Bus					Subway/Metro/Rail					Ferry	Remarks	Output	Work on tOD estimation related to
		ID	BS	AS	BT	AT	ID	BS	AS	BT	AT	All			
(Gordon et al., 2013)	London, UK		y*		y			y	y	y	y		* Only route number is recorded in AFC data. iBus (AVL) data are used to estimate the bus stop.	Intermodal journeys inference.	Trip chaining and rule-based transfer inference
(Jun and Dongyuan, 2013)	Nanning city, China	y	y		y									OD inference from smartcard	The pattern of transit commuter in morning and evening peak
(Munizaga et al., 2014)	Santiago, Chile	y	y†		y		y	y		y			† Boarding stop for bus is divided into two parts, 1. transactions made on the devices installed in the bus, 2. Transaction made on the devices installed in bus station. The later can give location directly.	Validation of the algorithm developed in Munizaga and Palma (2012) is performed.	Trip chaining
(Nunes et al., 2016)	Porto, Portugal	y	y		y			y		y			AVL data are integrated with AFC data.	Calculation of OD matrix in the entry-only system and distance-based fare	Trip chaining
(Kumar et al., 2018)	Minneapolis-Saint Pauls, US	y	y*	y*	y*	y*	y	y*	y*	y*	y*		* Dataset either contains boarding or alighting information. Also, AVL data are used to find the exact stop of the service taken.	Calculation of OD matrix based on either boarding or alighting	Trip chaining
(Chen and Fan, 2018)	Guangzhou, China	y			y									Boarding stop estimation	Rule-based boarding stop and bus direction detection
(Yan et al., 2019)	Shenzhen, China	y*			y								* Bus ID and route of the bus is recorded. GPS (AVL) data are integrated to get the boarding stop location.	Alighting stop estimation	2-step algorithm comprised of trip chaining and machine learning (Markov chain model)
(Zhao et al., 2019)	Nanjing	y			y		y		y		y			Metro to bus transfer inference	Association rule learning and k-means clustering
(Huang et al., 2020)	Suzhou, China	y	y**		y								** Boarding stop is estimated using GPS data	Estimation of OD matrices	Trip chaining and use of GPS location to infer origin
Entry-exit system OD inference															
(Assemi et al., 2020, Alsger et al., 2016, Nassir et al., 2015, He et al., 2015, Alsger et al., 2015)	South East Queensland, Australia	y	y	y	y	y		y	y	y	y	y		1-2. Alighting stops inference, 3. stop-to-stop OD matrix and short activity detection, 4. and 5. Validation of trip chaining assumptions and sensitivity of parameters used in transfer detection	1. Neural network application (2, 4, and 5) Trip chaining, 3. short activity detection
(Cheng et al., 2020)	Guangzhou, China							y	y	y	y			Alighting stop inferenc	Latent dirichlet allocation
(Jung and Sohn, 2017)	Seol, Korea	y	y	y	y	y		y	y	y	y		Only bus mode is used in the analysis	Alighting stop inference	Deep learning architecture

ID=Vehicle ID, BS=Boarding Stop location, AS=Alighting Stop location, BT=Boarding Time, AT=Alighting Time, All=Variables including all fields (i.e., ID, BS, AS, BT, AT)

2.2 Data cleansing

Data cleansing requires information about the sources and type of errors to refine the data. Equipment failure and human error are two reported sources of errors. Equipment failure can be due to smartcard reader, desynchronised time clocks of the data collecting devices, installed GPS, and or the overall system. Human errors may include but not limited to forget to tap-off the card, tapping a wrong card, etc. Such inconsistent transactions may contribute to as many as two percent of the total transactions (Translink, 2016). Overall, above failures introduce the following errors in smartcard data; no record for boarding/alighting time and/or location, the recorded boarding time and/or location is equal to the alighting time and/or location (in an entry-exit system only), alighting time is earlier than boarding time for the same trip (in an entry-exit system), missing smartcard ID, duplication of an event, transactions on an untraceable stop (i.e., stop location cannot be traced), etc. In the later stage of analysis, trips can be found with higher or lower than typical travel time, for instance, a trip of several hours in a medium-sized city (Luo et al., 2017). This error may arise due to faulty recording of boarding and or alighting time as discussed previously.

Besides, the data description in Assemi et al. (2020) shows a minimum travel time and distance of 0.18 minutes and 0.03 km, respectively, which may suggest boarding on the wrong bus and alighting upon realisation without riding, fare evasion, or other human error (e.g., tapping a card multiple time where passenger thinks that tap-in went wrong). It is required to create an upper and a lower limit of travel time and travel distance of a trip-legs to be considered for analysis. Egu and Bonnel (2020) considered a lower limit of 10 minutes between the boarding (or tap-in at a station) and alighting (or tap-out at the same station) from the same vehicle, while Luo et al. (2017) chose the lower limit for the same duration as 1 minute. The former can potentially possess problem since a shorter trip can be less than 10 minutes in an urban area. The authors suggest using the network analyst experience to decide the threshold value because it may be site-specific, and different values may be suitable for different areas.

The type of error and its correct interpretation is very critical in tOD estimation. The wrong interpretation of an error may lead to inaccurate tOD. It is possible that various researchers use a contrasting approach to interpret the same ambiguity in the data. For example; Nunes et al. (2016) consider that second consecutive transaction down the line on the same route represents duplication. Therefore, the subsequent transaction is excluded from analysis while other researchers assume the subsequent event as a separate trip. In another study, smartcards with a high number of trips are deleted, arguing that it may represent the transit agency employees (Yan et al., 2019). Consequently, the data must be checked for all possible inconsistencies that must be removed before and during analysis, if found.

Once the smartcard data are refined and error-free, it can be used for stOD estimation.

2.3 Estimation of unknowns

As summarised in Table 1 and discussed earlier, the number of unknowns depend on the type of data and attributes recorded corresponding to a transaction made on PT. Most of the studies have focused on the destination location inference. However, some of the systems also lack the boarding location information. The challenges corresponding to the estimation of unknowns and approaches to handle such problems by various researchers are discussed in the following sub-sections.

2.3.1 Boarding location estimation

For tOD estimation, quality of data plays a vital role. Table 1 summarises the variables available to the researchers in smartcard data. In literature, most of the studies are performed on the entry-

only system, while studies conducted on the entry-exit system merely validate the models developed for the entry-only system. Moreover, rail/metro passengers are required to tap-in at a subway or rail station entrance. If more than one train serves that station, the train boarded, and travel direction cannot be directly inferred since passengers are served by one entrance for both the directions and requires appropriate assumptions.

Based on the available fields in the recorded data of an entry-only system, the boarding stop estimation can be divided into three categories, shown in Table 2. The first category is when both boarding stop and time are recorded in the data; therefore, it can be directly used for the application. The second category is when boarding time is recorded, but boarding stop information is missing. In this case, the required field can be inferred by fusing smartcard data with other data sources (such as AVL data, schedule data, etc.), if available. Usually, it is uncommon to come across the third category of smartcard system, where neither boarding time nor place is recorded. However, one of the studies (Lahat et al., 2015) is conducted on a similar system. The study fused AFC, APC and GTFS (Google Transit Feed Specification) data and reported 92% of success rate for boarding location inference. Boarding stop estimation is generally not a research problem because almost all the agencies record the boarding stop/station (Alsger et al., 2016, Alsger et al., 2015, Chu and Chapleau, 2008, Hofmann and Mahony, 2005, Jun and Dongyuan, 2013, Li et al., 2011, Nassir et al., 2015, Nassir et al., 2011, Nunes et al., 2016, Trépanier et al., 2007).

The AFC data employed in the study Kumar et al. (2018) has two types of trips data, i.e., one set of data include boarding information with no alighting information, and the other dataset have alighting information with missing boarding information. For the latter case, to determine boarding information, the same logic discussed below can be applied, if found missing. Also, in the case of the entry-exit system, if either boarding or alighting information is missing, it can be found by methods discussed in this section and section 2.3.2, respectively.

Table 2 Categories of boarding stop estimation problem based on available features in the smartcard data

Boarding time	Boarding stop	Remarks
✓	✓	No calculation required to estimate the boarding location
✓	×	Boarding location can be calculated by integrating smartcard data with another dataset, such as AVL, scheduled data, etc., if available
×	×	Boarding location can be inferred with the availability of GTFS, APC, and AFC data

Suppose AFC data do not provide the location of the boarding transaction directly (second category). In that case, AVL and AFC data sets are integrated based on the timestamp and route number (Farzin, 2008, Gordon et al., 2013, Sánchez-Martínez, 2017, Wang et al., 2011). Alternatively, instead of route number, vehicle ID can also serve the purpose (Tu et al., 2018). AVL data usually contain the location of the transit vehicle with the timestamp. Due to the location data errors (mostly due to GPS errors), it is hard to infer 100% of transactions. Zhao et al. (2007) reported that 5.4% of the total inference failure is due to AVL and GPS.

In the study of Cui (2006), the equipment ID of farebox installed in a bus is linked to bus ID. Afterwards, AVL and AFC data are integrated to estimate the boarding stop. In Munizaga and Palma (2012) study, the origin is directly available using metro and bus station (stationary points) database, except for fare boxes that are installed on the bus (dynamic points), AVL data are used to infer the origin. While integrating AFC and AVL datasets, transactions made within a radius of 110 m are assumed to have the same origin stop (i.e., all the stops within 110m are believed to be same) (Cui, 2006, Farzin, 2008). This assumption is justifiable for big bus stops where the number of stopping bays exceeds five (assuming 20 m bay; for six bays, length of a single stop would be 120m). Kumar et al. (2018) utilised AFC, AVL and GTFS data to find the closest stop

of the transit route for each trip leg. The study reports a mean error of 17 m (52 ft) in the GPS locations recorded in AFC data. Furthermore, in a recent study, Huang et al. (2020) employed the DBSCAN algorithm for spatial clustering and matching of AVL data and transit stops.

In a dense Beijing network, Ma et al. (2012) proposed a method to find the boarding location of flat fare buses where available fields are only route number and driver ID. The study develops a methodology to infer the boarding stop information without using dataset other than smartcard. The smartcard transactions are first grouped by vehicle ID and boarding time. The boarding stop is inferred by comparing the observed travel time of grouped transactions and calculated travel time between stops. The study created a Markov Chain based Bayesian decision tree algorithm to estimate the probability of a set of transactions belongs to a stop. The study reported accuracy of approximately 70% when applied on two routes. Later, Chen and Fan (2018) proposed an improvement in transit direction detection. However, the proposed methodology has great potential for improvement, specifically if applied with scheduled data. Nevertheless, the current method's accuracy is also unknown, as the study does not provide any validation (Chen and Fan, 2018).

There are instances when boarding transaction time is logged instead of the boarding time. For example, in a gated entrance (usually in a metro station), it is required to find the passenger train boarding time (also known as passenger-to-train assignment), while the recorded time corresponds to the transaction time. This problem is similar to case three in Table 2, where the smartcard data can be integrated with AVL or scheduled data to find the actual boarding time (Zhang et al., 2016).

Integration of AFC and AVL (GPS) datasets seem to be a practical solution to estimate the boarding location, hence used by researchers subjected to its availability. Scheduled data can also be utilised for the same purpose if AVL data are not available. Though, it will exclude more transactions from the analysis due to discrepancies in the scheduled and operational headway. The authors believe that before using scheduled data for inference of boarding location, the overall headway adherence (or other time performance indicator) of the transit system could be calculated to increase the confidence on estimated boarding location.

2.3.2 Alighting location estimation (Destination inference)

This section first provides general details on the alighting location estimation methodology, which will develop a basic understanding regarding the problem in hand. Afterwards, the details on how various researchers have estimated alighted location are provided. In general, alighting location is determined using the trip-chaining method, described below.

2.3.2.1 General description of the trip-chaining method

In trip chaining method, where only the boarding locations are known, the alighting locations are inferred from the boarding location of the successive trip. Here, if we chain the entire day's boardings of the user in a cycle manner, then the alighting location of a trip should be spatially constrained to the boarding location of the successive trip. Mathematically it can be expressed as:

$$A_j = k \rightarrow \begin{cases} \min\limits_k d\left(B_{j+1}, k\right) & j < n \\ \min\limits_k d\left(B_1, k\right) & j = n \end{cases}$$
$$k \in S_j^r \, such \; that \begin{cases} d\left(B_{j+1}, k\right) \le d_{max} & j < n \\ d\left(B_1, k\right) \le d_{max} & j = n \end{cases} \quad (1)$$

A_j is alighting point of jth trip-leg, B_{j+1} is the boarding point of jth +1 trip-leg, n is the total number of trip legs of a passenger in a day, S_j^r is the set of stops downstream to the boarding stop of jth trip-leg (B_j) for the route r, $d(a,b)$ is (Euclidean or walking) distance between point a and b, and d_{max} is the maximum threshold walking distance. k is the candidate stops within the set of stops S_j^r having distance lower than d_{max} from B_{j+1} (or B_1 in case of $j=n$).

Figure 2 illustrates typical scenarios in trip chaining method in which a transit user makes various trips in a day. Let's say, a transit user first travels from his home on route R_1 from boarding stop B_1 to office. Later, the user takes R_2 from stop B_2 and reach a mall for shopping or any other activity. From here, the user takes route R_3 from stop B_3 and arrives home.

As per Equation (1), alighting stop A_1 must lie within distance d_{max} from stop B_2, i.e., $d \leq d_{max}$. Further, the equaiton determines the closest stop to B_2 out of list of candidate stops k. Similarly, to infer alighting stop for trip 2 (A_2), the boarding stop of trip 3 (B_3) can be utilised. Furthermore, to determine the alighting location of last trip (A_3), boarding location of days' first trip (i.e., B_1) can be employed (Figure 2).

A_n is the alighting stop of last trip of the day, B_1 is boarding the day's first trip, and S_n^r is the set of downstream stops of last boarding stop B_n of route r. Usually, the threshold distance for intermediate and last trips is considered dissimilar, mainly due to the nature of the connected activity. In literature, various researchers have opted for different values for the two distances; the divergence, rationale, and suitability of which is discussed in the next sub-section.

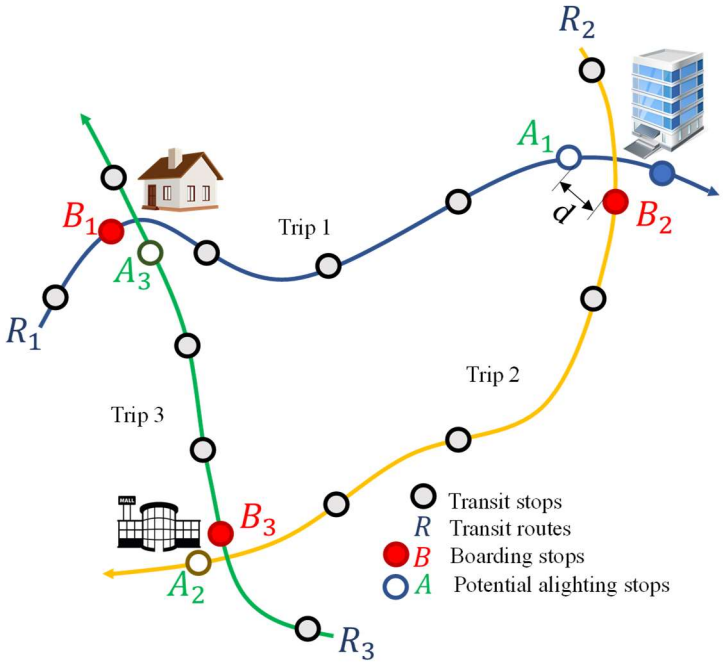

Figure 2 Graphical representation of the simplest scenarios of trip chaining method to infer the alighting location

2.3.2.2 Research related to alighting location estimation

Most of the researchers have used trip chaining method to estimate alighting location and or time of a trip made in an entry-only system, as discussed above. Here, the discussion is extended on the adoption of various rules and values for trip chaining. Trip chaining is first proposed by Barry et al. (2002) by recommending two assumptions to link trips made by a passenger using a

particular card. The assumptions are:

(1) A high percentage of passengers begins their next trip close to the finishing point of the last trip; and
(2) Day's first trip starting station and day's last trip ending station is the same.

The first assumption (Equation (1)) is also known as the continuity assumption. It implies that a passenger would not opt for another mode during his/her all-day journeys, i.e., non-usage or non-utility of modes other than PT (Trip 1 and Trip 2 in Figure 2). In other words, for a passenger, the destination stop of the previous trip can be found within walking distance of the next boarding stop. Threshold walking distance values used by different researchers are summarised in Table 3. The continuity assumption narrows down the number of candidate-stops for estimation of the previous alighting stop. The number of candidate stops is further filtered by selecting only those stops which serve the last boarded route. Therefore, easing the process of identifying previous alighting stop.

The second assumption (Equation (1)) is also known as the day's symmetry trip assumption. It can be understood by the fact that a passenger first trip boarding stop and last trip alighting stop of a day generally is close to his/her resident (Trip 3 in Figure 2). As the study carried by Barry et al. (2002) is for subway tOD matrix estimation only, this assumption remained true for as many as 90% of the cases when validated through the New York Metropolitan Transportation Council (NYMTC). In the later studies, this assumption is relaxed by stating that "Day's first trip starting stop and day's last trip ending stop are close to each other (i.e., within some radius of threshold value)" to estimate the tOD for buses. This concept was introduced by Trépanier et al. (2007) and later used by many other researchers who worked on the tOD estimation problem primarily involving bus mode.

The main shortcoming of trip chaining method is that destination of all transactions with exactly one trip in a day cannot be inferred using one-day smartcard data. Trépanier et al. (2007) utilised longitudinal smartcard data (one-month data) to minimise the impact of this shortcoming. They searched a single entry card number on the analysis day in other days of the month. The destination was inferable if a card has a similar transaction (records with the same route, and a minimum boarding time difference) within a whole month. Trépanier et al. (2007) claimed successful destination inference of an extra 13% of the smartcard holders having a single transaction in a day. Another study quoted successful destination inference of 80% of non-inferred transactions due to single entry in a day by considering the next day first trip location as the destination (Kumar et al., 2018). Besides, building on the previous work, He and Trépanier (2019) further improved the alighting location algorithm by proposing Kernel density-based estimation of candidate locations' spatial and temporal probability. The study claims to have soundly inferred the alighting location of 91% of total trips.

Similarly, Jun and Dongyuan (2013) addressed the error source by considering transactions from 5 working days to estimate tOD. The authors claim that with these values, the accuracy of the adopted model is 83%. The shortcoming of this study is that it only calculates the tOD for trips made in morning and evening peaks leaving all the transactions which are made in non-peak time. Therefore, the total number of cards analysed in this study is 40-52%. All those transactions having $K_t = 2$, $M_t = 1$, and $N_t = 1$ in five working days are considered for analysis. Where K_t is the total number of records of a smartcard in the morning and the evening peak period; M_t and N_t are respectively, the number of records in the morning peak and evening peak.

In addition, to decrease the number of single transaction cards, Barry et al. (2009) observed that at midnight (12 am) there is still some activity going on and touches its minimum value around 3 am. Therefore, for analysis, Barry et al. (2009) recommended and used a virtual day starting from 3 am of one day to the same time of the next day. Following this, Munizaga et al. (2014) considered 4 am and Nunes et al. (2016) 5 am as the start of the day. Adaptation of virtual day has enabled researchers to decrease the number of entries with a single transaction. This idea may not be suitable for small to medium-sized cities where there is least (or no) PT activity at midnight.

Still, it can be more suitable for metropolitan cities where nightlife exists.

As stated earlier, a threshold walking distance value is set to estimate the previous trip alighting point from next trip boarding point. Cui (2006) set the threshold to ≈ 1100 m (0.01° latitude) and estimated OD matrix for the rail system only. To determine the alighting station, Munizaga and Palma (2012) introduced generalised time *Tg*, which includes the walking distance and is represented by equation (2).

$$Tg_i = t_i + f_w . \frac{d_{i-post}}{s_w} \tag{2}$$

Where t_i is the initial time, s_w is average walking speed, f_w refers to penalisation factor or disutility of walking time, and d_{i-post} is the distance between point '*i*' and next boarding. The optimisation problem is to minimise Tg_i, which is generalised cost while keeping $d_{i-post} < d = 1000m$. The generalised cost was introduced to handle the problem that arose from two-way routes. In Munizaga and Palma (2012) study, the origin of bus trips is not available directly from the data. Therefore, if a passenger taps the card in the bus station, it is required to infer bus and its routes as well. In another study, Munizaga et al. (2014) used the threshold walking distance (d_{max}) of 1000 m. Also, for London, Wang et al. (2011) and Gordon et al. (2013) used the d_{max} of 1000 m (12-minute walk, assuming walking speed as 5 km/h). In a recent study, Nunes et al. (2016) considered a cut-off distance of 640 m. While, using an entry-exit system data from South-East Queensland (SEQ), Australia, Alsger et al. (2015) performed a sensitivity analysis of the distance travelled between boarding stop and previous alighting stop. The study showed that 82% of the consecutive transactions are made within 400 m. It also showed that the number of successive transactions does not improve if the d_{max} exceeds 800 m. Further, Alsger et al. (2016) proposed an improved algorithm and used d_{max} of 530 m claiming that this improves OD matching from 66% to 72%.

Table 3 portrays the crucial studies which have used distinct d_{max} for tOD estimation. In all studies mentioned in Table 3, d_{max} is the Euclidean (aerial) distance except Nassir et al. (2011), where the study calculated walking distance by multiplying a factor of $\sqrt{2}$ with Euclidean distance to accommodate for actual walking distance. Euclidean distance may not represent the exact distance walked by a passenger since the neighbourhood's street connectivity can affect the distance travel by passenger to access public transport. Use of Euclidean distance can sometimes cause underestimation of the distance intended to consider, and therefore can lead to the false estimation of alighting stop.

d_{max} can depend on many other variables, for instances, mode of PT involved in analysis, the architecture of the city in analysis, demographics, terrain, whether a geographic barrier in commuting is present or not, etc. As recommended by TCQSM (2013), riders are willing to walk more distance (800 m) to access rail or subway station as compared to the bus (400 m). Also, the pattern of a city (grid or radial) can have a strong influence on the walking distance travelled by the passenger. In the grid-patterned city, one-way streets can cause the passenger to travel more than usual. A highly walkable neighbourhood is defined by a variety of land use, high residential density, and street connectivity. Ghani et al. (2016) concluded that walking pattern is not the same for all the neighbourhoods to access transport. Moreover, Ghani et al. (2018) consider age as a critical factor in walking for transport. The terrain is also a key factor which can cause the difference in walking behaviour of a city. For example, passengers are willing to walk more on flat terrain as compared to rolling and mountainous terrain (Ceder et al., 2015). Another study supplemented these results reported that elevated stations cause a reduction in ridership (Zhao et al., 2013).

Table 3 Walking distance used for destination inference in various studies

Distance (m)	Studies	City	PT mode	City type	Geographic barrier within the city	Terrain
530	(Alsger et al., 2016)†	SEQ, Australia	Bus, rail, ferry	Radial	Yes, river	Rolling
640	(Nunes et al., 2016)	Porto, Portugal	Bus	Radial-Grid	Yes, river	Rolling
800†	(Alsger et al., 2015, Nassir et al., 2011)	SEQ, Australia	Bus, rail, ferry	Radial	Yes, river	Rolling
750	(Gordon et al., 2013)	London, UK	Bus & subway	Radial	Yes, river	Rolling
1000	(Gordon et al., 2013, Munizaga et al., 2014, Munizaga and Palma, 2012, Wang et al., 2011, Yan et al., 2019)	Shenzhen, China	5‡: Bus	Grid/Mix	No	Rolling
		Santiago, Chile	2 & 3: Bus	Radial	No	Mountainous
		London, UK,	1: Bus & subway 4: Bus	Radial	Yes, river	Rolling
1110	(Cui, 2006)	Chicago, US	Bus	Grid	No	Flat terrain

† This is the recommended value of walking distance apart from other values used for analysis in Alsger et al. (2015). ‡ 1 (one) corresponds to study appears first in the list, while 5 is the last study, i.e., (Yan et al., 2019)

Besides, the increased usage of other transport modes (e.g., ride-hailing services, bicycle, e-scooters, etc.) for short trips in an urban area may affect the continuity assumption. For instance, Lime (2018) reported 27% of all trips made are to access or egress PT. It suggests that users can avail modes other than walking for intermediate trips by which they can cover more distances.

Instead of using walking distance as an indicator for alight stop inference, Sánchez-Martínez (2017) used the disutility for different stages in a PT trip, i.e., walking time (at entry and exit of the station), in-vehicle time, transfer time, and walking time from the station/stop to the final destination, and minimises the cost function. Equation (3) is employed to calculate relative cost at different paths of PT trip.

$$V = \theta_e t_e + \theta_w t_w + \theta_v t_v + \pi_t n_t + \theta_t t_t + \theta_a t_a \quad (3)$$

Where n_t and π_t is the number of transfer and disutility of each transfer. t_e, t_w, t_v, t_t, and t_a is walking time at entry or exit, waiting time, in-vehicle time, transfer time, and walking time to the final destination in minutes, respectively. θ_e=1, θ_w=2, θ_v=1, θ_t=10, and θ_a=5 is disutility of associated PT trip stage, respectively (Sánchez-Martínez, 2017). Although the author claims a total destination inference of 73% (including paper ticket user and transactions with one entry) is achieved, the study does not count for the waiting time due to in-vehicle congestion (denied boarding). The numerical values assigned to the disutility of PT trip stages and validation of the proposed method are yet to perform.

The second trip chaining assumption defined earlier is used in two different forms in literature. Some studies (Barry et al., 2009, Barry et al., 2002, Cui, 2006, Farzin, 2008, Li et al., 2011, Nassir et al., 2011) assume that last stop/station of a passenger is the same as day's first stop/station.

† Studies are performed on entry-exit system.

However, other studies considered that the last stop/station is within walking distance radius from the day's first stop/station. The latter proposition seems more realistic in the case of bus OD estimation, primarily because of having different stop number/location in a different direction (inbound/outbound). Studies such as Cui (2006), and Farzin (2008) assumed all the stops within a certain radius to be a single stop. These studies took the tolerance as 111 m (3.5" or 0.001° of latitude). The use of a cluster of stops as a single stop also copes with the lack of GPS precision. Further, this assumption is correct when there exist longer stops (more than 100 m) in a transit network (Farzin, 2008).

The threshold value for the second assumption was set by Trépanier et al. (2007) as 2000 m. Using entry-exit system data, Alsger et al. (2015) reported that 88% of the passengers return to within 800 m of origin of the day. The study concluded that the average distance between the day's first and last stop is more than 5 km, showing that the remaining 12% of the passengers have very high distance between days' first and last stop (up to 36 km). Another study reported that 72.6% of the total erroneously estimated tOD trips are because of high distance values (more than 800 m) between boarding stop of days' first trips and last stop of the days' last trip (Alsger et al., 2016).

In a subway or rail, the day's symmetry assumption may be valid for many passengers because rail passengers tend to engage in long-distance trips. In contrast, bus passengers have relatively shorter distance trips due to its comparatively local function (Kieu et al., 2015b). More often, bus users may like to go for shopping, buy grocery, or see a friend on the way from work to home and return home using another mode of transport (for instance, walking, car-sharing, para-transit, etc.).

Besides rule-based algorithms, recently, the probabilistic or machine learning models are adopted for alighting inference. The main objective is to relax the trip chaining assumptions due to transit passenger behaviour's complex nature. Cheng et al. (2020) utilised Latent Dirichlet Allocation, a probabilistic model to estimate the alighting stop. The model is trained to predict the destination stop based on the time and stop of origin by employing 3-month smartcard data. The study claims a 2% improvement in the result based individual passenger analysis. Jung and Sohn (2017) used supervised machine learning by employing rectified linear unit with two hidden layers to estimate alighting location. This study used 27 variables related to the smartcard transaction and land use (Table 4). Yan et al. (2019) proposed a two-step algorithm for destination inference – trip chaining, and machine learning. The transactions with non-inferred alighting information from trip-chaining are estimated using diverse machine learning techniques. Following, Assemi et al. (2020) proposed a methodology involving neural network for the same purpose. Usually, the motivation to use different methods for the same approach is to increase the algorithm's accuracy. For instance, Assemi et al. (2020) reported an accuracy of 79.5% of the neural network model compared to a contemporary rule-based model with 72.2% accuracy. More details regarding the validation are given in Table 7.

Table 4 Studies based on machine learning along with the variables used

Study	Method	Variables used
(Jung and Sohn, 2017)	Supervised machine learning	**Transaction related variables:** current and next boarding time, number of transfers, network and Euclidean distance to all candidate stops, inter-transaction time, bus stop densities at upstream and downstream, generalised travel time, and average travel speed **Land use variables:** residential, commercial, cultural, and office floor area in the 500m radius of origin stop; residential, commercial, cultural, and office floor area in the 500m radius of next trip origin stop; residential, commercial, cultural, and office floor area in the 500m radius of candidate alighting stop
(Yan et al., 2019)	Naïve Bayesian, support vector machine, decision tree, random forest, and k-nearest neighbour algorithm	Boarding location and time, number of point of interest and their distribution, and transit route number
(Assemi et al., 2020)	Neural network	Boarding location and time, number of stops (and distance) between the boarding and potential alighting stops, if a transaction is last trip of the day, and estimated trip duration (and distance) between the boarding and alighting at a stop.

The usage of machine learning approaches and probabilistic models is increasing for smartcard data application such as alighting stop estimation, pattern mining, short and long term ridership prediction, etc., for their enhanced predictive capabilities. The machine learning approaches are generally data-hungry, and computationally expansive; however, they do not impact the above smartcard applications since the real-time application is not involved here except for short term ridership prediction (Toqué et al., 2017, Yang et al., 2021).

2.4 Transfer detection

Generally, a card needs to be tapped every time a passenger boards a bus, generating a transaction in the system. In most of the rail or subway PT, the card needs to be tapped while entering the station only and not while changing the train or at the transfer point. Therefore, a single transaction from bus denotes a trip (single boarding and alighting from PT), while that from rail/subway represents a possible journey which can include single or multiple trips.

2.4.1 Description of typical Rules for transfer inference

Researchers have devised guidelines to separate a transfer from activity, of which most widely used are outlined with the help of a figure. Figure 3 depict five scenarios needed to distinguish between the transfer and activity. The bifurcation between transfer and activity is done by applying spatiotemporal and the last transit route constraints. The temporal constraint is known as the maximum transfer time (MTT), and the spatial constraint is termed as maximum transfer distance (MTD). MTT is defined as the time difference between two consecutive alightings and boarding of a smartcard user within a day. MTD is the maximum walking distance between two consecutive alightings and boarding during which a passenger is assumed to have taken transfer from one transit service to another.

If the time duration (t) is greater than MTT or distance between stops (alighting and next boarding stop) (d) is greater than MTD, it will be labelled as an activity. The last transit route constraint is applied by matching the transit route taken on previous and current trip-leg. If the transit route is the same for successive trip-legs, it is considered as an activity without concerning its direction, MTT and MTD. Nevertheless, this rule may not be suitable for ring routes, alternate short and full service, skip-stopping (or limit stop) operation, and during disruptions where users have to change the vehicle without being involved in an activity.

$$\begin{matrix} \textit{Activity and} \\ \textit{transfer inference} \end{matrix} = \begin{cases} t \le MTT, d \le MTD, R_i \ne R_{i+1}, \textit{Transfer} \\ \textit{else,} \qquad\qquad\qquad\quad \textit{Activity} \end{cases} \tag{4}$$

Here, R_i is the transit route taken for trip-leg 'i'. The scenarios portrayed in Figure 3 can be matched with Equation (4) to decide between activity and transfer. Whereas, out of the five scenarios depicted in Figure 3, the first four cases show activity (no-transfer) and the last case represents the transfer.

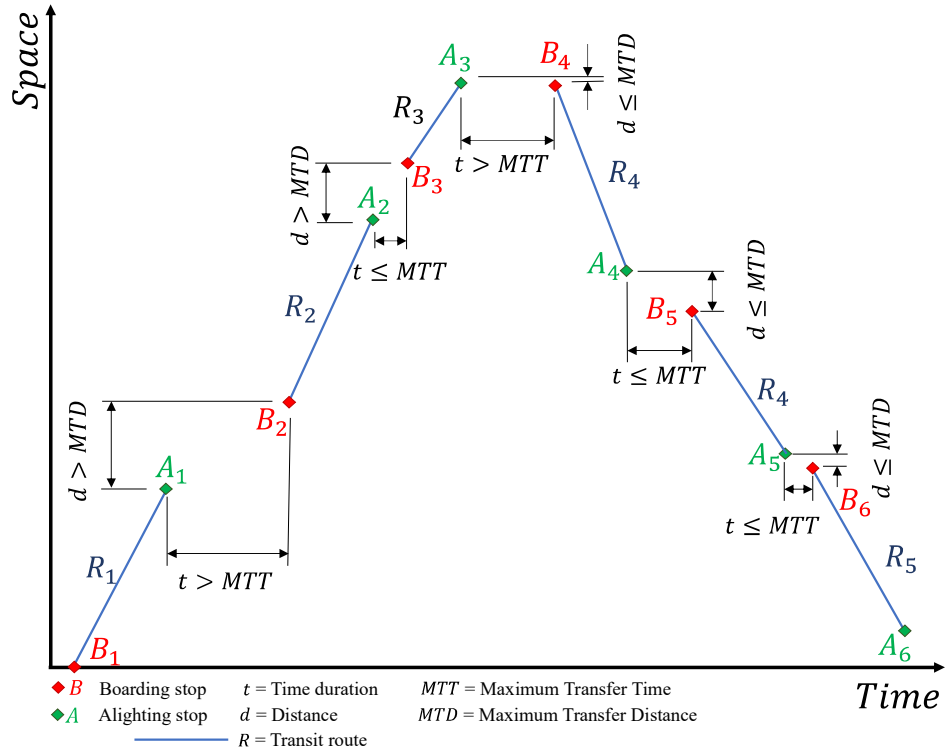

Figure 3 Illustration of various scenarios for an activity or transfer detection employed in literature

Various definitions and values used in literature for MTT and MTD are explained in the following sub-sections.

2.4.2 Maximum transfer time (MTT)

In some studies, MTD is also referred to as Inter-Transaction Time (ITT) because distance can be converted to time by assuming a suitable walking speed value. As shown in Table 5, MTT ranges from 18 minutes to 90 minutes. Huang et al. (2020) adopted the MTT as time less than the transit frequency. Nassir et al. (2011) proposed MTT as a minimum of 90 minutes or time required for walking from alighted station to boarding station plus time needed for a minor activity. If the time between successive alighting and boarding is less than minor activity time (30 minutes), it is labelled as transfer without applying any other rule. The same principle is also employed by Kumar et al. (2018).

Instead of using a single value of MTT, Chu and Chapleau (2008) converted the Euclidean distance between stops to time by dividing the distance with walking speed (taken as 4.3 km/hr). In continuation of this work, Gordon et al. (2013) used MTT corresponding to maximum Euclidean distance of 750 m (assuming the walking speed of approx. 3 km/hr) plus an allowance of 3 minutes. Maximum waiting time for the bus is taken as 45 minutes. Gordon et al. (2013)

further considered if a passenger has not taken the first available transit service; it is labelled as an activity. Later, Yan et al. (2019) and Yap et al. (2017) also utilised the same rule. This assumption is valid in most cases; however, this assumption may lead to underestimating transfers in the case of denied boarding (in-vehicle congestion) and transit disruption. The denied boarding can often occur in peak hours on busy routes and during disruption (Durand et al., 2018). Thereby, Yap et al. (2017) integrated the AFC and AVL to get each trip's occupancy of all transit services. Afterwards, the spare capacity of the first available transit service is calculated. If it does not have enough spare capacity, it is marked as denied boarding, and the event will not be marked as an activity. Since various transit operators tackle disruptions in different ways, therefore, it is required to come up with a more robust methodology to infer transfer in disruption scenario (Sun et al., 2016).

Seaborn et al. (2009) considered a range of transfer time for transfer within several PT modes. The study used MTT of 15-25 minutes, 30-50 minutes, and 40-60 minutes for underground-to-bus, bus-to-underground, and bus-to-bus transfer, respectively. The transfer within the underground can not be inferred as passenger taps-on only when they enter a station and tap-off while they leave the station, i.e., no transaction is recorded when a passenger changes train.

Table 5 Maximum Transfer Time used by various researchers

Time (minutes)	**Studies**
18	(Barry et al., 2009)
30	(Ali et al., 2016, Bagchi and White, 2005, Liu et al., 2019, Munizaga et al., 2014, Munizaga and Palma, 2012)
60	(Alsger et al., 2016, Alsger et al., 2015, Mosallanejad et al., 2019, Nassir et al., 2015, Yan et al., 2019)
90	(Hofmann and Mahony, 2005, Kumar et al., 2018, Nassir et al., 2011)
Variable	(Chu and Chapleau, 2008, Gordon et al., 2013, Seaborn et al., 2009, Yap et al., 2017)
< transit frequency	(Huang et al., 2020)

It is worth noting that MTT depends on the maximum time headway of route under consideration, and the distance between stops. For example, if a next bus route has a headway of 60 minutes, using 30-minute MTT value would exaggerate the number of transfers. Also, special consideration can be given to routes that only operate in morning and evening peaks because passengers may like to wait more to get a particular route bus. Hence, the authors suggest that a reasonable value of MTT must include the headway of transit service taken and the time taken to walk from alighted stop to the next boarding stop.

2.4.3 Maximum transfer distance (MTD)

Li et al. (2011) defined MTD as the standard maximum transfer distance and asserted that it is dependent on the economic status of the city, PT coverage, etc. Various values for MTD used by researchers are summarised in Table 6.

Alsger et al. (2016) used an entry-exit system data from SEQ, Australia. The study estimated the number of trips by taking a transfer distance of 400, 800, 1000, and 1100 m by utilising tap-on data only. The number of trips estimated is compared with the actual number of trips from entry-exit system data. The authors reported that by using an ITT of 60 minutes, the percentage of tOD matching is 72% for 530 m and 86% for 800 m of maximum transfer distance (Alsger et al., 2016). The study also identified two sources of error in the estimation of tOD using an entry-only system that is lack of at least two trips in a day by a passenger, and the distance between alighting and next boarding stop has a considerable effect on the tOD estimation.

Table 6 Maximum transfer distance values adopted in literature

Distance (m)	Studies
400	(Liu and Zhou, 2019, Liu et al., 2019, Mosallanejad et al., 2019, Nassir et al., 2015, Yap et al., 2017, Zhao et al., 2007)
530	(Alsger et al., 2016)
750	(Gordon et al., 2013)
800	(Alsger et al., 2015)
1000	(Munizaga et al., 2014, Yan et al., 2019)
1500	(Huang et al., 2020)

Table 6 reveals that MTD has a minimum of 400 m and a maximum of 1000 m value in literature. As discussed for alighting stop inference (section 2.3.2.2), the maximum transfer distance would depend on many variables like the city's architecture, terrain, demographics, etc. The transfer distance would also depend on PT mode in question (based on transfer between bus-bus, bus-rail, and rail-rail in line with the TCQSM (2013) recommendations). Seaborn et al. (2009) and Gordon et al. (2013) have used variable transfer time for transfer within different modes. However, none of the studies so far used variable transfer distance for transfer between modes. The analysis from Munizaga et al. (2014) shows that the walking distance is dependent on the land use surrounding of stop/station. It would be worth estimating tOD using different transfer distance for transfer within PT modes and validate the methodology by endogenous or exogenous technique to quantify its impact on the tOD estimation. Validation of all the studies quoted in Table 6 is presented in section 2.6 (Table 7).

In addition to abovementioned three rules for activity detection Figure 3, more rules are found in the literature. Details of those rules and guidelines to use them are presented below.

2.4.4 Additional rules for transfer detection

Some studies used the assumptions set by the respective PT agency to infer transfer primarily for fare deduction. For example, in the case of Cui (2006), while transferring, transit passenger used two distinct cards, a standard AFC card, and a transfer card. When a transfer card is tapped, it generated a different type of transaction for a maximum of three legs; hence the study employed the same data. Alsger et al. (2015) used different walking distance and ITT to observe their effect on the tOD matrix. Here, the referenced number of transfer and subsequent tOD is determined by applying the definition for transfer detection set by TransLink (PT agency of Brisbane, Australia). TransLink assumes a transfer if the ITT is less than 60-minutes, and a maximum of three consecutive transfers can be taken. Similarly, Munizaga and Palma (2012) study for Santiago, Chile, reported that the PT agency allows three transfers in 2-hour span. Since the riders know the transfer fare rule in advance, adoption of the same rules for transfer detection can lead to underestimating the actual number of transfers. Application of contrasting rules for transfer inference will lead to different tOD. Thereby, it is crucial to develop, test, and employ a robust and realistic set of transfer detection assumptions.

Gordon et al. (2013) proposed three tests for transfer detection: binary test, temporal test and spatial distance based on simple logic, and temporal and spatial constraints. Binary test separates the transactions which are a final transaction of the day, transactions with no subsequent inferred origin, or no inferred destination because the activity cannot be detected in all such cases. Temporal constraints include the maximum interchange time, which is reliant on the Euclidean distance between two stops and maximum waiting time or the maximum headway. The spatial limit tests consist of the maximum transfer/interchange distance, circuity (the ratio of distance travelled in PT and the Euclidean distance between the start and end for multiple stage journey), and the cumulative angular difference over two or more trips are less than a specific value. Once all three tests are passed, the transaction is considered as a separate journey (activity). Otherwise, it is linked to the same card's previous transaction (i.e., inferred as a transfer). This method's

application gave a transfer ratio of 22% over buses without considering transfers in the subway. The study reported that the percentage of passengers having two transfers in a day is much lower than the estimated when compared with the local travel demand survey. Seaborn et al. (2009) also concluded similar results, where different transfer times, e.g., 15-25 minutes, 30-50 minutes and 40-60 minutes were used for the underground-to-bus transfer, bus-to-underground transfer, and bus-to-bus transfer, respectively.

In the continuation of Gordon et al. (2013) study, Nassir et al. (2015) presented the two-stage transfer detection algorithm. In the first stage, relatively relaxed general constraints are applied, namely, the route taken (for two consecutive trips), temporal filter (60-minute of ITT), and the spatial filter (400 m of maximum transfer distance). The second stage algorithm is based on the following five spatial and temporal criteria, which is claimed to be capable of determining short activity:

- *gap (S_1),* defined as the time gap between alighting and next boarding (less than or equal to 20 minutes);
- *gap ratio (S_2),* the ratio of time gap to total travel time (less than or equal to 0.4);
- *off-optimality (S_3),* travel time difference between the observed trajectory and the fastest path (less than or equal to 20 minutes);
- *off-optimality ratio (S_4),* the ratio of off-optimality to total travel time (less than or equal to 0.5); and
- *circuity (S_5),* the ratio of the sum of the Euclidean distances of each trip leg in the journey to the Euclidean distance between the origin and the destination (less than or equal to 1.7).

The value assigned to each criterion is calibrated using a set of smartcard data of South-East Queensland (SEQ), Australia. A transaction is considered as transfer by calculating:

$$T_f = S_4 \cup \left\{ \left(S_1 \cup S_2 \cup S_3 \right) \setminus S_5 \right\} \tag{5}$$

The fifth criteria (S_5) is previously used by Munizaga et al. (2014) by considering a cut-off value of 2 for on-route distance covered $\left(f_{on-route}\right)$, and Euclidean distance $\left(f_{Euclidean}\right)$ ratio. Nassir et al. (2015) used smartcard data from SEQ, Australia, which is an entry-exit system and validated the proposed algorithm using the Household Travel Survey (HTS) data from 2009. The HTS data contain 290 cases with one or more transfer in a day out of 983 cases of the interchange. The authors claimed an overall transfer detection accuracy of stage 1 and 2 as 99.8%. The study claimed to give better results as compared to other studies; however, the generalisation of the study's results is yet to be done. The tOD estimation model, in general, is area-specific and needs recalibration if intended to be used for another area. Besides, the criterion S_3, S_4, and S_5 can give misleading results in case of transit disruption events such as railway track closure, broken vehicle, etc. (Durand et al., 2018, Yap et al., 2017). Hence care must be exercised when applying these constraints for transfer detection.

In addition to above rule-based transfer detection studies, Liu et al. (2019) employed integer programming and convex quadratic programming optimisation technique to estimate transfer in an entry-only system by utilising tap-only smartcard data and GTFS data. The study compared the results with output from trip chaining method. The results showed r-squared of 0.92 among the stOD matrices estimated by employed integer programming and trip chaining method. Association rule learning and k-mean clustering are also employed for transfer inference (Zhao et al., 2019), where the only variables used in the analysis is ITT. The study reported a median transfer time of fewer than 20 minutes. However, no information is stated on the upper level of ITT. While optimisation and machine learning techniques are not widely used to find transfers, the method is yet to be rigorously validated by an exogeneous dataset.

The literature on smartcard data is still evolving. It is evident from the above discussion that the evaluation of many factors that are vital for transfer or short activity detection is still missing. For example, due to crowding and bus bunching the user behaviour towards route choice may change (Arriagada et al., 2019, Fourie et al., 2016, Yap et al., 2020), as a transit user may opt for next less crowded service, instead of first crowded transit service. Likewise, the effect of transit service travel time reliability on transfer detection is unexplored. Though research is available on the estimation of travel time reliability from smartcard data (Dixit et al., 2019, Lee et al., 2014, Liu et al., 2020), its effect on transfer detection (more specifically, MTT and MTD) is yet to be quantified. On the same line, more comprehensive studies are required to fully identify the effects of various type of disruptions, ring-lines, stop-skipping operations, and alternate short and full line service on the transfers inference rules. Therefore, it is concluded that there is room for improvement in the simple to complex algorithms proposed by researchers.

2.5 Zone to zone OD (ztOD) estimation

First, to continue with this section, stOD, ztOD, and population or scaled tOD are defined. A stop-level OD refers to an OD matrix having stop-to-stop transit trips information on a transit network, while a ztOD have zone-to-zone transit trips information. A population tOD refers to an OD matrix having 100 percent trips made, i.e., there is no missing trip.

AFC data provide information regarding origin stop (in case of the entry-only system), or origin-destination stops (in case of the entry-exit system) of the rider. In both the systems, the AFC transactions record stop-to-stop trip/journey for buses and station-to-station for rail or subway. More specifically, the smartcard data lack the knowledge about how riders access the stop (or station), and how riders reach their destination (absence of first and last mile information).

A few studies in the literature have worked on the inclusion of first and last mile in tOD estimation from smartcard data. Studies (Alsger et al., 2016, Alsger et al., 2015, Assemi et al., 2020, Barry et al., 2009, Farzin, 2008, Li et al., 2011, Munizaga and Palma, 2012, Zhou et al., 2019) have applied basic heuristics to convert stOD matrix to ztOD matrix which may be suitable in the case where walking is done to access/egress a stop.

Assemi et al. (2020), Ali et al. (2016), Munizaga and Palma (2012), Li et al. (2011), and Farzin (2008) assigned a TAZ zone to a stop based on its physical existence, i.e., if a stop lies in the boundary of a zone, it is presumed that all the trips started and ended here belong to the same zone. This assumption works reasonably well for most of the stops in a city, except those stops that lie exactly on the zones' boundary. In most cases, main roads serve as a boundary line for zoning purpose, including TAZ. For example, Figure 4(a) shows a portion of bus stops and rail stations and Brisbane Strategic Transport Model (BSTM) zones. From the figure, it is apparent that most of the stops lie exactly on the dividing line of BSTM zones, which can lead to the false estimation of the ztOD matrix. Figure 4(b) further strengthens this point where it can be seen that buffer area (400 m radius) of most of the stops lies in more than one BSTM zone. Alsger et al. (2016) and Alsger et al. (2015) used the BSTM zone of Brisbane, Australia to convert stOD matrix to ztOD matrix, but the studies lack the details on the method adopted for aggregation.

In addition to the above studies, Amaya et al. (2018) estimated the origin zone of frequent users by calculating the centroid of all the stops where the first transaction of day is made during one week. The results showed over 70% accuracy when validated against an exogenous dataset.

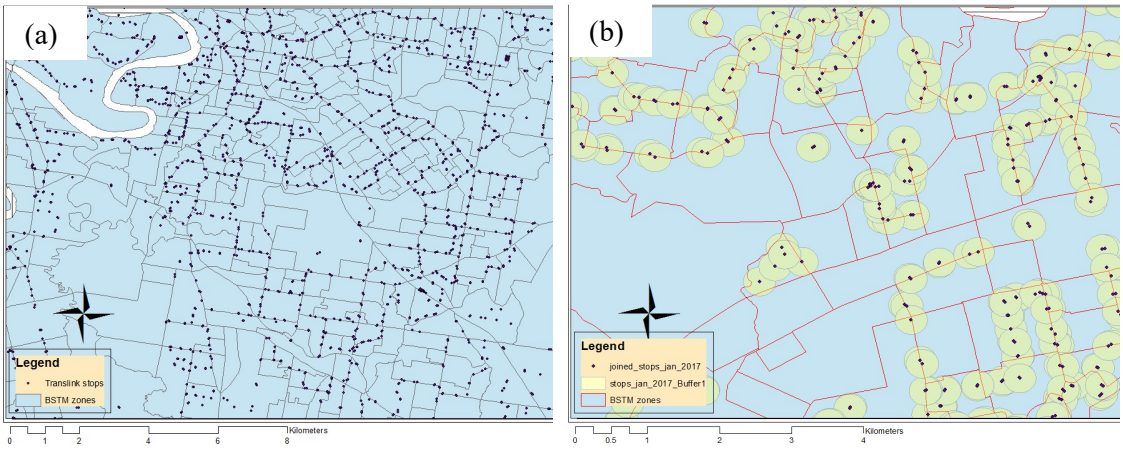

Figure 4 Sample of BSTM zones and TranLink stop of Brisbane (a) Geographical location of stops and BSTM zone boundaries, and (b) buffer of 400 m around stops

Furthermore, few other researchers have provided rationales to group the transit stops. For instance; McCord et al. (2012) proposed an algorithm to aggregate/disaggregate a transit route stop-level OD matrix based on the flow probability. The objective of the study is to reduce tOD matrix size by preserving the OD flow pattern. These methods are devised to combine stops of a single route; their application at a system or network level is not easy. To overcome this problem, Lee et al. (2012) aggregated bus stops from GTFS by incorporating distance between stops, text-similarity in stop names, and land use in the catchment of a stop. The study aims to combine stops at a micro-level and do not aggregate stops at zone-level.

Other studies used more advanced clustering techniques, such as DB-SCAN, k-means clustering, etc., to aggregate the stops at a coarser level. Luo et al. (2017) grouped the transit stops by applying k-means clustering technique. The study minimises the distance between stops within a group and maximises intra-groups flow. By this method, stops can be aggregated efficiently; however, it does not solve the primary task of assigning trip proportion from a stop to all the near zones.

The main focus of the formulation adopted in the above-cited articles is to find ways to group or cluster the transit stops. Notwithstanding, the studies mentioned above provide a heuristic to reduce the tOD matrix dimension for helpful visualisation and other possible applications, like transit route designing, identification of low and high PT usage areas, etc. The proposed methodologies are not able to infer the fraction of trips from the smartcard to predefined origin and or destination zones (such as TAZ) when a stop lies at the border of two or more zones. It is essential to recognise the trips origin-destination on predefined zones to be able to integrate the tOD with other pre-existing strategic and operational models.

Till date, Barry et al. (2009) considered variable other than spatial relation of stop and zone. The authors assigned zones to trip by logit allocation technique and used variables such as walking distance, population or employment and time of the day. But the detail on the methodology of logit allocation technique is not presented in the article. Moreover, Tamblay et al. (2016) presented a zonal inference model which considers distinct variables for trip generation and attraction in a zone. The variables considered are land use (commercial, educational, residential, industrial, offices, and health) and a cost function to access or egress from a stop. Building on previous work, Tamblay et al. (2018) proposed the use of calibrated constants for the aforementioned cost function with the addition of route choice model by employing the demographic variables and trip patterns. Computed ztOD is compared with OD surveys, which shows absolute percentage error of a zone for origin inference range 45-62% while for the destination it ranges 56-60%.

It is apparent that attempts are made to aggregate stOD to ztOD. However, there is a need to develop more robust methodologies incorporating first and last mile problem collectively. The

first and last mile may include but not limited to the effect of park & ride facilities, kiss & ride, free on-street parking, use of any other non-integrated public transport mode on aggregated ztOD.

Once the ztOD is developed, the last step in estimating the final tOD is to compute population tOD. The tOD from smartcard data, usually, does not represent the population tOD. There are various reasons due to which the penetration rate of smartcard is not 100%. For instance: errors in the smartcard data discussed in section 2.1; passengers who do not tap their card while boarding (and alighting) a transit service, e.g., those having a monthly pass, and fare evasion; and those who use a medium other than smartcard for fare payment (e.g., paper tickets). Research studies can be found in the literature, which addresses the scaling problem of tOD matrices estimated from smartcard data. The research is primarily built on the iterative proportional fitting (IPF) method (Ben-Akiva et al., 1985) and doubly-constrained growth factor method.

To determine population tOD, IPF requires a base OD matrix and passenger-counts for each origin and destination. The tOD from smartcard data serves as base OD matrix, and passenger counts are obtained from either on-board surveys, APC, or passenger count surveys at entry and exit of transit facilities (bus stops and train stations). Ji et al. (2015a), Ji et al. (2015b), and Mishalani et al. (2011) presented the application of IPF method by employing the boardings and alightings at transit (bus) stops (APC data) along with other proposed method to find the population flows. The study estimated transit ridership at the transit-route level. Scaled trips are then aggregated for all transit routes operating between two OD pairs (stops). The main disadvantage of the study is that this method does not incorporate transfers. Thereby, the tOD is based on passenger-trips instead of passenger-journeys. To overcome this shortcoming, Gordon et al. (2018) modified IPF method to develop a journey-based tOD matrix. The stOD matrix is scaled using the count data from bus farebox, and station gates.

Recently, Egu and Bonnel (2020) applied a list of rules to produce population ztOD matrix. Trips related to transactions – where the origin is known, but the destination is unknown – are distributed among the destinations with the same proportion as inferred journeys. Further, all flows in the ztOD matrix are enhanced by a factor observed in a local fare evasion survey to incorporate the journeys not recorded by smartcard (for reasons given above). Afterwards, the IPF method is applied on the ztOD where counts are taken from APC data for selected bus and tram stops, and train stations to get the population ztOD.

In addition, the distribution of trips from paper ticket or freedom passes over OD pairs requires further investigation. Currently, they are distributed equally among all the OD pairs (Egu and Bonnel, 2020). However, the literature suggests that the origin/destination of the non-smartcard user may be different than smartcard users (Graham and Mulley, 2012, Tran, 2012).

2.6 Validation of proposed algorithms

One of the most challenging tasks in computing reliable tOD matrix from smartcard data is the validation of the proposed algorithm for tOD estimation. Table 7 summarises how researchers have worked the validation problem so far, where all the related studies are grouped based on the type of validation performed. The alighting only validation type group represent the studies where validation is shown on the alighting location inference only, and so on.

The tOD estimation and validation accuracy are dependent on the input data type (an entry only or an entry-exit system data), and fields available in the smartcard data. As grouped in Table 7, various studies have employed validation on different phases of the tOD estimation process. The data cleansing is part of all the phases. In a study where only transfer detection algorithm is proposed and validated, the error would be the cumulation of error in the data cleansing and the transfer/activity inference phases. For studies where the alighting stop is inferred along with transfer detection, the total error would be the sum of error in data cleansing, transfer/activity detection, and alighting stop inferencing. As a consequence, the combined transfer and alighting stop inference may have low accuracy.

Furthermore, it is essential to understand the omitted data not included in the accuracy measurement. For instance, while working on destination stop inference, it is vital to know if the accuracy measurement includes all the transactions or excludes passengers with only one transaction (trip-leg) per day. The former case may have low accuracy; however, it may be of more interest due to the inclusion of more data in the analysis.

tOD matrix estimation algorithms are validated using endogenous and exogenous validation. Endogenous validation is defined as the validation based on the same (or part of the same) dataset used for algorithm development. Exogenous validation is based on the dataset from an external source, i.e., dataset other than used for algorithm development. The proposed algorithm can be validated by both the methods in an entry-only system. Nonetheless, only the endogenous method is applied in the entry-exit system due to the nature of data involved in the process. It can be noted from Table 7 that most of the studies (Barry et al., 2009, Barry et al., 2002, Farzin, 2008, Gordon et al., 2013, Munizaga et al., 2014, Nassir et al., 2015, Seaborn et al., 2009, Wang et al., 2011, Zhao et al., 2007) used exogenous validation. The endogenous validation is first used by Devillaine et al. (2013) portraying that validation can be done by exploring the derived variables like speed, distances, travel times, and trip stage time from smartcard followed by other studies (Alsger et al., 2016, Alsger et al., 2015, He et al., 2015, Munizaga et al., 2014).

Table 7 Validation methods adopted for various steps of tOD estimation algorithm

Type of validation	Study	Details of validation	Dataset used for validation	Sample size	Results of validation
Alighting only	(Barry et al., 2002)	Trip chaining assumption	Travel dairy of NYMTC[I]	250 rider's data	Assumptions are valid for 90% of the subway users.
	(He et al., 2015)	Alighting location estimation	Smartcard data (SCD) (entry-exit system)	One-month SCD from Brisbane	79% accuracy for alighting location estimation
	(Jung and Sohn, 2017)	Alighting location estimation	SCD (entry-exit system)	20% of one-day transaction (≈165,000)	Testing phase accuracy is 60% (87.5 % with 1-stop tolerance)
	(Yan et al., 2019)	Alighting location estimation	SCD (entry-exit system)	Two-weeks Beijing, China bus SCD	74.4% accuracy for regular users 70.2% accuracy for irregular users
	(Assemi et al., 2020)	Alighting location estimation	SCD (entry-exit system)	Same day SCD from Brisbane	79.5% accuracy for all transactions 95.8% accuracy with an accepted error of one-stop
Transfers only	(Seaborn et al., 2009)	Multi-modal OD inference model	LTDS[II]	12000 individuals	A minimum variance of +3% and the maximum variance of -20% is found between the number of transfers in LTDS data and estimated data.
	(Gordon et al., 2013)	Inference of OD matrix for intermodal journeys	LTDS	NA	Reported to be same as that for (Seaborn et al., 2009)
	(Nassir et al., 2015)	Transfer detection	HTS[III]	1693 journeys	99% accuracy for transfer detection
	(Alsger et al., 2016)	Transfer walking and time	SCD (entry-exit system)	One day SCD 0.18 million transactions	86% accuracy using 800m of MTD and 60 MTT
	(Liu et al., 2019)	Transfer detection using optimization techniques	SCD (trip-chaining)	One-month weekday SC transactions ≈76500	0.92 r-squared between modelled and smartcard OD
Transfers and alighting only	(Farzin, 2008)	Zone-to-zone bus OD matrix	OD Household Survey	N/A	Significant differences between destinations and trip pattern. Passengers with transfers are 50% as compared to the system average of 34%.
	(Alsger et al., 2015)	Transfer walking and time, and days' symmetry assumption	SCD (entry-exit system)	One day SCD (0.5 million transactions)	88% of transfers are detected using 60-min MTT, 800m MTD. 88% riders return to within 800-m radius from the origin.
Boarding and alighting only	(Wang et al., 2011)	Inference of bus OD matrix (route-wise)	BODS[IV]	Route 185 ≈14000 boardings recorded	66% and 65% of BODS surveyed destinations are inferred for each direction.
	(Huang et al., 2020)	Inference of bus OD matrix (route-wise)	Ticket recycling survey	100% data for 72 round trips	72-85% for boarding inference, 70-86% for alighting inference
OD based	(Zhao et al., 2007)[††]	tOD matrix	CTA[V] Customer OD survey	N/A	with a 95% confidence interval, the estimated tOD values are not different than 10%.
	(Barry et al., 2009)	Multi-modal OD inference model	Entrance and exit counts in the subway, ride check data for bus	100% on station entrance & exit, a couple of routes for bus	No quantitative analysis is done.
	(Kumar et al., 2018)	OD estimation	On-board surveys	NA	Graphical results are presented for boarding and alighting comparison.

Type of validation	Study	Details of validation	Dataset used for validation	Sample size	Results of validation
	(Egu and Bonnel, 2020)	OD estimation, transfer algorithm, and OD scaling	HTS and onboard OD surveys (ODS)	HTS includes 164000 households, ODS ≈ 30%	The error in mean trip-legs per journey of AFC—HTS, and AFC—ODS is 2.1% and 8.4%, respectively. The error in total trips is correspondingly 27% and -5.5%. The percentage difference for AFC—HTS in 1-leg journeys, 2-leg journeys, 2+ leg journeys is respectively -1%, -1, and +2%, and that for AFC—ODS is respectively, -8%, +5%, and +3%. The RMSE‡ between AFC—HTS and AFC—ODS scaled tOD is correspondingly, 39% and 21%.
Boarding, transfers, and alighting	(Munizaga et al., 2014)	OD estimation and transfer algorithm	OD metro surveys	601 individuals	Correct estimation of boarding stop is 99%, alighting stop is 84% and 67% for route choice of only metro users.
			Volunteers	586 trips	Transfer inference is 90% correct
			SCD	One-week SCD (35 million transactions)	Suggestions are made based on sensitivity analysis of SCD

†† Partial validation performed, I New York Metropolitan Transportation Council, II London Travel Demand Surveys, III Household Travel Survey, IV Bus passenger Origin-Destination surveys, V Chicago Transit Authority, ‡Root Mean Square Error

The assumptions proposed by Barry et al. (2002), discussed in alighting location estimation section, are validated by using the information from the travel diary of NYMTC. Two hundred and fifty riders having a total of 595 trips are selected for validation in which 100 riders have exactly two trips per day, and 150 riders have more than two trips. The results showed matching of 90% between the proposed assumptions for trip chaining and trip information from travel dairy of NYMTC. The final OD is also compared with tally data (from gates at the station entrance) with a maximum of 4% difference with the estimated passengers. The validation showed promising results, though, the study's limitations include the non-availability of a large dataset for validation from travel dairy data and inherent problems in the tally data. Also, this study is conducted only for subway riders.

Zhao et al. (2007) partially validated the proposed methodology by using extensive Chicago Transit Authority (CTA) customer OD survey. With a confidence interval of 95%, the difference between the estimated passenger mile travelled and calculated from the CTA OD survey is less than 10%.

Farzin (2008) used HTS data from 1997 for validating the proposed procedure. Validation is done based on destination distribution of riders, trip patterns and the total number of transfers. Although conformity between two OD matrices (using HTS survey data and estimated tOD) is not convincing, it is an early attempt to validate the proposed methodology. According to study, the possible reasons for differences are (a) due to the time lag between survey data and current data (estimated OD is for the year 2006, while OD from household survey is for 1997), significant changes may have occurred to the network, (b) the results are for fare card users only, and (c) predefined fare rules for transfer. Apart from reasons given in the study, trips are not separated from journeys based on the time and space definition discussed in transfer/activity detection section; instead, the transactions are directly used which could have led to an overestimation of transfers (Farzin, 2008).

Barry et al. (2009) validated the algorithm using the entrance and exit counts in subway stations and ride check data for bus system. The article does not provide a further discussion on validation. Seaborn et al. (2009) employed LTDS 2006 data for validation by comparing the total number of journeys made in a day and the number of transfers per card per day. The difference in the number of total journeys per weekday is within 5%. PT journeys per passenger per day is 2.05 from survey while that from the proposed algorithm using various MTT values, it is between 2.23 to 2.33. It portrays that the proposed algorithm is overestimating the number of journeys. Also, the difference of the number of transfers per card per day between smartcard data and survey data are found to be +6%, -20%, +7%, +3% for 1, 2, 3, and 4 transfers, respectively. This non-linear trend of difference (positive difference for one, negative for two, and again positive for three and four transfers) between smartcard data and survey data is not explained in the study (Seaborn et al., 2009).

Nonetheless, the same problem is later faced by Gordon et al. (2013) where the results are compared with LTDS and found that transactions with two transfers are overestimated from the adopted method (Gordon et al., 2013). From the non-linear trend of error in transfers, it is concluded that transfer detection is not only dependent on time spent between consecutive alighting and boarding. Instead, investigation of dependency of other variables like the distance between stops, land use, etc. on transfer is inevitable.

Wang et al. (2011) used the Bus passenger Origin-Destination Surveys (BODS) data to validate the tOD estimation algorithm. Out of 5 routes studied, one transit route is validated against BODS data. Difference between the number of boarding in BODS and smartcard data is found between 7-8%. Percentage distribution of alighting pattern from inferred alighting is matched with BODS data, and the results are within a range of ±2% except for one stop which is shopping centre (Wang et al., 2011). Such a finding reflects that land use significantly affects the riders' boarding and alighting distribution.

Munizaga et al. (2014) validated the OD inference algorithm developed in a previous study by Munizaga and Palma (2012). As shown in Table 7, three types of validations are performed, two are exogenous, and one is endogenous. For exogenous validation, the datasets used are from OD metro survey and a group of volunteers. Endogenous validation presents the sensitivity of parameters/assumptions used in the alighting stop inference and transfer detection, i.e., walking distance, the first transaction of the day, single transaction in a day, MTD, and use of a virtual day. Munizaga et al. (2014) proposed using unlike distance values based on land use near a stop to infer alighting stop. Table 7 outlines the findings for the exogenous validation done on OD metro survey data and data based on volunteers. OD metro survey data results may be biased because the survey is conducted in metro stations only, therefore neglecting bus-bus transfer. In volunteers' data, the significant proportion is of students, which can also possibly introduce biases in the results.

Alsger et al. (2015) validated the tOD estimation algorithm developed for in an entry-only system by employing data from an entry-exit system of SEQ, Australia. The study provides a detailed sensitivity analysis of various assumptions, which were proposed for an entry-only system. Three assumptions are tested namely MTT, MTD, and days symmetry assumption. The study revealed that 88% transfers occur using 60 minutes MTT, and 800 m MTD and the value does not improve beyond 800 m MTD. The study further reported that 82% of riders return to their first point of the day, while 88% of riders return to within 800 m of origin (Alsger et al., 2015). Almost, the same results are presented in another study by Alsger et al. (2016). For comparison, the total number of transfers are assumed to be same as recorded by the fare system (considering MTT of 60 minutes), which may not indicate the actual number of transfers. This problem always occurs when performing endogenous validation for transfer detection; therefore, the results cannot be generalized.

Nassir et al. (2015) validated the proposed algorithm using HTS. The proposed two stages algorithm performs reasonably well to detect transfer, as the overall claimed accuracy is above 99%. Transferability of the proposed algorithm is yet to perform because there are five filters in the second stage of the algorithm (discussed in detailed in section 2.4), which need calibration using the smartcard data. Also, the major problem with the validation is the small sample size. Out of 1693 journeys, only 290 data points, which are actual transfers, are used for validation.

Huang et al. (2020) adopted ticket recycling method to validate the OD inference methodology. The study reported a 100% sample size for four routes with 72 round trips consisting of 10,551 trip-legs. The study claims an accuracy of 71.7% to 85.4% for boarding stop inference while for alighting stop the accurateness is 70% to 86.7% calculated. Though the ticket recycling method is an effective way of estimating stop-to-stop OD matrix, it has some inherent problems, such as the transfer detection method, which cannot be tested due to the nature of collected data. It is costly and increases the boarding and alighting time of passengers. Egu and Bonnel (2020) compared their results with HTS and ODS data. As presented in Table 7, the higher trip-legs per journey from AFC data as compared to HTS and ODS methods portray that the current rule-based approach still lacks in detecting the smaller activities.

From the above discussion, it is evident that numerous researchers have used different datasets for validation of their proposed algorithm. HTS or local travel demand surveys are frequently used for this purpose. Due to the high cost of such surveys, the number of available data points for validation is limited. It is required to develop a validation strategy independent of costly data sources to standardized and evaluate the proposed algorithm for tOD estimation.

One of the potential data sources with low cost and high sample size could be the registered users' riders database. Many agencies do register their smartcards; hence at least origin can be robustly estimated with high accuracy. Such data can be used in all origin estimation studies, including the first-mile problem, and destination estimation for the last trip of the day in an entry-only system. Registered users, when combined with users whose smartcard is linked with their credit card, can result in very high sample size. Additionally, some transit agencies do allow credit cards to be directly used in public transit for fare collection purposes. However, it is challenging to get that

data due to apparent security and privacy reasons.

Other heterogeneous big data such as Global System for Mobile Communications (GSM) data (White and Wells, 2002), high-frequency mobile phone location data (Calabrese et al., 2011), and social media data, e.g., Twitter data (Lee et al., 2015) are rarely explored for the validation purpose of tOD estimation (Liu and Zhou, 2019). There are studies which have loosely coupled smartcard and mobile phone data for the tOD analysis, such as Holleczek et al. (2014), and Regt et al. (2017) but their work doesn't directly build on the tOD estimation or validation problem. Studies such as Gu et al. (2017), where smartphone data is used to detect short activities, can be potentially used to validate the transfer inference algorithm. Also, latterly, smartphone location is used to create travel diary, which can be possibly used to validate the transfer inference algorithm, and boarding and alighting information inference algorithms (Imani et al., 2020). However, such data have an inherent problem of low sample size. Nonetheless, these datasets are widely used for various application in transportation; their potential to validate the tOD from smartcard data is yet not extensively reconnoitred.

Moreover, except Liu et al. (2019) and Egu and Bonnel (2020), all studies mentioned in Table 7 performed micro-level validation, i.e., validation is performed on individual data points. Nonetheless, a macro-level validation may provide more robust validation results that can be used to compare the accuracy of various approaches. Besides, it is understandable that macro-level validation requires a large dataset, which may not be available to the researchers in most cases.

3 Future research needs

Currently, the tOD from smartcard data is used by many PT agencies worldwide for planning purpose. However, it is not 100% correct at this point in time (Egu and Bonnel, 2020, Spurr et al., 2018); therefore its usage as an alternative or integration to a more traditional type of modelling, such as steps in four-step modelling needs to be investigated (Harrison et al., 2020).

The proposed future extension works are categorised as the problems with conversion of stop level OD to zonal level, issues with transfer detection algorithm, and some assorted research problems.

3.1 Conversion of stOD to ztOD

To improve the quality of ztOD estimation by aggregating stODs, following research questions need to be addressed:

(1) How to quantify and evaluate the impact of park & ride, kiss & ride, and free on-street parking on the ztOD?

Park & ride play a vital role in the PT operations. There are 18 park & ride sites in Chicago alone, which accommodates more than 6000 spaces. In London and New York there are more than 50 and 20 park & ride sites operating, respectively. The inclusion of park & ride component is crutial for accurate tOD estimation. A park & ride facility can be accessed through motorised modes from a long distance. A trip actual origin/destination zone may be different than the one recorded in smart card data for stations with Park & Ride and similar facilities (provision of free on-street parking in zones (near CBD), Kiss and Ride.

The existing literature on the first and last mile problem should be explored to look at ways to solve the above issue. Furthermore, different data sources to estimate the true origin/destination of such facility users should be explored. For instance, smartcard transaction data generally has the information on the registered users addressed, though such information is not shared due to privacy concerns. Nevertheless, the information at the residential zone of the user can be shared. Alternatively, other datasets such as dedicated surveys and number plate matching in these facilities can be conducted to understand their origin/destination choice, trip-purpose, and reasons

for selecting a facility.

(2) How to quantify and incorporate the induced transit demand into tOD from smartcard data?

The tOD estimated from smartcard data is highly biased towards the transit supply, i.e., there may exist higher PT demand in a zone where the observed trips may be limited to the non-availability of transit service (Lu et al., 2020). Thereby, the tOD from smartcard data gives the served demand. Total transit demand is the sum of served demand (tOD from smartcard) and induced (potential) demand in a zone (Hussain et al., 2020a), where the later is generally unknown.

One potential solution to identify induced demand is to divide the whole study area zones into two: high demand zones and low demand zones. A model for high demand zones can be developed by employing socio-economic and demographic characteristics. The calibrated model can then be utilised to predict all low demand zones' induced demand by assuming that high demand zones have no potential demand.

3.2 Transfer detection and trip chaining assumptions

To distinguish between transfer and activity following research questions are identified. It also includes future research questions related to trip chaining.

(1) Can variable transfer time and variable transfer distance instead of static MTT and MTD, respectively, enhance the performance of transfer detection heuristic?

Most of the studies have employed static MTT to detect transfer except Seaborn et al. (2009) and Gordon et al. (2013). Some studies modelled variability in MTT by considering transfer type (train to bus, bus to train, or bus to bus), and headway of the transit service. Further research is needed to define variable MTT based on other factors such as terrain (flat, mountainous, undulating), type of stop/station (at grade, underground, elevated), headway of buses (preferably the actual arrival time of transit service), actual distance between stops, land use (e.g. residential, commercial, educational, recreational), type of city planning (grid, radial, mixed), trips with a tentative purpose (as dissimilar purpose have different activity time)., along with transfer type, and transit service headway. Application of dynamic MTT in tOD estimation becomes inevitable specifically in cases, where the transfer stations in the multi-modal network are not linked. For instance, Fairfield station, Brisbane, Australia has a minimum distance of 450 m to the nearest bus stop. While Roma Street station, Brisbane, Australia has integrated stops and can be accessed by changing the platform only. Hence, both the station may define separate criteria for MTT values. Likewise, similar heuristics can be delineated for MTD.

(2) What is the effect of considering a subset of total available PT modes or agencies (penetration rate and spatial availability of transactions, respectively) on the first assumption of trip chaining, and transfer detection (i.e., MTT and MTD)?

Studies on a multi-modal network where analysis is carried for any one mode (or does not consider one or more transit modes for analysis) defies first assumptions of trip chaining, which states that travellers do not use another mode of transport. Violation of trip chaining assumption may lead to a wrong inference of the results. The same logic can be applied on transfer detection algorithm for a person who can travel more distance in lesser time by using non-integrated transit services, without being involved in an activity. Hence, in future, it is worth exploring the effect of one modal (or one agency) on the maximum distance assumed for the first assumption of trip chaining, and MTT and MTD of transfer detection. Also, free transit services, for instance, free loop (route 30, 40, and 50 in Brisbane, Australia) can lead the false transit OD within its operational radius, because the passengers using these services cannot be tracked in smartcard data. Alsger et al. (2017) calculated tOD at various penetration rate and spatial distribution of the transactions. The study concluded that while penetration rate is significant for tOD calculation; it

becomes less vital when increased more than 60%. On the other hand, there is a considerable effect of spatial uniformity of smartcard transactions on tOD estimation. However, the sensitivity of using fewer modes or agency data on an individual component of tOD is yet to be investigated.

(3) How to quantify the impact of ridesourcing services usage on the MTT and MTD in transfer detection, and on the maximum distance between the alighting stop of the previous trip and next boarding stop for alighting stop inference in trip chaining method?

In last five years or so, globally, there is an exponential growth in the demand for ridesourcing services, such as Uber, DiDi, etc., and e-scooter companies, such as Lime, Lyft, etc. The increased usage of such services can pose questions on the assumptions made about the maximum walking distance in trip chaining and transfer detection. For instance, in the case of Lime e-scooter, the average trip-length is reported as 1.66 km (1.06 mile) where 27% of total 6 million trips are made globally to access or egress public transport (Lime, 2018). Likewise, independent transit services, such as Airlift, which is not integrated into the transit smartcard also lie in the same category. With the increased use of such services, the adopted values of MTD for transfer detection in section 2.4.3 may not give accurate results. Since users can travel a longer distance in less time, hence wrongly inferring an activity. Besides, these services also impact the assumption related to the disutility of non-public transit modes in trip chaining (the first assumption in section 2.3.2). Therefore, to estimate true tOD from smartcard data, it is imperative to find ways to quantify its effect on MTT and maximum distance between alighting and next boarding stop.

For this purpose, it may be effective to know the public transit passenger behaviour towards the utility of ridesourcing services in a local condition enabling the researcher to modify the current rules (or threshold values of currently implemented rules). It can be done by conducting longitudinal surveys to track users for a longer duration. Alternatively, other big data sources such as GSM data and/or smartphone-based data can be integrated with smartcard data, giving individual trip-leg mode (Huang et al., 2019, Nikolic and Bierlaire, 2017).

(4) How to quantify and validate the effect of planned and unplanned disruptions, road congestion, and in-vehicle congestion on transfer detection algorithm?

The existing literature on the tOD estimation from smartcard data ignores the increased waiting time in peak hour due to congestion, and limited capacity or denied boarding. This high waiting time can increase passenger transfer time without being involved in any activity. There exists a study which takes into account the transit disruption in the transfer detection algorithm (Yap et al., 2017); however, various type of disruptions can have a distinct impact on transit users. Also, various transit operators deal differently with a separate type of disruptions. For example, in case of signal failure on rail routes, Yap et al. (2017) reported that users take an alternate route to reach their destination, while if that happens in Brisbane, Australia, buses (called train buses) are deployed between the specified stations (Deng et al., 2018, Translink, 2020). Therefore, it is vital to develop and validate more robust transfer detection algorithm that can be equally applied in all (or most of the) planned and unplanned transit disruptions events.

To solve the above issue, it may be helpful to integrate the smartcard data with another dataset such as AVL data, archived GTFS-live data, or any other dataset providing information regarding disruptions and congestions. At the time and place where disruption/congestion is identified, a relaxed or modified criterion for transfer detection must be introduced.

3.3 Miscellaneous research problems

Potential miscellaneous research problems that may need further exploration are as follows:

(1) What is the effect of paper ticket user or token users on the overall tOD matrix?

In tOD estimation, paper ticket user or token user's travel pattern is considered as same as that of card users. This assumption needs more evaluation as travel behaviour of paper ticket users may

significantly be different than those using smartcard as evident from literature (Graham and Mulley, 2012, Tran, 2012).

For this purpose, the destinations highly likely visited by paper ticket users such as tourist hotspots can be assessed and compared with other destinations (e.g., shopping centres) to quantify differences in the paper ticket and smartcard users' travel pattern.

(2) Can the inclusion of land use of the study area enhance the algorithm prediction rate for alighting stop estimation?

Until recent, land use is not considered to estimate alighting stop as highlighted by Li et al. (2018), Faroqi et al. (2018) and Munizaga and Palma (2012), which can potentially decrease the error in tOD estimation.

(3) How to quantify the effect of assumption of single card association with one user on tOD matrix estimation from smartcard data?

Condition of using the same smartcard by a single person throughout the day, which is a hidden assumption, is not tested except by Chu and Chapleau (2008). The study reported that the smartcard includes the cardholder's photo, which means that a single user uses a particular card. In Brisbane's, Australia case, only cards with concessions can be checked to validate the identification and entitlement to use the card. Hence, it is worth quantifying the error introduced in tOD due to the aforementioned assumption.

(4) Can fusing various big datasets, such as GSM data, smartphone-based high location data, loop detector data, etc. with smartcard data help in providing better tOD estimation and validation?

Recently, researchers are moving from trip-chaining method to supervised machine learning techniques to infer alighting location mainly because of high predictability of machine learning methods and to relax the assumptions of trip chaining. It is expected that attributes from smartcard data when combined with the features of other big datasets, such as smartphone location data, will provide more accurate results. For instance, Wu et al. (2018) proposed a method to estimate travel demand using data from various sources (smartphone type devices, HTS, GPS, and sensors). Likewise, Harrison et al. (2020) also provide guidelines to improve the transportation system a whole by integrating various datasets. Apart from tOD estimation, the same datasets can also be explored for validation of tOD algorithm. This area of research is yet to explored and is anticipated to improve the existing tOD estimation and validation algorithms.

4 Conclusion

This paper outlines the current knowledge and research gaps in the literature on the use of smartcard data to estimate ztOD. While most of the ztOD estimation process may be considered as well accepted among the researchers throughout the world, there is a need to present the process framework in a more robust and understandable way to be used by practitioners. The main topics covered are data cleansing; estimation of unknowns (i.e., boarding/alighting location estimation, and transfer detection); validation of proposed algorithms, and conversion of stOD to ztOD. Following are the significant findings and conclusions of the article.

- It is essential to clean the big data for inconsistencies due to equipment and human errors as inconsistent data may contribute up to two percent.
- To estimate alighting location in an entry-only system, selection of threshold distances for underlying assumptions of trip chaining model, i.e., the disutility of non-PT modes and days' symmetry assumption, need verification for local conditions. In literature, these values range between 530 – 1100 m and 0 – 2000 m, respectively.

- For transfer detection, threshold values of MTD and MTT used by various researchers vary and range between 400 – 1500 m and 18 – 90 minutes, respectively. Therefore, calibration of these values will better represent the actual condition, hence producing more realistic tOD.
- There is a need to formulate and test a more robust and cost-efficient method for validating overall proposed tOD estimation algorithms.
- The current method of ztOD estimation from stOD is not satisfactory and may cause a considerable error. Therefore, further research is needed to refine the aggregation process.
- The broad issues that are recommended for further investigation include i) conversion of stOD to ztOD, ii) transfer detection, iii) assumption about paper ticket user and single-person use of a smartcard, iv) inclusion of land use for destination inference, v) the effect of ridesourcing and e-scooter utility on assumptions of trip chaining and transfer detection, vi) integration of other big datasets for tOD estimation, and vii) the effect of planned and unplanned disruptions on tOD estimation.

5 References

Ali, A., Kim, J. & Lee, S. 2016. Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering,* 20**,** 1532-1539.

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,* 68**,** 490-506.

Alsger, A., Mesbah, M., Ferreira, L. & Safi, H. 2015. Use of smart card fare data to estimate public transport origin–destination matrix. *Transportation Research Record: Journal of the Transportation Research Board***,** 88-96.

Alsger, A., Tavassoli, A., Mesbah, M. & Ferreira, L. 2017. Evaluation of Effects from Sample-Size Origin-Destination Estimation Using Smart Card Fare Data. *Journal of Transportation Engineering, Part A: Systems,* 143**,** 04017003.

Amaya, M., Cruzat, R. & Munizaga, M. A. 2018. Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. *Journal of Transport Geography,* 66**,** 330-339.

Arriagada, J., Gschwender, A., Munizaga, M. A. & Trépanier, M. 2019. Modeling bus bunching using massive location and fare collection data. *Journal of Intelligent Transportation Systems,* 23**,** 332-344.

Assemi, B., Alsger, A., Moghaddam, M., Hickman, M. & Mesbah, M. 2020. Improving alighting stop inference accuracy in the trip chaining method using neural networks. *Public Transport,* 12**,** 89-121.

Bagchi, M. & White, P. R. 2005. The potential of public transport smart card data. *Transport Policy,* 12**,** 464-474.

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,* 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. *Transportation Research Record: Journal of the Transportation Research Board,* 1817**,** 183-187.

Behara, K. N. S., Bhaskar, A. & Chung, E. 2020. A Novel Methodology to Assimilate Sub-Path Flows in Bi-Level OD Matrix Estimation Process. *IEEE Transactions on Intelligent Transportation Systems*, 1-11.

Ben-Akiva, M., Macke, P. P. & Hsu, P. S. 1985. Alternative methods to estimate route-level trip tables and expand on-board surveys. *Transportation Research Record*, 1-11.

Bracher, B. & Bogenberger, K. 2018. Modelling the Long-Term Modal Effects of a Dynamic Congestion Charging Zone. *Transportation Research Board 97th Annual Meeting.* Washington DC, United States.

Calabrese, F., Lorenzo, G. D., Liu, L. & Ratti, C. 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing,* 10**,** 36-44.

Cats, O., Vermeulen, A., Cebecauer, M., Jenelius, E. & Susilo, Y. Generating network-wide travel diaries using smartcard data. Transit data 2019, 5th Internaitonal workshop and symposium, 8-10 July 2019 Paris, France.

Cats, O., Wang, Q. & Zhao, Y. 2015. The identification and classification of urban centres using public transport passenger flows data. *Journal of Transport Geography,* 48**,** 10-22.

Ceder, A., Butcher, M. & Wang, L. 2015. Optimization of bus stop placement for routes on uneven topography. *Transportation Research Part B: Methodological,* 74**,** 40-61.

Chapleau, R., Trépanier, M. & Chu, K. K. 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, 2008. 25-31.

Chen, Z. & Fan, W. 2018. Extracting bus transit boarding stop information using smart card transaction data. *Journal of Modern Transportation*.

Cheng, Z., Trépanier, M. & Sun, L. 2020. Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation*.

Choudhury, C. F., Ben-Akiva, M., Rapolu, S. R., Emmonds, A. & Rajiwade, S. S. 2011. Evaluating the impact of interventions on network capacity. *Transportation Research Board 90th Annual Meeting.* Washington DC, United States.

Chu, K. A. & Chapleau, R. 2008. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board,* 2063**,** 63-72.

Cipriani, E. & Fusco, G. 2004. Combined signal setting design and traffic assignment problem. *European Journal of Operational Research,* 155**,** 569-583.

Cui, A. 2006. *Bus passenger Origin-Destination Matrix estimation using Automated Data Collection system.* Masters in Transportation Engineering, Massachusetts Institute of Technology.

Deng, Y., Ru, X., Dou, Z. & Liang, G. J. S. 2018. Design of bus bridging routes in response to disruption of urban rail transit. 10**,** 4427.

Devillaine, F., Munizaga, M., Palma, C. & Zúñiga, M. 2013. Towards a Reliable Origin-Destination Matrix from Massive Amounts of Smart Card and GPS Data: Application to Santiago. *Transport Survey Methods: Best Practice for Decision Making.* Emerald Group Publishing Limited.

Dixit, M., Brands, T., van Oort, N., Cats, O. & Hoogendoorn, S. 2019. Passenger Travel Time Reliability for Multimodal Public Transport Journeys. *Transportation Research Record,* 2673**,** 149-160.

DoD, U. 2008. Global positioning system standard positioning service performance standard, 4th Ed. *Assistant secretary of defense for command, control, communications, and intelligence*.

Durand, A., Van Oort, N. & Hoogendoorn, S. 2018. Assessing and Improving Operational Strategies for the Benefit of Passengers in Rail-Bound Urban Transport Systems. *Transportation Research Record: Journal of the Transportation Research Board,* 2672**,** 421-430.

Egu, O. & Bonnel, P. 2020. How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? An empirical investigation in Lyon. *Transportation Research Part A: Policy and Practice,* 138**,** 267-282.

Ellison, R. B., Ellison, A. B., Greaves, S. P. & Sampaio, B. 2017. Electronic ticketing systems as a mechanism for travel behaviour change? Evidence from Sydney's Opal card. *Transportation Research Part A: Policy and Practice,* 99**,** 80-93.

Faroqi, H., Mesbah, M. & Kim, J. 2018. Applications of transit smart cards beyond a fare collection tool: a literature review. *Advances in Transportation Studies,* 45.

Farzin, J. 2008. Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. *Transportation Research Record: Journal of the Transportation Research Board,* 2072**,** 30-37.

Fourie, P. J., Erath, A. L., Ordóñez Medina, S. A., Chakirov, A. & Axhausen, K. W. 2016. Using smartcard data for agent-based transport simulation. *Public Transport Planning with Smart Card Data.* CRC Press.

Ghani, F., Rachele, J. N., Loh, V. H. Y., Washington, S. & Turrell, G. 2018. Do differences in built environments explain age differences in transport walking across neighbourhoods? *Journal of Transport & Health*.

Ghani, F., Rachele, J. N., Washington, S. & Turrell, G. 2016. Gender and age differences in walking for transport and recreation: Are the relationships the same in all neighborhoods? *Preventive Medicine Reports,* 4**,** 75-80.

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,* 2343**,** 17-24.

Gordon, J. B., Koutsopoulos, H. N. & Wilson, N. H. M. 2018. Estimation of population origin–interchange–destination flows on multimodal transit networks. *Transportation Research Part C: Emerging Technologies,* 90**,** 350-365.

Graham, P. & Mulley, C. 2012. Public transport pre-pay tickets: Understanding passenger choice for different products. *Transport Policy,* 19**,** 69-75.

Gu, W., Zhang, K., Zhou, Z., Jin, M., Zhou, Y., Liu, X., Spanos, C. J., Shen, Z.-J., Lin, W.-H. & Zhang, L. 2017. Measuring fine-grained metro interchange time via smartphones. *Transportation Research Part C: Emerging Technologies,* 81**,** 153-171.

Guozhen, T., Lidong, L., Fan, W. & Yaodong, W. Dynamic OD estimation using Automatic Vehicle Location information. 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference, 20-22 Aug. 2011 2011. 352-355.

Han, G. & Sohn, K. 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transportation Research Part B: Methodological,* 83**,** 121-135.

Harrison, G., Grant-Muller, S. M. & Hodgson, F. C. 2020. New and emerging data forms in transportation planning and policy: Opportunities and challenges for "Track and Trace" data. *Transportation Research Part C: Emerging Technologies,* 117**,** 102672.

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.

Hensher, D. A. & Button, K. J. 2008. *Handbook of transport modelling*, Emerald Group Publishing Limited.

Hofmann, M. & Mahony, M. O. Transfer journey identification and analyses from electronic fare collection data. Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005., 13-15 Sept. 2005 2005. 34-39.

Holleczek, T., Yu, L., Lee, J. K., Senn, O., Ratti, C. & Jaillet, P. 2014. Detecting weak public transport connections from cellphone and public transport data. *Proceedings of the 2014 International Conference on Big Data Science and Computing.* Beijing, China: Association for Computing Machinery.

Huang, D., Yu, J., Shen, S., Li, Z., Zhao, L. & Gong, C. 2020. A Method for Bus OD Matrix Estimation Using Multisource Data. *Journal of Advanced Transportation,* 2020.

Huang, H., Cheng, Y. & Weibel, R. 2019. Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies,* 101**,** 297-312.

Hussain, E., Behara, K. N., Bhaskar, A. & Chung, E. 2020a. A Framework for the Comparative Analysis of Multi-Modal Travel Demand: Case study on Brisbane Network. *Submitted to IEEE Transaction on Intelligenet Transportation Systems*.

Hussain, E., Bhaskar, A. & Chung, E. 2020b. A novel origin destination based transit supply index: Exploiting the opportunities with big transit data. *Submitted to Journal of Transport Geography*.

Hussain, E., Bhaskar, A. & Chung, E. 2020c. Zone prioritization for transit improvement using potential transit demand estimated from smartcard data , QUT eprints.

Imani, A. F., Harding, C., Srikukenthiran, S., Miller, E. J. & Habib, K. N. 2020. Lessons from a Large-Scale Experiment on the Use of Smartphone Apps to Collect Travel Diary Data: The "City Logger" for the Greater Golden Horseshoe Area. *Transportation Research Record: Journal of the Transportation Research Board***,** 1-12.

Ji, Y., Mishalani, R. G. & McCord, M. R. 2015a. Transit passenger origin–destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. *Transportation Research Part C: Emerging Technologies,* 58**,** 178-192.

Ji, Y., You, Q., Jiang, S. & Zhang, H. M. 2015b. Statistical inference on transit route-level origin–destination flows using automatic passenger counter data. *Journal of advanced transportation,* 49**,** 724-737.

Jun, C. & Dongyuan, Y. 2013. Estimating Smart Card Commuters Origin-Destination Distribution Based on APTS Data. *Journal of Transportation Systems Engineering and Information Technology,* 13**,** 47-53.

Jung, J. & Sohn, K. 2017. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems,* 11**,** 334-339.

Kieu, L.-M., Bhaskar, A. & Chung, E. 2015a. A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data. *Transportation Research Part C: Emerging Technologies,* 58**,** 193-207.

Kieu, L.-M., Bhaskar, A. & Chung, E. 2015b. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems,* 16**,** 1537-1548.

Kieu, L. M., Bhaskar, A. & Chung, E. 2015c. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems,* 16**,** 1537-1548.

Kumar, P., Khani, A. & He, Q. 2018. A robust method for estimating transit passenger trajectories using automated data. *Transportation Research Part C: Emerging Technologies,* 95**,** 731-747.

Kusakabe, T. & Asakura, Y. 2014. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies,* 46**,** 179-191.

Lahat, D., Adali, T. & Jutten, C. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE,* 103**,** 1449-1477.

Lee, A., van Oort, N. & van Nes, R. 2014. Service Reliability in a Network Context: Impacts of Synchronizing Schedules in Long Headway Services. *Transportation Research Record: Journal of the Transportation Research Board,* 2417**,** 18-26.

Lee, J. H., Gao, S. & Goulias, K. G. Can Twitter data be used to validate travel demand models. 14th international conference on travel behaviour research, 2015.

Lee, S. G. & Hickman, M. 2014. Trip purpose inference using automated fare collection data. *Public Transport,* 6**,** 1-20.

Lee, S. G., Hickman, M. & Tong, D. 2012. Stop Aggregation Model: Development and Application. *Transportation Research Record: Journal of the Transportation Research Board,* 2276**,** 38-47.

Li, D., Lin, Y., Zhao, X., Song, H. & Zou, N. Estimating a Transit Passenger Trip Origin-Destination Matrix Using Automatic Fare Collection System. 2011 Berlin, Heidelberg. Springer Berlin Heidelberg, 502-513.

Li, T., Sun, D., Jing, P. & Yang, K. 2018. Smart Card Data Mining of Public Transport Destination: A Literature Review. *Information,* 9**,** 18.

Lime 2018. One Year Report, https://www.li.me/hubfs/Lime_Official_One_Year_Report.pdf, Accessed on: 8/7/2020.

Liu, J., Schonfeld, P. M., Peng, Q. & Yin, Y. 2020. Measures of Travel Reliability on an Urban Rail Transit Network. *Journal of Transportation Engineering, Part A: Systems,* 146**,** 1-14.

Liu, J. & Zhou, X. 2019. Observability quantification of public transportation systems with heterogeneous data sources: An information-space projection approach based on discretized space-time network flow models. *Transportation Research Part B: Methodological,* 128**,** 302-323.

Liu, X., Hentenryck, P. V. & Zhao, X. 2019. Optimization Models for Estimating Transit Network Origin-Destination Flows with AVL/APC Data. *arXiv preprint arXiv:.05777*.

Lu, K., Liu, J., Zhou, X. & Han, B. 2020. A Review of Big Data Applications in Urban Transit Systems. *IEEE Transactions on Intelligent Transportation Systems***,** 1-18.

Luo, D., Cats, O. & van Lint, H. 2017. Constructing Transit Origin–Destination Matrices with Spatial Clustering. *Transportation Research Record: Journal of the Transportation Research Board,* 2652**,** 39-49.

Ma, X.-l., Wang, Y.-h., Chen, F. & Liu, J.-f. 2012. Transit smart card data mining for passenger origin information extraction. *Journal of Zhejiang University-Science C-Computers & Electronics,* 13**,** 750-760.

Ma, X., Liu, C., Wen, H., Wang, Y. & Wu, Y.-J. 2017. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography,* 58**,** 135-145.

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,* 36**,** 1-12.

McCord, M. R., Mishalani, R. G. & Hu, X. 2012. Grouping of Bus Stops for Aggregation of Route-Level Passenger Origin–Destination Flow Matrices. *Transportation Research Record: Journal of the Transportation Research Board,* 2277**,** 38-48.

Mishalani, R. G., Ji, Y. & McCord, M. R. 2011. Effect of Onboard Survey Sample Size on Estimation of Transit Bus Route Passenger Origin–Destination Flow Matrix Using Automatic Passenger Counter Data. *Transportation Research Record,* 2246**,** 64-73.

Morency, C., Trépanier, M. & Agard, B. 2007. Measuring transit use variability with smart-card data. *Transport Policy,* 14**,** 193-203.

Mosallanejad, M., Somenahalli, S., Vij, A. & Mills, D. 2019. An Approach to Distinguish Destination from the Alighting Stop based on Fare Data. *Journal of the Eastern Asia Society for Transportation Studies,* 13**,** 1348-1360.

Munizaga, M. A., Devillaine, F., Navarrete, C. & Silva, D. 2014. Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies,* 44**,** 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,* 24**,** 9-18.

Nassir, N., Hickman, M. & Ma, Z.-L. 2015. Activity detection and transfer identification for public transit fare card data. *Transportation,* 42**,** 683-705.

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***,** 140-150.

Nikolic, M. & Bierlaire, M. 2017. Review of transportation mode detection approaches based on smartphone data. *17th Swiss Transport Research Conference (STRC).* Monte Verità / Ascona.

Nunes, A. A., Dias, T. G. & Cunha, J. F. e. 2016. Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation. *IEEE Transactions on Intelligent Transportation Systems,* 17**,** 133-142.

Pell, A., Nyamadzawo, P. & Schauer, O. 2016. Intelligent transportation system for traffic and road infrastructure-related data. *International Journal of Advanced Logistics,* 5**,** 19-29.

Pelletier, M.-P., Trépanier, M. & Morency, C. 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies,* 19**,** 557-568.

Rao, W., Wu, Y.-J., Xia, J., Ou, J. & Kluger, R. 2018. Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transportation Research Part C: Emerging Technologies,* 95**,** 29-46.

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S. & Waller, T. S. 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies,* 75**,** 197-211.

Regt, K. d., Cats, O., Oort, N. V. & Lint, H. v. 2017. Investigating Potential Transit Ridership by Fusing Smartcard and Global System for Mobile Communications Data. *Transport Research Record: Journal of the Transportation Research Board,* 2652**,** 50-58.

Sánchez-Martínez, G. E. 2017. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board,* 2652**,** 1-7.

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,* 2121**,** 55-62.

Spurr, T., Leroux, A. & Chapleau, R. 2018. Comparative Structural Evaluation of Transit Travel Demand using Travel Survey and Smart Card Data for Metropolitan Transit Financing. 2672**,** 807-816.

Stopher, P. R. & Greaves, S. P. 2007. Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice,* 41**,** 367-381.

Sun, H., Wu, J., Wu, L., Yan, X. & Gao, Z. 2016. Estimating the influence of common disruptions on urban rail transit networks. *Transportation Research Part A: Policy and Practice,* 94**,** 62-75.

Tamblay, S., Galilea, P., Iglesias, P., Raveau, S. & Muñoz, J. C. 2016. A zonal inference model based on observed smart-card transactions for Santiago de Chile. *Transportation Research Part A: Policy and Practice,* 84**,** 44-54.

Tamblay, S., Muñoz, J. C. & de Dios Ortúzar, J. 2018. Extended Methodology for the Estimation of a Zonal Origin-Destination Matrix: A Planning Software Application Based on Smartcard Trip Data. *Transportation Research Record: Journal of the Transportation Research Board,* 2672**,** 859-869.

Tang, T., Liu, R., Choudhury, C. & Society 2020. Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. *Sustainable Cities,* 53**,** 101927.

Tavassoli, A., Mesbah, M. & Hickman, M. 2018. Application of smart card data in validating a large-scale multi-modal transit assignment model. *Public Transport,* 10**,** 1-21.

TCQSM 2013. Transit Capacity and Quality of Service Manual. *TCRP report 165.* 3rd ed. Washington, D.C.

TCSMS 2017. TransLink Customer Satisfaction Monthly Snapshot "Translink, Transportation and Main Roads, South East Queensland (SEQ), Australia".

Toqué, F., Khouadjia, M., Come, E., Trepanier, M. & Oukhellou, L. Short & long term forecasting of multimodal transport passenger flows with machine learning methods. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 16-19 Oct. 2017 2017. 560-566.

Tran, W. Analysis of the differences in travel behaviour between pay as you go and season ticket holders using smart card data. 1st Civil and Environmental Engineering Student Conference, 2012.

Translink, T. M. R., South East Queensland (SEQ), Australia 2016. Translink Tracker October-December 2016 Q2.

Translink, T. M. R., South East Queensland (SEQ), Australia 2020. https://translink.com.au/service-updates/rail-replacement-bus-stops, Accessed on: 8/7/2020.

Trépanier, M., Tranchant, N. & Chapleau, R. 2007. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems,* 11**,** 1-14.

Tu, W., Cao, R., Yue, Y., Zhou, B., Li, Q. & Li, Q. 2018. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *Journal of Transport Geography,* 69**,** 45-57.

Wang, W., Attanucci, J. P. & Wilson, N. H. 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation,* 14**,** 7.

Wang, Z. j., Li, X. h. & Chen, F. 2015. Impact evaluation of a mass transit fare change on demand and revenue utilizing smart card data. *Transportation Research Part A: Policy and Practice,* 77**,** 213-224.

White, J. & Wells, I. 2002. Extracting origin destination information from mobile phone data. *IET Conference Proceedings* [Online]. Available: https://digital-library.theiet.org/content/conferences/10.1049/cp_20020200.

Wu, X., Guo, J., Xian, K. & Zhou, X. 2018. Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. *Transportation Research Part C: Emerging Technologies,* 96**,** 321-346.

Yan, F., Yang, C. & Ukkusuri, S. V. 2019. Alighting stop determination using two-step algorithms in bus transit systems. *Transportmetrica A: Transport Science,* 15**,** 1522-1542.

Yang, X., Xue, Q., Ding, M., Wu, J. & Gao, Z. 2021. Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data. *International Journal of Production Economics,* 231.

Yap, M. & Cats, O. 2020. Predicting disruptions and their passenger delay impacts for public transport stops. *Transportation***,** 1-29.

Yap, M., Cats, O. & van Arem, B. 2020. Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science,* 16**,** 23-42.

Yap, M. D., Cats, O., Oort, N. v. & Hoogendoorn, S. P. 2017. A robust transfer inference algorithm for public transport journeys during disruptions. *Transportation Research Procedia,* 27**,** 1042-1049.

Zhang, F., Zhao, J., Tian, C., Xu, C., Liu, X. & Rao, L. 2016. Spatiotemporal Segmentation of Metro Trips Using Smart Card Data. *IEEE Transactions on Vehicular Technology,* 65**,** 1137-1149.

Zhang, L., Zhao, S., Zhu, Y. & Zhu, Z. Study on the method of constructing bus stops OD matrix based on IC card data. 2007 International Conference on Wireless Communications, Networking and Mobile Computing, 2007. IEEE, 3147-3150.

Zhao, D., Wang, W., Li, C., Ji, Y., Hu, X. & Wang, W. 2019. Recognizing metro-bus transfers from smart card data. *Transportation Planning and Technology,* 42**,** 70-83.

Zhao, J., Deng, W., Song, Y. & Zhu, Y. 2013. What influences Metro station ridership in China? Insights from Nanjing. *Cities,* 35**,** 114-124.

Zhao, J., Rahbee, A. & Wilson, N. H. M. 2007. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering,* 22**,** 376-387.

Zhou, J., Sipe, N., Ma, Z., Mateo-Babiano, D. & Darchen, S. 2019. Monitoring transit-served areas with smartcard data: A Brisbane case study. *Journal of Transport Geography,* 76**,** 265-275.

Zhou, X., Qin, X. & Mahmassani, H. S. 2003. Dynamic Origin-Destination Demand Estimation with Multiday Link Traffic Counts for Planning Applications. *Transportation Research Record: Journal of the Transportation Research Board,* 1831**,** 30-38.