Merging smart card data and train movement data: How to assign trips to trains?

Daniel Hörcher*

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

- TECHNICAL NOTE -

Abstract

This report explains the assignment method applied to link trips compiled in smart card data to train movements recorded in the signalling system. Particular attention has been paid to (1) origin-destination pairs with multiple potential route options, (2) peak-hour trips delayed by difficulties in boarding crowded trains at the origin station, and (3) trips originating or ending on rail lines not included in the train movement dataset.

In the current version of this paper the metro network on which the method has been applied is anonymised.

*Electronic address: d.horcher@imperial.ac.uk; Corresponding author

1 Objectives

The method described in this technical note is based on two datasets: a smart card dataset that includes all trips performed in the system with the time and location of check-in and check-out transactions, and a train movement dataset that contains each train's arrival and departure time at all stations. The two datasets have been recorded in the same time period. Our goal is to assign smart card trips to trains in the train movement data. This is not a straightforward exercise when multiple trains travelled on an OD pair between a passenger's check-in and check-out times. However, we can use statistical tools to realise the assignment with reasonable reliability.

Passenger assignment uncovers a number of important information about train operations and travel behaviour. Our short-term goal is to use this assignment as an input for crowding analysis. If all passengers are recorded in the dataset[1] and assigned to a train, then we gain information about the occupancy rate pattern of trains, and we can analyse the effect of crowding on passengers' various travel decisions. In the current projet phase of research our train movement dataset does not include all lines of the rail network under investigation. Throughout the rest of this paper the term *urban lines* refers to lines included in the dataset, while *suburban lines* are not, reflecting the usual setting in many large metropolitan areas that suburban rail lines are part of an integrated tariff system, but their operational practices hugely differ from urban metros.

Section 2 explains the typology of smart card trips based on the number of transfers, the number of feasible trains in the train movement data that could have been taken by the actual passengers, the ambiguity of route choice and whether the traveler used a suburban line. Section 3 presents intermediate results illustrating the most crucial parts of the assignment process. Finally, Section 4 discusses our experiences with computation and presents a snapshot of the results.

2 Trip types and related strategies

A graphical summary of trip typoligy is provided in Figure 1. Subsequently, Figure 2 gives an overview how access, transfer and egress time distributions of these trips are used in the assingment process.

[1] All stations in our experimental area are fenced and therefore the dataset should contain all passenger movements in the network.

A **Single trips with only one feasible train**

Trips within a single metro line (no transfers). Only one feasible train means that there was only one train leaving the origin station after the check-in time and arriving to the destination station before the check-out time. In other words, the previous train leaves somewhat earlier than the passenger checks in, and the next train arrives to the destination station somewhat later than the passengers taps out.

In this case the assignment does not require any assumption, we can directly link trips to the only feasible train.

B **Single trips with more than one feasible trains**

The trip has been performed within a single line without transfers, but more than one train travelled between the check-in and check-out time. At this stage even if a train left the origin one second after the traveller checked in, we consider it as a feasible train for that trip. We can calculate the access and egress times for each feasible trains and assign the trip to the train for which the corresponding access and egress times have the highest likelihood.

The basic assumptions here are the following. The distribution of access times of type A and B trips may be different, because the reason why B trips have multiple feasible trains may be that they did not board the first train due to crowding, or simply that a train left while they walked to the platform or took the escalators. We do not know exactly how the access time was shared by walking to the platform, waiting for the first train, and possible waiting for another train if board was not successful the first time.

However, we assume that the egress time has the same distribution for B and A trips. That is, leaving the station takes the same time no matter if the passenger arrived with the first feasible service or not. For type B trips at a destination station we can use the egress time distribution of type A trips arriving to the same station, and assign type B passengers to feasible trains based on the likelihood of egress times of alternatives.

One possible opposing argument against our method is that type A passengers are systematically faster in walking, and this is why they have shorter access and egress times so that only one train travelled during their trip. If this statement was true, then there would be correlation between access and egress times, representing individual characteristics related to the ability of faster walking speed

(age, health, reasons to hurry, etc.). However, the correlation between access and egress times in our dataset (more specifically among type A passengers for whom we can be sure about access and egress times), the correlation between the two variables is almost zero. Therefore we reject this opposing argument.

C **One-transfer trips with only one feasible combination of trains**

Trips that include exactly one transfer between metro lines, but for which we find only one feasible combination of connecting trains, allowing all access, transfer and egress times to be anything above zero. In this case the assignment is exclusive, just like in case of type A trips.

D **One-transfer trips with multiple feasible trains**

Multiple lines, multiple candidate trains. The reason behind the uncertainty can be that a train left during the passenger walked to the platform at the origin station or at the transfer station, or that she was unable to board the first train at either the origin or at the transfer.

We can safely assume again that the egress time of type D passengers has the same distribution as any other types, most importantly type A trips and C. Therefore we can treat the last leg of the trip separately and assign trips by comparing the probability of egress times of competing alternatives.

In the next step we assume that the access time at a specific station (in a specific time period) of type D has the same distribution as type B at the same station and the same time. For these two types the access time has the same components: walking to platform, waiting for the first train, and possibly waiting for subsequent trains if boarding the first services is impossible due to crowding. There is no reason to assume that these two types have different chances to board the first train, all other things being equal. However, we cannot use the transfer time distribution of type C now, because type C definitely didn't have to skip the first train, which cannot be outruled for type D. Therefore type D trips should be assigned to trains on the first leg of their journey based on the access time distribution of type B only.

Note that as a result of type D assignment we gain information on the transfer time distribution when the possibility of missing the first train at the transfer station is not excluded like in case of type C. We will use this distribution later on.

E **Multiple-transfer trips with only one feasible combination of trains**

The assignment in this case is straightforward again, however the occasions when multiple-transfer trips have only one feasible combination of trains are quite rare.

F **Multiple-transfer trips with multiple feasible trains**

The first and the last lags of the journey can be assigned the same way as in case of type D trips, using the access and egress time distributions of types B, and A as well as C, respectively. On the middle section(s) of the trip we assume the transfer times have the same distribution as for type D trips at the same stations. Thus, after we identified all feasible trains we have to choose one based on the joint probability of transfer times at the first and second transfer stations (see illustration below).

If the trip includes more than two transfers, then the likelihood of feasible train combinations on intermediate journey lags depends on the joint distribution of more than two random transfer time variables.

G **Trips departing from/arriving to suburban railway lines**

Our train movement dataset includes the urban lines of the experimental network, while the smart card system is extended to some other 'suburban' railway lines as well. Platforms are fenced along these lines so all passengers are registered who enter the network and included in our smart card dataset. However, without train movement data we cannot assign them to specific trains.

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 have train movement 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 official timetable's travel time, and replace the check-in or check-out time when the passenger may have arrived to the transfer stration. Then we reassign the trip to types A to E, dependending on the remaining transfers and feasible trains.

H **Trips with multiple feasible routes**

Complications may arise in the trip assignment if not only the choice of train, but also the choice of route in the network is unclear from the data. It may

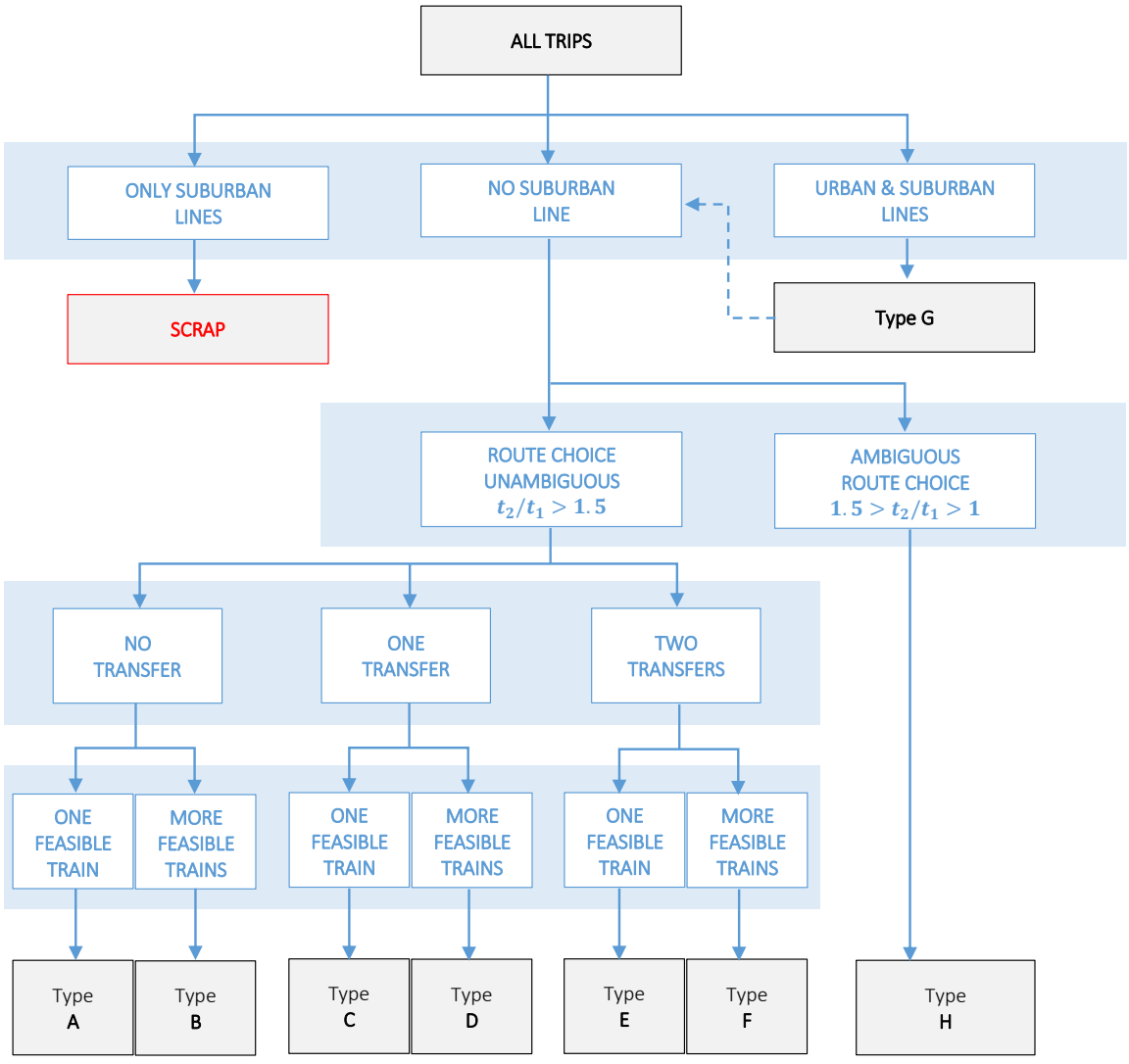


Figure 1: Trip typology based on lines, route choice, transfers and timetable unanimity

be possible that two alternative routes feature different number of transfers, and therefore we have to compare the likelihood of service combinations of different trip types.

There are two possibilities for dealing with this issue. First, we can separate route choice from train assignment. Based on travel times (and possibly other attributes like crowding) on alternative routes first assign the trip to the most likely route. Then identify possible trains on this route and assign the trip to the most likely feasible train (combination), as detailed above. Second, we can identify all feasible trains (or combinations) on all feasible routes connecting the OD pair, and based on access, egress and transfer times choose the most attractive service(s). Note that in the assinment process trip types should be assigned to trains in the fixed order detailed above, so trips with multiple potential routes and thus multiple potential route types should be assigned separately, after all unambiguous trips (types A-E) are assigned. This will increase computation time.

We chose the second method to improve the reliability of the assignment, with one limitation: we only considered potential train combinations on the first and the second shortest paths only. The reason for this was to keep computation time within a reasonable range. In addition, given that our experimental network is relatively simple, it is unlikely that the third shortest path is still competitive compared to the first. We also set a threshold level of travel time ratios between the 1st and 2nd shortest paths below which route choice can be traated as ambiguous: we picked $t_1/t_2 \leq 1.5$ on an intuitive basis.

Figure 1 summarises the typology of trips used in the assignment process. Figure 2 illustrates how access, egress and transfer time distributions have been derived and recycled in subsequent stages.

3 Intermediate results

After completing the straightforward trip assignment to types A and C, the distribution of egress times can be derived for each station as the difference between the assigned train's arrival time and the passenger's check-out time. Figure 3 plots these distributions in the same graph. It is clearly visible that egress times do differ station by station. Another way to improve the process would be to further differentiate these

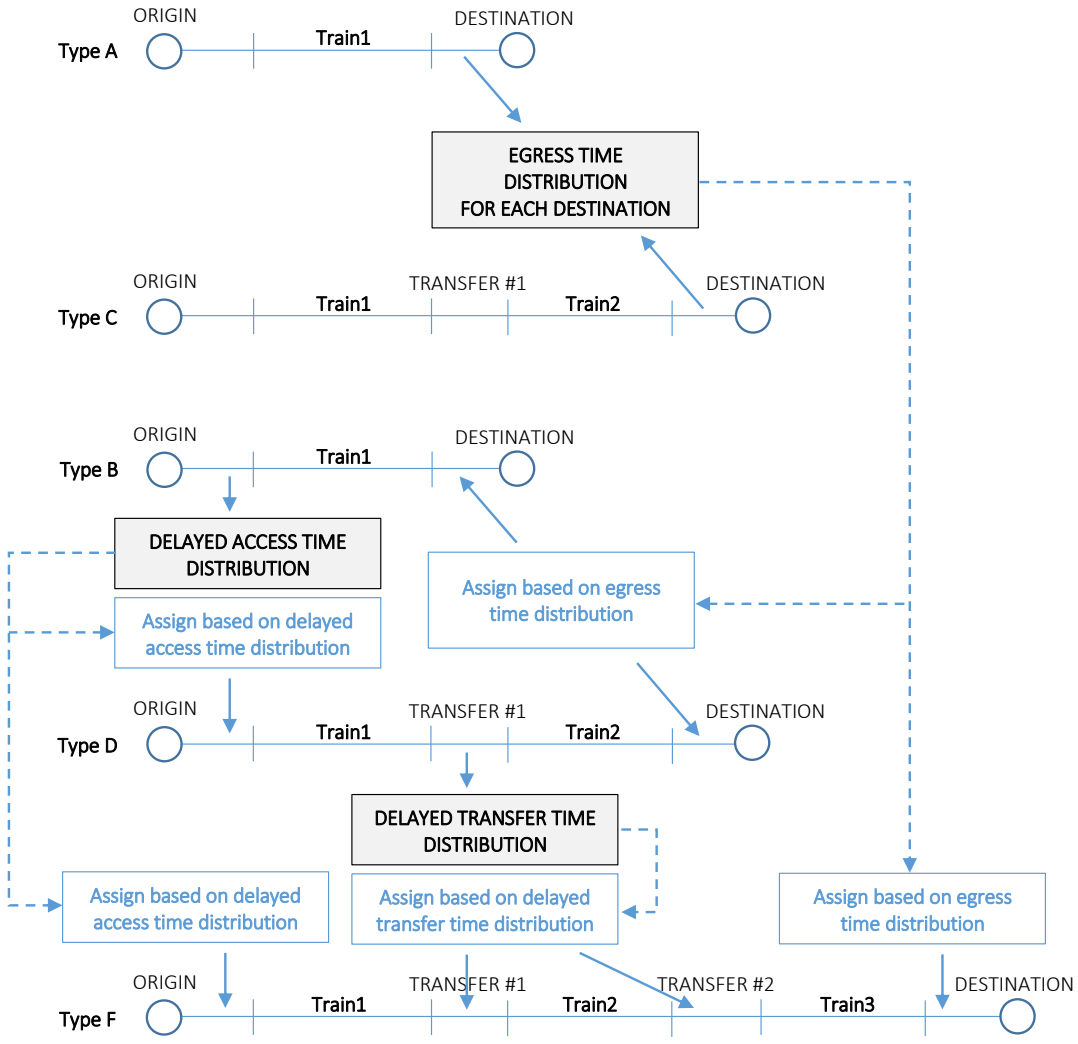


Figure 2: Schematic overview of trip assignment based on access, transfer and egress time distributions

distributions by platform, because at several stations platforms are located at different distance from the fare gentries (e.g. they can be beneath each other). Moreover, differentiation could be made by time of day or day of week, or any other proxy of station crowding, as passenger congestion may have an impact on the speed of leaving the station.

The first case when we apply the egress time distributions is the assignment of type B passengers. Recall that they made no transfer, but multiple feasible trains can be extracted from train movement data that were available within the time frame bounded by the check-in and check-out times. Thus, in this stage we pick the train with the most likely egress time, assuming that access times may have been affected by the inability to board the first train, but egress times have the same distribution.

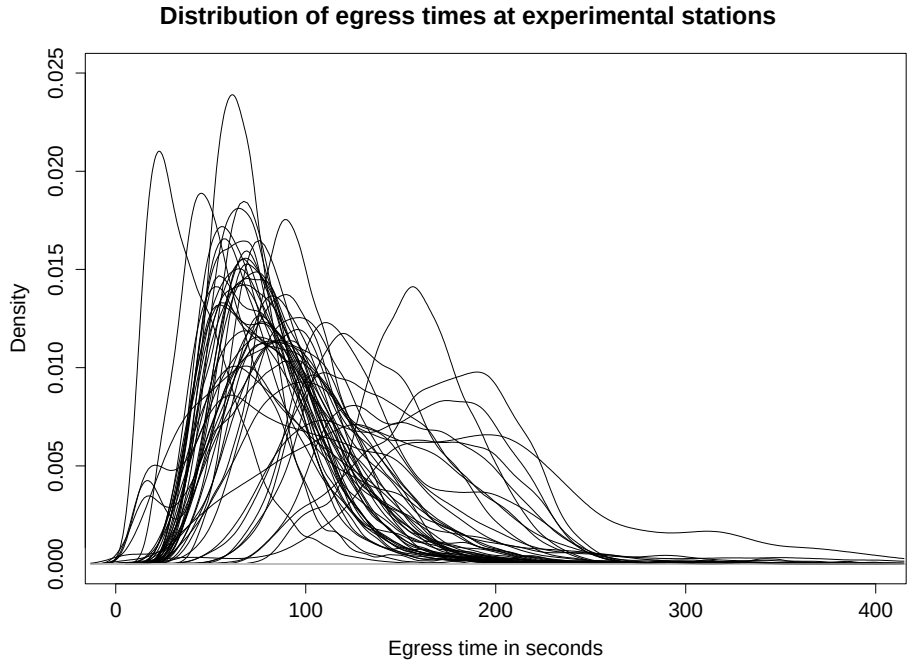


Figure 3: Egress time distribution. Each line represents a separate station's distribution

Let us illustrate the method on an existing trip. We consider a passenger who checked in at station A at 8h58'53" and traveled to Station B[2], where she tapped out at 9h25'31". Between these two points in time three trains passed along the Island line, so the assignment is ambiguous (this is why the trip has been put in group B). Table 1 shows what would be the egress time if the passenger took each of the three possible

[2]The schematic layout of this and subsequent illustrative study cases are depicted in Figure 7

trains: with service 202 it is 582 seconds, service 237 implies 314 seconds, while with the last feasible train, service 276 the egress time would be just 42 seconds.

We define t_k, $k = (1, ..., K)$, as the possible discrete values that egress time T can take and $\theta = (\theta_1, ..., \theta_K)$ as the associated probabilities for vector $t = (t_1, ...t_K)$ such that $P(T = t_k|\theta) = \theta_k$, thus effectively treating the data as a sample from a multinomial distribution of egress times.

In the last column of table 1 we calculated the probability of choosing the trains, conditional on the potential egress times associated with trains that actually travelled, based on the relative magnitude of density values:

$$Prob(T_i) = \frac{\theta_i}{\sum_{j=1}^{3} \theta_j}$$

In the final step of the assignment the algorithm chooses one of the trains randomly, using the probability values as weights. In our case, it is most likely that the passenger traveled with service 276. From Figure 4 we see that 42 seconds of egress time is relatively low. However, it is still more likely than spending as much as 314 seconds (more than 6 minutes) in the station, or 582 seconds in case of service 202.

Table 1: Feasible trains between Station A and Station B in our study case

#	Service ID	Egress time (s)	Density value	Probability
1	202	582	9.355e-06	0.003
2	237	314	6.205e-04	0.224
3	276	42	2.134e-03	0.772

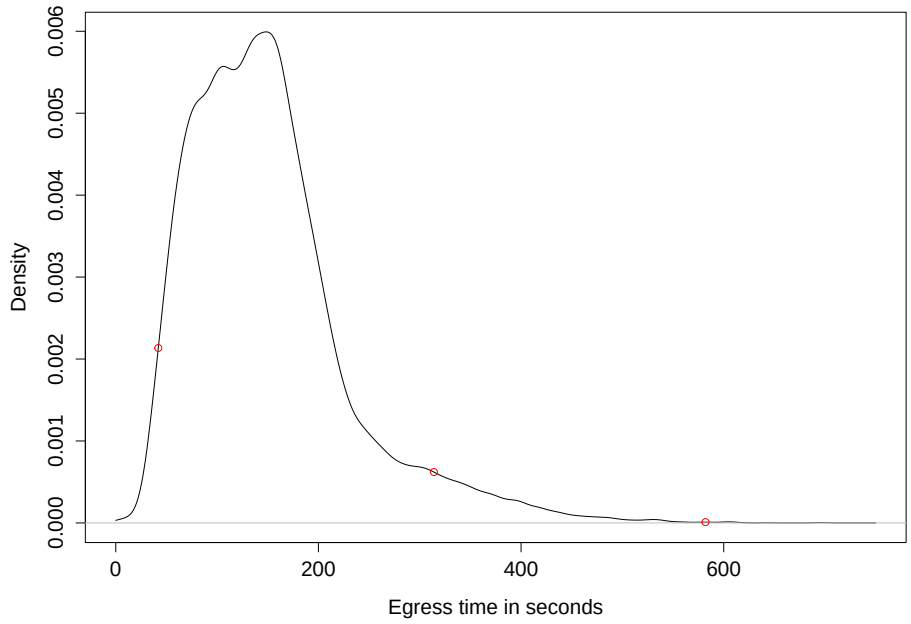


Figure 4: Egress time distribution at Station B with egress time densities for the three feasible trains highlighted

Proof: Derivation of train assignment from egress time PDFs

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. We know the p.d.f. of egress times at the station we are interested in.
- Event A: the occurance of a specific set T of candidate egress times among which the true egress time is. (We extract this information from train movement data.)
- Event B_i: egress time i is the true one, so the passenger traveled with the train associated with egress time i.

Assumption

Trains arrive randomly to the station, so the egress times included in T are independent from each other. That is, even if we know that one train arrived 60 seconds before the passenger checked out, the remaining potential trains in T could have been late or

early, and even the number of additional potential trains is random. As a consequence,

$$P(A|B_i) = P(A|B_j) \qquad \forall\, i \,\text{and}\, j \in T. \tag{1}$$

In practice we build this assumption on the fact that train movements are completely unrelated to egress times, i.e. leaving the train is something that happens after the train's arrival and has no feedback effect on train movements. In fact, A depends on service frequnecy and (any) turbulances causing delays.

Derivation

First of all let us express that the true egress time has to be among the set of candidate egress times:

$$\sum_{j \in T} P(B_j|A) = 1 \tag{2}$$

Using Bayes' Theorem

$$\sum_{j \in T} P(B_j|A) = \sum_{j \in T} \frac{P(A|B_j)P(B_j)}{P(A)} = 1. \tag{3}$$

As $P(A)$ is independent of j, after rearrangement

$$P(A) = \sum_{j \in T} P(A|B_j)P(B_j). \tag{4}$$

The probability that train i has been used, given the information we have about the set of potential trains, can be expressed applying Bayes' Theorem again as

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} \qquad \forall i \in T. \tag{5}$$

After replacing $P(A)$ with equation 4 we get

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j \in T} P(A|B_j)P(B_j)} \qquad \forall i \in T. \tag{6}$$

Given equation 1 in the assumption, $P(A|B_j)$ is independent of j and equals to $P(A|B_i)$, so equation 6 simplifies to

$$P(B_i|A) = \frac{P(B_i)}{\sum_{j \in T} P(B_j)} \qquad \forall i \in T. \tag{7}$$

That is, we can assign the relative density ratios of candidate egress times as probabilities of choosing the associated trains. □

When all type B passengers are assigned to trains, we can derive their acces time distributions, which is expected to be different from type A and C, because for type B we allow for the possibility of failing to board the first train. Figure 5 plots the density distributions of access times. It is worth noting two interesting observations. Many distributions have two local maxima around 150 and 250 seconds. This can be attributed to failed boardings; in this case passengers had to wait about another 2 minutes for the following train. There are some outliers among the stations with significantly longer access times. These are terminal stations where many people prefer to wait longer and have a guaranteed seat.

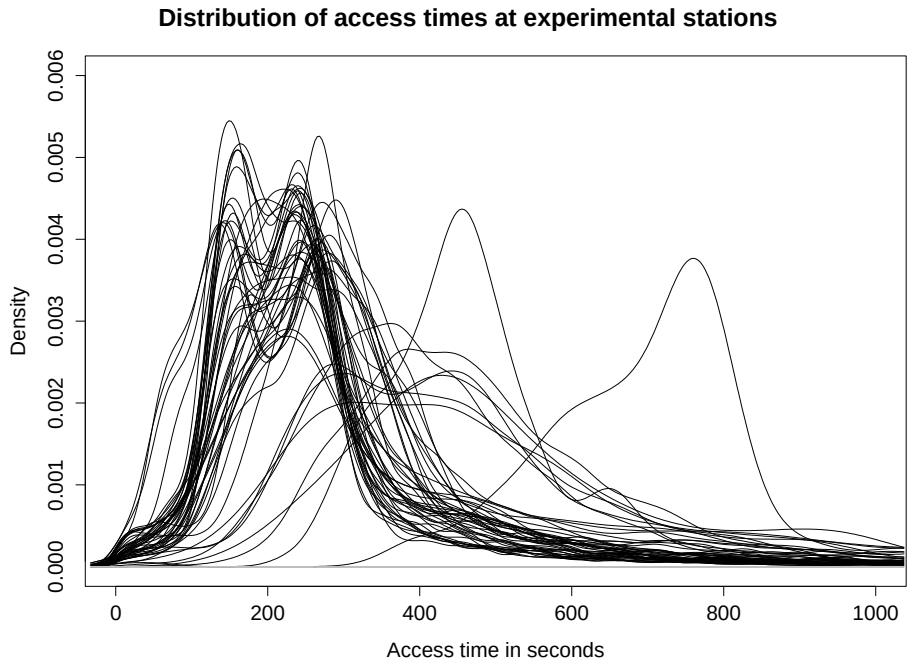


Figure 5: Acces time distribution. Each line represents a separate station's distribution

Having information on access and egress time densities, now we can turn to type D, i.e. one-transfer trips with multiple feasible train combinations. Let us again illustrate the calculation through an example. Our passenger departs from Station X at 18h50'11" and after a transfer at Station Y she taps out at Station Z at 19h09'45". In this case we can safely assume that the transfer station was Station Y, because the second shortest path, i.e. a long detour, would imply three times longer travel time according to the official timetable. Of course, we do not know when she arrived in Station Y and when she boarded the train to Station Z.

Let us therefore collect all trains between Stations X and Y on Line 1, and between Stations Y and Z on Line 2. Table 2 shows that we found three possible trains on Line 1 (with train IDs 296, 325 and 366) that provide transfer at Station Y to three Line 2 trains (79, 112 and 132). The latter two Line 2 services could have been reached by multiple Line 1 trains, so we have to evaluate six possible combinations. The egress time distribution at Station Z, plotted in Figure 6 clearly indicates that only train 132 can be reasonably considered on the second leg of the journey. In case of access times, the most likely Line 1 train was 325 with 277 seconds access time, but train 366 cannot be excluded either with its access time of 478 seconds.

Table 2: Feasible train combinations between Stations X and Z in our study case

#	ID 1	ID 2	Access (s)	Egress (s)	Transfer (s)	Density	Probability
1	296	79	37	527	136	0	0.000
2	296	112	37	302	348	4.506e-10	0.000
3	325	112	277	302	105	4.900e-08	0.002
4	296	132	37	113	540	1.618e-07	0.007
5	325	132	277	113	297	1.759e-05	0.716
6	366	132	478	113	96	6.763e-06	0.275

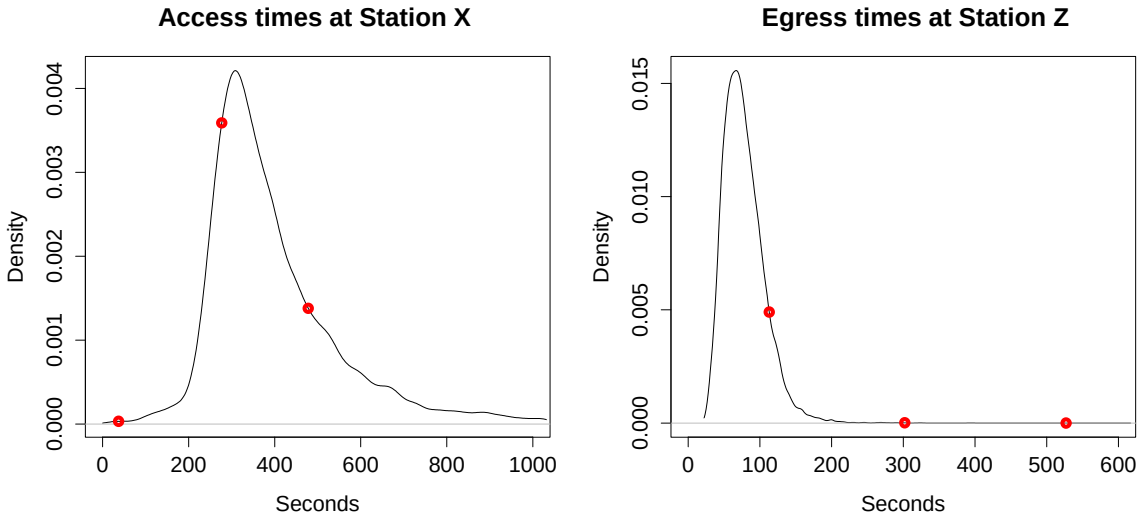


Figure 6: Access and egress times of feasible trains for an example transfer trip between Stations X and Z, with an interchange at Station Y

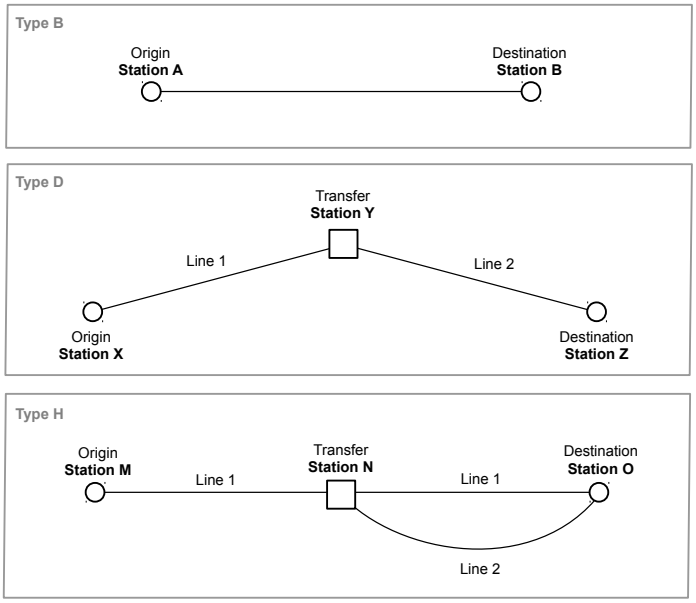


Figure 7: Schematic layout of explanatory study cases

As access times at Station X and egress times at Station Z are mutually independent random variables, the probability of choosing a train combination is simply the product of the respective densities. Thus, we assign probabilities to train combinations using

$$Prob(C_i) = \frac{AD_i \cdot ED_i}{\sum_{j=1}^{6}(AD_j \cdot ED_j)}.$$

The resulting probabilities are provided in the last column of Table 2. In line with our earlier intuitive predictions, the most likely trains with 71.6% probability were 325 with 132, but 366 with 132 also have 27.5% chance.

As all type D passengers are now assigned to trains, we can extract the distribution of transfer times at various stations. Note, that type C trips also have a transfer time distribution, but in that case there was only one feasible train combination, which outrules the possiblity that the passenger could not board the first crowded train. Therefore we focus on type D.

Figure 8 shows the resulting transfer time distribution at some of the most frequently used interchanges. In case of Stations 1 and 2, the distribution of transfer times is relatively flat. These are large stations with several platforms, so regular patterns remain hidden and very long transfer times (around 15 minutes) are not atypical at all. By contrast, transfer stations 3 and 4, have much simpler design allowing all passengers to switch train on the same platform. This is a possible explanation of the

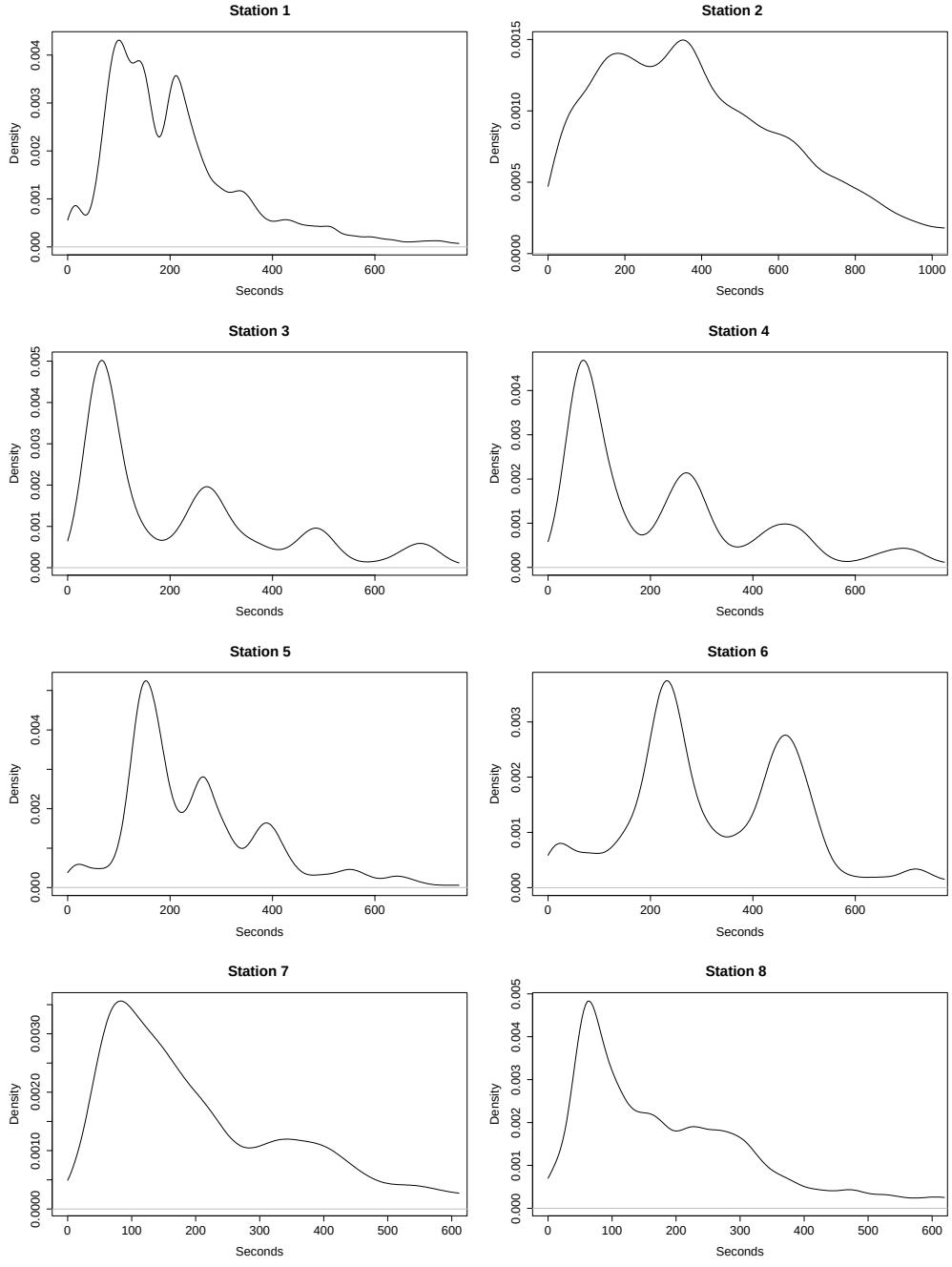


Figure 8: Transfer time distributions at the most densely used transfer stations between urban metro lines

fact that at these stations transfer times follow a regular pattern with a decreasing number of people waiting one, two or even three additional trains before being able to board. It may be another precondition for regular transfer time patterns to have constant headways between consecutive trains in the most crowded periods.

Similar phenomena can be observed at Stations 5 and 6. The only difference is that the former features three local peaks, while the latter has only two, suggesting that it is less usual that passengers have to wait three trains at Station 6 before boarding. At Stations 7 and 8 the secondary peaks disappear, from which one may assume that overcrowding is less severe at these transfer stations.

Note that in the current experiment we aggrageted all transfer times performed in a day at a particular station. Nevertheless, it would be possible to differentiate transfer times at separate platforms and even by time of day. This possibility is a low hanging fruit providing several additional insights.

Based on the assigment method used for type B and D, it straightforward how type F can be treated. Type F trips have two transfers with multiple feasible train combinations in most cases. We applied the same method:

1. Extract from the train movement dataset all trains that traveled in the given timeframe;
2. Combine feasible trains that provide trainsfers with each other;
3. Calculate the resulting hypothetical access, egress and transfer times and the densities corresponding to these time values;
4. Assign probabilities to train combinations after multiplying the densities of travel time components;
5. The algorithm chooses a combination randomly, using the probabilities we derived as weights.

It is more interesting to discuss type H, for which the route choice is ambiguous, because the second shortest path takes no longer than 1.5 times the journey time on the shortest path. If it was guaranteed that the number of transfers on the two routes are the same, we could use the same method as before: collect all feasible train combinations from the timetable and evaluate them based on the hypothetical access, egress and transfer times. But if the number of transfers is different, that the more transfers a route has, the lower the product of densities will become, simply because we include more density values in the multiplication. Therefore in this case we took the

geometric mean of travel time densities to calculate an option's 'aggregate' probability density:

$$D_{i,j} = \left(\prod_{l=1}^{n_i} d_{i,j,l} \right)^{\frac{1}{n_i}}, \qquad Prob(C_{i,j}) = \frac{D_{i,j}}{\sum_{k \neq \{i,j\}}^{N} D_k}$$

where n_i is the number of travel time components on route i, N is the number of feasible combinations. For example, in case of one-transfer trips, $n = 3$. Train combinations on route i are indexed with j, and k can represent any train combination on both routes except j on route i. Finally, $d_{i,j,l}$ is the density of individual access, egress and transfer times for train combination j on route i. This way alternatives on parallel routes with different number of transfers become comparable. This method is again based on the assumption that travel time components are independently distributed.

We illustrate the assignment on a study case (see Figure 7). Let us consider a trip between Station M and Station O. Check-in time was 14h45'35" and the tap-out has been registered at 15h17'29". There are two potential routes on this OD pair: a direct trip on Line 1 or a one-transfer journey with an interchange at Station N to Line 2. Transferring to Line 2 offers a shortcut, as the direct tirp has 1.3 times longer travel time according to the official timetable, including transfer time. However, is it easily possible that the inconvenience of transfering convinces some passengers to accept the time loss and travel directly. Therefore we extract from the train movement data all feasible train combinations on the two alternatives. Table 3 summarises them.

Table 3: Feasible train combinations on two routes between Stations M and O in our study case

#	ID 1	ID 2	Access (s)	Egress (s)	Transfer (s)	Density	Probability
1	526	–	28	31	–	2.189e-04	0.051
2	11	53	234	260	263	2.665e-03	0.620
3	37	53	438	260	54	1.090e-03	0.254
4	526	53	28	260	478	3.089e-04	0.072
5	526	193	28	671	71	1.366e-05	0.003

What we can see is that service 526 left the origin 28 seconds after the check-in and arrived to the destination station 31 seconds before check-out, so this is a feasible schenario. However, from service 526 the passenger could have switched to two possible trains on Line 2 at Station N: the first is 193 which implies 671 seconds egress time, and the second is 53 which arrived 260 seconds before the tap-out. Another difference

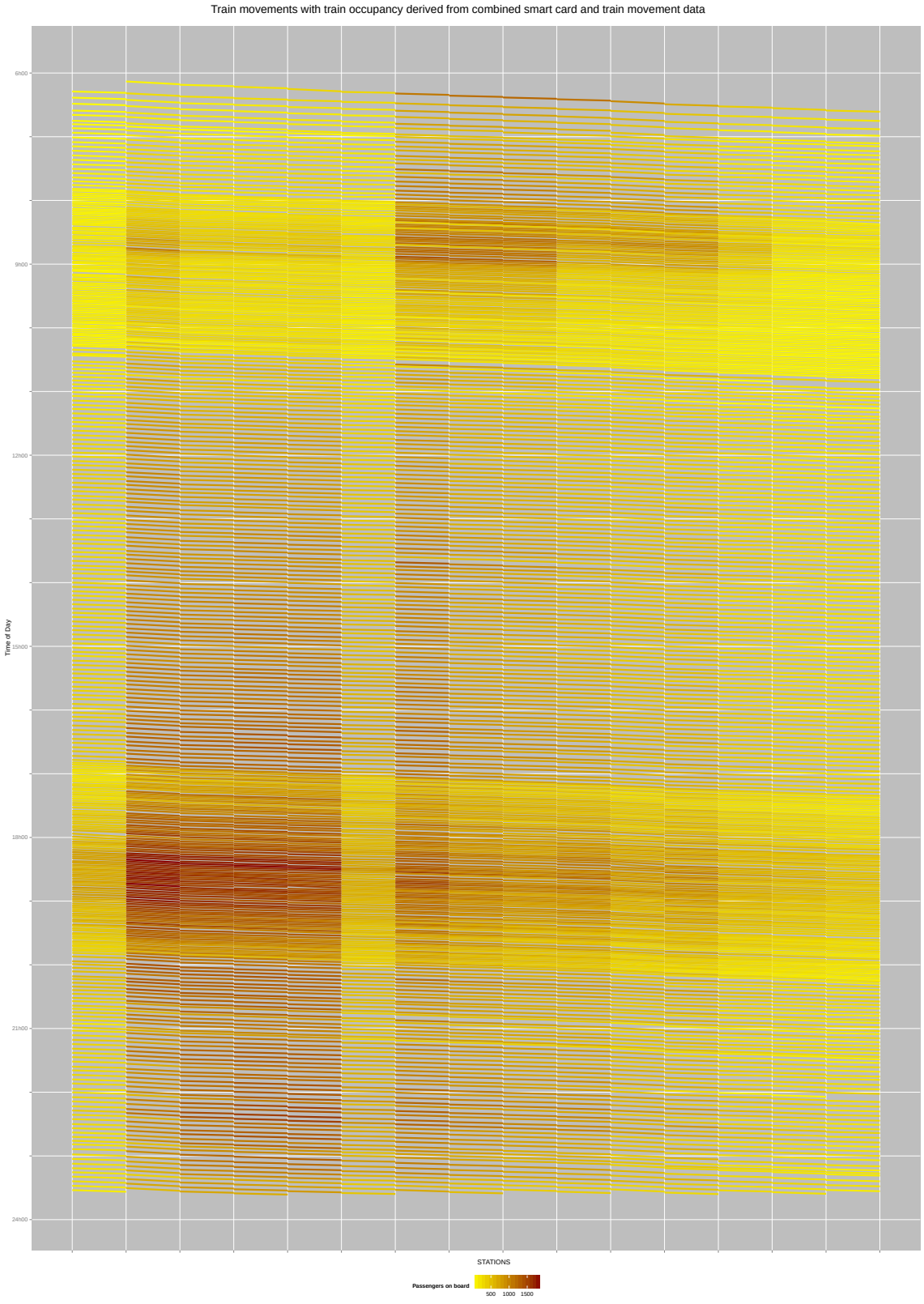
between them is the transfer time, for which we also have a probability distribution at Station N. In addition, train 53 on the Line 2 could have been reached by two other Line 1 trains departing later than train 526: these are 11 and 37, with 234 and 438 seconds of assume access times, respectively.

Based on the travel time distributions the calculation concludes that the direct trip has only 5 percent chance as opposed to three more likely transfer trip combinations. The last option in Table 3 can be outruled, because it would have too short access and transfer times, and 11 minutes to leave the destination station is also hard to explain.

4 Computation time and a snapshot of final results

We realised the assignment algorithm using the R programming environment. To derive shortest paths and official travel times, we relied on the *igraph* package of R. As the assignment process is relatively complicated and requires a number of internal decisions during computation, at the current stage it seems inevitable to process the smart card dataset with loops in the script for each trip. Given that our datasets contain around 5-7 million trips per day, computation time becomes a relevant issue, at least on ordinary PCs. Based on our experience computation times can reach two days on a PC featuring 3.40 GHz CPU and 16 GB RAM.

The figure below depicts the results of the assignment process on one of the urban metro lines of the experimental network. This is a graphical representation of train movement data for a single day. Each downward sloping line links the departure and arrival times of a train between two consecutive stations. Line colours show the number of passengers on board on each interstation. As all trains have the same capacity, passenger numbers are proportional to the average density of crowding. The relatively lower crowding density at the middle of the line are not surprising on this line, the pattern can be explained by significant tranfer flows at these stations.


Train movements with train occupancy derived from combined smart card and train movement data
6h00
9h00
12h00
15h00
18h00
21h00
24h00
Time of Day
STATIONS
Passengers on board
500
1000
1500