

How to use APC data to model passenger movement on-board? An application to Paris suburban train network

Rémi Coulaud^(1,2) & Mathilde Vimont⁽²⁾

¹ *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

² *Transilien, SNCF Voyageurs, 10 rue Camille Moke, 93220, Saint-Denis, France*

Extended abstract submitted for presentation at the 8th International Conference on Transport Network Reliability (Stockholm, 16-18 June, 2021)

1. Introduction

The continuous increase of passengers in Île-de-France transportation network attracts attention on non-uniform passenger distribution among coaches which reduces both passenger comfort and carrying capacity. The COVID-19 epidemic increases even more this need for precise information on passenger load inside trains (Tirachini & Cats 2020), especially if we consider the relatively high impact of passenger information on on-board crowding distribution (Zhang et al. 2017). Yet, allowing passengers to choose a boarding coach with respect to crowding can be done only if reliable on-board load information is available.

Numerous tools exist to estimate on-board crowding. Most of the literature relies on weight measurement in the air suspension system of the rolling stocks that allows a direct load measure (Jenelius 2019, Pefitsi et al. 2020). In some papers, the load measure is also obtained through infra-red or video sensors positioned on top of the train doors, as it is the case in Munich (Khomchuk et al. 2018). Though Automatic Passenger Counting (APC) gives an indirect measure of on-board crowding at a large scale (i.e, consist or train scale), it is difficult to get a reliable estimation at a smaller scale (i.e, coach scale) without taking into account passenger movement on-board.

In our case, passengers can board each coach through doors equipped with infra-red APC systems counting the number of alighting and boarding passengers at each stop. Within a consist (see Figure 2), coaches communicate to allow passengers to spread more uniformly between coaches. To our knowledge, few research studies have been led in such a context. Indeed, a large part of transportation literature studies the motivation behind boarding a specific coach as in Kim et al. (2014) or behind waiting for trains at a specific position along the platform (Hänseler et al. 2020). Schöttl et al. (2019) got closer to our problem and analysed the passenger seating strategy, yet not studying how to model their movements between coaches. There is also a large part of the pedestrian literature which took over quite similar issues, especially to understand alighting and boarding times thanks to either experimentation (Daamen et al. 2008) or automate cellular model (Seriani & Fujiyama 2019).

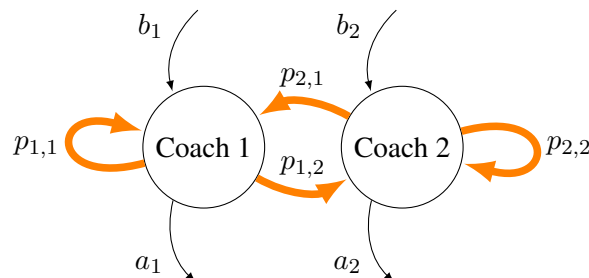


Figure 1: Illustration of the problem for a consist with two coaches

We propose to discretize a consist through a graphical model where each node represents a coach and each link between nodes is pondered by the probability to board one node and alight from the other node (see Figure 1). It naturally leads us to model passenger movement from one coach to another through a

multinomial probability distribution. Our model finds itself right in the middle of microscopic pedestrian simulation for platform-train or station pedestrian movement (Tang et al. 2017, Seriani & Fujiyama 2019) and macroscopic passenger movement in a transportation network through Origin-Destination matrix estimation (Van Zuylen & Willumsen 1980, Kuusinen et al. 2015).

2. Problem formulation

We recall that all doors of each consist are equipped with APC systems measuring the number of alighting and boarding passengers. For each coach $i = \{1, \dots, C\}$, with C the maximum number of coaches per consist, let us define $b_i^{k,s}$ and $a_i^{k,s}$ the number of boarding and alighting passengers from coach i at station s for trip k ¹. We write the number of boarding and alighting passengers for trip k at station s as follows: $b^{k,s} = \sum_{i=1}^C b_i^{k,s}$ and $a^{k,s} = \sum_{i=1}^C a_i^{k,s}$.

S_k is the set of all stations during a trip k . Based on the conservation flow property of Kuusinen et al. (2015) our data-set is corrected such that there is no measurement error at the consist scale at the end of trip k :

$$\sum_{s \in S_k} b^{k,s} = \sum_{s \in S_k} a^{k,s}. \quad (1)$$

We also define the cumulative quantities by trip for coach i : $b_i^k = \sum_{s \in S_k} b_i^{k,s}$ and $a_i^k = \sum_{s \in S_k} a_i^{k,s}$, (see Table 1). The random variable $X_{i,j}$ is the number of passengers boarding coach i and alighting from coach j . Seemingly, $p_{i,j}$ is the probability for passengers to board coach i and alight from coach j . We interpret X margins as follows:

- $X_{i,\cdot} = (X_{i,1}, \dots, X_{i,C})$ corresponds to the destination coaches (alighting coaches) vector for passengers that boarded coach i ;
- $X_{\cdot,j} = (X_{1,j}, \dots, X_{C,j})$ corresponds to the origin coaches (boarding coaches) vector for passengers that alighted from coach j .

Table 1: Notations and variables description

Notation	Description
i,j	coach number among $\{1, \dots, C\}$ with C the maximum number of coaches
k	a unique trip id defined by the triplet: (consist (lead:1, rear:2), train number, day)
s	a station id
$b_i^{k,s}$	number of passengers boarding coach i for trip k at station s
$a_i^{k,s}$	number of passengers alighting from coach i for trip k at station s
b_i^k	total number of passengers boarding coach i for trip k
a_i^k	total number of passengers alighting from coach i for trip k
$b^{k,s}$	number of boarding passengers for trip k at station s
$a^{k,s}$	number of alighting passengers for trip k at station s
$X_{i,j}^k$	number of shifted passengers i.e, passengers moving from coach i to coach j for trip k

¹Defined in Table 1

We used k in notations for the sake of completeness but we will only use it if needed in the following paragraphs. Contrary to the models used to estimate OD matrices, passengers move either to the left or the right side of the consist such that each coach j is accessible when boarding coach i . We suppose that passenger movement is modelled by a multinomial probability distribution conditionally to the number of boarding passengers for each door such that $X_{i,\cdot} \sim \mathcal{M}(b_i, p_{i,1}, \dots, p_{i,C})$. The associated distribution function is defined $\forall x_{i,1}, \dots, x_{i,C} \in (0, b_i)$ such that $\forall i \in \{1, \dots, C\}, \sum_{j=1}^C x_{i,j} = b_i$:

$$\mathbb{P}(X_{i,1} = x_{i,1}, \dots, X_{i,C} = x_{i,C} | b_i) = \frac{b_i!}{x_{i,1}! \dots x_{i,C}!} p_{i,1}^{x_{i,1}} \dots p_{i,C}^{x_{i,C}}.$$

We model passenger movement for each coach i such that we obtain a transition matrix given by, $p \in \mathbb{R}^{C \times C}$:

$$p = \begin{pmatrix} p_{1,1} & \dots & p_{1,C} \\ \vdots & \ddots & \vdots \\ p_{C,1} & \dots & p_{C,C} \end{pmatrix} \quad (2)$$

It verifies for all $i \in \{1, \dots, C\}, \sum_{j=1}^C p_{i,j} = 1$ and for all $i, j \in \{1, \dots, C\}^2 p_{i,j} \in [0, 1]$. The conditional expectancy of a single element of a multinomial random variable is $\mathbb{E}[X_{i,j} | b_i] = b_i p_{i,j}$. Each observation $k \in K$ corresponds to a unique trip. A classical estimator of $p_{i,j}$ is the maximum likelihood estimator defined as follows, $\forall i, j \in \{1, \dots, C\}^2$:

$$\hat{p}_{i,j} = \frac{1}{K} \sum_{k=1}^K \frac{x_{i,j}^k}{b_i^k}.$$

This equation is used by Krstanoski (2014) to estimate passenger distribution on platform and by Ben-Akiva et al. (1985) to estimate OD matrices. However in both cases they observed $x_{i,j}^k$ while in our case there is no available measure of passenger movement inside consists. To overcome this difficulty, we generalise the conservation flow property (1) at the coach scale such that the total number of alighting passengers from coach i for a trip is supposed to be equal to the total number of passengers which moved to i :

$$\forall j \in \{1, \dots, C\}, \quad a_j - \sum_{i=1}^C x_{i,j} = 0. \quad (3)$$

This leads us naturally to the following objective function defined by the conditional expectancy of the Euclidean distance between the total number of alighting passengers a and the random number of shifted passengers X :

$$\mathbb{E} \left[\sum_{j=1}^C \left(a_j - \sum_{i=1}^C X_{i,j} \right)^2 \middle| a, b \right],$$

of which an empirical version would be:

$$\frac{1}{K} \sum_{k=1}^K \sum_{j=1}^C \left(a_j^k - \sum_{i=1}^C x_{i,j}^k \right)^2.$$

However, as we cannot observe $x_{i,j}^k$, we replace it by its theoretical conditional expectancy $b_i^k p_{i,j}$. The number of shifted passengers from coach i to j corresponds to a fixed proportion of the number of passengers having boarded i . It converts the problem into a general least square model under constraints:

$$\begin{aligned}
\min_p \quad & \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^C \left(a_j^k - \sum_{i=1}^C b_i^k p_{i,j} \right)^2 \\
\text{s.t} \quad & \forall i, j \in \{1, \dots, C\}^2, 0 \leq p_{i,j} \leq 1 \\
& \forall i \in \{1, \dots, C\}, \sum_{j=1}^C p_{i,j} = 1
\end{aligned} \tag{4}$$

We easily find an analytical solution to this problem since it is equivalent to a regression problem where parameters are optimised under constraints.

To sum up the reasoning behind this model:

- we have observations of the number of boarding and alighting passengers for each door i but we do not observe passenger movement between communicating coaches;
- we assume that this passenger movement is driven by a multinomial probability distribution and we verify the generalised conservation flow property (1) at the coach scale;
- passenger movement is simplified as a parametric equation using transition matrix and boarding passengers by door.

We then introduce an additional assumption based on the willingness of passengers to move from their boarding coach. Indeed, it is classic to consider that passengers prefer to stay near their boarding coach (Kim et al. 2014). To take into account this hypothesis on passengers' behaviour, we add the following constraint to the optimisation problem (4), $\forall i, j \in \{1, \dots, C\}^2$ we have $\alpha_i \in [0, 1]$ such that:

$$p_{i,j} = p_{i,i} \times \alpha^{|i-j|}. \tag{5}$$

α embedded the passengers' willingness to move. In this case, we loose the convexity of the problem and thus need to use a non convex optimisation solver.

To briefly conclude, we will compare three models:

1. A free movement model based on equation (4);
2. A constrained movement model based on equations (4) and (5);
3. A naive model, also called no-movement model, based on the identity transition matrix, which corresponds to a situation where all passengers stay in their boarding coach.

3. Case study

We applied this methodology framework to lines H and L of the Paris suburban train network (France). On those lines, trains are composed of two consists, each separated in either seven (line L) or eight (line H) connected coaches (see Figure 2).

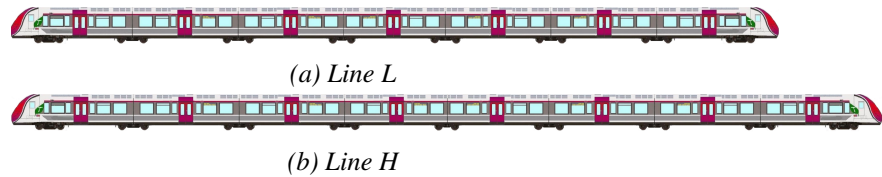


Figure 2: Consist models for line L and H. Each coach is defined by a door (in purple)

Regarding line H, we only considered trains running on its western section, composed of 14 stations located between Paris Gare du Nord and Pontoise. Similarly, we only considered trains running on the

southern section of line L, composed of 16 stations located between Paris Saint-Lazare and Versailles-Rive-Droite (see Figure 3).

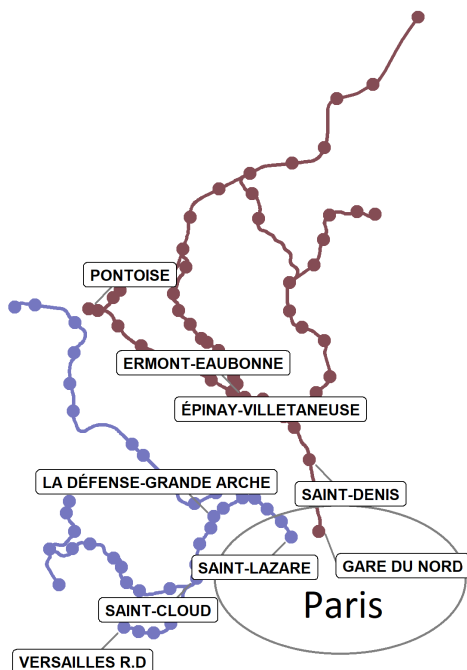


Figure 3: Geographic representation of line H and line L. Main stations of the considered sections are highlighted

We used APC data available for each coach at each train stop and aggregated the countings by trip. We respectively have data from the 1st of September 2018 to the 31st of August 2019 for the train data-set (line H \approx 21,000 trips vs. line L \approx 21,600 trips), and from the 1st of September 2019 to the 30th of November 2019 for the test data-set (line H \approx 5,100 trips vs. line L \approx 8,100 trips). This leads to a train/test ratio of 80/20. A sample data-set can be found in Table 5.

Transition probabilities are estimated on the train data-set, then we use the test data-set to evaluate models' performances. To estimate these probabilities, we also wanted to take stations' layout into consideration, as it has been shown to have an impact on passengers choices regarding boarding and alighting coaches (Kim et al. 2014, Fang et al. 2019). But as our objective function is defined at the trip scale, we hardly can take into consideration stations' layout as such in the optimisation problem. We partly overcame this issue by separating each data-set according to train ways of circulation (even: from suburb to Paris, odd: from Paris to suburb) and consist position (rear or lead consist) to estimate transition probabilities. Indeed, passengers' behaviour may differ between consists since they may be located differently relatively to platforms' entrances and exits. Similarly, passengers circulating on one way do not board and alight at the same stations and thus may behave differently within the train compared to the other way.

4. Results and conclusions

In this section, we present three main preliminary results:

- On-board passenger movement models decrease the difference between total alighting passengers and shifted passengers by coach;
- Passengers on both H and L lines rarely move by more than one or two coaches, which justifies the simpler constrained movement model;
- Some specific behaviours are revealed through the analysis of transition matrices.

Global performances and on-board behaviour

Naive, free and constrained movement models are evaluated through their objective function value (4) which is the difference between total alighting passengers (a) and total shifted passengers ($\hat{p}b$). We relied on the root of this quantity (i.e, RMSE) for ease of interpretation. They are gathered in Table 2.

Firstly, both movement models perform better than the naive model for the two lines, though the free movement model is the best, with an error divided by two for line H (**8.33** vs. 19.56) and almost by three for line L (**7.63** vs. 21.31). The improved performances of the free movement model compared to the constrained one are consistent with the more numerous parameters it relies on, thus allowing it to better fit the data.

Table 2: Performances of the three passenger movement models studied

Models	RMSE	
	Line H	Line L
Free movement	8.33	7.63
Constrained movement (α)	8.61	8.13
Naïve	19.56	21.31

We observe that performances of free and constrained movement models are very similar, with an error difference of only 0.28 and 0.50 for lines H and L respectively. The transition probabilities estimated through the free movement model also highlight that the probabilities of moving near the boarding coach (i.e, staying in the boarding coach or switching to one of the neighbour coaches) are high, with a probability averaged over all coaches comprised between 0.69 and 0.87 (see Table 3). This illustrates the passengers willingness to stay relatively close to their boarding coach.

Table 3: Probabilities of moving near the boarding coach (i.e staying in the boarding coach or switching to one of the neighbour coaches). To recall the circulation way, even: from suburb to Paris and odd: from Paris to suburb.

	Line H				Line L			
	Consist 1		Consist 2		Consist 1		Consist 2	
	Even	Odd	Even	Odd	Even	Odd	Even	Odd
Coach 1	0.90	0.88	0.70	0.73	1.00	0.94	0.95	1.00
Coach 2	0.73	0.75	0.83	0.90	0.97	0.74	0.91	0.91
Coach 3	0.75	0.75	0.58	0.71	0.85	0.83	0.82	0.90
Coach 4	0.72	0.65	0.79	0.75	0.86	0.72	0.82	0.79
Coach 5	0.72	0.62	0.70	0.74	0.86	0.81	0.81	0.87
Coach 6	0.52	0.85	0.89	0.79	0.95	0.74	0.89	0.93
Coach 7	0.55	0.59	0.90	0.80	0.58	0.48	0.56	0.40
Coach 8	0.59	0.66	0.26	0.27	-	-	-	-
Mean Coach	0.69	0.72	0.71	0.71	0.87	0.75	0.82	0.83

Specific movement strategies

If we take a closer look at these probabilities, we can spot a singularity affecting the rear coaches of trains circulating on both lines. Regarding line H, we notice a very low probability of staying near the 8th coach of the second consist for both ways of circulation (0.26 for odd trains vs. 0.27 for the others). Figure 4 shows how passengers boarding this coach in trains circulating in the odd way tend to propagate towards the front of the consist. Indeed, transition probabilities for this coach are uniform (0.126 ± 0.031).

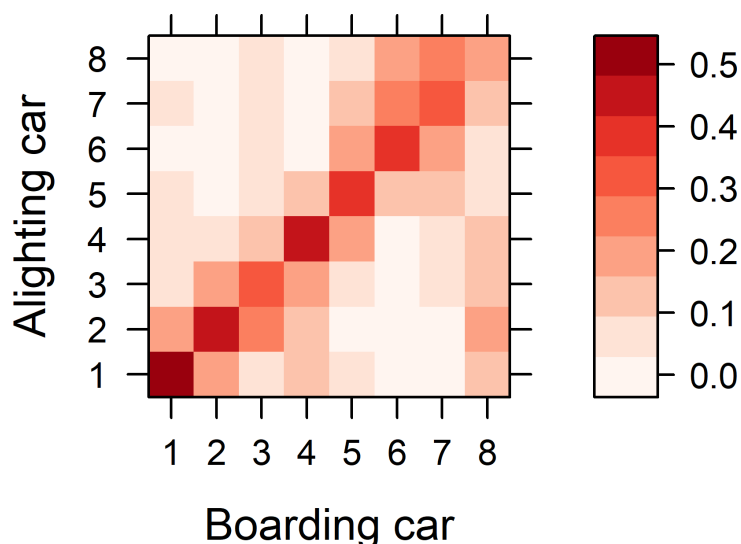


Figure 4: Transition matrices estimated with the free movement model for rear consists circulating on line H in the odd way

Yet, most of the passengers on this line section board at Paris Gare du Nord station, which is composed of a major platform entrance located near the rear of the train as shown in Figure 5. Passengers on this section also tend to alight at multiple stations along the trip, each having different layouts when it comes to platform exits.

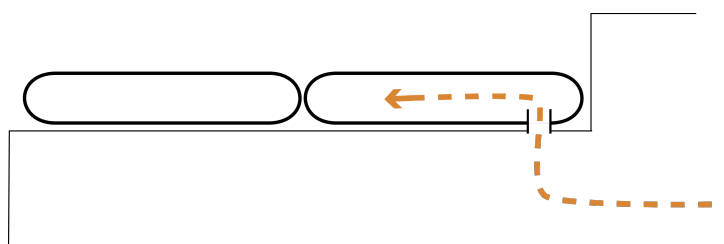


Figure 5: Drawing of Paris Gare du Nord platform station

The mean shares of boarding passengers in each coach at Paris Gare du Nord station presented in Figure 6 clearly show that most passengers board the 8th coach of one of the two consists. Thus, the singular aforementioned probability could be due to passengers boarding the closest coach to the entrance platform at Paris Gare du Nord and then propagating themselves in the train, searching either for an available sit or a place near destination exit. This behaviour would be consistent with the literature that exists on passenger strategy in choosing boarding and alighting coaches (Kim et al. 2014, Fang et al. 2019). Also, the strong impact of Paris Gare du Nord station layout on estimated transition probabilities on line H highlights the relevance of trying to go one step further and to find a way of directly including stations' layout in the optimisation problem.

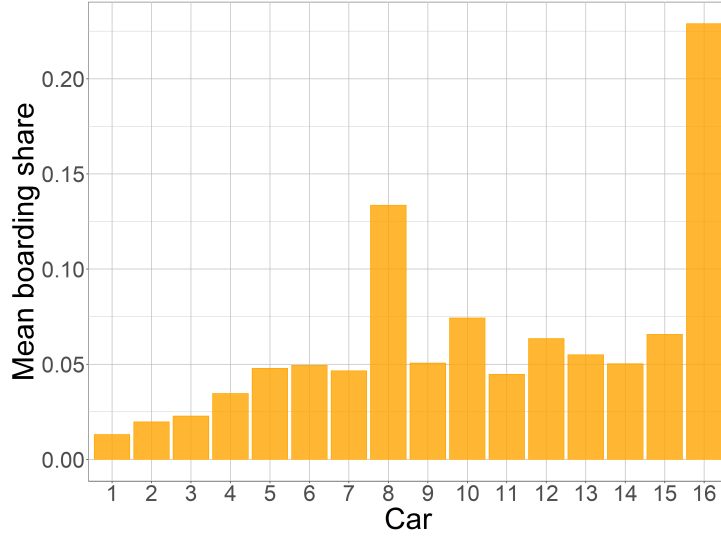


Figure 6: Mean boarding shares in each coach at Paris Gare du Nord station

Local performances

As the free movement model reduces the most the error made on the actual position of passengers within the train, we went further and studied how this model may correct for inconsistent load observations during the trip. More precisely, we noticed that not taking into account passenger movement within the consist led to numerous negative and extreme coach loads² during the trip, as well as plenty of non-null coach loads at terminus (i.e, the conservation passenger flow at the coach scale isn't satisfied). Those occurrences for both naive and free movement models are gathered in Table 4.

Table 4: Inconsistent coach loads observed during the trips on lines H and L for naive and free movement models

		Line H		Line L	
		Naive	Free movement	Naive	Free movement
Before terminus	Negative loads occurrences (%)	10.9	4.8	12.1	5.6
	Extreme loads occurrences ²	4294	1	3346	231
At terminus	Non null loads occurrences (%)	85.0	83.0	83.4	81.3

One of the main improvements provided by the use of the model, is the almost disappearance of extreme loads for both H and L lines. Regarding line H, it is easily explained by the fact that most of these extreme loads occur at origin station Paris Gare du Nord for trains circulating in the odd way and affect the 8th coach of the rear consist. The aforementioned study of transition matrices showed a clear effect of our model that is, the strong propagation of passengers from this coach to the rest of the consist. It is also consistent with the platform layout in Paris Gare du Nord.

²Above 120% of the coach capacity which is 138 passengers for line H and 130 passengers for line L

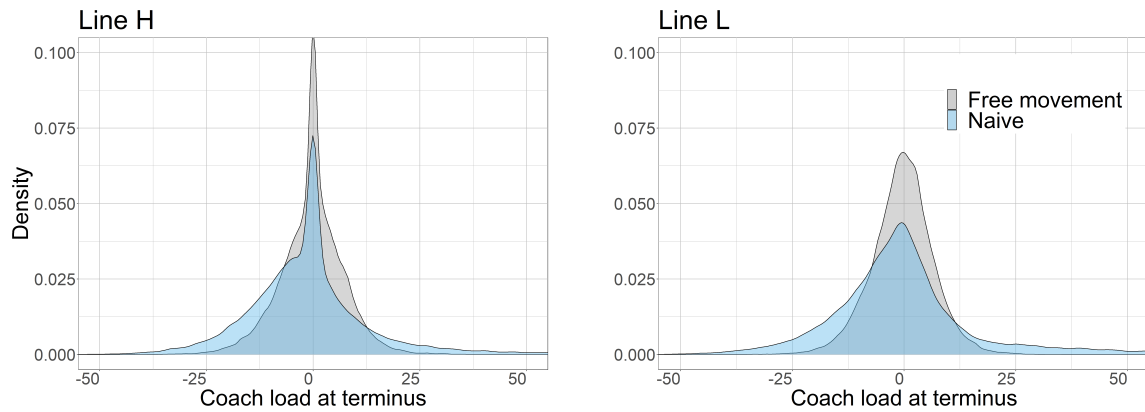


Figure 7: Distribution of coaches load at terminus for H and L lines

Our model also allows to improve the load estimation by reducing the number of negative coach loads during the trip by more than two for both lines: 4.8% vs. 10.9% occurrences in line H, and 5.6% vs. 12.1% occurrences in line L. Also, though the decrease in the number of non-null coach loads observed at terminus is quite weak (around 2% for both lines), Figure 7 shows that these loads are less scattered when the free movement model is used. This is consistent with the effect of the model on extreme loads, since reducing occurrences of extreme loads during the trip relies on homogenising the load along the consist.

References

- Ben-Akiva, M., Macke, P. P. & Hsu, P. S. (1985), *Alternative methods to estimate route-level trip tables and expand on-board surveys*, number 1037.
- Daamen, W., Lee, Y.-c. & Wiggendaad, P. (2008), ‘Boarding and alighting experiments: Overview of setup and performance and some preliminary results’, *Transportation Research Record* **2042**(1), 71–81.
- Fang, J., Fujiyama, T. & Wong, H. (2019), ‘Modelling passenger distribution on metro platforms based on passengers’ choices for boarding cars’, *Transportation Planning and Technology* **42**(5), 442–458.
- Hänseler, F. S., van den Heuvel, J. P., Cats, O., Daamen, W. & Hoogendoorn, S. P. (2020), ‘A passenger-pedestrian model to assess platform and train usage from automated data’, *Transportation research part A: policy and practice* **132**, 948–968.
- Jenelius, E. (2019), ‘Data-driven metro train crowding prediction based on real-time load data’, *IEEE Transactions on Intelligent Transportation Systems* **21**(6), 2254–2265.
- Khomchuk, P., Tuladhar, S. R. & Sivananthan, S. (2018), ‘Predicting passenger loading level on a train car: A bayesian approach’, *arXiv preprint arXiv:1808.06962*.
- Kim, H., Kwon, S., Wu, S. K. & Sohn, K. (2014), ‘Why do passengers choose a specific car of a metro train during the morning peak hours?’, *Transportation research part A: policy and practice* **61**, 249–258.
- Krstanoski, N. (2014), ‘Modelling passenger distribution on metro station platform’, *International Journal for Traffic & Transport Engineering* **4**(4).
- Kuusinen, J.-M., Sorsa, J. & Siikonen, M.-L. (2015), ‘The elevator trip origin-destination matrix estimation problem’, *Transportation Science* **49**(3), 559–576.

- Peftitsi, S., Jenelius, E. & Cats, O. (2020), 'Determinants of passengers' metro car choice revealed through automated data sources: a stockholm case study', *Transportmetrica A: Transport Science* **16**(3), 529–549.
- Schöttl, J., Seitz, M. J. & Köster, G. (2019), 'Investigating the randomness of passengers' seating behavior in suburban trains', *Entropy* **21**(6), 600.
- Seriani, S. & Fujiyama, T. (2019), 'Modelling the distribution of passengers waiting to board the train at metro stations', *Journal of Rail Transport Planning & Management* **11**.
- Tang, T.-Q., Shao, Y.-X. & Chen, L. (2017), 'Modeling pedestrian movement at the hall of high-speed railway station during the check-in process', *Physica A: Statistical Mechanics and its Applications* **467**, 157–166.
- Tirachini, A. & Cats, O. (2020), 'Covid-19 and public transportation: Current assessment, prospects, and research needs', *Journal of Public Transportation* **22**(1), 1.
- Van Zuylen, H. J. & Willumsen, L. G. (1980), 'The most likely trip matrix estimated from traffic counts', *Transportation Research Part B: Methodological* **14**(3), 281–293.
- Zhang, Y., Jenelius, E. & Kottenhoff, K. (2017), 'Impact of real-time crowding information: a stockholm metro pilot study', *Public Transport* **9**(3), 483–499.

Appendix

Table 5: First rows of train dataset on line H

Consist	Train Number	Date	b ₁	a ₁	b ₁	a ₂	b ₃	a ₃	b ₄	a ₄	b ₅	a ₅	b ₆	a ₆	b ₇	a ₇	b ₈	a ₈	Way
1	123400	05/03/2019	28	51	29	32	27	29	34	39	50	54	36	30	73	46	49	45	0
1	123409	22/02/2019	11	13	3	9	11	19	22	30	18	17	34	24	27	50	77	41	1
1	123419	16/12/2018	2	2	1	0	8	8	6	6	4	5	14	13	13	20	16	10	1
1	123499	14/06/2019	15	24	25	20	43	45	32	23	30	46	52	39	51	64	97	84	1
1	123543	16/08/2019	17	24	16	17	20	20	12	28	41	26	21	29	29	42	81	51	1
2	123400	06/12/2018	49	42	31	34	31	22	38	41	28	30	24	29	34	35	20	22	0
2	123402	15/05/2019	28	32	39	43	26	23	20	21	34	27	23	25	26	27	8	6	0
2	123424	18/03/2019	90	63	45	57	47	45	44	54	52	39	53	57	39	60	33	28	0
2	123490	28/12/2018	93	69	53	58	50	47	27	39	27	25	16	17	26	34	14	17	0