Crowding on Trains and Platforms: A New Approach to Big Data

Chinh Ho · Loan Ho

Abstract This paper develops a new approach to use open source data and public transport smart card data for establishing crowding on trains and platforms. Every day, an automatic fare collection (AFC) system registers millions of transactions, recording when and where train users get on and off the train network; however, very little is known about the specific services passengers use and where they transfer. This paper develops a data-driven algorithm to identify the most likely service(s) each passenger uses, given their observed tap-on and tap-off data. Using Sydney Trains as a case study, crowding profiles for all services are established and validated.

Keywords: Crowding on Trains · Crowding on Platforms · Opal Data · Smart Card Data · API · Sydney Trains · Train Load

1 Introduction

Crowding can be inconvenient, uncomfortable and leads to unpleasant experience, especially when passengers must stand very close together for prolonged periods. When passenger numbers exceed seating capacity by a certain percentage, crowding can also make passengers feel unsafe and dwelling time prolonged, which can impact on-time running. Although there is no legal limit on the number of passengers that can travel in any given train, public demand for better and safer train services has seen many transport authorities, such as Transport for London,
Transport for New South Wales (TfNSW), and Public Transport Victoria, started monitoring train crowding levels.

Given the growing interest in understanding and monitoring crowding-related problems in network planning and pricing policy, it is critical to develop empirical methods to recover crowding levels on the entire network and quantify the discomfort that crowding may cause to passengers. Recently, we have seen a growing number of choice models developed to evaluate passenger’s willingness to pay for lower crowded trains, either with a longer travel time or higher transport fare (Alejandro Tirachini et al., 2013, Marco Batarce et al., 2016, Daniel Hörcher et al., 2017); however, translating these model results into advancing public transport planning and rail timetabling is still challenging as this requires crowding profiles, in addition to how sensitive passengers are to different crowding levels. The adoption of smart cards for ticketing has opened great opportunities to recover crowding levels both on trains and platforms. The next section describes the challenges in using smartcard data and our proposed method.

2 Problem Description and Model Formulation

2.1 Problem description

To recover crowding on trains and platforms, we need to know three elements (i) the travel demand matrix, commonly referred to as origin-destination matrix or O-D matrix for short, (ii) the service/train each passenger use and where they transfer (if any), and (iii) the stopping pattern of each service. The first information can be obtained from smart card data if passengers are required to tap their smart cards on the card readers at both ends (i.e., origin and destination). Smart card data may also contain information on the service each passenger uses, depending on where smart card readers are deployed. If smart card readers are deployed inside the vehicles such as Sydney Buses, the service each passenger used is included in the smartcard data via bus number and vehicle identifier; however, for most train and metro systems, smart cards are deployed at stations, and thus smart card data do not contain the service number or train identifier.

Added to the challenge is that most train networks are designed as closed systems where passengers only leave traces at boarding and alighting stations. This means that very limited knowledge is known about which train lines a passenger used to travel for the observed O-D pair if this has multiple route options. Also, another challenge specific to a complex train network is that trains on the same line may have different stopping and continuing patterns. In many cases, stopping pattern for each train service (i.e., express, limited stops or all stops service) is available in the timetable, but this is often not the case for continuing pattern. That is, whether a train will proceed to another service/line when it arrives at the final stop such that the service identifier changes while passengers can remain on board and continue
their journeys. This operational complication of the train network is distinct from most metro networks in which the train stops at every metro station along the track and it serves that track/line only. This complication, together with limited information registered to the smart card systems, make it more challenging to reveal the crowding profile for each train service and station. We tackle this challenge with a data-driven method to assign passengers to train. This process is described below.

2.2 Passenger to train assignment method

Two steps are required to build a passenger-to-train assignment model. First, all feasible itineraries must be recovered and evaluated for each passenger, given their tap-on and tap-off data (i.e., when and where they check-in and check-out the train network). Second, an itinerary choice model needs to be developed and validated to assign each passenger to a specific service/train.

Feasible choice set

As for the identification of feasible itineraries, few algorithms have been developed in the literature, including entry and exit time constraints (Elizabeth Cheriyamadam Paul, 2010, Feng Zhou and Rui-hua Xu, 2012), k-shortest-paths (Daniel Hörcher et al., 2017), link elimination and Brute-Fore-Search algorithm (Lijun Sun et al., 2015). Except for the work of Hörcher et al. that generates feasible itineraries for a rather complex rail network of Hong Kong using the igraph package of R (Gabor Csardi and Tamas Nepusz, 2006), other studies either deal with a simplified metro network (Lijun Sun et al., 2015), or consider only OD pairs with one feasible itinerary (Elizabeth Cheriyamadam Paul, 2010), or those with no transfers at all (Yiwen Zhu, 2014, Yiwen Zhu et al., 2017a, b). A common reason for these implications, at least for model application, is to avoid complications caused by an exponential increase in the number of feasible itineraries, especially for a large train/metro network like Sydney Trains where multiple services share the same track and many journeys, defined by tap-on and tap-off times, involve one or more transfers. Thus, these methods are not suitable to reproduce the crowding pattern in an entire train network since transfer passengers, as can be seen in the result section, contributes significantly to the crowding levels on trains and platforms.

For a complex network like Sydney Trains with 308 stations spread over 2,242 km of track shared by multiple services, each of which having different stopping and continuing patterns, the number of feasible itineraries generated by any single algorithm mentioned above will be very large. For example, the Brute-Fore-Search algorithm used by Lijun Sun et al (2015) may have advantages in generating choice sets in short time for a simple network but the authors noted that an alternative method is required to support the modelling of shared track network (i.e., multiple services share the same track), and that special attention is required to remove irrational alternatives such as those with many transfers or those with 1.5 times
longer than the shortest path. The algorithm used by Daniel Hörcher et al (2017) employs the latter rule together with the number of transfers and feasible itineraries to classify journeys into different typologies in order to apply a sequential assignment process. The process uses information (i.e., egress time distribution) derived from journeys with single feasible itinerary for the assignment of non-transfer journeys with multiple feasible itineraries, and then the assignment of transfer journeys. While these strategies are necessary to simplify the models, the assumptions introduced may not necessarily justified (Yiwen Zhu et al., 2017b). On the one hand, introducing arbitrary rules such as maximum number of transfers to limit the number feasible itineraries may not reflect the real choice sets faced by the traveller since some may be willing to transfer to save travel time. On the other hand, classifying journeys based on the number of transfers and applying different assignment methods for these journeys implicitly assumes that the choice between services with and without transfer is not available. While this assumption may be valid in a metro network where trains stop at every station along their route, this is not the case in the Sydney context where many services are express or limited stops, especially during the peak hours where travellers may have to choose between an all-stop train without a transfer and an express train with a transfer to go to a local station being skipped on express services.

To address these issues without the need to introduce arbitrary rules, we develop an API program (i.e., Application Programming Interface) that queries the official Trip Planner webpage (https://transportnsw.info/) that many Sydney Trains users rely on for planning their journeys and returns all feasible itineraries for each observed tap-on and tap-off data point. The API program, in essence, combines entry and exit constraints with the k-shortest-paths algorithm in generating a full set of feasible itineraries, ensuring that the set of feasible itineraries are collectively exclusive and none of these are illogical in reality since information on the trip planner webpage is live with all service disruptions updated. The API program follows the following steps:

Step 1. Extract the full OD demand matrix from the smart card data and use as an input file for the search process.
Step 2. For each record in the input file (i.e., OD pair), retrieve the origin and destination station names.
Step 3. Call the API search for station names to get the station IDs at 12:00 AM on the specified date.
Step 4. Make API call to search for journeys from origin to destination.
Step 5. Process response and iterate through all the possible journeys, legs and stops to get data.
Step 6. Go through all journeys and check for the latest departure schedule and set it as the time for the next iterative search. If no later journey is found in that search, increase time by 30 minutes and search again.
Step 7. End search on that input record when all the train schedules from specified start date to specified end date have been collected and go to next record of the input file.

Step 8. End the process when we have completed the total records of the input file.

This process returns all possible journeys for each observed OD pair in the smart card data where journeys could be an alternate route going from origin to destination or a later departure/arrival time. Each journey comprises of one or more trip legs, defined as a series of stops/platforms that the train passes on its route. If a journey involves one or more transfers, the API program returns all possible journeys with different combinations of trip legs using the Trip Planner options of normal walking speed and 20 minutes maximum walking time (see https://transportnsw.info).

Of the possible journeys returned by the API program for each OD pair, some are identified as feasible itineraries while some are not, depending on the observed tap-on and tap-off times. An itinerary is feasible for the observed tap-on tap-off data if the schedule departure time is after tap-on time, and schedule arrival time is before tap-off time, assuming train services are punctual to timetables embedded in the Trip Planner webpage. To account for some minor delays to train services that are not necessarily updated in the Trip Planner, we allowed train services to be late up to 5 minutes; however, this condition is used only when requiring strict conditions of tap-on before boarding and tap-off after alighting results in no feasible itinerary for that journey. That is, only journeys with no feasible itinerary will enter the second round of search where the punctual assumption is relaxed so that train services can be late up to 5 minutes from the information published in the Trip Planner webpage. We selected this threshold based on the fact most service headways in Sydney is usually longer than 5 minutes. The role and necessity of this assumption is discussed in the result section.

Assignment method
Several methods have been developed to assign observed demand to the available capacity but in general these methods can be classified into a schedule-based or a frequency-based approach (Daniel Hörcher et al., 2017, Yiwen Zhu et al., 2017b). Which approach is more appropriate for establishing crowding statistics will depend on the available information on train movements and the purpose of the analysis. If the only available information is the average frequency of services along a certain train line for a given time of day, then the frequency-based approach can be used to estimate the average train load for that line in that time period. When the entire schedule of trains, including planned departure and arrival time at each station, is available, then the schedule-based approach can be used to assign passengers to trains. The assignment method could be deterministic or probabilistic, depending in the assumptions underlying the assignment process. For instance, the assignment model is deterministic if we assume that the passengers were served by the first train.
that arrived at the origin (Feng Zhou and Rui-hua Xu, 2012), or by the last train that arrived at their destinations (Kevin Buneman, 1981), or by the shortest connection with minimum access time, minimum egress time, and minimum transfers (Takahiko Kusakabe et al., 2010). Conversely, assuming that the egress time (Daniel Hörcher et al., 2017), access time (Elizabeth Cheriyamadam Paul, 2000), left-behind probability (Yanshuo Sun and Paul M. Schonfeld, 2015) for each station, or walking speed (Yiwen Zhu, 2014) follows a certain distribution, such as those observed from the manual passenger movement surveys or from journeys with one feasible itinerary, will result in a probabilistic model that predicts the probability that a given passenger takes a specific train among the set of feasible itineraries.

The process of deriving the probabilistic assignment method also relies on other assumptions. For example, (Daniel Hörcher et al., 2017) assumes that having information about one of the feasible itineraries provide the same knowledge about the rest of the feasible set as another feasible itinerary. While this assumption may be valid for a Metro network where services are run at regular headways (and thus from one train’s arrival time one may infer that other trains may have arrived one headway earlier or later), we found this assumption does not hold for the Sydney context at least for two reasons. First, many services share the same track but some are express, some are limited stops and some are local, with arrival time at large stations served by these services being scheduled to ensure a safety gap as opposed to a regular service headway. Second, trains may share part of the track, regardless of whether they are on the same line (see Figure 1), and this results in many stations being served at irregular headways, usually short enough (8 mins or less) to not bother the train users.

Another assumption previous studies usually make in deriving the probabilistic assignment method relates to the walking speed or walking distance of the passenger. Elizabeth Cheriyamadam Paul (2010) assumed that walking speed is an individual characteristic which is the same when the passenger accesses as when they egress from the train. By contrast, Yiwen Zhu (2014) assumed the walking distance for each passenger, ignoring their strategic behaviour when choosing a boarding location along the platform so that the distribution of walking speed amongst passengers were considered explicitly. Sung-Pil Hong et al. (2016) assumes the first come first served queuing discipline amongst the passengers which translates into an assignment method that requires a non-overlap between alighting intervals and amongst multiple boarding intervals. Although these studies need strong assumptions, the resulting models have an advantage of decomposing access time into walking and waiting time components.

The methods proposed here allows for more flexibility in terms of irregular service headways with different stopping patterns, overlapping boarding, alighting intervals, and numbers of transfers, in a random utility maximisation framework. It is simpler than Elizabeth Cheriyamadam Paul (2010), Yiwen Zhu (2014), and Yiwen Zhu et al
(2017b) in the sense that we do not attempt to decompose access time into its components or explicitly consider the impact of failed boarding with left-behind probability. Following Daniel Hörcher et al. (2017), we combine walking and waiting time at the origin station into an access time instead. Our model is also simpler than Lijun Sun et al (2015) in the sense that we do not attempt to estimate the route choice model parameters endogenously. Rather, as Elizabeth Cheriyamadam Paul (2010) and Yiwen Zhu (2014) suggested, we use a separate random route choice model formulating around the concept of generalised time and multipliers derived from the literature. We then conduct a sensitivity analysis to identify the best parameters for recovering crowding profiles by comparing the model results with onboard headcount. In general, our objective to develop a simple assignment method which can recover the crowding pattern for an entire shared track network where trains are run with different stopping patterns, using as few as possible assumptions on service headways, boarding and alighting intervals.

Given the set of feasible itineraries the API program identifies for each passenger journey observed in the smartcard data, we denote the utility obtained by passenger \( n \) from each itinerary \( j \) available in their feasible set with \( U_{nj} \) for \( j = 1, \ldots, J \). We do not observe this utility but do observe some attributes of the alternate itinerary such as travel time, number of transfers and characteristics of the passengers through the type of the card they use (e.g., adult, children or pensioner). The utility is therefore postulated to have both observable component, \( V_{nj} \) and unobserved component, \( \epsilon_{nj} \), given as:

\[
U_{nj} = V_{nj} + \epsilon_{nj}, \quad j = 1, \ldots, J
\]

(1)

Under the assumption of utility-maximising behaviour, a passenger \( n \) chooses itinerary \( i \) if and only if \( U_{ni} > U_{nj} \) for \( j \neq i \). The probability that the passenger \( n \) chooses itinerary \( i \) is therefore:

\[
P_{ni} = P(U_{ni} > U_{nj}, \forall j \neq i) = P(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}, \forall j \neq i)
\]

\[
= P(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i)
\]

(2)

Assuming that each random component \( \epsilon_{nj} \) is an independently and identically distributed extreme value (or \( iid \) Gumbel and type I extreme value). The probabilistic choice system in Equation (2) results in the following closed-form model, known as multinominal logit model (Train, 2009):

\[
P_{ni} = \frac{\exp(V_{nj})}{\sum_j \exp(V_{nj})}
\]

(3)
In this paper we specify the observable utility as a function of delayed access time, egress time, journey time, and number of transfers as follows:

\[ V_{nj} = \theta \left( \alpha \times \text{AccTime}_j + \beta \times jnyTime_j + \gamma \times \text{EgrTime}_j + \delta \times \text{Transfers}_j \right) \]  

(4)

Where, \( \text{AccTime}_j \) is the time required to access service \( j \) at the origin, defined as the duration time between the passenger’s tap on time and the schedule departure of the first train on itinerary \( j \);

\( jnyTime_j \) is the journey time of itinerary \( j \), defined as the duration between the schedule departure time of the first leg and the schedule arrival time of the last leg of itinerary \( j \);

\( \text{EgrTime}_j \) is the egress time at the destination associated with itinerary \( j \), defined as duration between the schedule arrival time of the last leg of itinerary \( j \) and the passenger’s tap-off time;

\( \text{Transfers}_j \) is the number of transfers involved on itinerary \( j \).

\( \theta, \alpha, \beta, \gamma, \delta \) are model parameters. Normalising \( \beta \) to 1, \( \alpha, \gamma, \) and \( \delta \) can be interpreted as access time, egress time, and transfer penalties (or multipliers) relative to journey time. We choose to use journey time instead of in-vehicle time because this is what the passengers see when they use the Trip Planner app/webpage. Parameter \( \theta \) then describes the passenger’s preference for generalised time in itinerary choice. As the random utility theory assumes a rational decision maker, parameter \( \theta \) should always be negative so that the passengers prefer a journey with shortest generalised travel time.

The model parameters could be estimated by maximum likelihood estimator if we observed the itinerary chosen by the passenger; however, this information is not available from the smart card data, and thus we derive these parameters from the literature and conduct sensitivity tests to determine how they influence the results and which values would be best in terms of reproducing the crowding levels observed in the manual surveys. Using different values for these parameters results in different assignment models found in the previous studies. For example, the probabilistic model (3, 4) will result in a deterministic model that assigns all passengers to the first train to arrive at the origin platform (e.g., Feng Zhou and Ruihua Xu, 2012) if we use a very large value for \( \alpha \). This is because the time required to access the service is penalised far more heavily than other attributes of the itinerary. As a result, all passengers who maximise their utility will select the first train (i.e., minimise the heavy loss associated with access time). Similarly, using a very large value for \( \gamma \) will result in a deterministic model which assumes that all passengers are
served by the last train that arrives at their destination (e.g., Kevin Buneman, 1981). The behavioural model proposed by Takahiko Kusakabe et al. (2010) can be replicated by using an equally large value for $\alpha$ and $\gamma$, a positive value for $\delta$, a negative value for $\theta$, and a zero value for $\beta$. Finally, an assignment method that is based on egress time only (Daniel Hörcher et al., 2017) would be equivalent to treating egress time as the only component of the observable utility (i.e., using a negative value for $\theta$ and a positive value for $\gamma$).

It is noted that our model does not separate walking time between the fare gentries and the platforms from first waiting time because of two reasons. First, the smart card data is precise up to one minute only while walking time between gentries and platforms in the case study of Sydney is usually less than one minute. Second, for few stations in Sydney where walking times between gentries and platforms are longer than one minute, the walking times are quite similar across feasible itineraries/platforms, resulting in walking time being cancelled out in the choice model. For a similar reason, the itinerary choice model does not consider fares since for each given OD pair where exists multiple itineraries, fares are determined by tap-on time (cf. departure time) and shortest track distance between origin and destination, and hence they are the same across multiple itineraries, and again cancel out in the model.

The itinerary choice model (1 – 4) delivers the probability that each passenger uses a particular service, given their smart card data. Aggregating these probabilities provides us the total number of passengers tap on/off, transfer on/off at each station along the itinerary. Note that we distinguish transfers from taps on/off so that the impact of transfers on crowding on platforms can be evaluated. Merging these aggregated probabilities with the timetable, either recovered from the API program or from GTFS data, gives the total number of people boarding and alighting at each station along each service. Crowding profile for each train service is simply the cumulative sum of boarding numbers (tap on + transfer on) less alighting passengers (including passengers who transfer to another service or tap off). Similarly, crowding profile of each station is a cumulative sum of passengers arriving at the station (tap on and waiting for boarding, transfer) less the number of passengers leaving the station. Since some trains may proceed to another service after finishing one service, we need to identify these services from the timetables and link them so that crowding statistics consider existing passengers from the previous train service.

3 Case study and results

3.1 The case study

To test our model, smart card data and GFTS data from Sydney Trains network is used. Sydney Trains is one of the world’s most complex systems which served
approximately 1.5 million journeys per weekday. The network includes seven train lines, labelled T1 to T7, and four intercity lines, many of which sharing the same track (see Figure 1). Overcrowding on Sydney Trains is reaching the breaking point (Matt O’Sullivan, 2017) and Transport for NSW (TfSNW) has started monitoring and addressing overcrowding on trains. For example, TfSNW run biannual train load surveys at selected stations on various train lines to support service planning and timetabling. Also, TfNSW has recently launched new timetables on the 26th December 2017 to tackle overcrowding with more services and more carriages (Transport for NSW, 2017). In addition, a few interchange stations such as Wynyard are under upgrade to accommodate increasing passenger numbers. These features make Sydney Trains an interesting case for model application.

![Fig. 1 Schematic overview of Sydney Trains network (source: Transport for NSW)](image-url)
3.2 Results

3.2.1 Itinerary matching results

Figure 2 shows the cumulative distribution of feasible itineraries and transfers per passenger journey observed in the smartcard system (known as Opal in New South Wales) on the Tuesday 1st March 2016.¹ This results from exposing the Opal data to the Trip Planner webpage using the API program with entry and exit time constraints. As can be seen in Figure 2, the API program can identify at least one feasible itinerary for all journeys observed, with the percentage of journeys having zero feasible itinerary being 0%. About three quarters of the journeys were identified with only one feasible itinerary. This means that once all possible itineraries for each journey have been identified (we use the API program described above), applying the time constraints (i.e., tap on before boarding with up to 5-minute adjustment for trains late, and tap off after alighting) results in 72% of the passenger journeys being linked to a unique service. In other words, we can identify precisely which services and stations the passengers used for 72% of the journeys, using only the smart card data and the API program. For the remaining 28% of the journeys with multiple feasible itineraries, we need an assignment method to link passengers to trains/services. Of these journeys, 21% have two feasible itineraries while 5% have three feasible itineraries, with a very small percentage of journeys having 4+ feasible itineraries. The feasible itineraries are distinguished by different departure/arrival times and/or different route options. Depending on the observed OD pair and time of day, some journeys may have many feasible itineraries (up to 64) but 99% of the observed journeys have four or less.

Turning to the number of transfers per journey, three quarters of the feasible itineraries involve no transfer while 22% of these feasible itineraries involves one transfer, with only 2.2% having two transfers. The percentage of feasible itineraries that involve 3+ transfers is very low, at 0.1%. These statistics confirming the principle employed for planning the Sydney Trains, which is to provide most customers with a one-seat journey. However, one quarter of feasible journeys involve at least one transfer, and thus the impact of transfer passengers on crowding on trains and platform is expected to be significant. We report these statistics for some key interchange stations in Sydney after conducting sensitivity analysis and evaluating the model against the manual headcount.

¹ While the Opal data are available for other dates, we select this date as a typical Tuesday for presenting and validating our model results due mainly to the availability of onboard headcount data in March 2016.
3.2.2 Sensitivity analysis and validation of route choice model

As shown in Fig. 2, a quarter of the passenger journeys have multiple feasible itineraries and a route choice model is used to predict the probability that each of the feasible itineraries was selected, without knowing the actual itinerary chosen by the passenger. Lijun Sun et al. (2015) and Daniel Hörcher et al. (2017) provide alternative ways in which one can estimate route choice model parameters with some caveats about the reliability of the parameter estimates. In this paper we draw on the literature to identify the likely value range of the various parameters that enter the route choice model (4) and conduct sensitivity tests to identify the best set of parameters that could recover the crowding levels at various stations during peak hour by line obtained from the TISNW on-board headcount survey conducted in the same period (Transport for NSW, 2016). Numerous studies have tried to quantify the impact of various attributes, such as travel time, waiting time, walking time, access time, egress time, transfers and crowding, in determining passenger route choice in urban train systems (Sebastián Raveau et al., 2014, Linjun Sun et al., 2015, Alejandro Tirachini et al., 2016, Daniel Hörcher et al., 2017); however, applying previous estimations directly to Sydney Trains network will raise the issue of parameter transferability (i.e., behavioural preferences vary significantly from city to city, and from one study to another). Thus, we start with the set of parameters/multipliers that are used in the Sydney authoritative transport model which evaluates one minute of auxiliary time (access/egress/waiting) at two minutes in-vehicle time, and each transfer at 5-minutes in computing a so-called “generalised time”.

Probabilistic assignment model based on access, journey time, and transfers
In this model, we focus on access time while ignoring egress time. Specifically, we fix \( \alpha \) at 2, \( \delta \) at 5, \( \beta \) at 1, \( \gamma \) at 0 in equation (4) and vary \( \theta \) between \(-0.02\) and \(-0.45\) to see how sensitive the model is to the generalised time parameter \( \theta \). Fig. 3 shows the difference in passenger number between two models with \( \theta = -0.020 \) vs. \( \theta = -0.45 \). Only stations where on-board headcount data are available are used for this sensitivity analysis with the aim to identify a set of parameters that produces the smallest difference between in crowding level between the model and on-board headcount. As can be seen, the crowding levels on train lines with simple operating pattern (e.g., Blue Mountain, Central Coast, Southcoast, T3, T5) are not very sensitive to the parameter of generalised time, reflected by the symmetrical and narrow distribution around zero. Conversely, the crowding levels on more complicated train lines such as T1, T2 and T4 are more sensitive to this parameter where the distributions of the difference are wider, sometime up to 150 (T2, T4) or 200 (T1) passengers per train. Note that the on-board headcounts were conducted during the peak hours (both AM and PM peaks) at busy stations, and thus a difference of less than 100 passengers at these stations is considered small.

![Fig. 3 Sensitivity tests of crowding levels to the parameter of generalised time](image)

The results that crowding levels are sensitive to generalised time parameter for complicated train lines but not so for simple lines could be explained by the number of feasible itineraries for passengers of these lines. Train users in the northwest of Sydney (e.g., Hornsby), for instance, have many options (express, limited stops via Chatswood, Macquarie University or Strathfield) to travel to and from the Sydney CBD, and thus the route choice model will influence crowding levels of T1 line the most. Bankstown or T5 lines, by contrast, have much fewer services, resulting in many users of these lines having only one feasible itinerary and thus crowding level is much less sensitive to the choice model that splits the probability amongst the feasible itineraries. This is illustrated in Table 1 which shows how the choice
probability that drives the crowding level is influenced by the value of parameter $\theta$ assumed for the generalised time. Using an example journey where a traveller has two feasible itineraries with 10 minutes apart, we replicate the model and show that the probability of choosing the first train vary from 50% to 100% depending on the value used for the generalised time parameter. When $\theta$ equals zero, the probability of choosing any train in the feasible choice set is completely random (i.e., equal probability for each route option). The larger (in magnitude) the value of parameter $\theta$, the higher the probability of choosing the first train, with a range of $\theta$ between $-0.15$ and $-0.45$ seeing the probability of choosing the first train increases to 95% and 100%, respectively. That is, considering access time, transfer and inter-platform time and ignoring egress time, a value of $-0.45$ for $\theta$ will result in a model in which all passengers board the first train arrives at their origin station.
We validate this model by comparing the model train load with on-board headcount for each station that enters the on-board headcount survey. Fig. 4 shows the histogram of the difference by train line, using three different parameters for generalised time. Three conclusions can be drawn from this comparison. First, the model consistently under-estimates crowding level for all train lines except for Central Coast. This is reflected in the majority of the distribution lying on the left of the vertical line at which the model predicts precisely the number of passengers on-board (i.e., zero difference). This is expected since on-board headcount includes all passengers, including a small percentage of passengers who still used paper tickets in March 2016 that were not registered to the smart card data the model used, and fare-evasion travellers that on average account for about 6% of train travellers, according to a recent fare compliance survey (Transport for NSW, 2018). Second, the model can replicate the crowding levels at the key stations quite well with most of the differences distributed around zero; however, in few cases the difference is as large as 500 or even 1000 passengers. This is likely to be a result of some variations to the working timetable between the 1st March 2016 and the dates when on-board headcounts were conducted or some incident such as previous train being cancelled, resulting in many people travelling on one sardine train. Third, although the difference produced by different values of \( \theta \) is small, a value of \(-0.025\) appears to produce the best results with closest matches to the on-board headcounts. Thus, we use this value for further sensitivity tests.

### Table 1 Sensitivity of itinerary choice model to the generalised time parameter: an example journey with two feasible itineraries

<table>
<thead>
<tr>
<th>Feasible Itinerary</th>
<th>Generalised time</th>
<th>Parameter ( \theta )</th>
<th>Utility ( V ) (Eq. 4)</th>
<th>( \exp(V) )</th>
<th>Probability (Eq. 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>0.00</td>
<td>0.00</td>
<td>1.0000</td>
<td>50.0%</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>0.00</td>
<td>0.00</td>
<td>1.0000</td>
<td>50.0%</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>-0.025</td>
<td>-0.63</td>
<td>0.3533</td>
<td>62.2%</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>-0.025</td>
<td>-1.13</td>
<td>0.3247</td>
<td>37.8%</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>-0.05</td>
<td>-1.25</td>
<td>0.2865</td>
<td>73.1%</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>-0.05</td>
<td>-2.25</td>
<td>0.1054</td>
<td>26.9%</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>-0.10</td>
<td>-2.50</td>
<td>0.0821</td>
<td>88.1%</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>-0.10</td>
<td>-4.50</td>
<td>0.0111</td>
<td>11.9%</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>-0.15</td>
<td>-3.75</td>
<td>0.0235</td>
<td>95.3%</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>-0.15</td>
<td>-6.75</td>
<td>0.0012</td>
<td>4.7%</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>-0.45</td>
<td>-11.25</td>
<td>1.3E-05</td>
<td>100.0%</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>-0.45</td>
<td>-20.25</td>
<td>1.6E-09</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
Fig. 4 Difference in passenger number between on-board headcount and train load produced by model assuming passengers board the first train

Probabilistic assignment model based on access, egress, journey times and transfers

In this assignment method, we consider egress time at destination stations in addition to access time at origin stations, inter-platform time and number of transfers. By using a positive value (cf. zero) for parameter $\gamma$, we consider egress times in the assignment of passengers to train. Still using a multiplier of 2 for access time ($\alpha = 2$), 5 for each transfer ($\delta = 5$), and fixing the generalised time parameter $\theta$ at $-0.025$, we conduct sensitivity tests on the egress time parameter $\gamma$ to identify the best value that could produce the crowding levels that match well with the on-board headcount data. Five values of $\gamma$ (0, 2, 5, 10, 15) are used for this exercise, with the value of 2 (i.e., $\alpha = \gamma$) resulting in a model that penalises egress time as heavily as access time. This means that the model will produce equal probabilities for each of the feasible itineraries that differ by departure and arrival times but otherwise match in terms of inter-platform time and number of transfers. Conversely, the model with $\gamma = 0$ ignores egress time entirely in assigning passengers to trains (results reported above) while $\gamma = 15$ will lead to a model that assumes all passengers were served by the last train that arrived at their destination stations. This is because every minute arrived early (or every minute longer in egress time) are penalised at the rate of 15 minutes longer in inter-platform time, and thus with a shortest headway of 3 minutes observed on the Sydney Trains network, this penalty rate will ensure that the last train has a superior utility, and hence the passengers.

Fig. 5 shows the sensitivity test results by train line. Overall, the model still underestimates the crowding levels on all train lines as compared to on-board headcounts; however, the differences are smaller with the mean of the difference closer to zero than before (i.e., the distribution of the difference between model and on-board headcount is shifting toward zero). That the model underestimates crowding levels on all train lines is expected since smart-card passengers made up a proportion of total passengers on-board that also include passengers travelling on paper-tickets or without a ticket, all were counted in the on-board headcount surveys. In March 2016 when we have the on-board headcount data to validate our
models, the overall take-up rate of Opal card was around 94% (but this varied significantly across regions with the lowest level of uptake at 73% in southwest of Sydney that is served by the T2, T3, T5 line and highest at 95.4%), and the fare evasion rate was around 6% of total travellers (Transport for NSW, 2018). The model is sensitive to the egress time parameter, with larger value of \( \gamma \) producing better results. This is reflected in the distribution of the difference where more services have predicted train loads that either match exactly with or are very close to the on-board headcounts. For example, the model that uses \( \gamma = 15 \) can predict exactly the crowding levels for 120 services out of 517 services on the T1 line that have headcount survey data. The figures for the models that use \( \gamma = 0 \) (ignore egress time information entirely) and \( \gamma = 2 \) are 108 and 109, respectively. For other train lines, the distributions of the difference between the model results and on-board headcount are mostly identical across all the values of \( \gamma \) tested.

Figures 4 and 5 shows how accurate the assignment models estimate the crowding level by train lines but this does not show how the model performs across stations along each train line. Fig. 6 looks at this by showing the extent to which the model under-estimate the crowding level at various stations where on-board headcount data were collected in order to understand where the model predicts the train load well and where it does not. Stations where on-board headcounts were conducted are shown in colours with the size of a point indicating the sample size of headcounts for each station. A surveyed station may belong to two different train lines, and in these cases the same sizes are split and shown for each train line. For example, Parramatta station in the west of Sydney is served by two train lines T1 and T5. Train loads at Parramatta station during the peak hours (AM and PM) were observed on 45 services on T1 lines and 7 services on T5 lines, and thus the same sizes of headcount for this station are represented by two bubbles: the big bubble represents T1 counts vs. the small bubble represents T5 counts. The colour of the bubble indicates the extent to which the model underestimates the average train load during peak hours at each station, with the grey dots indicating that on-board
headcounts were not conducted at these stations by TfNSW, and thus comparison is not possible.

As can be seen, the differences in passenger number between model results and on-board headcounts are quite small for all train lines except for T5 (the purple line) and T2 (the green line) via Revesby, both are highlighted in Figure 6. The fact that the model significantly under-estimates crowding levels on these two lines is in line with the lower rate of Opal uptake in the regions these two lines serve. This can be seen clearly in Fig. 6 where the West and Southwest of Sydney have an Opal uptake rate in March 2016 of 60% to 70%, compared to nearly 95% uptake rate in the inner
and the North/Northwest of Sydney which are served by T1 lines and T2 line via Granville. All stations along these lines have estimated crowding levels that match well with on-board headcounts. Given that the model is well validated with all the differences in crowding level between the model and the on-board headcounts justified, we can use the model to obtain insights into crowding on trains and platforms. These are presented in the next section.

3.2.3 Crowding on train

Visualising crowding on trains
The assignment model provides the probability that a particular train was used by the passenger. Combining this with the journey stage of passengers and the stopping patterns of the train, we develop an animation as a way of visualising crowding levels on train for each service across the entire day so that the crowding patterns (i.e., how crowding levels change across temporal-spatial dimension) can be extracted and used for service planning purposes. Fig. 7 shows a snapshot of this animation during the AM-peak. The animation shows where each train is at 08:11:00 and its crowding level in terms of both passenger numbers and train load with a load factor of 100% meaning that there is a seat for each passenger. Train loads are colour-coded while the number of passengers on-board is represented by the size of the rectangles/trains. Thus, redder and bigger rectangles indicate crowded trains. It can be seen clearly that the T1 Western line was the worst for overcrowding with many trains to the Sydney CBD in the morning peak already over-crowded from the West, leaving Inner-West passengers with standing room only. Other lines of the T1 (Northshore via Gordon and Northern via Strathfield or Macquarie Park) were also over-crowding during the peak hours. Passengers of the T4 Illawarra line were also suffered from overcrowding trains when trains were already full as early as Mortdale/Penshurst station in the South which is about 17 km from the City where many people get off, and hence train load reduces. These results are consistent with the recent statistics released by transport authority (Matt O’Sullivan, 2018).
Crowding profile for each service

While the animation is useful for identifying the train lines that currently face overcrowding problems, this information is considered high level which confirms what most train users in Sydney already knew, which was that demand for services was booming and that action was needed. Thus, to be useful for service design, insights at the lower level are needed since crowding levels change as trains proceed with some segments of the network being suffered from overcrowding more than others. In this section, we look at the crowding profile for some typical services and relate this to the service design in order to identify areas where improvements may be needed.
Fig. 8 shows a crowding profile of an example service on the T4 line in the morning peak, connecting Southern and Eastern suburbs of Sydney. As can be seen, crowding levels vary substantially along the journey with passengers travelling on the segment between Wolli Creek and Town Hall stations experienced overcrowding where passenger numbers in excess of seating capacity by 135% – a load factor considered as ‘crush load’ by TfNSW. The result lends support to the design of express service between Hurstville and Wolli Creek stations where crowding starts building up.
Fig. 8 Crowding profile of a Waterfall – Bondi Junction service on T4 line in the morning peak in March 2016

3.2.3 Crowding on stations
From the quadruple of the smart-card data (origin, destination, tap-on time and tap-off time) and the probability that the passengers took a certain train together with the departure and arrival times for each leg, we can estimate when the passengers
arrive and leave each station, including transfer stations, on their journey. We use this information to estimate and visualise station loads at any time during the day. This section first looks at the impact of transfers on key interchange stations in Sydney and then presents the results of crowding on stations/platform.

Impact of transfers on station crowding

Fig. 9 shows the total number of passengers that use the key interchange stations in Sydney as a transferring point. For graphing purposes, transfer passengers are aggregated by 10-minute block, and thus the 2,500 transfer passengers at Central station at 8:00 indicate that during the 10-minute period between 8:00 and 8:10 there were 2,500 passengers transfer at Central station, adding to the crowd of passengers who use Central as the origin or destination that are captured in the smart-card data. Compared to Central station, Redfern station is much smaller in terms of station capacity/floor area, but it received comparable transfer passengers during the am-peak. Clearly, transfer passengers add more burden on crowding at Redfern station than at Central station.

Fig. 9 Number of transfers at key interchange stations by 10-minute block: 22 March 2017

Crowding profile for each station

Combing the number of passengers who use a station as a transfer point with those who use it as the first or last station (origin or destination) we develop a station load animation to show the total number of people present at each of the 308 stations across the Sydney suburban train network. Fig. 10 shows a snapshot of this animation at 8:27. As can be seen, Redfern station is in the top five busiest stations during the AM peak, after Town Hall, Central, Wynyard and North Sydney; however, Redfern station significantly less room for passengers than the other four stations and the government is currently upgrading Redfern station to improve station accessibility and safety of the passengers. This lends credit to the current study.
4 Conclusions

This paper aims to quantify not only crowding inside the vehicle (train) but also on the platforms/stations. The latter is very timely for decision making relating to which stations will need an upgrade to accommodate increasing passenger follow. In the context of Sydney, there are some ways to approximate crowding inside the train using on-board headcount survey or train weight, but this is not the case for crowding at station. With the adoption of Opal smart cards for automatic fare collection purposes, it becomes possible to obtain these performance metrics without the need to conduct manual surveys. For instance, TfNSW used to conduct manual counts of passengers at the barriers to obtain station usage, but these statistics can be now easily obtained from the smart card data by aggregating the number of tap-on and tap-off at each station; however, obtaining crowding levels on trains and platforms/stations from the smart-card data is still a challenge because we do not know from the data which services the passengers use and where (if any) they transfer. This research has developed and validated an algorithm that can be used to identify the most likely path the passengers used from their smart card data. The model outputs are used to supplement the information recorded to the smart card data in order to identify the crowding levels on trains and platforms/stations,

Fig. 10 Visualising crowding at stations: a snapshot of station load animation
taking into account transfer passengers and how long they remain at the station/train on their journey.

Using on-board headcounts at many different stations and services across the train network, we have compared different assignment models used in the previous studies and benchmarked them against this ground truth. We found through the case study of Sydney establishing a set of feasible itineraries that are realistic for each passenger in terms of reflecting what choices they had at the time of travel is much more important than using sophisticated models to assign passengers to trains. Our sensitivity analyses suggested that although egress-time-based assignment models are better than access-time-based assignment models, the difference in average crowding level between the two types of models is small. Admittedly, the small difference is due to a large proportion of train journeys in Sydney that have only one feasible itinerary. These journeys are not influenced by the selected model which is designed to predict the most likely train the passengers used if they have more than one option. In other train/metro systems with a larger proportion of journeys with multiple feasible itineraries, the assignment method will certainly play a bigger role and this calls for a more sophisticated assignment model such as those that are able to account for failed boarding (Yiwen Zhu et al., 2017b).

This research has a practical impact on decision making to tackle over-crowding on trains and stations through timetable planning, station upgrade, and/or change to policy such as off-peak discount to encourage travellers to switch time. It clearly has a potential to change the current practice of conducting the train load surveys for monitoring over-crowding. Pre-Opal, crowding on trains statistics have been obtained by deploying surveyors to count the passengers on board at some stations and services during the peak hours across the Sydney Trains Network. This survey is very time-consuming and costly. With this research, we can establish the level of crowding on each train and each station for any given day or any time by merging smart-card data with journey planner/automatic vehicle location data.

Future research includes the extension the assignment model to introduce a capacity constrain to prevent trains from carrying more passengers than it was observed to carry. This extension is important to reflect the travel behaviour when passengers react to an over-crowding train by not boarding it but waiting for the next services. Another direction for future study relates to the application of the validated model to identify how the new timetable introduced in December 2017 – February 2018 by Transport for NSW has altered the choice the passengers have and to what extent this new timetable address over-crowding on trains and platforms/stations.
References


Sung-Pil Hong, Yun-Hong Min, Myoung-Ju Park, Kyung Min Kim, and Suk Mun Oh (2016). Precise estimation of connections of metro passengers from Smart Card data. Transportation, 43, 749-769.


