Crowding cost estimation with large scale smart card and vehicle location data

Daniel Hörcher*, Daniel J. Graham, and Richard J. Anderson

Railway and Transport Strategy Centre 
Department of Civil and Environmental Engineering 
Imperial College London

Submitted for Transportation Research Part B
First submission: 8 June 2016 
Revised version: 17 September 2016 
Final version: 27 October 2016

Abstract

Crowding discomfort is an external cost of public transport trips imposed on fellow passengers that has to be measured in order to derive optimal supply-side decisions. This paper presents a comprehensive method to estimate the user cost of crowding in terms of the equivalent travel time loss, in a revealed preference route choice framework. Using automated demand and train location data we control for fluctuations in crowding conditions on the entire length of a metro journey, including variations in the density of standing passengers and the probability of finding a seat. The estimated standing penalty is 26.5% of the uncrowded value of in-vehicle travel time. An additional passenger per square metre on average adds 11.9% to the travel time multiplier. These results are in line with earlier revealed preference values, and suggest that stated choice methods may overestimate the user cost of crowding. As a side-product, and an important input of the route choice analysis, we derive a novel passenger-to-train assignment method to recover the daily crowding and standing probability pattern in the metro network.

Keywords: public transport, crowding, revealed preference, smart card data, AVL data

*Corresponding author – Electronic address: d.horchert@imperial.ac.uk
1 Introduction

Public transport services are congestible: as the number of users increases relative to the available capacity, the attractiveness of the service drops and customers become less willing to use it or pay the same fare. This reduction in willingness to pay can be considered as a user cost, just like the value of time that passengers have to allocate for travelling. However, the cost of crowding is much more difficult to measure than other consumption externalities in transport, e.g. the time lost in road congestion. In the latter case there is a physical relationship between the density of traffic and the average speed of vehicles. In public transport travel time is almost unaffected by the density of passengers\(^1\). The user cost of crowding appears in the form of discomfort caused by the physical proximity of fellow travellers\(^2\), less personal space and limited access to certain amenities of the vehicle, such as preferred seats, fresh air, handle bars, or quick access to doors.

The main planning and policy areas where crowding cost measurements are utilised are demand modelling, investment appraisal as well as service quality and tariff optimisation. Considering that the cost of discomfort in very high crowding may reach the uncrowded travel time cost of the same trip (Whelan and Crockett, 2009), congestion has been shown to be an important factor of demand, and therefore congestion-relieving investments do provide important welfare benefits for society (Cats et al., 2016, Haywood and Koning, 2015, Prud’homme et al., 2012). Crowding costs have important implications on the optimality of supply-side decisions too (de Palma et al., 2015, Tirachini et al., 2013). As crowding is a consumption externality, the optimal fare should reflect the marginal external cost that an additional trip imposes on fellow passengers, when capacity cannot be adjusted. Thus, adapting methods originally developed for road congestion pricing in public transport is clearly a relevant challenge on the research agenda. Given the rising interest in understanding crowding-related problems in public transport planning and policy, it is crucial to improve the empirical methods that we apply to quantify the discomfort that crowding causes.

Decades after the appearance of the first speed-flow functions, advances in discrete choice modelling finally made the quantitative measurement of crowding disutilities possible. Wardman and Whelan (2011) and Li and Hensher (2011) provide comprehensive reviews of the evolution of crowding cost estimation. The vast majority of crowding disutility measurements were performed using stated preference (SP) techniques. Revealed preference experiments are less widely used by researchers\(^3\), mainly because it is difficult to collect data and control

---

1 increased boarding and alighting times and other congestion-related delay factors are partly or fully compensated by shorter headways and less waiting time in peak periods. The actual travel time that passengers experience may remain independent of crowding, especially in case of high capacity urban rail systems.

2 for a detailed review on the psychological aspects of crowding discomfort see Evans and Wener (2007) and Thomas (2009).

3 Some of the innovative exceptions include Kroes et al. (2013) and Tirachini et al. (2016).
for all travel attributes in real choice situations, or there is no sufficient variation between alternatives in crowding levels. Our goal is to illustrate that surveying and SP methods are no longer needed to measure the cost of crowding when such data are not available, and a comprehensive revealed preference experiment can be conducted instead using automated fare collection (AFC) and vehicle location (AVL) data.

The revealed preference approach that we propose here provides a number of advantages compared to traditional SP methods. Most importantly, data collection is cheaper, faster, more reliable and reproducible. Also, limitations in sample size can be relaxed significantly. Choices are realistic in the sense that we observe how passengers behave in their everyday life, under ordinary conditions. Additionally, the in-vehicle travel environment often varies during a public transport journey – in our RP framework these variations can be controlled for, while in a survey it is hard to describe and explain them for respondents. In general, the range of retrievable choice attributes is wider, given the rich information content of AFC and related datasets.

However, to enjoy these advantages we have to cope with multiple challenges:

1. We have to recover passengers’ in-vehicle crowding experience. This can be solved by merging AFC and AVL data in a passenger-to-train assignment process.
2. The actual route chosen is not recorded in the data, it has to be inferred. We present two alternative methods to identify route choices.
3. Crowding varies across different segments (links or inter-station legs) of alternative routes. On the other hand, what we observe is a choice between two routes, i.e. two sets of links. Therefore we have to find a way to aggregate non-additive link-level attributes in the choice model.
4. In reality passengers are not fully informed about travel conditions on alternative routes, especially not prior to the trip; their decisions are based on expectations, that we have to model with reasonable assumptions.

We address all these challenges in the paper. The ultimate goal of the experiment is to estimate a crowding-dependent travel time multiplier as a function of crowding density and the probability of standing.

The paper is structured as follows. Section 2 reviews the earlier literature of crowding-related empirical analyses. We devote separate sub-sections for passenger-to-train assignment techniques, as well as the evolution of stated preference crowding cost estimation models and recent contributions with revealed preference experiments. Section 3 is the backbone of this paper: after a data description we present the passenger-to-train assignment methodology, which is a key input for the subsequent behavioural analysis. Section 3.3 details the discrete choice model, and the way in which route choices and route attributes are inferred from the data. Finally, Section 4 delivers the estimation results and Section 5 concludes.
2 Previous Literature

The main contribution of this paper is a revealed preference crowding cost estimation experiment. However, in order to analyse customers’ decisions we have to recover the travel conditions that they experienced and expected to experience on competing routes – most importantly, the density of in-vehicle crowding. Therefore the passenger-to-train assignment method, discussed in Section 3.2, is a crucial preparatory part of the analysis. In Section 2.1 we review previous attempts to merge smart card and metro vehicle location datasets, while Section 2.2 investigates the literature of crowding cost estimation.

2.1 Inferring in-vehicle crowding

Smart card data (SCD) refers to the information retrieved from automated fare collection systems widely used in major urban public transport networks. Electronic ticketing gained popularity because of its obvious advantages versus paper tickets in the form of a faster, cheaper, safer, more transparent and more comfortable mean of tariff enforcement. However, as Bagchi and White (2005) envisioned in an early contribution, SCD is an important information source for travel behaviour analyses as well. Chu (2010) and Pelletier et al. (2011) published comprehensive summaries of recent experiences with SCD in demand modelling and transport planning.

An important step in data-based crowding analysis is to measure train loads by assigning the demand pattern to the available capacity. The quality of this assignment depends on the available information on train movements. In the simplest case what we know is the planned average frequency of trains (Spiess and Florian, 1989), and therefore only the average train load can be extracted for a given period of time. When the entire schedule of trains, including departure and arrival times at each station, is available for the researcher, then in a schedule-based assignment passengers can be linked to unique trains (Frumin, 2010, Kusakabe et al., 2010, Nuzzolo and Crisalli, 2009), which leads to a more disaggregate measure of crowding in the network. By replacing published schedules with actual train positions recorded in the signalling system, the accuracy of the assignment further improves, because delays are taken into account and the implied duration of passenger movements (i.e. access and egress times) can be measured with seconds precision.

In an urban rail context the first publicly available attempts to merge AFC and AVL data were described by Paul (2010) using London Underground data and Zhu (2014) for Hong Kong MTR. Paul (2010) recovered the distribution of egress times from smart card trips with only one feasible itinerary, and derived the access time distribution for each station assuming that the scaling factor is the same as the ratio of average access and egress times in manual passenger movement surveys. Furthermore, she assumed that walking speed is a constant individual characteristic of passengers, and a trip can be assigned to feasible
itineraries such that the implied access time should be in the same percentile as the egress time at the destination. Although Paul (2010) needed strong assumptions on the reliability of individual walking speed, it is clearly an advantage of her method that all passengers were assigned to a train, so approximating crowding patterns became possible for an entire metro line. Zhu (2014), by contrast, treated the distribution of walking speed among passengers explicitly; a valuable side product of her experiment was actually that she estimated the parameters of the distribution of walking speed. To do so, however, assumptions had to be made about the walking distance which is strongly affected by passengers’ strategic behaviour when choosing a boarding location along long platforms. Zhu (2014) applied her method for single trips with no transfers at all, outside the peak period to avoid complications caused by station congestion. Therefore the method in its current format is not suitable to reproduce the crowding pattern in an entire network.

The most recent assignment method has been published by Hong et al. (2015). They assumed that the check-out time intervals of passenger groups alighting from subsequent trains can never overlap with each other, and in case of departure stations no more than two check-in time intervals may overlap. We found that these assumptions, especially at large stations with multiple access ways, do not hold in the Hong Kong context at least for two reasons. Peak headways may be less than two minutes, which is much lower than 3.5 minutes in Seoul, according to Hong et al. (2015), and often less than the observed range of egress times. Also, failed boardings are very usual in peak periods, which implies that three or even more boarding groups may accumulate and mingle at busy stations. Failed boarding may occur not only at the origin station, but also during transfers. As a result, the number of feasible itineraries (train combinations) may explode in case of multi-transfer trips.

The method proposed in Section 3.2 allows for more flexibility in case of overlapping boarding, alighting and transfer time intervals, in a fully probabilistic framework. It is simpler than Paul (2010) and Zhu (2014) in the sense that we do not attempt to decompose access times into walking and waiting time elements, including potential delays after failed boarding. We measure the distribution of the aggregate access time instead. As opposed to what Paul (2010) and Zhu (2014) suggest, we derive route choice probabilities from the full set of feasible itineraries on alternative routes, without a separate random utility route choice layer. In general, our objective is to develop a simple assignment method which is able to recover the crowding pattern in a heavily congested metro network based on AFC and AVL datasets only, without additional information from manual counts and surveys.

2.2 Measuring the user cost of crowding

Crowding as a travel demand determinant has been identified around the late 1980’s, much later than travel time related trip attributes. The subsequent evolution of crowding cost measurements has been reviewed and summarised by two often cited articles, Wardman and
Whelan (2011) and Li and Hensher (2011). Instead of listing all related studies, we focus on major methodological improvements. Crowding cost functions express the monetary or travel time value or crowding disutilities in function of a physical representation of passenger density. One can identify six major stages in the evolution of crowding cost functions:

1. Measurement of the standing penalty, without considering the actual or expected density of standing passengers (Fowkes and Wardman, 1987).
2. Load factor based crowding cost functions (Accent Marketing and Research and Hague Consulting Group, 1997) that allow the cost of discomfort to be expressed as a continuous function of demand relative to seat supply. Disadvantage: load factor depends strongly on the vehicle’s interior layout, and therefore results cannot be applied for different rolling stock design.
3. Crowding representation based on average passenger density per square metre (MVA Consultancy, 2008). This alleviates the disadvantages of load factor based metrics.
4. Value of time multipliers (Whelan and Crockett, 2009). The user cost of a unit of in-vehicle travel time depends on the level of crowding. Thus, crowding multipliers can be applied across populations with different baseline uncrowded values of travel time, as long as they are equally sensitive to crowding.
5. Controlling for seat preferences (Wardman and Murphy, 2015) and additional nuisance factors such as noise, smell, lost and wasted time, and the risk of injury and robbery (Haywood et al., 2015).
6. Controlling for consumer heterogeneity with respect to market segments (Faber Maunsell and Mott MacDonald, 2007) as well as income and geographic location (Whelan and Crockett, 2009).

The vast majority of the papers and consultancy reports available in the literature are based on stated preference experiments. The limitations of surveying in the form of sample size, potential response biases and implementation costs are well-known (Wardman, 1988). The main reason why revealed preference studies are scarce in this area is that it is hard to observe choice situations where the trade-off between crowding and other trip attributes is visible and measurable for the researcher with sufficient variation between alternatives.

There are exceptions, however, like LT Marketing (1988) who recorded at Seven Sisters station in London whether passengers chose crowded express trains or a parallel stopping service with more available capacity. This is an obvious trade-off between discomfort and travel time. Kroes et al. (2013) realised a similar exercise in Paris, at the junction of a heavily used and an uncrowded branch of a suburban railway line heading to the city centre. After comparing their RP results to a similar SP choice situation they found that in practice a lower percentage of passengers wait for the next train than they state in the survey. The authors’ explanation was that in the RP choice there is uncertainty about the crowding level.
on the next train. We have to point out a limitation of the data collection of this experiment though: they only observed crowding at the boarding station, while passengers' decisions are affected by the overall expected crowding experience throughout the entire trip.

Batarce et al. (2015) also combined RP and SP methods to measure the cost of crowding in Santiago, Chile, and found comparable results with respect to the rest of the literature. They did not make any assumption about the functional form of the crowding cost function, they used categorical variables instead for three levels of crowding to capture nonlinearity in the cost function. The source of 'load profiles in the different metro lines', on which their crowding measurements are based, is not specified. Kim et al. (2015) found evidence in a smart card RP route choice setting, using the path detection algorithm of Hong et al. (2015), that train occupancy does affect the route choice preferences of metro users in Seoul. However, they could only express crowding discomfort as an additive utility component, i.e. the parameters in the travel time multiplier specification had remained insignificant. Finally, Tirachini et al. (2016) used the fact that some passengers are willing to travel a couple of stations in the opposite direction of a metro line in order to secure a seat in the direction of their actual destination. They used smart card data from Singapore, but they only measured the difference between standing and seated disutility in function of crowding density (in other words, a crowding-dependent standing penalty function), which is again a limitation.

We conclude this section with a summary of potential improvements identified in the literature:

1. The crowding experience varies between different legs of a public transport journey. Choice models have to be accommodated for varying conditions.
2. In an urban rail context passengers have no certain information, only expectations about whether they will stand or sit. Therefore choice models should control for the expected probability of finding a seat, which also varies during the journey.
3. Crowding cost estimations published so far are based on stated preferences, or revealed preferences with limited information about the expected crowding experience for the entire journey.

This paper shows that combined smart card and vehicle location data can be used to tackle these deficiencies in a revealed preference framework.

3 Data and Methodology

Our experiment has two main parts from a methodological point of view: the passenger-to-train assignment (Section 3.2) and the revealed preference route choice model introduced in Section 3.3. As in a revealed preference experiment attribute levels on competing alternatives are not provided from the same data source, their recovery may require additional efforts.
Section 3.4 explains how link and route level attributes were extracted from our AFC and AVL datasets merged in the assignment process.

3.1 Data sources

The AFC and AVL datasets were provided by Hong Kong MTR, the urban and suburban rail operator of Hong Kong and a member of the Community of Metros facilitated by the Railway and Transport Strategy Centre (RTSC) at Imperial College. Metros around the world are increasing interested in exploiting the information content of the AFC and AVL data they collect. However, these datasets have to comply with certain requirements to become suitable for the analysis presented in this paper.

First of all, all trips as well as trains have to be registered in the datasets in order to derive crowding patterns reliably. This condition is satisfied in our case, as in the MTR network all stations are fenced and our AFC datasets contain all trips performed in the network. By contrast, in certain metros the introduction of smart cards is in a transitional stage, so that not all ticket types are registered at the fare gates. In other cases stations are not fenced due to restrictions in station design, which implies that, depending on the rate of fare evasion, some passengers may enter the metro facilities without being registered. Travel pass holders not being obliged to check in is another threat for data completeness. Obviously, the quality of data also has to be satisfactory, i.e. the number of false smart card transactions and physically impossible trips have to remain under an acceptable threshold.

Second, the AFC data has to contain information about transactions at both the departure and destination stations. Many metros, especially those with a flat fare policy, do not register passengers at the destination\(^4\). Pelletier et al. (2011) reviewed a number of experiments that were able to estimate the destination of trips with high reliability. However, without the check-out time it becomes very difficult to assign passengers to trains, especially during rush hours when failed boardings are usual. We avoided these difficulties as in Hong Kong fares are determined on an origin-destination basis, so passengers always have to check out.

Third, the AVL datasets also have to be comprehensive for the experiment, at least in terms of precise departure and arrival times at each station, for each train. A usual challenge

\(^4\)The availability of check-out data can be a major concern hindering the direct applicability of our method in other metro networks. To assess the severity of this concern we made a short survey among the 31 members of the CoMET and Nova metro benchmarking groups managed by the RTSC. We found that with the exception of Berlin and Barcelona, 29 out of 31 metros use smart card or magnetic ticket systems as the predominant mean of fare collection. Slightly more than half of them, 16 metros do collect check-out data for each smart card trip. Check-out data collection is strongly linked to the fare structures: metros with flat fare tariffs usually do not require passengers to check-out, while for a distance or zone based network it is inevitable that smart cards have to be registered at exits as well. As a rule of thumb we identified that most Latin American members do not have check-out data, the picture is mixed in Europe and North America, while most of the rapidly developing Asian metros apply usage dependent fares, so checking out is usually obligatory for passengers.
is that metros may have a wide variety of signalling systems in their network, so that AVL data is not collected for all parts of the network in an integrated database. As we discuss later in this paper, missing data on a small number of lines can be handled at the expense of assignment reliability. In many cases movements are registered in on-board information systems, but then these records are not compiled into a single dataset for the entire fleet. Moreover, research attempts may be hindered if the AFC and AVL clocks are not synchronised precisely. In the relatively recently built and highly standardised urban metro network of Hong Kong MTR all these threats are mitigated.

3.2 Passenger-to-train assignment

Our assignment methods are different for various trip types, depending on the number of feasible train itineraries, transfers and route options. Section 3.2.1 explains the criteria we used to classify trip types, while Section 3.2.2 presents the core of the assignment methodology.

3.2.1 Trip typology

In the rest of this paper we refer to trip types as Figure 1 depicts. The majority of trips belongs to types A to F, where the only difference is the number of transfers (no transfer for A and B, one transfer for C and D, and two or more transfers for E and F), and the number of feasible itineraries extracted from the AVL data. An itinerary for a transfer trip is feasible if the first train’s departure time is higher than the check-in time, the last train arrives earlier than the passenger taps out, and all transfer times are greater than zero. For all these trip types route choice is unambiguous, i.e. the scheduled travel time on the second shortest path is more than 1.5 times longer than on the first shortest path. Sun et al. (2015) called this value as the path superiority coefficient. Type H trips do not satisfy this condition, so potential itineraries on multiple alternative routes have to be considered in their case. We discuss the role and necessity of this intuitively selected threshold after the assignment methodology in Section 3.2.2. Shortest routes and scheduled travel times were evaluated with the igraph package of R (Csárdi and Nepusz, 2006, Kolaczyk and Csárdi, 2014).

At the initial phase of this project three suburban rail lines were not included in the AVL dataset, and therefore we have to treat suburban trips separately. Passengers taking suburban trains only are completely removed from the analysis. Transfer trips using urban as well as suburban lines are classified as type G.

We treat type G trips in the following way: we calculate the shortest path between its origin and destination and identify the transfer station where they entered or left the urban metro network, for which we do have AVL data. We neglect the suburban part and replace the suburban origin or destination with the transfer station. Accordingly, we deduct the time the passenger supposedly spent on the suburban part based on the published travel time, and
Figure 1: Trip typology based on lines, route options, transfers and the availability of train itineraries
replace the check-in or check-out time with the time when the passenger may have arrived to the transfer station. Then we reassign the trip to types A to H, depending on the remaining number of transfers and feasible itineraries.

3.2.2 Assignment methodology

The assignment strategy is summarised in Figure 2. For type A, C and E trips the assignment is straightforward: there is only one feasible itinerary. These trips provide a distribution of egress times for each station. It is likely that most passengers with multiple feasible itineraries are peak travellers, while those who used the only feasible itinerary were off-peak passengers. The main source of delay in peak periods is that passengers may fail to board the first or even the second arriving train at the origin station due to overcrowding. We assume that delays at the destination due to station congestion are much smaller in magnitude compared to the time lost after failed boarding. Therefore it is safe to assume that type B passengers, for example, have the same egress time distribution as types A, C and E, but their access time distribution is different, with larger mean and spread. This assumption is reinforced by the fact that correlation between access and egress times is very low, 0.097 for the entire population of our assignment dataset.

Based on the assumption of identical egress time distributions we can link probabilities to candidate itineraries based on the likelihood of the implied egress times at the destination. Using these probabilities we assign type B trips, which provides a delayed access time distribution for each origin station, including the effect of congestion on access times.

For type D trips the egress time distribution is not sufficient to infer the likelihood of candidate transfer itineraries, because the train taken on the first leg of the journey is independent of the egress time. Therefore in this case we evaluate potential itineraries based on the likelihood of the implied access and egress times. Thus, after this step we gain information about the transfer time distribution at line intersections, including the effect of failed boardings when transferring. Finally, we assign type F trips using both the access, transfer and egress time distributions at the relevant stations. In the next paragraphs we derive the assignment probabilities.

Type B: single trips, more feasible itineraries

To derive passenger-to-train assignment probabilities we rely on the following definitions:

- Egress time: the time spent between the moment when the train stopped at the platform and the passenger checked out at the fare gentries. For each destination station we define vector $\mathbf{E} = (E_1, ..., E_M)$ as the possible discrete values that egress time $E$ can

---

5 For the sake of simplicity we do not differentiate stations in the notation of this section, i.e. the derivations are applicable for a single OD pair.
Figure 2: Schematic overview of trip assignment based on access, transfer and egress time distributions
take and $e = (e_1, ..., e_M)$ as the associated probabilities, such that $P(E = E_m|e) = e_m$, thus effectively treating the egress time observations of type A and C passengers as a sample from a multinomial distribution of egress times.

- Event $\mathcal{I}$ represents that we observe set $S$ of candidate train itineraries for a particular type B trip. Candidate train $i \in S$ is associated with egress time $E_i$, which is the difference between the time when the passenger under investigation checked out and train $i$ arrived to the platform. This information, extracted from vehicle location data, narrows down the set of possible egress times from the entire range of $E$ to a limited number of candidate values.

- Event $C_i$: egress time $E_i$ is the true one, so the passenger traveled with the train associated with egress time $E_i$. In the method derived below $P(C_i)$ will be the choice probability that we assign to candidate train itinerary $i$.

For type B passengers the assignment is based on egress times only. In other words, $P(C_i^B) = P(E = E_i|e) = e_i$. Furthermore, we assume that headways (i.e. the time between two consecutive train arrivals) are random variables and the egress times included in $S$ are independent from each other. That is, having information about one of the candidate trains provides the same knowledge about the rest of the itinerary set as another potential train:

$$P(\mathcal{I}|C_i) = P(\mathcal{I}|C_j) \quad \forall i \text{ and } j \in S. \quad (1)$$

We build this assumption on the fact that train arrival times are determined by the scheduled service headways and some random noise in actual train movements. Therefore from one train’s arrival time one may infer that other trains may have arrived one headway earlier and later, but their precise arrival time is uncertain. We assume here that the uncertainty is identical no matter which train we have information about.

It is certain by definition that the true egress time has to be among the set of candidate egress times. Thus, using Bayes’ Theorem,

$$\sum_{j \in S} P(C_j|\mathcal{I}) = \sum_{j \in S} \frac{P(\mathcal{I}|C_j)P(C_j)}{P(\mathcal{I})} = 1. \quad (2)$$

From this equation we can express $P(\mathcal{I})$ and plug into the conditional probability of train $i$ of the itinerary set being the true one:

$$P(C_i|\mathcal{I}) = \frac{P(\mathcal{I}|C_i)P(C_i)}{P(\mathcal{I})} = \frac{P(\mathcal{I}|C_i)P(C_i)}{\sum_{j \in S} P(\mathcal{I}|C_j)P(C_j)} \quad \forall i \in S. \quad (3)$$

Given the assumption in equation (1), the conditional probability for type B trips simplifies to

$$P(C_i^B|\mathcal{I}) = \frac{P(C_i^B)}{\sum_{j \in S} P(C_j^B)} = \frac{e_i}{\sum_{j \in S} e_j} \quad \forall i \in S. \quad (4)$$

13
This implies that the probabilities that we have to assign to each train in the set of candidate itineraries equal to the relative magnitude of probabilities in the discretised distribution of egress times.

In practice, we fit a non-parametric kernel probability density function from type A and C egress time data for each station, and discretise it in seconds intervals to derive $e$ values. What we actually get is a probability mass function of egress times in integer seconds. Note that both the smart card and train location data are recorded with seconds precision, so our original egress time observations are also discretised.

Types D and F: Transfer trips, one feasible route

Intuition suggests that type A and B passengers may have different access time distributions, simply because the former group definitely did not fail to board the first arriving train, which cannot be excluded for the other. After the assignment of type B passengers we can derive a delayed access time distribution, that does take into account the delay effect of congestion. We use the delayed access time distribution in the assignment of type D and F trips.

As in case of egress times, we define for each origin station vector $\mathbf{A} = (A_1, ..., A_L)$ as the vector of possible discrete values that access time $A$ can take, and $\mathbf{a} = (a_1, ..., a_L)$ as the associated probabilities, such that $P(A = A_i | \mathbf{a}) = a_i$. In other words, we treat the access time observations of type B passengers as a sample from a discretised multinomial distribution of delayed access times.

Without any information on a subset of feasible itineraries, the probability of the occurrence of any type D itinerary becomes $P(\mathcal{C}_D^i) = P(A = A_i \land E = E_i | \mathbf{a}, \mathbf{e}) = a_i \cdot e_i$, where symbol "\land" is the conjunction operator. For type D passengers the feasible itineraries in set $S$ consist of a train on each leg of the journey with positive access, egress and transfer times. Following the same logic that led to equations (3) and (4), the probability that itinerary $i \in S$ has been the one taken by the passenger is

$$P(\mathcal{C}_D^i | \mathcal{I}) = \frac{P(A = A_i | \mathbf{a})P(E = E_i | \mathbf{e})}{\sum_{j \in S} P(A = A_j | \mathbf{a})P(E = E_j | \mathbf{e})} = \frac{a_i \cdot e_i}{\sum_{j \in S} a_j \cdot e_j}, \forall i \in S. \quad (5)$$

The assignment of type D trips delivers a distribution of delayed transfer times at line intersections. Again, these transfer time variables may differ from the transfer times of type C trips, because passengers may had failed to board the first train at subsequent legs of their journey. We treat transfer time variables at each station the same way as access and egress times: $\mathbf{T}^r = (T^r_1, ..., T^r_N)$ denotes the vector of discrete values that transfer time $T^r$ may take at station $r$, while $\mathbf{t}^r = (t^r_1, ..., t^r_N)$ is the vector of associated probabilities.

Type F trips were performed with two or more transfers; $\tau$ denotes the set of transfer

\[\text{Note that the expected access time is not necessarily longer in the peak, because the congestion effect may be compensated by shorter headways.}\]
stations for a particular journey. Thus, without any limitation on the number of transfers, the probability of an arbitrary feasible access, egress and transfer time combination becomes

\[ P(C^F_i) = P(A = A_i | a) P(E = E_i | e) \prod_{r \in \tau} P(T^r = T^r_i | t^r) = a_i \cdot e_i \cdot \prod_{r \in \tau} t^r_i, \quad (6) \]

and in case of \( I \), so that we know a set \( S \) of feasible itineraries among which the true itinerary is,

\[ P(C^F_i | I) = a_i \cdot e_i \cdot \prod_{r \in \tau} t^r_i \sum_{j \in S} (a_j \cdot e_j \cdot \prod_{r \in \tau} t^r_j) \quad \forall i \in S. \quad (7) \]

**Type H: Transfer trips, more feasible routes**

Now we turn to the case when, given a particular origin and destination station, route choice is ambiguous. Let us define \( K \) as the set of feasible routes for a particular type H trip. Event \( R_k \) denotes that route \( k \in K \) has been chosen by the passenger. Route choice and itinerary choice probabilities on each route have to satisfy

\[ \sum_{k \in K} P(R_k) = 1 \quad \text{and} \quad \sum_{i \in S_k} P(C_i | R_k, I) = 1, \quad (8) \]

where \( S_k \subset S \) is the set of feasible itineraries on route \( k \). If competing routes feature different number of transfers, so that equation (7) cannot be applied directly, we have to split the assignment problem into a route choice level and an itinerary choice level:

\[ P(C^H_{i \in S_k} | I) = P(R_k) P(C_{i \in S_k} | R_k, I). \quad (9) \]

The itinerary choice level, i.e. \( P(C_{i \in S_k} | R_k, I) \), can be replaced by either (4), (5) or (7), depending on the number of transfers on route \( k \). For the route choice problem, i.e. \( P(R_k) \), Paul (2010) and Zhu (2014) suggested that a traditional random utility discrete choice demand model should be applied. We propose an alternative approach based on the available access and egress times of itineraries on potential routes. This approach may be particularly useful when the researcher, as in our case, does not have information about route choice preferences (e.g. the value of time) in the experimental area. Even though for type H passengers multiple feasible routes are available for a specific trip, all feasible itineraries have the same origin and destination stations. We use that the potential access and egress times on competing routes may tell a lot about the likelihood of choosing one or another route. For example, if there was no train travelling on one of the competing routes, we can certainly exclude that route without considering user preferences in route choice.

Let us define \( \sigma \) as the set of feasible departure and arrival itineraries represented by the feasible combinations of the implied access and egress times. For itineraries in \( \sigma \) only the first and last legs of the journey matter, and itineraries in the original \( S \) that only differ in
the train(s) used for the middle leg(s) are not differentiated. Furthermore, \( \sigma_k \subseteq \sigma \) is the set of departure and arrival itineraries on route \( k \). We define the probability that route \( k \) has been chosen as

\[
P(R_k|I) = \frac{\sum_{i \in \sigma_k} P(A = A_i|a) P(E = E_i|e)}{\sum_{j \in \sigma} P(A = A_j|a) P(E = E_j|e)} \quad \forall k \in K.
\]  

Using these route choice probabilities we can derive equation (9) for each feasible itinerary of a trip and the passenger-to-train assignment is complete.

The critical reader may ask why we set the threshold level of the path superiority coefficient to 1.5 for type H trips and why do not we treat all origin-destination pairs with multiple route options in group H, no matter how long the second shortest path is. The answer is simply that we need a reasonable number of type A, B, C and D trips to extract reliable egress, access and transfer time distributions for each station. By setting the threshold level to 1.5 we implicitly assume that crowding and other user costs can never compensate for a fifty percent difference in travel times between the two most attractive route alternatives. In other words, the difference in crowding multipliers can never be greater than 0.5. Figure 3 plots the cumulative distribution of the superiority coefficient, from which we learn that for the vast majority of trips route choice is not ambiguous in this relatively simple network.

With the current critical superiority coefficient 22.4% of the trips have been assigned to type H.

The reader may also ask why do not we consider the third and fourth shortest paths for certain origin-destination pairs. There is no methodological burden to apply equation (10)
for more than two alternative routes. However, in the relatively simple metro network where we applied the method there is no reason to do so, i.e. none of the OD pairs have more than two reasonable alternative routes. □

We realised the assignment algorithm in the R programming environment. As feasible itineraries have to be recovered and evaluated for each trip, and our Hong Kong datasets contain around 5-7 million trips per day, computation time became a relevant issue, at least on ordinary PCs. Our experience shows that computation times can reach two days on a PC featuring 3.40 GHz CPU and 16 GB RAM. The authors agree with one of the referees who suggested that future research efforts should focus on improved algorithms that may reduce computation time.

From the assignment results we derived the following summary statistics for the inner-city ring depicted in Figure 4. Considering each interstation section served by a train as a separate observation, the mean crowding density is 0.732 standing passenger per square metre. This drops to a weighted mean of 0.704 when we use running time on each section as weighting factor. The distribution of crowding density is heavily skewed to the right: the quartiles are 0, 0.428 and 1.214 passengers per square metre\(^7\). Maximum density is slightly above 5.1 traveller on a square metre. Note, however, that these are average figures for relatively long trains, i.e. the actual density may go way above what we measured at certain parts of the vehicle, especially close to doors.

3.3 Modelling route choice preferences

This section presents the revealed preference route choice experiment based on the train load pattern derived in Section 3.2. As fares in Hong Kong only depend on the origin and destination of a trip, and therefore passengers do not make trade-offs between crowding and the fare paid in a route choice situation, our objective will be to capture the trade-off between travel time and crowding. In other words, we apply the travel time multiplier approach reviewed in Section 2.2.

3.3.1 Experimental design

The MTR metro network features a downtown loop where some cross-harbour passengers have to choose between an Eastern and a Western route. There are no other reasonably

\(^7\)Zero standing density implies that all passengers can be seated
competitive paths in the network\(^8\), so the choice set is fully defined. Figure 4 depicts the network layout. We included four plus four stations in the middle sections of the Kwun Tong Line (North) and the Island Line (South); between certain OD pairs involved scheduled travel times on the two routes are exactly the same, while others have slightly different competing travel times to ensure sufficient variance in route attributes. Luckily, all transfer stations allow on-platform interchange in both directions. The interior design of trains is highly standardised in the network, so that passengers have almost identical travel experience using any rolling stock types of the fleet. We can safely assume that transfer station design and train types do not have significant dynamic effect on route choice preferences, while static effects can be controlled for in alternative specific constants.

\[\text{Figure 4: Schematic network layout of the downtown metro loop in Hong Kong}\]

The Western route via the Tsuen Wan Line is usually more crowded than the Eastern, the difference in crowding density may go above 1.5 passenger per square metre. As we include 32 origin-destination pairs in the analysis, there is variability in the difference in travel times as well. This ensures that for many passengers there is no dominant alternative in the route choice situation.

\(^8\)As a matter of fact, a combination of the heavy rail lines via Hung Hom may be a substitute for the Western route in our model. Please note that passengers transferring at Tsim Sha Tsui and East Tsim Sha Tsui have to tap out and tap in again (the underpass between the two stations is not part of the tariff zone), so in the smart card dataset they appear as two separate trips. In other words, we surely do not observe any users in the experiment taking the heavy rail lines via Hung Hom. Also note that we only included stations to the East from Kowloon Tong station. It takes 19 minutes to travel from Lok Fu to Admiralty, with one interchange between the Kwun Tong Line and the Tsuen Wan Line, most probably at Mong Kok. The heavy rail option needs two more transfers (three in total at Kowloon Tong, Hung Hom and Tsim Sha Tsui), and the scheduled travel time is 30 minutes. This is why we do not consider the heavy rail option as a reasonably competitive alternative in the choice model.
We intend to explain the observed route choices with the following attributes that may vary on each link (interstation section) of a route\textsuperscript{9}:

- Travel time on the link, including half of the dwell time at the two stations that the link connects\textsuperscript{10};
- Density of standing passengers;
- Probability of standing.

Furthermore, some additional choice attributes are measured for the entire route:

- Time spent at transfer stations;
- Reliability (variability) of transfer times;
- Which train arrives first at the origin station? – A binary variable only applicable for origin stations with a central platform that allows passengers to take the train that arrives first\textsuperscript{11}.

To account for all link- and route-specific attributes we developed a random utility discrete choice model. First, let us assume that utility on link \( l \) is determined by

\[
v_l = v_l(t_l, c_l, p_l) = \alpha t_l [1 + m(c_l, p_l)],
\]

where \( t_l, c_l \) and \( p_l \) are the expected travel time, crowding density and standing probability, respectively. The disutility of a unit of uncrowded in-vehicle travel time is captured by \( \alpha \), which is further increased by the crowding multiplier, \( m(c_l, p_l) \). We define the multiplier as a linear function of the crowding variables:

\[
m(c_l, p_l) = \beta_c c_l + \beta_p p_l.
\]

Our goal is to estimate the value of \( \beta_c \) and \( \beta_p \). In this specification the link-level utility function becomes

\[
v_l = \alpha t_l + \gamma_c t_l c_l + \gamma_p t_l p_l,
\]

where the coefficients of the interaction terms are \( \gamma_c = \alpha \beta_c \) and \( \gamma_p = \alpha \beta_p \). Crowding density and standing probability are not link-additive attributes. Therefore we add them to the utility function in interaction with the link-level travel time, which allows to aggregate them for the entire route. The utility function for route \( r \) thus becomes

\[
U_r = ASC_r + \sum_{l=1}^{n_l} v_l + \gamma_w t_r^{bw} + \gamma_{wvar} t_r^{wvar} + \gamma_d \delta_{plat} \delta_{carr} + \varepsilon_r,
\]

\textsuperscript{9}Clarification of terminology: in our experiment passengers have to choose between two routes. Each route consists of three legs with two transfer stations between the legs. Finally, each leg includes multiple links or interstation sections. Some travel attributes may differ on each link, some others can only be defined for the entire route. We do not distinguish leg-specific attributes. In this paper transfers are referred to as all activities between alighting the train and boarding the next one at transfer stations. Itineraries are the set of train movements performed on journey legs that would have allowed the passenger to reach her destination station between the tap in and tap out times recorded in the smart card dataset.

\textsuperscript{10}Except transfer stations where the train’s dwell time is included in the transfer time

\textsuperscript{11}Real-time information on train arrivals is not provided for passengers in the stations concerned.
where

\[ \sum_{l=1}^{n_l} v_l = \alpha \sum_{l=1}^{n_l} t_l + \gamma_c \sum_{l=1}^{n_l} t_l c_l + \gamma_p \sum_{l=1}^{n_l} t_l p_l, \]

and \( n_l \) is the number of links on route \( r \), \( ASC_r \) is an alternative specific constant, and \( t^w_r \) is the expected total waiting time at transfer stations. Finally, \( \delta_r^{\text{plat}} \) and \( \delta_r^{\text{arr}} \) respectively are dummies set to one when at the origin station of route \( r \) trains in both directions can be accessed from the same central platform, and the first train arriving to the station serves route \( r \). In the unusual but not negligible case when both trains were already at the platform when the passenger was expected to arrive there, we set \( \delta_r^{\text{arr}} \) equal to zero for both directions.

### 3.3.2 Identifying route choices

What we can learn from smart card records is only the entry and exit stations of a trip. There is no explicit information, however, about the actual route chosen when multiple feasible routes exist between the origin and destination. Note that the passenger-to-train assignment for type H trips, introduced in Section 3.2.2, provides an estimate of route choice implicitly. The reliability of these estimates can be further improved with two alternative methods. In this section we summarise all three route choice identification methods we applied.

**Probabilistic assignment based on access, egress and transfer times**

In the base case we rely on the route choice estimated in the assignment process. That is, the assignment is based on the access time distribution at the origin station, the egress time distribution at the destination, and transfer time distributions at intersections. For the OD pairs under investigation this means that candidate trains have been collected from three different lines, which often makes the number of potential itineraries very high. As a consequence, especially in peak periods when services follow each other at high frequency on both routes, this method may not give a completely reliable estimate of route choice, because the second and third most likely trains may also receive considerable probability in the assignment.

**Assignment based on egress times only**

In this approach we focus on egress times at destination stations only. We improve the assignment method by deriving directionally differentiated egress time density functions from single trips with only one feasible train. Directional differentiation can be important at stations where platforms for the two directions are located beneath each other, so that the expected walking time to fare gentries may be different. Station design suggests that this may be an issue at 3 out of the 8 stations in our experiment. At the remaining five stations the two directions share the same platform.
Based on the egress time distributions we derive for each trip probability densities for all trains arriving no earlier than 10 minutes before the check-out occurred. Then we assign probability weights to candidate trains based on the relative magnitude of density values, as in equation (4). Finally, we assign the trip to the most likely arriving train if that train has more than 75% probability. As a consequence, many trips are left without assigned trains, but those that are assigned can be considered even more reliable.

Assignment based on arrival waves at fare gentries

The last method is based on smart card data solely. We neglect the actual train arrival and departure times. As Hong et al. (2015) observed and used extensively in their assignment method, passengers alighting from the same train normally arrive to the fare gentries in bunches. Figure 5 provides an illustration of such arrival waves at a busy station on the Island line. About many passengers in an arrival wave we can tell which direction they arrived from. For example, the direction of arrival is trivial for passengers who checked in on the same line and travelled without transfers. Hong et al. (2015) referred to these users as reference passengers.

In Method 3 we count the number of such passengers in ten-second intervals at each station in the experiment and identify those time periods when more than 80% arrived from...
the same direction. Then we assign this direction to trips in the RP experiment checking out in the same time period. The 80% threshold is only applied if more than six no-transfer travellers checked out in a particular interval to avoid errors caused by a couple of randomly arriving “leftover” passengers. Thus, Method 3 is not able to assign route choices to all trips either, but for a subset of them the identification is even more convincing.

We decided to create a subset of observations for which the three route identification methods vote for the same route, as the available sample size was really not a limiting factor. Ridership on the 32 experimental OD pairs is 18,401 trips per day in total – after subsetting 3,529 remained available, which appeared to be a reasonable sample size for a route choice model. Note that for the majority of the scrapped trips the three methods did not provide conflicting results, but either Method 2 or 3 gave no result at all, due to the relatively high threshold percentages we prescribed for the assignment in both cases.

As the validity of our experiment strongly depends on the quality of the dependent variable, it is important to quantify the precision of the methods presented above. For certain passengers in the dataset there is no feasible train itinerary on one of the alternative routes, and therefore route choice is absolutely unambiguous; we can use them as a control group to validate Methods 2 and 3. In the original dataset 6,832 trips (around 37%) were identified in the control group. We found that Method 2 produced correct route choice observations for 92.55% of the experimental trips, while Method 3 performed 86.00%. When both two methods are taken into account, 94.79% of the observations are correct. When all three methods are considered, the precision of route choice inference is expected to be even higher.

Why did not we use simply the control group for the RP experiment as well? The reason is that in the control group there is no sufficient variance in travel times because the route chosen is generally shorter than the one with no feasible itinerary. As a consequence, we could not measure the trade-off between time and crowding discomfort. However, 48.43% of the final experimental subsample of 3,529 trips does belong to the control group.

3.4 Link and route attributes

This section details how in-vehicle travel time, crowding density, standing probability, waiting time, waiting time reliability and train arrival priority variables were generated.

3.4.1 In-vehicle travel time

The expected travel time on links, i.e. $t_l$ in equation (12), can be easily retrieved from the AVL data. In practice, for a specific link, we collect all train travel times throughout the day. Dwell times at intermediate non-transfer stations are split into half and added to movement.

\[12\text{The same does not apply for Method 1, because in that case the estimated route choice can never be wrong if there is no feasible itinerary on any of the alternative routes.}\]
times on neighbouring links. Then we fit a locally weighted scatterplot smoothing curve on the data. Thus, the expected travel time can be estimated based on the time of day. The same process had to be repeated for all links in the experimental area.

Figure 6 illustrates the method for a relatively long interstation section that can also be considered as a major bottleneck in the network. Note that due to the automated operation in the metro system expected travel times do not vary significantly throughout the day. Waiting times at transfer stations are much more important factors in journey time reliability.

![Figure 6: Travel times of trains in function of time of day at a cross-harbour section](image)

### 3.4.2 Crowding density

By merging smart card data with vehicle location data, the number of passengers on board each train becomes measurable for each link it passes through. After subtracting the number of seats and dividing by the available floor area for standees, we get the density of crowding in terms of standing passengers per square metre. We calculate this value for all trains on each link and fit again a locally weighted smoothing curve on the observations, as Figure 7 illustrates.
Figure 7: Crowding density in terms of standing passengers per square metre, in function of time of day at a cross-harbour section. The same smoothing curves were fitted for all links in the experiment.

We found that even in rush hours the density of crowding may significantly fluctuate within a short period of time. Note that what we calculate is an average density, while the distribution of crowding may vary along the train, so at certain parts of the train crowding may exceed the highest values in our graphs. Also, minor delays and deviations in headways may cause crowding in low-demand periods as well. Figure 7 shows that crowding significantly drops in the peak shoulders, which can be explained by the fact that the operator increases service frequency well before demand reaches its peak. By applying locally weighted averages we assume that passengers are experienced and aware of temporal fluctuation patterns in crowding conditions.

Ultimately, the value of $c_l$ in equation (12) is calculated based on the time when the passenger was expected to arrive to link $l$, and the estimated crowding density in that period.

### 3.4.3 Standing probability

The goal in this section is to derive the probability that the average passenger on a particular OD pair travels seated on a particular line section and include this link-level travel attribute in the utility function as variable $p_l$ in equation (12).

It is obvious that the earlier someone boards the train, the higher the probability of finding a seat until all seats are occupied. After that, the number of alighting passengers and the share of those among them who had a seat become important. The likelihood of being seated increases with journey length assuming that being seated is always preferred over standing, because alighting passengers increase the chance of finding an empty seat. The importance of the chance of finding a seat from a passenger point of view had already been discovered and considered in dynamic transit assignment by Leurent (2006), Schmoecker (2006), Sumalee et al. (2009) and Hamdouch et al. (2011). We use their assumptions with
minor modifications: passengers prefer to be seated over standing; boarding travellers have
equal chance to find a seat (there is no FCFS mechanism or queuing), but after alighting at
a station standing on-board passengers have a priority in occupying empty seats over those
who just board the train. We neglect all kinds of strategic behaviour in seeking for seats as
well as consumer heterogeneity apart from trip origin and destination.

Notation for the standing probability derivation is summarised in Table 1. Let us focus on
the first station of a line where there are no alighting or dwelling passengers,
\( A_1 = W_1 = 0 \), and thus \( N_1 = B_1 \). The number of boarders is the sum of OD volumes departing at the
first station: \( B_1 = \sum_{j=2}^{n} D_{1j} \). The likelihood of finding a seat among them depends on their
number relative to the seat capacity of the train. Therefore, assuming that all passengers
have equal chance to find a seat no matter the OD pair they belong to, we can write that
\[
P_{1j,1} = \min(1, s/B_1), \quad \text{and} \quad S_{1j,1} = P_{1j,1}D_{1j}. \tag{15}
\]
The number of seats being occupied after leaving the first station directly comes as
\( \xi_1 = \sum_{j=2}^{n} S_{1j,1} = \min(s, B_1) \).

At intermediate stations we have to account for alighting and dwelling passengers as well.
These volumes can be calculated by summing up respective origin-destination demand values:
\[
B_k = \sum_{j=k+1}^{n} D_{kj}, \quad A_k = \sum_{i=1}^{k-1} D_{ik}, \quad W_k = \sum_{i=1}^{k-1} \sum_{j=k+1}^{n} D_{ij}. \tag{16}
\]

Boarding and alighting values have to sum up to the difference between consecutive link

---

**Table 1: Notation in Section 3.4.3**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_{ij} )</td>
<td>Inelastic demand between ( i ) and ( j ), where ( i &lt; j )</td>
</tr>
<tr>
<td>( P_{ij,k} )</td>
<td>Probability that ( ij ) passengers have a seat after leaving station ( k ), ( i \leq k &lt; j )</td>
</tr>
<tr>
<td>( S_{ij,k} )</td>
<td>Number of ( ij ) passengers having a seat after leaving station ( k ), ( i \leq k &lt; j )</td>
</tr>
<tr>
<td>( T_{ij,k} )</td>
<td>Number of ( ij ) passengers standing after leaving station ( k ), ( T_{ij,k} = D_{ij} - S_{ij,k} )</td>
</tr>
<tr>
<td>( N_k )</td>
<td>Number of passengers on board after leaving station ( k )</td>
</tr>
<tr>
<td>( B_k )</td>
<td>Number of passengers boarding at station ( k )</td>
</tr>
<tr>
<td>( A_k )</td>
<td>Number of passengers alighting at station ( k )</td>
</tr>
<tr>
<td>( W_k )</td>
<td>Number of passengers dwelling at station ( k )</td>
</tr>
<tr>
<td>( s )</td>
<td>Number of seats in train</td>
</tr>
<tr>
<td>( \phi_k )</td>
<td>Number of seats freed up by passengers alighting at station ( k )</td>
</tr>
<tr>
<td>( \omega_k )</td>
<td>Number of seats remained occupied after alighting at ( k ), ( \omega_k = s - \phi_k )</td>
</tr>
<tr>
<td>( \xi_k )</td>
<td>Number of seats being occupied after leaving station ( k )</td>
</tr>
<tr>
<td>( n )</td>
<td>Number of stations, ( \max(k) = n - 1 )</td>
</tr>
</tbody>
</table>
loads, so that \( N_k = N_{k-1} - A_k + B_k \). Obviously, the number of occupied seats remains \( \xi_k = \min(s, N_k) \). A crucial element of the calculation is to express how many seats are freed up by alighting travellers. This equals to the sum of seats occupied by alighting passengers after leaving the previous station:

\[
\phi_k = \sum_{i=1}^{k-1} S_{ik,k-1}.
\] (17)

Therefore the number of seats still used after alighting passengers leave their seats, accounting for the fact that not obviously all seats have been occupied when the train arrived to \( k \), equals to

\[
\omega_k = \xi_{k-1} - \phi_k.
\] (18)

Consequently, \( s - \omega_k \) seats are to be redistributed among standing and boarding passengers.

At this point we employ the assumption also used by Leurent (2006) and Sumalee et al. (2009): standing passengers already on-board have a priority over boarding peers, so that those who just board the train can find a seat only if there are less standing passengers and empty seats. On the other hand, standees already on-board have equal chance to find a seat, so we neglect any differences in strategic behaviour between OD groups. Thus, after leaving station \( k \), the number of seated users on a particular OD pair, and the corresponding probability of being seated are

\[
S_{ij,k} = S_{ij,k-1} + T_{ij,k-1} \cdot \min\left(1, \frac{s - \omega_k}{W_k - \omega_k}\right); \quad P_{ij,k} = \frac{S_{ij,k}}{D_{ij}}, \quad \forall i < k < j
\] (19)

Note, that \( W_k - \omega_k \) is the sum of all standing passengers staying on board after the train stops. Equation (19) implies that the probability of finding a seat monotonically increases during the trip for all origin-destination markets. The actual value of \( S_{ij,k} \) can be calculated by plugging equations (15) to (18) into (19) and building up the probabilities from the very first station to \( k \).

Now we turn to those just boarding the train at an intermediate station, so that \( i = k \). We assume that they compete for the remaining empty seats (quantified by \( s - W_k \)) with equal opportunity. Considering that the share of \( kj \) passengers among all boarders at station \( k \) equals to \( D_{kj}/B_k \), we find that \( S_{kj,k} = \frac{D_{kj}}{B_k} \cdot \max(0, s - W_k) \), and finally \( P_{ij,k} = \frac{S_{ij,k}}{D_{ij}} \).

Figure 8 illustrates how the above calculation works in practice. Let us consider a simple OD pair with a distance of only four stations on the same line. We calculated the probability of standing for all trains connecting this OD pair throughout the whole day, for all four inter-station line sections. Then we fitted a locally weighted smoothing curve on the observations to derive the expected travel conditions. It is clearly visible in figure 8 that the probability of standing drops for subsequent links, as some alighting passengers free up occupied seats. For
Figure 8: Standing probability for passengers travelling through four selected consecutive links of a metro line. The grey vertical line belongs to a sample passenger who checked in at the first station at 8h00.
example, a passenger checking in at 8h00 (see the grey vertical line in the plots) had basically no chance to sit on the first link, but on the fourth link she had 50% probability already to travel seated. (Obviously, she arrived to subsequent links with certain delay relative to the check-in time.)

3.4.4 Transfer times

As a result of the assignment process we gain information on how much time passengers spent at transfer stations. These data can be used to estimate what waiting time passengers expect on alternative routes. As all the experimental passengers travelled through two transfer stations (i.e. Mong Kok and Admiralty in the West and Yau Tong and North Point in the East), we considered the joint distribution of transfer times based on all trips that we assigned to routes using the same two transfer stations. In this case our observations are individual passengers, of which tens of thousands may have been recorded at transfer stations daily. Fitting locally weighted non-parametric curves is challenging with this sample size. Therefore we fitted a simple cubic spline to derive the expected transfer time in function of departure time at the origin station, as Figure 9 depicts for the two directions of the same pair of transfer stations.

![Figure 9: Observed and expected waiting times for a large sample of passengers at one of the transfer stations in the experiment](image)

Note that transfer times in off-peak are clustered around the multiples of around 4-5 minutes. A potential explanation is that these clusters correspond to feasible itineraries: the first bunch always reached the first arriving train, in the second bunch passengers missed one train, etc. As frequencies are constant between the two peaks, transfer times are very similar within the clusters. Also note that we defined transfer time as the time spent between the moment when the arriving train stopped and the departing train left the platform. Therefore, passengers using the same train had exactly the same transfer time in the dataset. As a result,
the expected waiting time (the fitted curve) is just slightly above the first cluster of waiting
times.

During the morning and afternoon peaks, when headways are just around 1-2 minutes on
both connecting lines, these clusters disappear. However, the expected waiting time is not
significantly shorter than out of the peak, due to queuing when boarding the trains. Even
though the probability that someone cannot board the first train due to crowding is higher
in the peak, headways are also shorter, and the interplay between the two factors keeps the
expected waiting time within a relatively narrow range. Figure 9 shows that passengers have
to expect more waiting in the morning peak towards the city centre, although frequencies are
equally high in both directions. This is not the case in the peak shoulders, when frequencies
are still very high, but crowding eases.

Figure 9 suggests that the spread of potential transfer times also varies along the day. This
may have an important effect on the reliability of travel times, and eventually on passengers’
travel behaviour as well. We include this attribute in the route choice model as equation
(13) shows. It is not straightforward, however, which measure of spread is the most suitable
to reflect user preferences. We found that our parameter estimates are strongly affected by
the choice of transfer time variability metrics. We tested eight metrics reviewed by Taylor
(2013):

• standard deviation (Bates et al., 2001),
• coefficient of variation (Herman and Lam, 1974),
• variance (Engelson and Fosgerau, 2011),
• and 90th percentile of transfer time (Lam and Small, 2001),
• the difference between the 80th and 50th percentiles (Van Loon et al., 2011),
• the buffer time and the Buffer Time Index13 with mean and median transfer time
  (FHWA, 2010).

The best performing metrics in terms of model fit and the significance of crowding pa-
rameters were the standard deviation, the variance, the 90th percentile of transfer time and
the buffer time. The estimated $\beta_c$ coefficients are in the range of 0.11 and 0.14, while $\beta_p$
varies between 0.227 and 0.284. In the rest of this paper we use the standard deviation, as
it is the most widely applied travel time variability metric in the literature, and it delivered
median crowding multiplier values among competing estimates. A detailed discussion of the
role of travel time reliability is out of the scope of this paper, but the authors are ready to
provide further details upon request.

13The buffer time is defined as the difference between the 95th percentile and the mean (or median), while
the Buffer Time Index is the buffer time divided by the mean transfer time.
4 Estimation Results

In this experiment it is assumed that passengers are fully informed about the expected link and route attribute levels at least in the period when they travel. They have to make a route choice decision at the origin station, so we cannot use directly the information about the actual travel conditions they experienced after the decision. Based on these assumptions we generated a dataset through the following steps:

1. List for each OD the links of the network that are used on the two alternative routes.
2. Based on the check-in time and the expected running times on links, estimate at what time was the passenger going to arrive to subsequent links.
3. Extract link-level crowding attributes from the non-parametric regressions and calculate their interactions with in-vehicle travel times. Sum up the interaction terms according to equation (14).
4. Add route-level attributes to the dataset, including the transfer time and its standard deviation as well as train arrival priorities.
5. Repeat the same process for all trips in the RP experiment.

We estimated the parameters in a logit model, assuming that the error term, \( \varepsilon_r \) in equation (13), is a random variable with extreme value type I (Gumbel) distribution. We used the mlogit package Croissant et al. (2012) developed for R. Estimation results are summarised in Table 2.

The negative signs of the estimated coefficients show that in-vehicle travel time, waiting time as well as waiting time variability cause disutility for passengers. The fact that one of the trains arrive earlier than the next in the other direction does provide an incentive for decision makers, which clearly indicates an opportunistic behaviour in route choice. All these route attributes are statistically significant and their inclusion improves model fit as well as the log-likelihood and likelihood ratio test \( \chi^2 \) values.

The estimated \( \gamma_p \) and \( \gamma_c \) coefficients in Model 4 are significant at the 95\% confidence level after including alternative specific constants for the morning and afternoon peaks. Their sign is negative, which implies given that \( \alpha \) is also negative that \( \beta_p \) and \( \beta_c \) in the crowding multiplier are greater than zero. With the estimated values of Model 4, \( \beta_p = 0.265 \) and \( \beta_c = 0.119 \), when crowding density is measured in passengers per square metre. Standard errors for the multipliers are recovered from the standard errors of \( \alpha \), \( \gamma_c \) and \( \gamma_p \) using the Delta method (Oehlert, 1992).
Table 2: Estimation results

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC\textsubscript{W}</td>
<td>0.144</td>
<td>0.772***</td>
<td>0.451**</td>
<td>1.706***</td>
</tr>
<tr>
<td></td>
<td>(0.176)</td>
<td>(0.207)</td>
<td>(0.186)</td>
<td>(0.268)</td>
</tr>
<tr>
<td><strong>ASC\textsubscript{W}</strong> in a.m. peak</td>
<td></td>
<td></td>
<td></td>
<td>−2.307***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.403)</td>
</tr>
<tr>
<td><strong>ASC\textsubscript{W}</strong> in p.m. peak</td>
<td></td>
<td></td>
<td></td>
<td>−0.811***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.152)</td>
</tr>
<tr>
<td>t</td>
<td>−0.00642***</td>
<td>−0.00545***</td>
<td>−0.00560***</td>
<td>−0.00502***</td>
</tr>
<tr>
<td>(in-veh. time)</td>
<td>(0.00023)</td>
<td>(0.00027)</td>
<td>(0.00027)</td>
<td>(0.00028)</td>
</tr>
<tr>
<td>t·c</td>
<td></td>
<td>−0.00080***</td>
<td></td>
<td>−0.00060***</td>
</tr>
<tr>
<td>(crowd density)</td>
<td></td>
<td>(0.00014)</td>
<td></td>
<td>(0.00022)</td>
</tr>
<tr>
<td>t·p</td>
<td></td>
<td></td>
<td>−0.00165***</td>
<td>−0.00133**</td>
</tr>
<tr>
<td>(standing prob.)</td>
<td></td>
<td></td>
<td>(0.00033)</td>
<td>(0.00052)</td>
</tr>
<tr>
<td>t\textsubscript{w}</td>
<td>−0.00963***</td>
<td>−0.01201***</td>
<td>−0.01116***</td>
<td>−0.00883***</td>
</tr>
<tr>
<td>(transfer time)</td>
<td>(0.00085)</td>
<td>(0.00096)</td>
<td>(0.00093)</td>
<td>(0.00113)</td>
</tr>
<tr>
<td>sd(t\textsubscript{w})</td>
<td>−0.00999***</td>
<td>−0.01624***</td>
<td>−0.01334***</td>
<td>−0.00623*</td>
</tr>
<tr>
<td>(tr. time reliability)</td>
<td>(0.00300)</td>
<td>(0.00316)</td>
<td>(0.00307)</td>
<td>(0.00367)</td>
</tr>
<tr>
<td>δ\textsubscript{plat δarr}</td>
<td></td>
<td></td>
<td></td>
<td>0.19320***</td>
</tr>
<tr>
<td>(arrival priority)</td>
<td></td>
<td></td>
<td></td>
<td>(0.07409)</td>
</tr>
</tbody>
</table>

**Multipliers**

| β\textsubscript{c}  | 0.1463*** | 0.1192*** |
|                      | (0.0303)  | (0.0457)  |
| β\textsubscript{p}  | 0.294***  | 0.2654**  |
|                      | (0.0674)  | (0.1067)  |

Observations 3,495 3,495 3,495 3,495
McFadden $\rho^2$ 0.441 0.449 0.447 0.464
Log Likelihood −1,127.72 −1,111.18 −1,115.28 −1,080.64
LR Test 1,776*** 1,809*** 1,801*** 1,870***
AIC 2263.442 2232.358 2240.555 2179.287
BIC 2288.078 2263.154 2271.351 2234.719

Note: Std. errors in parantheses
*p<0.1; **p<0.05; ***p<0.01
Figure 10: Value of time multiplier in function of the probability of standing and on-board passenger density, according to revealed route choice preferences

In Model 4 beside the above mentioned trip attributes we included three alternative specific constants (ASCs): a general ASC for the Western harbour crossing route, and two others for the morning and afternoon peaks. The role of the ASCs is to control for any fixed route characteristics beyond the attributes that we are able to observe. These fixed effects may change by time of day, for example due to unobserved variations in station conditions, and because of potential variations in the taste of the representative passenger. Travel conditions in the morning peak are certainly different from the afternoon peak, due to heavy directional demand imbalances and the intensity and spread of the peak. This gives an intuitive justification for applying distinct ASCs for rush hours in the morning and afternoon.

We chose the final specification after a series of tests with other model specifications with different sets of ASCs, including

- no constant at all,
- a single ASC for the Western route,
- an additional ASC for peak periods, with no differentiation between morning and afternoon,
- dummies for three-hour intervals,
• dummies for each hour.
• In the morning peak most metro lines are crowded in the direction of Hong Kong Island, and the opposite demand pattern appears in the afternoon. Therefore we tested a model with active ASCs from North to South in the morning peak, and from South to North in the afternoon.

The estimated crowding multipliers are similar in magnitude with varying statistical significance. As it is expected, models with many fixed effect variables achieve better fit at the expense of model simplicity. In light of the principle of parsimony, one may compare the models using the Bayesian information criterion (BIC) that awards model fit (i.e. log-likelihood) and penalises for the number of covariates (Schwarz, 1978). This statistic suggests that our final specification with separate morning and afternoon ASCs is the best performing model.

The linear multiplier surface (11) estimated in Model 4 is visualised in Figure 10. In fact, \( \beta_p \) can be interpreted as the standing penalty – due to the linear specification of the multiplier the standing and seated crowding cost functions are parallel in this model. Similarly, \( \beta_c \) indicates that the disutility caused by an additional passenger per square metre on average is equivalent to the value of 11.92% of the travel time. At six passengers per square metre and no chance to find a seat, the value of time is more than 98% higher than in uncrowded conditions, so effectively it doubles.

Our travel time multipliers are comparable in magnitude but somewhat lower than earlier stated preference results. Whelan and Crockett (2009) and the meta-analysis of Wardman and Whelan (2011) resulted in similar values for seated passengers, but the standing penalties they found are significantly higher: while in our case it is just 1.265, Whelan and Crockett (2009) measured 1.53. Wardman and Whelan (2011) as well as Batarce et al. (2016) concluded that the standing multiplier may go above 2.5 in the worst conditions, while the highest multiplier in our experiment remained under 2.

Our results are very similar to the ones measured by Kroes et al. (2013) in a combined SP and RP experiment from Paris, which reassures the authors’ hint that SP methods may overestimate the user cost of crowding. Our results clearly resemble the combined SP and RP crowding multipliers of Batarce et al. (2015) as well. Using data from Santiago de Chile they found that the marginal disutility of travel time at 6 passengers per square metre is twice as the marginal disutility at the lowest crowding level.

5 Conclusion

This paper presents a comprehensive roadmap to derive the user cost of crowding in a revealed preference route choice framework. Our method relies on raw AFC and AVL datasets only, thus providing an easily reproducible technique for researchers who do not have additional data on station design, passenger characteristics, manual counts, survey results, etc., for
the metro network of interest. The estimated standing multiplier is 1.265. An additional passenger per square metre on average adds 0.119 to the crowding multiplier. These results are in line with earlier revealed preference values, and suggest that stated choice methods may overestimate the user cost of crowding. As a key input for the route choice experiment we derived a passenger-to-train assignment method that recovers the crowding density and standing probability pattern in the entire metro network. The assignment is based on the likelihood of access, transfer and egress times associated with feasible itineraries in the AVL dataset.

In case of the revealed preference experiment we would like to highlight two main contributions to the literature of crowding cost estimation. First, our method controls for the variability of crowding levels during a metro journey, evaluating crowding on each link of a trip. Second, we control for the probability of standing, which is clearly an important determinant of comfort that passengers cannot anticipate with certainty prior to the travel decision. These two features cannot be incorporated in a traditional stated preference experiment, nor in a revealed preference setting when researchers do not observe travel conditions during the entire journey.

Our methods, both the passenger-to-train assignment and the crowding cost estimation, can be extended in multiple directions. In the assignment it is appealing to measure and apply more disaggregate movement time distributions, e.g. separate access time distributions for peak periods and individual transfer time distributions for different pairs of platforms in a large station. In the discrete choice model the statistical treatment of potential measurement errors in the assignment data may improve the reliability of our results. The relatively simple expectation formation assumption applied in this paper could be improved by recovering earlier travel experiences of smart card users. Also, future research efforts may focus on the nonlinearity of crowding cost functions, allowing for imperfect substitution patterns between standing probability and crowding density, and address individual heterogeneity in crowding avoidance preferences.

Crowding in a densely used urban rail network is costly for society, and comparable in magnitude to the user cost of time loss in road congestion. Urban transport policy often neglects the adverse effects of discomfort in public transport when arguing that modal shift eliminates the social burden of congestion. Overall, we believe that an easily implementable crowding cost estimation method may raise more attention to the appropriate management of crowding externalities in public transport, including the need for crowding-dependent pricing. This method allows public transport operators to develop better business and economic cases for service changes, investments and pricing policies.
Acknowledgement

The authors are grateful to Hong Kong MTR Corporation Ltd., the data provider of this research project. We very much appreciate the recommendations of two anonymous referees that helped improve this paper significantly. We also thank Jos van Ommeren, Eric Kroes, Fabien Leurent and Jan-Dirk Schmöcker for their valuable comments. The usual disclaimers apply.

References


Croissant, Y. et al. (2012), ‘Estimation of multinomial logit models in R: The mlogit Packages’. R package version 0.2-2. URL: [http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf](http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf)


Faber Maunsell and Mott MacDonald (2007), ‘Rail overcrowding, reliability and frequency demand impacts project’. Consulting report prepared for Centro.

FHWA (2010), ‘Travel time reliability: Making it there on time, all the time’. US Department of Transportation, Federal Highway Administration.


URL: https://ssrn.com/abstract=2707043


MVA Consultancy (2008), ‘Valuation of Overcrowding on Rail Services’. Report prepared for the UK Department of Transport.


