THE USE OF SMART CARD DATA TO ANALYSE PLATFORM CROWDING

Loan T.K. Ho, Chinh Q. Ho

Introduction
Crowding is one of the important factors affecting mode choice in public transport as well as contributing to overall public transport service satisfaction alongside the reliability of the service. According to the literature review of Li and Hensher (2011), willingness to pay (WTP) for crowding reduction is as beneficial as travel time savings. For example, in NSW, Australia, WTP for travel time saving in an uncrowded train is estimated at $8.45 per hour (2003 dollars), but the WTP to save travel time for a crowded train was estimated at $9.92, or a crowding mark-up of $1.47 per hour (Douglas & Karpouzis, 2006). In terms of service satisfaction, if the level of crowding satisfaction becomes 10% better, this results in a 4.6% increase in overall rail service satisfaction (Thompson et al., 2012).

Crowding can be inconvenient and uncomfortable and leads to unpleasant experience, especially when passengers must stand very close together for prolonged periods. Haywood et al. (2017) summarised the main reasons of this discomfort as being caused by three drivers: dissatisfaction with not having seats during the travel journey, less space and comfort to utilise the time productively on public transport and distress caused by traveller physical closeness. When passenger numbers exceed seating capacity by a certain percentage, crowding can also make passengers feel unsafe (Bel, 2013), and dwell time can be prolonged, which can impact on-time running, leading to influence on rail line capacity (Lam et al., 1999; Vuchic, 1981).

The remainder of this chapter comprises four sections. The first section briefly reviews the literature, with a specific focus on methods used to estimate the crowding on platforms. This is followed by a description of the Opal smart card system used in the Sydney Trains network and the structure of the Opal smart card data and the methodology to estimate crowding levels on platforms from this smart card data. The third section discusses key findings, and the chapter ends with a conclusion.

Literature review on platform crowding
Crowding is perceived in two aspects: objective (density measurement) and subjective (the perception and tolerance of passengers based on the objective levels of crowding). The subjective aspect can lead to changes in mode and path choice behaviour (Kim et al., 2015) as crowding
levels change, although behavioural heterogeneity is observed, and this source of preference heterogeneity stems from passenger’s emotion and experience (Li & Hensher, 2011).

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 identify crowding levels at different points on the entire network and quantify the discomfort that crowding may cause to passengers. Recently, a growing number of choice models have been developed to evaluate passengers’ willingness to pay for less crowded trains, either in terms of a longer travel time or by a higher transport fare (Batarce et al., 2016; Hörcher et al., 2017; Tirachini et al., 2013); however, translating these model results into advancing public transport planning and rail timetabling is still challenging, as this requires crowding profiles, in addition to knowledge about how sensitive passengers are to different crowding levels.

Before smart card data were widely used as inputs for estimating levels of crowding (both at train stations and platforms), surveys were used to acquire a count of the number of passengers at the stations, together with information on trains scheduled and real-time timetables and station layouts. Lam et al. (1999) assessed crowding levels at two downtown light rail transit (LRT) station platforms in Hong Kong using stated preference surveys in the morning peak on the LRT platforms (number of users departing, arriving, boarding and aligning) and interviewing passengers for their travel behaviour and routine. The crowding information was divided into various levels of congestion both on the train and on the platform as inputs to regression models measuring dwell time delays and discomfort levels caused by crowding. Their model revealed that half of the passengers surveyed preferred to choose the least crowded platform condition (on the scale from 1.2 m²/pedestrian to <0.28 m²/pedestrian), with a dwell time delay of up to 2.7 minutes added to an average waiting time of 3 minutes for an average in-vehicle travel time of 15 minutes. This stated preference survey study also found that train users similarly perceived crowding on the platform with the same waiting time at different stations. However, passengers with shorter travel-time trips are more tolerant of crowding on the train.

Crowding in rail stations can be studied by image processing techniques from surveillance data at the stations. Surveillance data are used to estimate passenger flow based on which crowding profile is established. Virgona et al. (2015) used data from three-dimensional cameras to sense people by detecting individual heads and shoulders. This study leverages information from sensor systems at train stations to mitigate the privacy concerns associated with the use of standard colour CCTV cameras. This approach, however, is not widely applied, mainly because the authors acknowledge that some technical issues, such as angle effect (front, side, back) and inconsistent shape and style of the human body (head size, hairstyle), should be dealt with to provide a more accurate passenger flow.

The application of smart cards for ticketing, also called automated fare collection (AFC) systems, offer great opportunities to monitor crowding levels on both trains and platforms by providing a big data source for different times of a day and different days of a week. This is also helped by the way similar data storing formatting in different systems allows similar data processing methods over multiple systems. Hong et al. (2016) estimated crowding on trains using smart card data and timetable data. These authors developed algorithms to identify boarding, transferring and alighting patterns of passengers based on their tap-on and tap-off times of certain pairs of origin (O) and destination (D) available in the smart card data. The trip patterns for specific entry and exit intervals at O-D stations are then derived from train timetables. This method is validated with an accuracy level of 92% for detecting unique trip sequences based on one week of smart card data including the weekend in Seoul, Korea. This method was also adopted by Kim et al. (2015), which used a logit model to examine how passengers changed their journey...
Smart card data and platform crowding

path so as not to encounter or to escape from the crowd at the station. The model is tested on the real trip sequences from smart card data acquired by Hong et al. (2016).

Smart card data, in conjunction with automated vehicle location (AVL) data, can be used to recover train schedules in real time. Koutsopoulos et al. (2017) used these data sources to investigate levels of congestion at 47 stations of two underground lines in London. The tap-on and tap-off times of passengers (demand side) were assigned to match with AVL data of the underground system (supply side) to find out their trip patterns using a three-step probability model. Predictive analytics were then applied to forecast crowding at each station platform.

This literature review suggests that only a handful of studies that estimate crowding at the platform level from smart card data can be found in the literature. This chapter contributes a case study using the Sydney smart card data and the next section provides the empirical context of this chapter.

The empirical setting

This case study chapter uses data from the Opal card, the smartcard used by travellers in Sydney, NSW, Australia. The Sydney Trains network is one of the world’s most complex systems, serving approximately 1.5 million journeys each weekday and about 1.1 million journeys each weekend day. The network included eight train lines, labelled T1 to T8 at the time of the data collection, and four intercity lines, many of which share the same track. Figure 34.1 shows a schematic overview of the Sydney Trains network as of August 2018 (Transport for NSW, 2018) when the smartcard Opal data used in this study were retrieved. Overcrowding on the Sydney Trains network, both on trains and platforms, was identified as reaching a breaking point (O’Sullivan, 2017). Transport for NSW (TfNSW) started to monitor and address overcrowding on trains from this time, running biannual train load surveys at selected stations on various train lines to support service planning and timetabling. Also, TfNSW recently attempted to tackle overcrowding with more services and more carriages (Transport for New South Wales, 2017) as well as providing major upgrades to accommodate increasing passenger numbers at a number of important interchange stations such as Wynyard and Redfern. These features make the Sydney Trains network an interesting case study.

To identify crowding on platforms, four elements are needed: (i) the travel demand matrix, commonly referred to as origin-destination matrix or O-D matrix for short; (ii) the service/train each passenger uses and where they transfer (if any); (iii) the stopping pattern of each service, including platform details; and (iv) the platform and station layouts. In the case of Sydney, the O-D matrix is obtained from the Opal data themselves, since train passengers are required to tap on and tap off in order to pay the correct fare for each journey and enter and exit some stations, especially those in the central business district (CBD). As with many train and metro networks around the world, the Sydney Trains network is designed as a closed system where passengers only leave traces at boarding and alighting stations. This means there is very limited knowledge about which trains/services a passenger uses to travel for the observed origin-destination (OD) pair, especially if the OD pair has multiple route options. A few approaches have been developed in the literature to assign passengers to trains in order to obtain this second piece of information. This case study uses a two-stage modelling method developed by (see Ho & Ho, 2018 for details) to obtain the information required in (ii) using an API that queries the official journey planners (www.transportinfo.nsw.gov.au) and returns all possible combinations of services that can take the passenger from the origin to the destination. Time constraints are then applied (i.e., tap-on must be before boarding and tap-off must be after alighting) to identify all feasible trains/services for each journey observed in the Opal
Figure 34.1  The Sydney Trains Network as of August 2018

Source: Transport for NSW, 2018
data. Once the choice set for each journey is formed (i.e., all feasible trains/services have been identified), a utility-based assignment method is used to compute the probability that each train/service was used.

The stopping pattern for each train service (i.e., express, limited stops, or all stops service) is available in the timetable via the GTFS data bundle. For the purpose of quantifying crowding at the platform level, it is necessary that the stopping patterns include platform information, in addition to stations where each train would stop along its journey. Alternatively, stopping patterns can be established from AVL data if these are available. For the case of Sydney, GTFS real-time data are used, which is in essence a combination of GTFS and AVL data. These datasets provide not only the stopping pattern but also the continuing pattern if a train proceeds to another service/line when it arrives at the final stop where the service identifier (i.e. trip_id) may change but passengers can remain on board and continue their journey. This operational complication of the Sydney Trains network is distinct from most metro networks in which the train stops at every metro station along the track and serves that track/line only. Finally, the layout of each station (including platforms) was obtained from a GIS database owned and maintained by Sydney Trains.

The high-level workflow is summarised in the following steps:

1. GTFS real-time data is merged with the Journey Planner data to identify the stopping and continuing pattern of each train. This step enriches the Journey Planner data with real-time information from the GTFS real time. In doing so, platform information is added to the route and any modifications to the scheduled departure and arrival times to the actual time for every stop to account for the cases where trains are not punctual to the timetable and/or some last-minute changes that were not updated in the Journey Planner. There were very few of the last-minute changes (<0.1%).

2. Opal data (Tap on/off) is then merged with the enriched Journey Planner data to identify feasible itineraries for each Opal tap on/off record. For a complex train network where trains share tracks, like the Sydney Trains network, generating the comprehensive set of trains for each O-D pair is the most challenging task, since the number of possible alternatives increase exponentially if traditional methods are used. As identified previously, an API was developed (see Ho & Ho, 2018) and followed by applying time-based constraints (i.e., tap-on must be before boarding and tap-off must be after alighting) to select only feasible trains from the comprehensive set of trains.

3. A passenger-to-train assignment model is applied to compute the probability that a user chose each itinerary from the set of feasible itineraries identified in step 2. Although different assignment methods can be used to assign passengers to trains, this study chose to use the utility-based assignment rule described in Ho and Ho (2018).

4. Sensitivity analysis is performed to identify the set of passenger-to-train model parameters that best replicate the crowding levels observed at various stations by the on-board headcount. The best set of parameters for generalised travel time, transfer, access, egress, inter-platform time, and overcrowding penalty to compute crowding profile for each train and each station is identified. This best set of parameters are used to re-compute the probability that each train in the feasible set was selected (step 3). This provides all necessary information for computing the number of people on each train, each station, and each platform at any time.

5. Station layout is then used to estimate the time required to reach or exit the platform from the previous activity, be it tap on or arriving for transfer or tap off.
6 Compute the total number of people on each platform based on the results of step 5 (time people arrived and left each platform). This is the cumulative sum of arriving passengers less the cumulative sum of departing passengers.

7 Convert the total number of people on each platform to its crowding level by time of day using the total area of each platform.

Results

Feasible trains

Figure 34.2 shows the cumulative distribution of feasible itineraries/routes and number of transfers per passenger journey observed in the Sydney Trains network on Tuesday 20 March 2018. This is the result of step 2 previously, after exposing the Opal data to the enriched Journey Planner data scraped from the official transport information webpage and applying the time constraints to require that tap-on must be before boarding and tap-off must be after alighting. As can be seen in Figure 34.2, around 65% of the observed journeys have only one feasible itinerary/route given the tap-on and tap-off times and locations. This highlights the difference in service headway between the Sydney Trains network and other metro networks such as Hong Kong and Seoul, where the vast majority of journeys observed in the smartcard data have multiple feasible trains. This difference means that for the case of the Sydney Trains network, it is much more important to have an accurate set of feasible trains than a sophisticated assignment method, since, once all possible itineraries for each journey have been identified, applying the time constraints results in 65% of the passenger journeys being linked to a unique service. In other words, the services and stations the passengers used for 65% of the journeys can be established with precision using only the smart card data and the enriched Journey Planner data. For the remaining 35% of the journeys with multiple feasible itineraries, an assignment method (step 3) is used to link passengers to trains/services.

Turning to the number of transfers per journey, about 75% of the feasible itineraries involve no transfer, while 23% of these involve one transfer, with only 2% having two transfers. The
percentage of feasible itineraries that involve 3+ transfers is very low, at 0.1%. These statistics confirm the principle employed for planning the Sydney Trains network, which is to provide most customers with a single-seat journey. However, more than one quarter of feasible journeys involve at least one transfer, and thus the impact of transfer passengers on crowding on train stations and platforms is expected to be significant. These statistics are reported for some key interchange stations in Sydney after validating the model against the manual headcount conducted in March 2016 (step 4).

**Crowding on stations**

Figure 34.3 shows crowding profiles of the top ten busiest stations in Sydney on the same Tuesday (20 March 2018). This is a result of applying a passenger-to-train assignment model (step 3) and aggregating the number of passengers present at each station by time interval (here it is every 5 minutes). The stations are ranked by daily passenger flows and the top ten stations selected for presentation. As can be seen, Central station is the busiest station in terms of passenger number, followed closely by Town Hall station. Central station is busier than Town Hall station during the morning peak, but the reverse is true during the afternoon peak. For example, during the 5-minute interval between 8:00 and 8:05, Central station received nearly 4,500 passengers, or an average of 900 passengers per minute. Between 18:00 and 18:05 on the same date, Town Hall station received 5,000 passengers, or an average of 1,000 passengers per minute. The remaining eight busy stations in the top ten received far fewer passengers, but sometimes some of these stations (e.g., Wynyard and Redfern) received more than 2,000 passengers over a 5-minute interval during the peak hours (7:00–9:00 and 16:00–18:30). It is noted that while station loads as presented in Figure 34.3 are useful to identify stations where overcrowding may be an issue, station loads are not the same as the crowding level experienced or perceived by train passengers. This is because some stations are bigger than others in terms of platform and concourse areas. Also, within the same station, some platforms may receive more passengers than others, depending on the timetable scheduling. The next section presents platform load results where pedestrian flows at the station level are converted to passenger flow at the platform level.

**Crowding on platforms**

Figure 34.4 shows the number of passengers present on each platform of Central station by time of day. To obtain this result, an average walking speed of 1 m/s inside the train station is adopted, taken from the research by Railway Technical Research Institute (2016) (Chapter 21), to estimate the time each passenger arrived and left the platform, given the walking distance from the previous location, be it an entry point (i.e., tap-on) or an alighting platform; the latter includes alighting for transfer and alighting for the final destination (i.e., tap-off). For the purpose of demonstration, this chapter assumes the same average walking speed for all passengers and that this walking speed is independent from the crowding density. In reality, however, walking speeds vary across passengers and reduce substantially as the crowding density increases (Bruno & Venuti, 2008). This limitation could be lifted by estimating individual walking speed from the smart card data, making the outcome a function of crowding density; however, this is beyond the scope of the current chapter.

As can be seen in Figure 34.4, passenger flows differ significantly across the platforms within the same station (i.e., Central) for the same time period. Central station is the most important interchange hub in the NSW Trains network, with all train lines except for the T5 line going through this station (see Figure 34.1). This station provides all train services, including urban,
Figure 34.3  Crowding on stations: top 10 busiest stations in Sydney on Tuesday 20 March 2018
Figure 34.4  Crowding on platforms at Central station on Tuesday 20 March 2018
intercity, rural, and regional. Platforms 1 to 12 are used for intercity and rural and regional trains (with lower service frequency), while Platforms 16 to 25 are used for urban trains that have a much higher service frequency. This explains the distinction in passenger flow between Platforms 1–11 and Platforms 16–25 observed in Figure 34.4. As can be seen in this figure, Platforms 18 and 19 are busiest in the afternoon (around 6 PM), while Platforms 16 and 17 are busiest in the morning (around 8:30 AM when most people arrive for work starting at 9 AM). Between 6 PM and 6:30 PM, Platform 18 constantly holds more than 1,250 passengers, while Platform 19 has around 750. Together, these two platforms receive around 2,000 passengers per minute. It is noted that Platforms 18 and 19 are island platforms, which is a station layout arrangement where a single platform is positioned between two tracks within a railway station. Central station was designed with many island platforms (e.g., Platforms 16 and 17, Platforms 8 and 9). With a size of 160 m long and 10 m wide (including staircase holes and on-platform train controller offices), this island platform has an average density of more than 2,000/1,600 = 1.25 people/m². Compared to the 6 level of services typically used in transport engineering (i.e., Level A = no passenger needs to sit next to another, while Level F = crush load), this platform density is considered Level D for waiting and Level E for walking area, according to Fruin (1971) and the Transit Capacity and Quality of Service Manual (Transportation Research Board of the National Academies (TRB), 2013). This level of service is worse than the Level C that RailCorp (the agency for rail property assets, rolling stock, and rail infrastructure) aims to provide, according to Douglas and Karpouzis (2005).

The operator, Sydney Trains, can use evidence provided in this section, particularly in Figure 34.4, to reduce crowding on the platform. For example, Platforms 18 and 19 of the Central Station are both busy during the afternoon peak, but they are much less busy in the morning peak. The reverse is true for Platforms 16 and 17, which are busy in the morning peak. Thus, if all trains that are scheduled to stop on Platform 19 in the current timetable are diverted to Platform 17 instead (and vice versa), the crowding on both Platforms 16/17 and 18/19 would reduce substantially because the total number of passengers on these island platforms are spread more evenly across the peak periods (i.e. one line is busy in the morning peak, one line is busy in the afternoon peak). The current timetable schedules trains that are busy during the same peak period to stop on the same island platform. This makes crowding on the platform worse when trains on these lines arrive at Central station at the same time. With this simple operational twist, passengers could be spread across multiple island platforms so as to reduce the level of crowding on platforms. The disadvantage of this approach is that transfer passengers would need to walk down the staircases and up again, but the transfer time is expected to be quite short, since the two island platforms (i.e., Platform 16/17 and Platform 18/19) are next to each other.

Conclusions

This chapter has shown that the crowding level on each train platform can be estimated using smart card data in conjunction with Open Data such as real-time GTFS (or real-time vehicle location data) and the Journey Planner API. This case study of the Sydney Trains network, which is one of the most complicated train networks in the world, where many trains share the same tracks, has demonstrated how this is done by first recovering feasible trains for each observed journey from the smart card data using time-based constraints and then applying a passenger-to-train assignment method to compute the probability that each train in the feasible set was selected. Detailed instructions for each step are provided to facilitate the application of the method developed herein to other train networks.
Obtaining crowding on platforms for each of the stations in any busy train network is important not only for the purpose of designing station layouts but also for train timetable planning and economic appraisal of initiatives aiming to improve customer experience by increasing station and/or platform capacity. Many studies have estimated customers’ willingness to pay (or marginal substitution rate) for reducing crowding, both on trains and on stations/platforms, but the application of these study outcomes (i.e., WTP) requires the crowding levels under the ‘status quo’ (or business as usual) and the ‘do something’ scenarios. This chapter has shown a method to obtain the latter so that economic benefits of improving pedestrian experience inside a train station can be computed by applying the corresponding value of waiting/walking time under different levels of crowding.

An addendum to this chapter, written before the COVID-19 pandemic, is the obvious question as to the extent which crowding on public transport, both on trains and stations, is a concern in the post-COVID-19 world. Many jurisdictions, including New South Wales, are reopening their economies after a lockdown, but physical distancing is expected to be observed in public places, including on public transport. It appears that the analyses presented herein are very useful to provide a foundation upon which various scenarios of travel in the ‘new normal’ world, including the staggered start and finish times, could be simulated and the reality of practicing physical distancing on public transport can be assessed. The reader is referred to Ho (2020, May 18) for a detailed analysis and discussion.

Acknowledgements: The authors thank Transport for NSW and Sydney Trains for providing necessary data for use in this chapter.

Note

1 Journeys missing a tap-on or a tap-off point will incur a default maximum fare, while journeys with a missing tap-on are considered fare non-compliant. When caught, non-compliant travellers will face a $200 fine.

References


