Dynamic Train Demand Estimation and Passenger Assignment

Yuming Ou[1], Adriana-Simona Mihaita[1], Fang Chen[1] .

Abstract— **Understanding real-time train occupancy is a critical problem for public transport management, especially in the service disruption scenarios. To address this problem, this paper proposes a public transport passenger assignment method for estimating the time-dependent train occupancy comprising of a three-step modelling approach. Firstly, we make use of train station tap-on and tap-off information collected by Automated Fare Collection systems to estimate the initial time-dependent Origin-Destination matrix (OD) of the train network. Secondly, we take advantage of real-time train scheduling data to calibrate the initial OD matrix according to travel time, transfer time and waiting times across train lines. Thirdly, the calibrated OD matrix together with train scheduling data are used to generate entire passenger travel trajectories from origins to destinations including all path segments, by following a probabilistic hybrid Markov-driven approach. Lastly, after knowing all passenger trajectories, we further estimate the passenger occupancy for every train in the entire network in a given short time window. The results are applied over the real Sydney train network in Australia, and showcase that the proposed method can accurately quantify time-dependent passenger flows at a station platform level of granularity.**

Index Terms— **train assignment, public transport, OD estimation.**

I. INTRODUCTION

Several large cities around the world rely on train, metro or subway systems to accommodate the large travel demand from continuously increasing population. To cite a few, Transport for London has suffered a 70% increase in public transport patronage over the last 20 years [2], the MRT railways system in Hong Kong 56% [3], while the Sydney Train Network in Australia has reached 377.1 million annual patronage in 2019 [4] which represents a further 17.7% since 2013 [5] raising the maintenance costs to almost 1.46 BilAUD due to extensive track line constructions and integration with new Metro line to keep the pace with travel demand. As a result of this increased demand, many rail systems operate almost at maximum capacity during the peak periods with passengers crowding on platforms and inducing event more delays in train operations.

Therefore, estimating a real-time passenger loading for trains across the entire networks represents a true challenge and open research problems due to many factors which can interfere such as: a) train line inter-connectivity, b) stochastic technical disruptions, c) public events or d) badly interconnected multi-modal public transport systems. The problem is manifold and extends from planning perspective to real-time train operations.

A. Paper Contributions

The proposed method in this paper takes advantage of real-time train scheduling data to calibrate an initial time-dependent OD matrix which is generated based on train station tap-on and tap-off automatic data collected. Further on, the calibrated time-dependent OD matrix which represents the refined travel demand across the train network is used to assign each individual passengers to the most likely path from origin to destination, by taking into consideration several factors such as travel time between stations, transfer and waiting time. The final steps represents a hybrid Markov modelling and probabilistic estimation of passenger occupancy for every train in the entire train network in a given time window. The entire methodology comprises of three major steps which are further detailed in Section II and applied to the Sydney train network case study detailed in Section III. The current approach makes use of the following data input: a) **the network layout** and connectivity between train tracks, b) **the timetable information**: the scheduling of train trips according a master timetable which includes departures and arrival of each train trips between any stations in the network and c) **tap-on and tap-off information**: the numbers of passengers entering and existing train stations collected by an Automated Fare Collection (AFC) system provided in an aggregated hourly fashion.

The available information for modelling poses the following challenges which we address in this paper:

1) the patronage numbers are aggregated for every 15-min of time interval for protecting passenger privacy; the exact tap-on and tap-off times, and the links between tap-on and tap-off are unknown; this requires a robust routing methodology for travel path generation;
2) there is an unknown association between tap-on and tap-off records: as the data contains no information on where specific passengers tap-on followed by a tap-off in the network for protecting passenger privacy, the OD matrix estimation becomes very challenging;
3) train stations currently have multiple platforms so passengers once entered a train station can direct themselves towards any train line and board any train trip; this becomes very challenging for large train stations such as Central station with more than 30 platforms;
4) in-station transfer: passengers can take various transfer options for travelling to the same destination and currently no information is recorded when passengers interchange at specific train stations to change their trips and transfer between various segments of their journey.

[1] All authors are with the University of Technology in Sydney, Faculty of Engineering and IT, School of Computer Science: 61 Broadway Ultimo, NSW, Australia. Corresponding authors contact: yuming.ou@uts.edu.au

The results presented in Section III-A are based on the main outputs of the methods which are summarized in: a) the initial train demand comprising the assigned passengers travelling between any two stations in the network, b) the re-calibrated and final dynamic demand comprising passengers between any OD pair regardless of their train trips by each 15-min time-interval, c) the dynamic number of passengers waiting on each train station platform based on preferred path choice and d) the number of passenger who alight from a train trip and board to another train trip inside the same train station (estimating the transfer passengers at each train station). Finally, Section IV concludes this paper.

B. Related works

In the literature, various approaches have be undertaken to tackle these challenges out of which we name a few of them in the following. Authors in [6] proposed a probabilistic passenger to train assignment model by matching the fare transactions to the automatic vehicle location from tracking system and taking into consideration the access/egress time. However, the case study exemplification has been only applied to few train stations without major interchanges in between and would need a further a scalability analysis. In our study we consider the possibility of passengers moving between any platforms, and transferring between any crossing stations.

The study presented in [7] developed a regression model based on automatic fare card data and the distances between origin and destination stations to decompose the gate-to-gate journey time and estimate the location of passengers inside the network. This is a similar approach to what we propose in this paper, but our methodology considers both physical distance and ideal travel times between train stations, and also the real planned and operational travel times based on the train scheduling data set.

A recent paper [8] focused on estimating the crowding penalty in a discrete route choice frame-work, by considering that: a) for single itineraries, the “delayed access time” as the time between tapping-in and boarding, b) for transfer journeys, the assignment was based on the delayed access time distribution at the origin station and egress time distribution at the destination station. This represents as well a simplified approach but can face difficulties if several itineraries might experience very similar delays/egress times.

[9] proposed a collaborative optimization for metro train scheduling and train connections combined with passenger flow control strategy applied to a single train line in Beijing which does not contain any transfers. The authors proposed a mixed integer non-linear programming model to realise the trade-off among the utilization of trains, passenger flow control strategy and the number of awaiting passengers at platforms, however for transfers train changes the scalability of the method would become a challenge.

[10] proposed an analysis of subway station capacity with the use of queueing theory for building a network analytical model and discrete-time Markov Chain (DTMC). The main limitation here is application to a single train station with a high number of decision variables when generalising the approach. In our study, we consider hybrid-state Markov chains which have both a continuous and discrete-time behaviour for modelling the transitions of passengers between platforms, and this is presented as a third approach for passenger assignment across across a larger interconnected train network.

[11] proposed an event-driven model that involves three types of events, i.e., departure events, arrival events, and passenger arrival rates change events by considering walking ans transfer times of passengers. They solve the non-linear non-convex problem by using evolutionary algorithms (EVs) applied to a case study area with only two transfer stations. This is an interesting approach which might be bounded by long computational times of the EVs.

[12] built a dynamic simulation model of passenger flow distribution on schedule-based rail transit networks with train delays, by considering the origin-to-destination matrices, the passenger’s alternative choices (by applying a stochastic dynamic user equilibrium), waiting time and switching to other transportation modes. While the approach can be used for validation purposes of analytical train passenger assignment, its application for fast real-time train network modelling can represent a limitation.

As a future extension of passenger assignment modelling approaches, mobile data can bring a solid benefit and this was shown by [13] who conducted experiments in the Paris Metro to assess the potential of using cellular phone data to infer travel times, train loads, and OD flows. The train trajectory and mobile phone trajectory events were linked. However, there is currently very limited access to mobile data and mapping it in high details for movements across platforms can be quite problematic.

II. Methodology

The methodology of the current paper is split in three main parts which represent the major contributions based on initial data constraints and requirements. The area of our case study analysis represents the city of Sydney, and the train network for which we apply the current methodology is represented in Fig. 1. In the rest of this paper we use the abbreviation STM to denote the Sydney train metropolitan network.

A. Initial OD matrix estimation

First part of contributions consists in a systematic modelling approach for estimating the initial Origin-to-Destination matrices $OD(T_r)$ containing the number of passenger trips assigned between any pair of two train stations in the network $\{S_i, S_j\}$ $i, j \in \{1, .., N\}$ during a time period T_r. All notations in use are provided in Table I. As stated previously, the main challenge in this first estimation step consists in the aggregated number of tap-on/tap-off numbers available in blocks of 15-minute time interval during a 24-time period (mainly due to privacy concerns of public transport users). The initial OD estimation methodology comprises the following steps:

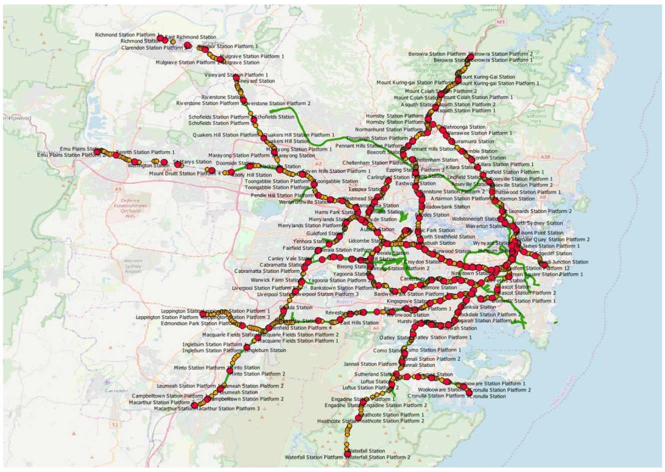
Richmond Station Platform 1
Richmond Station
East Richmond Station
Clarendon Station Platform 2
Windsor Station Platform 1
Mulgrave Station Platform 1
Mulgrave Station Platform 2
Mulgrave Station
Vineyard Station Platform 1
Vineyard Station
Riverstone Station
Riverstone Station Platform 1
Riverstone Station Platform 2
Schofields Station Platform 1
Schofields Station Platform 2
Schofields Station
Quakers Hill Station Platform 1
Quakers Hill Station
Marayong Station Platform 1
Marayong Station Platform 2
Marayong Station
Emu Plains Station
Emu Plains Station Platform 2
Penrith Station Platform 1
Werrington Station
St Marys Station
Doonside Station
Blacktown Station
Mount Druitt Station Platform 4
Rooty Hill Station
Seven Hills Station Platform 1
Toongabbie Station Platform 1
Toongabbie Station Platform 2
Toongabbie Station
Pendle Hill Station Platform 1
Wentworthville Station
Westmead Station
Parramatta Station
Harris Park Station
Merrylands Station
Merrylands Station Platform 1
Guildford Station
Yennora Station
Fairfield Station
Canley Vale Station
Cabramatta Station
Cabramatta Station Platform 1
Warwick Farm Station
Liverpool Station Platform 1
Liverpool Station
Liverpool Station Platform 3
Casula Station
Leppington Station
Leppington Station Platform 1
Leppington Station Platform 3
Leppington Station Platform 4
Edmondson Park Station Platform 1
Glenfield Station Platform 4
Macquarie Fields Station
Macquarie Fields Station Platform 2
Macquarie Fields Station Platform 1
Ingleburn Station Platform 1
Ingleburn Station Platform 2
Ingleburn Station
Minto Station Platform 1
Minto Station
Minto Station Platform 2
Leumeah Station Platform 1
Leumeah Station Platform 2
Leumeah Station
Campbelltown Station Platform 1
Campbelltown Station Platform 2
Macarthur Station Platform 2
Macarthur Station
Macarthur Station Platform 3
Berowra Station Platform 2
Berowra Station
Berowra Station Platform 1
Mount Kuring-Gai Station
Mount Kuring-gai Station Platform 2
Mount Kuring-gai Station Platform 1
Mount Colah Station Platform 2
Mount Colah Station
Mount Colah Station Platform 1
Asquith Station
Asquith Station Platform 2
Asquith Station Platform 1
Hornsby Station Platform
Normanhurst Station Platform
Wahroonga Station
Warrawee Station Platform 1
Thornleigh Station Platform 2
Turramurra Station
Pennant Hills Station Platform 1
Pennant Hills Station
Beecroft Station
Pymble Station
Gordon Station
Cheltenham Station Platform
Cheltenham Station
Killara Station Platform 1
Killara Station
Epping Station Platform 3
Epping Station
Carlingford Station
Lindfield Station Platform 1
Lindfield Station
Eastwood Station
Roseville Station Platform 1
Roseville Station
Roseville Station Platform 2
Telopea Station
Denistone Station Platform 2
Chatswood Station Platform 1
Artarmon Station Platform
Artarmon Station
Meadowbank Station
Camellia Station
St Leonards Station Platform 2
Rhodes Station
Rydalmere Station
Rydal Station
Wollstonecraft Station
North Sydney Station
Auburn Station
Waverton Station
Olympic Park Station
Milsons Point Station
North Strathfield Station
Circular Quay Station Platform 2
Lidcombe Station
Homebush Station
Wynyard Station
St James Station Platform 1
Berala Station Platform
Berala Station
Burwood Station
Museum Station
Edgecliff Station
Chester Hill Station
Croydon Station
Bondi Junction Station
Birrong Station
Birrong Station Platform 2
Redfern Station Platform 12
Yagoona Station
Yagoona Station Platform 2
Newtown Station
Green Square Station Platform 1
Canterbury Station
St Peters Station
Bankstown Station Platform 1
Belmore Station
Mascot Station
Bardwell Park Station Platform 1
Mascot Station Platform 2
Kingsgrove Station
Domestic Station Platform 1
Riverwood Station Platform 1
Banksia Station
Revesby Station
Riverwood Station
Rockdale Station Platform 2
East Hills Station
Hurstville Station
Kogarah Station Platform 4
Allawah Station
Oatley Station
Oatley Station Platform 1
Como Station
Como Station Platform 1
Jannali Station Platform 2
Jannali Station Platform 1
Jannali Station
Sutherland Station
Gymea Station
Loftus Station
Loftus Station Platform 2
Woolooware Station Platform 1
Woolooware Station
Cronulla Station Platform 2
Cronulla Station Platform 1
Cronulla Station
Engadine Station Platform 1
Engadine Station
Engadine Station Platform 2
Heathcote Station Platform 1
Heathcote Station
Heathcote Station Platform 2
Waterfall Station
Waterfall Station Platform 1
Waterfall Station Platform 2

Fig. 1: Network layout of the Sydney train network.

TABLE I: Notations in use for initial OD estimation.

Variable	Definition
N	the total number of train stations in the network,
$S_i, i \in \{1,..,N\}$	station ID,
$d_{i,j}$	distance between stations $i, j \in \{1,..,N\}$,
$\mathbf{D} = \left[d_{i,j}\right]_{i,j\in\{1,..,N\}}$	matrix of distances between any 2 stations i, j,
$\vec{D}_{S_i} = \left[min(d_{i,j})...max(d_{i,j})\right]$	the vector of all distances from station i to any station j in the network ordered from the closest to the farthest station,
$\mathbf{TT} = \left[tt_{i,j}\right]_{i,j\in\{1,..,N\}}$	matrix of travel time between any 2 stations i, j,
$\vec{TT}_{S_i} = \left[min(tt_{i,j})...max(tt_{i,j})\right]$	the vector of all crescent travel times from station i to any station j, using maximum speed of each train segment
$Np^{S_i}_{T_{on}}(t)$	total number of passengers entering station S_i at time t, regardless of stopping station
$Np^{S_i}_{T_{off}}(t)$	total number of passengers exiting station S_i at time t, regardless off departing station
$\vec{r}_{i,j} = \left[seg^1_{i,j},...seg^K_{i,j}\right]$	the route trip between stations i, j which contains several train segments $seg^K_{i,j}$, where K is the total number of train segments between stations $\{i, j\}$
$Td_{\vec{r}_{i,j}}$	the departure time of a route $\vec{r}_{i,j}$,
$\vec{tt}_{\vec{r}_{i,j}} = \left[tt^1_{i,j},...tt^K_{i,j}\right]$	the travel time of each segment in a route trip $\vec{r}_{i,j}$ recorded when travelling between stations $\{i, j\}$,
$E_{T_{off}}(\vec{r}_{i,j})$	the estimated time off for a route trip $\vec{r}_{i,j}$,
T_r	the 15-min time interval during a 24-hour period where $r \in \{1...96\}$;
$w_{\vec{r}_{i,j}}(T_r)$	the weight associated with a routing path between stations i, j calculated at time interval T_r,
$OD(T_r)$	the Origin-Destination matrix containing the number of assigned passengers between any pair of train stations estimated at time T_r,
$\hat{Np}^{S_i,S_j}_{T_{off-i}}(T_r)$	total number of passengers exiting station S_j during time interval T_r after departure from S_i

1) Given any two stations S_i, S_j in the STM network, calculate the matrix of distances denoted $\mathbf{D} = [d_{i,j}]_{i,j\in\{1,..,N\}}$, where by distance we refer to the physical distance in meters calculated between train stops.
2) For each of the train stations S_i, calculate the vectors of all distances between S_i and any station S_j in the network ordered from the closest to the farthest station $\vec{D}_{S_i} = [min(d_{i,j})...max(d_{i,j})]$. This is essential in determining the shortest path finding between the station at later stage.
3) For each station S_i, build random samples from the total number of entering/exiting passengers using aggregated data sets detailed previously (denoted $Np^{S_i}_{T_{on}}(t)$, $Np^{S_i}_{T_{off}}(t)$ respectively), at 1-min time-interval frequency and batched as well in 15-min time interval blocks. The final purpose is to able to assign/distribute the entering number of passengers across all train stations, by following an initial assignment method based on route weighted detailed in steps 4-7 below.
4) Calculate and store all the possible paths between any pair of two stations and rank them from shortest to longest path based on their estimated end of trip time. This translates in:
   a) calculate all route trips $\vec{r}_{i,j} = \left[seg^1_{i,j},...seg^K_{i,j}\right]$ between stations S_i, S_j, containing several train segments $seg^K_{i,j}$, each having its own travel time denoted as $\vec{tt}_{i,j} = \left[tt^1_{i,j},...tt^K_{i,j}\right]$ (calculated by using the maximum speed defined by Sydney train network and the total distance between origin and destination),
   b) for each route trip between two stations, calculate the estimated time-off, denoted $E_{T_{off}}(\vec{r}_{i,j})$, where

$$E_{T_{off}}(\vec{r}_{i,j}) = Td_{\vec{r}_{i,j}} + tt_{i,j} \tag{1}$$

   c) for each $E_{T_{off}}(\vec{r}_{i,j})$, identify the 15-min time interval in which the trip will finish, noted as T_r, where $r \in \{1...96\}$ (as there are $4*24 = 96$ time intervals in a day). For example, T_1 represents the time interval from $12:00AM - 00:15AM$ and T_{96} the time interval from $11:45PM - 12:00AM$. An example of such ranking and estimated time of arrival at this stage is provided in Table II here below.

TABLE II: Routing and estimated time of arrival example.

Time	Entries	OD	Route	Estimated arrival time	Allocated time interval
$Td_{\vec{r}_{i,j}}$	$Np^{S_i}_{T_{on}}(t)$	$S_i - S_j$	$\vec{r}_{i,j} = \left[seg^1_{i,j},...seg^K_{i,j}\right]$	$E_{T_{off}}(\vec{r}_{i,j}) = Td_{\vec{r}_{i,j}} + tt_{i,j}$	$T_r = r,\ r \in \{1,..96\}$
$07:01$	120	$S_1 - S_2$	$\vec{r}_{1,2} = \left[seg^1_{1,2}, seg^2_{1,2}, seg^3_{1,2}\right]$	$E_{T_{off}}(\vec{r}_{1,2}) = 07:09(07:01+8')$ AM	$T_r = 29$, (07:00-07:15 AM)
$07:01$	200	$S_1 - S_3$	$\vec{r}_{1,3} = \left[seg^1_{1,3}, seg^2_{1,3}\right]$	$E_{T_{off}}(\vec{r}_{1,3}) = 07:16(07:01+15')$ AM	$T_r = 30$, (07:15-07:30 AM)
...	...	...	...	...	
$07:01$	450	$S_1 - S_{250}$	$\vec{r}_{1,250} = \left[seg^1_{1,250},...seg^4_{1,250}\right]$	$E_{T_{off}}(\vec{r}_{1,250}) = 07:46(07:01+45')$ AM	$T_r = 32$, (07:45-08:00 AM)
$07:01$	60	$S_2 - S_1$	$\vec{r}_{2,1} = \left[seg^1_{2,1}, seg^2_{2,1}\right]$	$E_{T_{off}}(\vec{r}_{2,1}) = 07:07(07:01+6')$ AM	$T_r = 29$, (07:00-07:15 AM)
...	...	...	...	...	
$07:01$	610	$S_2 - S_{250}$	$\vec{r}_{2,250} = \left[seg^1_{2,1}..seg^6_{2,250}\right]$	$E_{T_{off}}(\vec{r}_{2,250}) = 08:01(07:01+60')$ AM	$T_r = 33$, (08:00-08:15 AM)
...	...	...	...	...	
$11:01$ PM	20	$S_{250} - S_1$	$\vec{r}_{250,1} = \left[seg^1_{250,1},...seg^3_{250,1}\right]$	$E_{T_{off}}(\vec{r}_{250,1}) = 11:23(11:01+22')$ PM	$T_r = 94$, (11:15-11:30 PM)
...	...	...	...	...	
$11:01$ PM	85	$S_{250} - S_{249}$	$\vec{r}_{250,249} = \left[seg^1_{250,249},...seg^8_{250,249}\right]$	$E_{T_{off}}(\vec{r}_{250,249}) = 11:48(11:01+47')$ PM	$T_r = 96$, (11:45PM-12:00AM)

5) By using initial aggregated Tap-off information sampled over 15-min time intervals, we calculate the total number of passengers ending their trips in each of the T_r time interval, at any station of the STM network ($S^{S_i}_{toff}(T_r)$) by using:

$$S^{S_i}_{toff}(T_r) = \sum_{i=1}^{N} Np^{S_i}_{T_{off}}(T_r) \tag{2}$$

We make the observation that due to the length and departure-time of each trip, the ending of the trips in each T_r produces an optimised and reduced data set of all possible tap-off journeys which is different than if we would have considered all passengers exiting the train network stations in each 15-min time interval.
6) For each route ending in a specific time interval T_r, we further calculate the weight associated with a routing path between stations S_i, S_j at time interval T_r, which we denote $w_{\vec{r}_{i,j}}(T_r)$, as follows:

$$w_{\vec{r}_{i,j}}(T_r) = \frac{Np^{S_i}_{T_{off}}(T_r)}{S^{S_i}_{toff}(T_r)} \tag{3}$$

which satisfies the condition that:

$$\sum w_{\vec{r}_{i,j}}(T_r) = 1 \tag{4}$$

7) By using Eq. 3, we finally calculate the initial assigned number of passengers tapping-off per time-interval as an expression of the total number of passenger entering the stations and their associated exiting weights:

$$Np^{S_i,S_j}_{T_{off-i}}(T_r) = w_{\vec{r}_{i,j}}(T_r) \times Np^{S_i}_{T_{on}}(Td_{\vec{r}_{i,j}}) \tag{5}$$

8) Calculate the error between the total assigned number of passengers across all stations, per each time-interval against the original sampled passengers exiting the stations at each time interval and re-iterate steps 1-7 if the error is more than a maximum error threshold which we establish at 15%:

$$Err(S_i) = \left| S^{S_i}_{toff}(T_r) - \sum_{j=1}^{N} Np^{S_i,S_j}_{T_{off-i}}(T_r) \right| \tag{6}$$

9) Results obtained in Eq. 5 are finally used to obtain the OD matrix of assigned number of passengers departing from any station as origin in the network; this is a time-dependent OD matrix and we estimate 96 matrices per each 24-h time interval as expressed below:

$$\mathbf{OD}(\mathbf{T_r}) = \left[Np^{S_i,S_j}_{T_{off}}(T_r) \right]_{i,j \in \{1,..,N\}} \tag{7}$$

This step ends the initial assignment and represents the entry point of more complex passenger assignment procedure detailed in the following two subsections.

B. Recalibration of passenger OD assignment

The second part of the current contributions translates into a recalibration of the initial OD matrices by taking into consideration more complex information of the train trip scheduling across the STM network, coupled together with planned real timetable information obtained from API connection to GTFS data stream on a daily basis.

TABLE III: Notations in use for OD recalibration.

Variable	Definition
$tt^{GT}_{i,j}$	the GTFS total travel time recorded between stations $\{S_i, S_j\}$,
$\vec{TT}^{GT}_{S_i}$	matrix of GTFS travel times recorded between any 2 stations $\{S_i, S_j\}$,
$TF_{\vec{r}_{i,j}}$	the recorded transfer time along a route between $\{S_i, S_j\}$,
S^{tr}_k	transfer station ID,
$WT_{\vec{r}_{i,j}}$	the waiting time before boarding a train,
$WT_{\vec{r}_{i,j}}$	waiting travel time for a route $\vec{r}_{i,j}$
$TF_{\vec{r}_{i,j}}$	transfer time between segments for a route $\vec{r}_{i,j}$
$E^A_{T_{off}}(\vec{r}_{i,j})$	expected time-off of a train (route) trip using $WT_{\vec{r}_{i,j}}$ and $tt^{GT}_{i,j}$
$E^B_{T_{off}}(\vec{r}_{i,j})$	expected time-off of a train (route) trip using $WT_{\vec{r}_{i,j}}$, $tt^{GT}_{i,j}$ and $TF_{\vec{r}_{i,j}}$.
$S^{S_i}_{toff-c}(T_r)$	the total number of recalibrated passengers ending their trip at S_i
$w_{\vec{r}_{i,j}}(T_{rA})$	the recalibrated weight of a route recorded in T_{rA}
$Np^{S_i,S_j}_{T_{off-c}}(T_r)$	the recalibrated number of assigned passengers tapping-off at each T_r
$\mathbf{OD^c}(\mathbf{T_{r-c}})$	the recalibrated matrix of passenger trips

The following steps are currently proposed and use the notations provided in Table III:

1) Given any two stations in the STM network, extract the total number of passengers initially assigned in the OD matrix obtained previously, for each T_r time interval.
2) Refine all previous paths that have been previously found at step I.4.a by reconstructing the entire possible journeys not just by using distance between stops and maximum speed, but information from real time-table scheduling, total number of transfers between each train segments, waiting time, etc. This step represents a further enhancement and refinement of Table II by adding more features of each possible route such as:
    a) real travel time of the entire journey from origin to destination extracted from train GTFS data specifications; let's denote this as $tt^{GT}_{i,j}$; the vector of all travel times of all possible paths from from station i to any station j will now be based on real time-table scheduling, and we denote it as $\vec{TT}^{GT}_{S_i}$,
    b) the recorded transfer time between platforms in each transfer station along the route, denoted as: $TF_{\vec{r}_{i,j}}$ by taking into consideration the synchronisation between the arrival of a train from origin station (S_i) to a transfer station (S^{tr}_k), and the scheduled departure or the interconnecting trains from S^{tr}_k to the final destination S_j.
    c) the waiting time which is determined from the tap-on time of passengers every minute until de departure of the next scheduled train trip which we denote as $WT_{\vec{r}_{i,j}}$.
3) Based on the previous measures defined above, for each possible routes between two stations we now define two different possibilities of calculating the estimated time-off for a trip, by further adjusting Eq.1 to take one of the following forms:

$$E^A_{T_{off}}(\vec{r}_{i,j}) = Td_{\vec{r}_{i,j}} + WT_{\vec{r}_{i,j}} + tt^{GT}_{i,j} \tag{8}$$

$$E^B_{T_{off}}(\vec{r}_{i,j}) = Td_{\vec{r}_{i,j}} + WT_{\vec{r}_{i,j}} + tt^{GT}_{i,j} + TF_{\vec{r}_{i,j}} \tag{9}$$

4) rank all possible routes (paths) from shortest to longest, which are obtained for each of the cases above (A,B) proposed above. Each solution will provide different paths/routes as being the preferred ones and these will be evaluated using various performance metric criteria detailed in the last step of this procedure.
5) for each $E^A_{T_{off}}$, $E^B_{T_{off}}$, identify the 15-min time interval in which the trip will finish, noted as T_{rA}, T_{rB}, where $r \in \{1...96\}$ as stated before.
6) By using initial aggregated Tap-off information sampled over 15-min time intervals, we calculate the total number of passengers ending their trips in each of the T_{rA}, T_{rB} time interval, at each station of the STM network by using:

$$S^{S_i}_{toff-c}(T_{rA}) = \sum_{i=1}^{N} Np^{S_i}_{T_{off}}(T_{rA}) \tag{10}$$

$$S^{S_i}_{toff-c}(T_{rB}) = \sum_{i=1}^{N} Np^{S_i}_{T_{off}}(T_{rB}) \tag{11}$$

7) For each route ending in a specific time interval T_{rA}, T_{rB}, we further calculate the weight associated this route, by using the following equations:

$$w_{\vec{r}_{i,j}}(T_{rA}) = \frac{Np^{S_i}_{T_{off}}(T_{rA})}{S^{S_i}_{toff-c}(T_{rA})} \tag{12}$$

$$w_{\vec{r}_{i,j}}(T_{rB}) = \frac{Np^{S_i}_{T_{off}}(T_{rB})}{S^{S_i}_{toff-c}(T_{rB})} \tag{13}$$

which satisfy the conditions that:

$$\sum w_{\vec{r}_{i,j}}(T_{rA}) = 1 \tag{14}$$

$$\sum w_{\vec{r}_{i,j}}(T_{rB}) = 1 \tag{15}$$

8) Finally, by using Eq. (14)-Eq. (15), we calculate the recalibrated number of assigned passengers tapping-off per time-interval as an expression of the total number of passengers entering the stations and their associated exiting weights:

$$Np^{S_i,S_j}_{T_{off-c}}(T_{rA}) = w_{\vec{r}_{i,j}}(T_{rA}) \cdot Np^{S_i,}_{T_{on}}(Td_{\vec{r}_{i,j}}) \tag{16}$$

$$Np^{S_i,S_j}_{T_{off-c}}(T_{rB}) = w_{\vec{r}_{i,j}}(T_{rB}) \cdot Np^{S_i,}_{T_{on}}(Td_{\vec{r}_{i,j}}) \tag{17}$$

9) In order to evaluate the effectiveness of each case (A,B), we compute various key performance indicators, by comparing results from Eq. (16)-Eq. (17) to original sampled tap-off passengers information:

$$R_A^2 = 1 - \frac{\sum_{i=1}^{N}\left(Np^{S_i,S_j}_{T_{off-i}}(T_{rA}) - Np^{S_i,S_j}_{T_{off-c}}(T_{rA})\right)^2}{\sum_{i=1}^{N}\left(Np^{S_i,S_j}_{T_{off-i}}(T_{rA}) - \frac{1}{N}\sum_{i=1}^{N} Np^{S_i,S_j}_{T_{off-c}}(T_{rA})\right)^2}$$

$$RMSE_A = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Np^{S_i,S_j}_{T_{off-i}}(T_{rA}) - Np^{S_i,S_j}_{T_{off-c}}(T_{rA})\right)^2}$$

$$SMAPE_A = \frac{100\%}{N}\sum_{i=1}^{N}\frac{2\cdot\left|Np^{S_i,S_j}_{T_{off-i}}(T_{rA}) - Np^{S_i,S_j}_{T_{off-c}}(T_{rA})\right|}{\left|Np^{S_i,S_j}_{T_{off-i}}(T_{rA})\right| + \left|Np^{S_i,S_j}_{T_{off-c}}(T_{rA})\right|}$$

Similarly, we calculate R_B^2, $RMSE_B$ and $SMAPE_B$. Based on final comparison between the performance of each approach, we will choose the best method achieving minimal results on all/most of the metrics. This approach will be the one defining the final calculation of the recalibrated OD matrix as presented in the next and final step.

10) Finally, we compute the recalibrated time-dependent OD matrix of assigned number of passengers departing from any station in the network as follows:

$$\mathbf{OD^c}(\mathbf{T_{r-c}}) = \left[Np^{S_i,S_j}_{T_{off-c}}(T_{r-c})\right]_{i,j\in\{1,..,N\}} \tag{18}$$

where $T_{r-c} \in \{T_{rA}, T_{rB}\}$ and is chosen based on final assessment from previous step.

C. Platform passenger assignment

As previously mentioned, the known variables that we have for each train stations are the total number of passengers tapping on and off at the main entrance in each station, together with the recalibrated numbers of passengers travelling between any two stations. However, this provides an overview of the train network performance and does not reflect the total number of passengers assigned to each train and each platform arriving/departing to/from a stations S_i, which would provide a higher level of granularity and insight regarding the overall train performance and passengers assignment across the entire train service.

Fig. 2 represents the modelling of passenger arriving and departing from a train station which contains several platforms, and for which the total number of passengers is known between specific time intervals. Table IV presents the notations used in this subsection which are detailed in the following as well.

TABLE IV: Notations in use for platform assignment.

Variable	Definition
M	number of platforms of a station S_i,
R	daily number of train trips arriving at a station S_i,
$Np^{inS_i}(t)$	number of passengers inside a station S_i at time t, including those transferring, waiting and remaining in train,
$Np^{S_i}_{a_k}(t), k\in 1,..M$	number of passengers arriving at a platform k belonging to a station S_i at time t,
$t+\Delta t_r$	scheduled departure time of a train trip $r\in\{1,R\}$
$Np^{S_i}_{d_k}(t+\Delta t_r), k\in 1,..M$	number of passengers departing from a platform k belonging to a station S_i after a scheduled departure time,
$T^r_{tripID}(t)$	the scheduled train trip ID arriving at S_i at time t,
$T^r_{tripID}(t+\Delta t_r)$	the scheduled train trip ID departing from S_i at time $t+\Delta t_r$

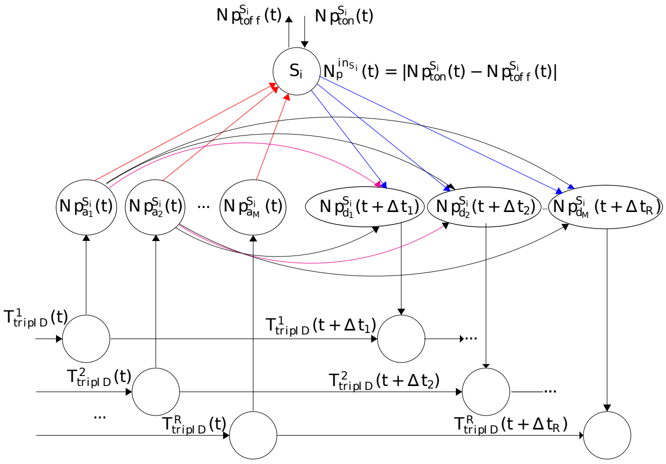

Fig. 2: Modelling schema of passengers arriving/departing from train stations.

We first start by describing the total number of passengers inside a station S_i at time t, including those transferring, waiting and remaining in train as:

$$Np^{inS_i}(t) = \left|Np^{S_i}_{T_{off}}(t) - Np^{S_i}_{T_{on}}(t)\right| \tag{19}$$

where $Np^{S_i}_{T_{off}}(t)$ is expressed as the total number of persons arriving at platform k of station S_i at time t from various trains and heading towards the exit:

$$Np^{S_i}_{T_{off}}(t) = \sum_{k=1}^{M} Np^{S_i}_{a_k}(t) \tag{20}$$

and $Np^{S_i}_{T_{on}}(t)$ becomes the total number of persons entering the station S_i and heading to departing from one of the platforms k at time $(t+\Delta t_r)$:

$$Np^{S_i}_{T_{off}}(t) = \sum_{k=1}^{M} Np^{S_i}_{d_k}(t+\Delta t_r), r\in 1,R. \tag{21}$$

These arrival/departing number of passengers are represented with red/blue arrows respectively in Fig. 2, which we will further identify as a hybrid Markov Chain model (HMCM) due to the dual continuous-discrete behaviour (continuous variables of passengers and discrete state representing each platform at a specific time). As an observation, we could have represented the states of arrival and departing from platforms as single states, described by two continuous-time variables (number of passengers arriving/departing at each platforms) and returning arcs for those staying on the same platform after getting off, but we wanted a clear separation of automata modelling based on train arrival/departing at each train platform, during a specific time interval $\Delta t_r, r\in\{1,R\}$.

By continuing our analysis we further express $Np^{S_i}_{a_k}(t), k\in\{1,..M\}$ as:

$$Np^{S_i}_{a_k}(t) = Np^{exit-S_i}_{a_k}(t) - Np^{remain-S_i}_{a_k}(t) + Np^{transfer-S_i}_{a_k}(t) \tag{22}$$

where $Np^{remain-S_i}_{a_k}(t)$ is the total number of passengers remaining in the same train which can be expressed as:

$$Np^{remain-S_i}_{a_k}(t) = Np^{S_i}_{a_k->d_k}(t+\Delta t_k) \tag{23}$$

and $Np^{transfer-S_i}_{a_k}(t)$ represents the total number of passengers randomly transferring from platform a_k to other platforms of departure d_t, after a specific time interval $\Delta t_r, r\in\{1,..R\}$ (all transfers possible from current platform):

$$Np^{transfer-S_i}_{a_k}(t) = Np^{S_i}_{a_1->d_2}(t+\Delta t_2) + ...Np^{S_i}_{a_1->d_M}(t+\Delta t_M) \tag{24}$$

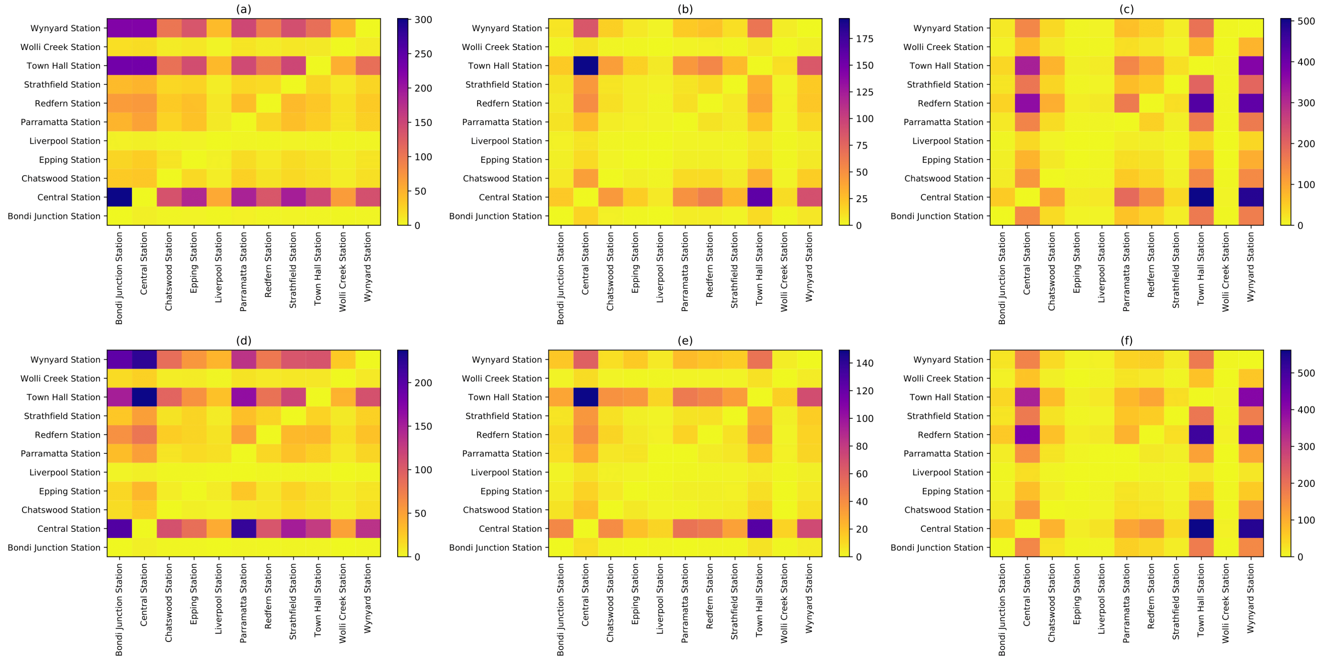

Fig. 3: Heat Map for selected stations as a) initial OD at 8AM, b)initial OD at 12PM, c) initial OD at 5PM, d) calibrated OD at 8AM, e) calibrated OD at 12PM and f) calibrated OD at 5PM.

Similarly, we express $Np^{S_i}_{d_k}(t+\Delta t_r)$ as:

$$Np^{S_i}_{d_k}(t+\Delta t_r) = Np^{enter-S_i}_{d_k}(t+\Delta t_r) - Np^{remain-S_i}_{d_k}(t+\Delta t_r) + Np^{transfer-S_i}_{d_k}(t+\Delta t_r). \tag{25}$$

where $Np^{remain-S_i}_{d_k}(t+\Delta t_r)$ is the total number of passengers remaining in the same train (not changing platforms and $Np^{transfer-S_i}_{d_k}(t+\Delta t_r)$ is the total number of passengers arriving at platform d_k for departure at time interval $(t+\Delta t_r)$ (all transfers possible to a current platform); similarly, these are expressed as:

$$Np^{remain-S_i}_{d_k}(t+\Delta t_r) = Np^{S_i}_{a_k->d_k}(t+\Delta t_k) \tag{26}$$

$$Np^{transfer-S_i}_{d_k}(t+\Delta t_r) = Np^{S_i}_{a_1->d_2}(t+\Delta t_2) + ...Np^{S_i}_{a_1->d_{k-1}}(t+\Delta t_{k-1}) \tag{27}$$

where

$$t+\Delta t_{k-1} \leq t+\Delta t_k, etc. \tag{28}$$

We make the observations that we can model the transitions between various platforms of station, by taking into considerations the transition probability matrix of the HMCM which need to satisfy the conditions across the above transition probabilities as follows:

$$Pr(Np^{S_i}_{a_1->d_2}(t+\Delta t_2)) + ...Pr(Np^{S_i}_{a_1->d_M}(t+\Delta t_M)) = 1$$
$$Pr(Np^{S_i}_{a_1->d_2}(t+\Delta t_2)) + ...Pr(Np^{S_i}_{a_1->d_{k-1}}(t+\Delta t_{k-1})) = 1.$$

III. Case Study

As shown in Fig. 1, our study has been applied over the Sydney train network in Australia. The entire train network expands over various states in Australia, but for the purpose of keeping the analysis concise, we only focus on the New South Wales and most specifically on the Sydney region which contains in total over 175 train stations with a total of 506 platforms.

A. Results

OD estimation: Firstly we followed the method detailed in Section II-A to estimate an initial OD matrix based on tap-on/tap-off data which contains 175x175x24x4=2.94 million OD pairs. The OD matrix covers every 15-min of time window from 00:00:00AM to 24:00:00AM and each pair represents the number of passengers travelling from one Origin Station to a Destination Station departing within a 15-min time interval. We run the OD estimation algorithm (multiple threads) on a machine with Intel i7 CPU and 16GB RAM which took almost 10 hours in terms of computational time. The considerable time cost is mainly due to the path routing computation. The error of initially estimated OD matrix is around *13.6%* calculated by using Eq. (6) which will be further reduced after applying the calibration in next step. To illustrate the OD matrix, Fig. 3 (a), (b) and (c) show a selection of OD matrix heat maps for the selected stations at 08:00AM, 12:00PM and 17:00PM respectively. The 11 selected stations consists of 3 major stations in Sydney CBD and 8 interchange stations outside CBD. We observe that already the OD matrix heat maps disclose a pattern that in the morning peak hours passengers are travelling from other stations outside Sydney CBD to the main Central station, Town Hall station and Wynyard station, whereas in the afternoon peak hours they are travelling from CBD to other areas. Due to lack of space in this paper we further provide three sample representations of the entire 175x175 OD matrices in the online supplement material provided at [1, Fig. 9 a, b, c] which showcase overall traffic patterns across the entire train network in the city.

TABLE V: Performance metrics of calibration approaches A,B

Calibration Approach	R^2	RMSE	SMAPE
A	0.8611	75.84	33.51%
B	0.8604	62.39	32.09%

OD calibration: After obtaining the initial OD matrix, we calibrated it using the proposed calibration schemas A and B, detailed in Section II-B. Their effectiveness is evaluated using R^2,

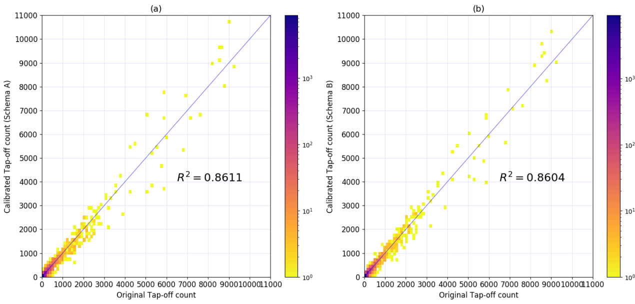

Fig. 4: Coefficient of determination of calibrated OD matrix for approach a) A and b) B.

RMSE and SMAPE which are shown in Table V. The performance of each approach seems to be very effective ($R^2 > 0.85$, a low $RMSE < 80$ and good SMAPE thresholds below 35%) with an evident improvement when applying method B; this indicates that for our case study in Sydney, including the waiting time in the route choice estimation will improve considerably the performance of the train passenger assignment by almost 17% ($RMSE$ is reduced to 62.39 from 75.84).

Further more, Fig. 4, Fig. 5 and Fig. 6 illustrate the R^2 distribution and box-plot performance of each train station (in terms of RMSE and SMAPE) after the calibration. The R^2 of recalibrated OD versus the initial one indicates a large number of passengers falling under the threshold of $2,000$ passengers per 15-min time interval with few outliers reaching $10,000 - 11,000$ passengers across highly circulated CBD train stations during peak hours. There are however no missing information or large number of small outliers being detected after the calibration procedure which reinforces the method efficiency. RMSE values indicate very good accuracy (majority fall below 100) for both A and B approaches. Similarly SMAPE values maintain strong records below 25% across majority of stations, including the busiest ones, for approach A, and below 20% for approach B. Epping Station is the one showcasing the highest variance of the SMAPE values, mostly due to the interchange nature of the station receiving high speed trains from cities located at North of Sydney.

Similarly, the heat maps of calibrated OD Matrix for 11 selected stations on 08:00AM, 12:00PM and 17:00PM are presented in Fig. 3 (d), (e) and (f) respectively revealing a slight re-distribution of passenger across Central station, Parramatta (in the west of the city) and Epping (to the North). Afternoon peak seem to the busiest across Town Hall, Central and Wynyard stations (inside CBD) maintaining the same trends as before. The entire recalibrated matrices can also be found in the online supplement at [1, Fig. 9 d, e, f] which reveal more pregnant morning and afternoon peak patterns across several stations in the network.

Platform passenger assignment: was lastly conducted for the entire network. Fig. 8 and Fig. 7 show the number of passengers in Central station and Town Hall station on 08:00AM, 12:00PM and 17:00PM respectively.

Both figures showcase the time-dependent evolution of: a) off board passengers arriving at each station and going towards the exit (as per Eq. (22)), b) the onboard passengers departing after entering the station (as per Eq. (25)) as well as the number of passenger transferring in and out of platforms (as per Eq. (24) and Eq. (27) respectively).

The passenger assignment results demonstrate a consistent pattern with the above OD matrix heat maps. There are more passengers entering in both Central station and Town Hall station in the morning peak hours than the afternoon peak hours and the reverse applies. It also can be observed that the number of passengers transferring between platforms in the two stations are significant and it is comparable with the number of passengers entering and

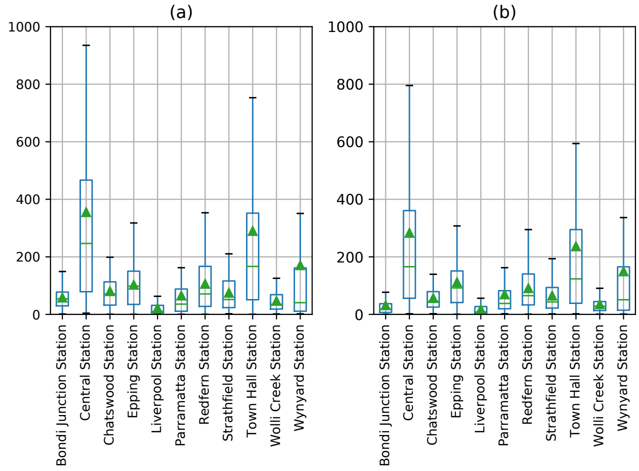

Fig. 5: RMSE values for selected stations for approach a) A and b) B.

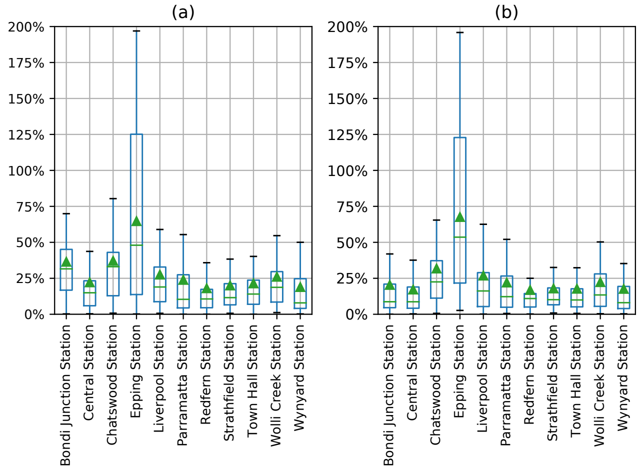

Fig. 6: SMAPE values for selected stations for approach a) A and b) B.

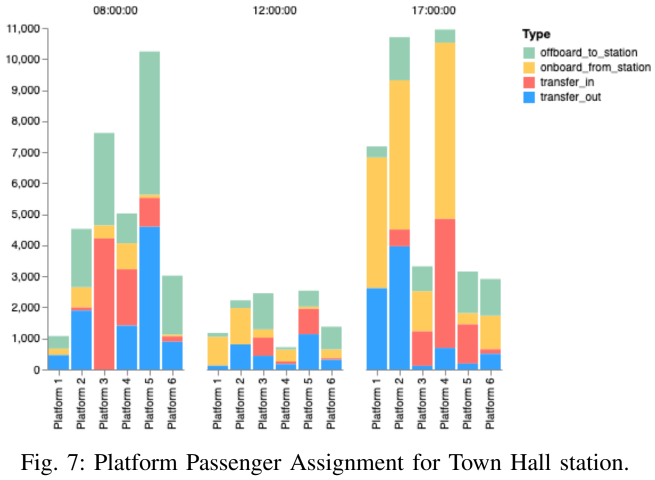

Fig. 7: Platform Passenger Assignment for Town Hall station.

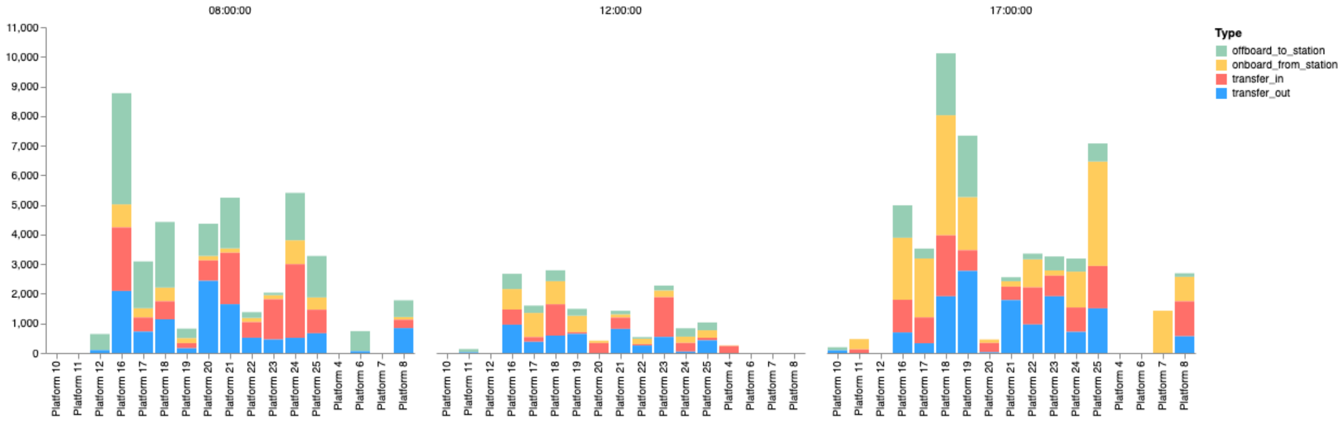

Fig. 8: Platform passenger assignment for Central station by each time interval, and a selected number of platforms (to fit the chart).

existing the stations, which implies that both stations are very circulated transport hubs. The passenger assignment can reveal large amounts of activities inside stations other than station tap-on/tap-off activities, which is helpful for understanding station performances and improving situation awareness. The current methodology and analysis provide as well a powerful insight into the implications of train disruptions and load impact across platforms, and the entire stations in general.

IV. Conclusion

This paper studied the train demand estimation and public transport passenger assignment problem, which are critical steps for any public transport management centres. To address these problems, we proposed a three-step modelling approach leading to the final estimation of train occupancy. The results are carried in a case study focusing on the entire Sydney region train network. The main contributions of this work consist of:

- a method for estimating the initial time-dependent OD matrix under data constraint circumstance,
- a method for calibrating the initial OD matrix using real-time train scheduling data,
- a method for platform passenger assignment to quantify passenger flow at platform level of granularity, and
- an application case study on a large scale train network in a real-life setting which adds up to almost 2.94 million of time-dependent OD pairs; this implies significant computational challenges and scalability which the current approach has demonstrated.

Future extensions of the current work include: a) showcasing the performance of the method on a high variety of train paths across the network before/after the calibration, b) considering more accurate real-time information such as train delay and modelling the impact of this delay on daily commuters and finally c) embedding the impact of large disruptions across the entire train network in order to estimate the most crowded platforms of each train stations pending on interconnection train lines. We are looking at further using mobile data for train passenger assignment refinement and validation. This however raises complexities in terms of geo-location, aggregation of passengers movement inside/outside of trains stations, etc.

References

[1] online supplement, “Appendix: Dynamic Train Demand Estimation and Passenger Assignment,” 2020, https://www.dropbox.com/s/0kqkflg07gm3cio/Supplement_Dynamic_Train_Demand_Estimation_and_Passenger_Assignment.pdf?dl=0.

[2] Mayor of London, “Travel in london,” 2015. [Online]. Available: http://content.tfl.gov.uk/travel-in-london-report-8.pdf

[3] MTR, “Hong kong patronage updates,” 2020, http://www.mtr.com.hk/en/corporate/investor/patronage.php#search.

[4] Sydney Trains, “Sydney train annual report,” 2019. [Online]. Available: https://www.transport.nsw.gov.au/news-and-events/reports-and-publications/sydney-trains-annual-reports

[5] Bureau of Transport Statistics, “Train statistics 2014,” 2014. [Online]. Available: https://www.transport.nsw.gov.au/sites/default/files/media/documents/2017/Train%20Statistics%202014.pdf

[6] Y. Zhu, H. N. Koutsopoulos, and N. H. Wilson, “A probabilistic passenger-to-train assignment model based on automated data,” *Transportation Research Part B: Methodological*, vol. 104, pp. 522 – 542, 2017.

[7] Y. Sun and R. Xu, “Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data,” *Transportation Research Record*, vol. 2275, no. 1, pp. 58–67, 2012.

[8] D. Hörcher, D. J. Graham, and R. J. Anderson, “Crowding cost estimation with large scale smart card and vehicle location data,” *Transportation Research Part B: Methodological*, vol. 95, pp. 105 – 125, 2017.

[9] R. Liu, S. Li, and L. Yang, “Collaborative optimization for metro train scheduling and train connections combined with passenger flow control strategy,” *Omega*, vol. 90, p. 101990, 2020.

[10] X.-y. Xu, J. Liu, H.-y. Li, and J.-Q. Hu, “Analysis of subway station capacity with the use of queueing theory,” *Transportation Research Part C: Emerging Technologies*, vol. 38, p. 28–43, 01 2014.

[11] Y. Wang, T. Tang, B. Ning, T. J. van den Boom, and B. D. Schutter, “Passenger-demands-oriented train scheduling for an urban rail transit network,” *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 1 – 23, 2015.

[12] W. Li and W. Zhu, “A dynamic simulation model of passenger flow distribution on schedule-based rail transit networks with train delays,” *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 3, no. 4, pp. 364 – 373, 2016.

[13] V. Aguiléra, S. Allio, V. Benezech, F. Combes, and C. Milion, “Using cell phone data to measure quality of service and passenger flows of paris transit system,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 198 – 211, 2014, special Issue with Selected Papers from Transport Research Arena.