Modeling and prediction of passenger boardings at transfer nodes of public transport networks

Federico Gallo*[1] and Francesco Corman[2]

[1]Dr., Department of Mechanical, Energy, Management and Transportation Engineering, University of Genoa, Italy
[2]Prof. Dr., Institute for Transport Planning and Systems, ETH Zurich, Switzerland

Short summary

Public transport transfer nodes represent challenging elements of a network when it comes to predicting vehicle occupancies, due to the high number of boarding and alighting passengers and the related uncertainties that are associated to these stops. For this reason, understanding the passenger dynamics at these stops can give a contribute in enhancing public transport occupancy predictions.
This work presents a methodology to model the passenger boardings at transfer stops, which considers both the passengers that reach the stop on their own from the external (e.g., by foot or bicycle) and passengers that reach the stop on board a public transport vehicle. The proposed methodology is applied to a case study related to one of the major transfer stops of the city of Zurich. Results could be used to improve boarding predictions and to enhance the understanding of the passenger transfers at such stops.
Keywords: Public transport; Passenger boarding predictions; APC data.

1 Introduction

The crowding of public transport (PT) vehicles significantly affects the attractiveness and the quality perceived by the users, as well as the efficiency of PT operations. In particular, high crowding levels can cause reliability problems (Drabicki et al. (2022)) and represent a major source of discomfort for the passengers, who have been found willing to choose less crowded trips even if this implies longer travel times, and even to pay more for traveling in less crowded services (Shelat et al. (2022)). Thus, a prior knowledge about PT crowding can help passengers in planning and making their trip (contributing in particular to a decrease of high crowding levels, Drabicki et al. (2021)), as well as the PT provider to better manage the available resources and to introduce additional services when and if these are needed.
Past studies have predicted PT occupancy using different data sources, like Smart Card data (Zhang et al. (2017)), text from the Web and Social Networks (Pereira et al. (2015); Rodrigues et al. (2017)) and, more recently, automatic vehicle location (AVL) and automatic passenger counting (APC) data, which gained importance thanks to their potential in providing real-time information (Jenelius (2020); Wood et al. (2023); Roncoli et al. (2023); Hoppe et al. (2023)). In addition, different methods have been tested and used, such as time series-based methods (Zhang et al. (2017)), neural networks (Rodrigues et al. (2017)), tree-based algorithms (Ding et al. (2016)), and regression (Jenelius (2020); Wood et al. (2023)).
That works mainly focused on predicting the occupancy for single lines, without considering the effect of multiple lines in the network. We started addressing this topic in (Gallo, Sacco, & Corman (2023)), which focused in particular on corridors served by multiple overlapping lines. In this paper, we focus specifically on the transfer nodes (i.e., stops crossed by multiple lines, where passengers can make transfers) of a network and on the effect other lines serving the node can have on a given line. As a matter of fact, transfer nodes are usually the key elements and the most crowded stops of a PT network; in real-time predictions, due to the related uncertainties associated to such nodes (and thus to the resulting boardings), they often act as a *barrier* that make it difficult to predict the occupancy after them. Thus, understanding the passenger dynamics and transfers at such nodes can increase the quality of the predictions made in all the PT network.

In this paper we propose a passenger boarding model capable of explicitly considering the effect, on the passenger boarding predictions, that other lines can have on a given one at transfer stops, and we apply it on one of the major transfer nodes of the PT network of the Swiss city of Zurich.

2 Methods and data

Data input

The considered dataset is related to the period from 10/12/2017 to 09/12/2018 and the city of Zurich, and includes four data types. First, the number of passengers getting on (boardings) and off (alightings) a certain tram vehicle at a certain stop, day and time (collected by on-board sensors, and available only in about 30% of the rolling stock). Second, the scheduled and the real departure times at each stop of a trip (available on the Zurich Open Data website https://data.stadt-zuerich.ch). Third, data related to the holidays in Zurich (collected from the website feiertagskalender.ch). Last, the rainfall data (amount of rain accumulated during a 10-minute measurement interval), which was collected from the website of the ETHZ Institute for Atmospheric and Climate Science (https://iac.ethz.ch/the-institute/weather-stations.html).

Boardings model

The aim is to model the number of passengers who board a vehicle providing service s of a line l at a PT transfer stop a. To this end, the general idea is that at such kind of PT stops, passenger arrivals can be divided into two groups:

- passengers arriving at the stop from the 'external' (e.g., by walking or cycling). For these passengers, assuming a Poisson distribution, we consider an arrival rate $\lambda_l^a(t)$;
- passengers arriving at the stop with PT (i.e., on board a PT vehicle), who will use the stop as transfer between two lines. These passengers do not arrive at the stop randomly, but they arrive at the time the vehicle they are on reaches the stop.

The following assumptions hold:

- the passenger arrival rate $\lambda_l^a(t)$ and the number of transfer passengers $p_{l,v}^a(t)$ is constant over the time period t;
- passengers always manage to board the first vehicle that arrives (i.e., there are not denied boardings) and are not influenced by the crowding of the vehicle itself.

Table 1: Summary of notation

Variable	Description
$\mathcal{V}^a$	set of lines whose path crosses stop a
$\lambda_l^a(t)$	arrival rate, at stop a and in time period t, of passengers of line l
$H_{s,l}^a$	headway of line l, at stop a, for the service s (i.e., time from the last departure of a vehicle of line l)
$t_{s,l}^a$	arrival time, at stop a, of service s of line l
$p_{l,v,d}^a(t)$	number of passengers who transfer from line v (which is traveling towards direction d) to line l at the stop a in the time period t
$x_{s,l,v,d}^a$	number of vehicles of line v and direction d arrived at stop a in the time interval $(t_{s-1,l}^a = t_{s,l}^a - H_{s,l}^a, t_{s,l}^a)$

Given the previous considerations, we propose the following formula to estimate the number of passengers boarding at stop a the PT vehicle providing service s on line l:

$$ON^a_{s,l} = \lambda^a_l(t) \cdot H^a_{s,l} + \sum_{v \in \mathcal{V}^a - l} \sum_{d \in \{1,2\}} p^a_{l,v,d}(t) \cdot x^a_{s,l,v,d}. \tag{1}$$

The first term of Equ. (1) represents the number of passengers, willing to board the vehicle, who arrived from the external; this fraction is function of the real line headway $H^a_{s,l}$, as the more service s is delayed, the more passengers have time to reach the stop. The second term of Equ. (1) represents the fraction of passengers who reach stop a with a PT vehicle. More specifically, it is the summation, over all the lines $\mathcal{V}^a$ operating at stop a in the two possible directions of travel 1 and 2, of the times $x^a_{s,l,v,d}$ a vehicle of line v (in direction d) arrived at stop a after the last departure of line l (i.e., in the time interval $(t^a_{s-1,l} = t^a_{s,l} - H^a_{s,l}, t^a_{s,l})$) multiplied by the number of passengers $p^a_{l,v,d}(t)$ willing to transfer, at stop a, from line v (traveling in direction d) to line l.

Equ. 1 can be written for each tuple $\left\{ON^a_{s,l}, H^a_{s,l}, x^a_{s,l,v,d}, v \in \mathcal{V}^a, d \in \{1,2\}\right\}$ of observed values. Thus, thanks to the assumption of constant arrival times and passenger transfers in a certain time interval t (e.g., an hour) and stop a, the problem can be formalized as

$$\Lambda^* = \text{argmin}\,(Y - X\Lambda)^2 \tag{2}$$

where $\Lambda := [\lambda^a_l(t) \quad p^a_{l,v,d}(t), v \in \mathcal{V}^a, d \in \{1,2\}]^T$ is the vector containing the arrival rates and transfers of the passengers at the stop to be estimated, X is the matrix collecting all the observed values of headways and line arrivals, and Y is the vector of the related observed boardings.

3 Results and discussion

Case study and descriptive analysis

We apply the methodology described in Sec. 2 to the PT network of the Swiss city of Zurich and, in particular, to the stop of Zürich, Bucheggplatz (Fig. 1). This is one of the most important transfer stops of the Zurich network, with many lines crossing it.

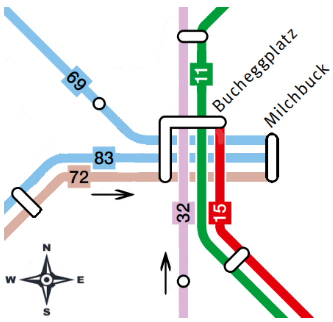

Figure 1: Scheme of the considered case study. The arrows indicate direction 1.

We select line 11 (North direction) as the line for which predicting the boardings: this line is operated by trams and is one of the most crowded in Zurich. Lines 32, 69, 72 and 83 are operated by buses or trolleybuses; since they cross the stop in two different directions, we distinguish between direction 1, which is the one indicated by the arrow (towards North-East), and direction 2, which is the opposite one (towards South-West). Apart from line 69, these lines are characterized by long trips crossing a large portion of the city. Line 15 is operated by trams, and is the only one that share a portion of its path with line 11; we assume no transfer of passengers between line 15 and line 11 at the stop of Zürich, Bucheggplatz because, due to the layout of the stop, passengers would have to walk to transfer between the two lines, whereas they can easily transfer between them in the previous stops, which are overlapped. Additionally, we assume that the other direction of line 11 does not interfere with the considered one, i.e., no passengers arrive with line 11 at the stop of Zürich, Bucheggplatz from North, and return back immediately by taking line 11 in the opposite direction (towards North). Consequently, we consider 9 independent variables: the real headway of line 11 ($H^a_{s,11}$) and the number of arrivals ($x^a_{s,11,v}$) of the other four lines considered, each of them with the two possible directions. For simplicity, the stop and time indexes will be dropped in the following notation.

As a preliminary analysis, we evaluate the correlation between the independent variables of the model in Equ. (1). Fig. 2 shows the resulting correlation coefficients (computed with the Pearson method); the color of each rectangle is proportional to the correlation level. Non significant values (p-value greater than 0.05) are not shown.

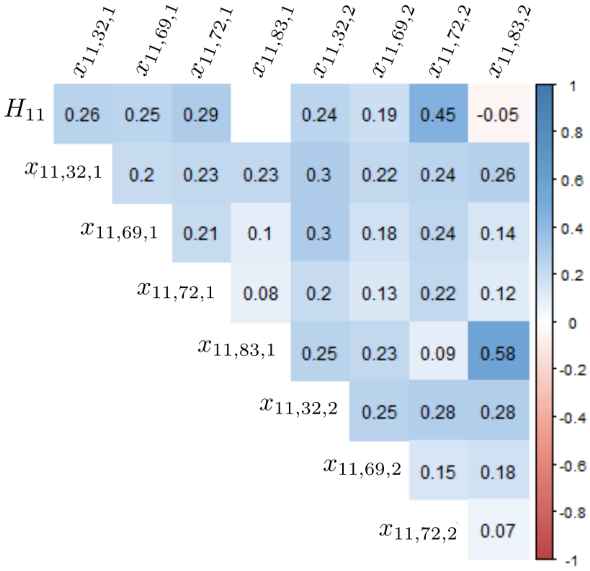

Figure 2: Correlation between the headway of line 11 and the number of arrivals of the other lines.

We highlight a low positive correlation between the headway of line 11 and the number of arrivals of the other lines, except for line 83. We can explain this by noting that the higher the headway of line 11, the more time the other lines have to arrive at the stop. In addition, as shown in Fig. 3b, the considered PT network is highly regular. In particular, the third quartile of the delay of all the arrivals at the addressed stop is $66s$, with an average delay of $24s$ for the tram line 11 and an average delay of the other lines of $51s$. Small delays can favor the above-mentioned correlation since the PT lines are more likely to arrive at the stop always in the same time interval $(t^a_{s-1,l} = t^a_{s,l} - H^a_{s,l}, t^a_{s,l})$. In addition, rare delays can favor correlation because, for instance, when the predicted line is delayed, the other lines are likely not, and this results in having more arrivals of other lines for an higher headway H_{11}.

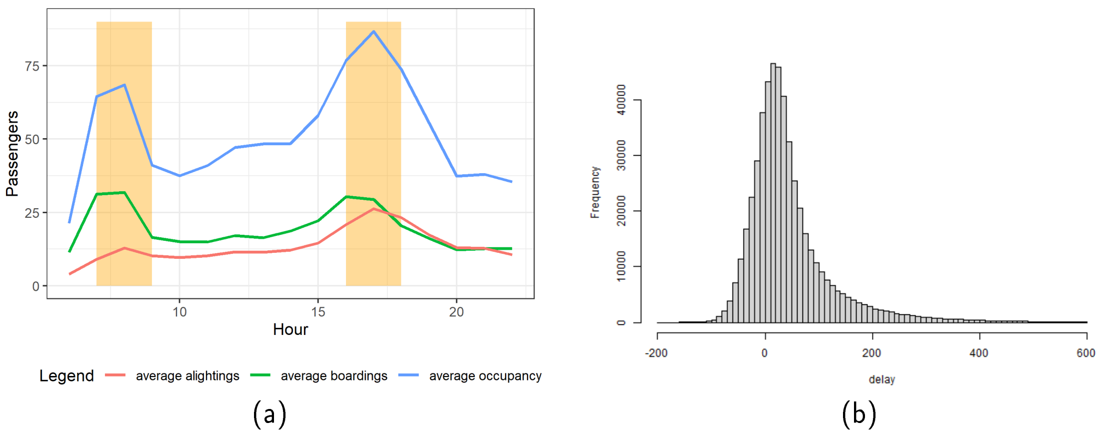

Figure 3: Average passenger boardings, alightings and occupancy per hour (of a non rainy working day) for the tram line 11 (a), and histogram of the delays of all the PT arrivals (b) at the stop of Zürich, Bucheggplatz.

In addition, we observe (Fig. 3a) two daily peaks of demand, one in the morning (7.00-9.00) and one in the evening (16.00-18.00), during which the average boardings have very similar values. For this reason, we will restrict the following analyses to such intervals, and we will consider a constant passenger arrival rate λ for each interval.

Model and prediction results

We solved the problem in Equ. (2) for the two considered time intervals (morning peak and evening peak). To eliminate the effect of the weather conditions (which can have an impact on the PT usage, as shown in Gallo, Spanninger, & Corman (2023)) and of the day type, we focused on the non rainy working days (from Monday to Friday, except for holidays). The resulting dataset consists of 435 observations for the morning peak time and 433 for the evening peak time.
Tab. 2 shows the results, in terms of passenger arrival rate (pax/s) from the external (second column), passenger transfers between one of the four considered lines and line 11 for each direction of travel (third to tenth column), multiple R^2 and global p-value (last two columns). Significant values (p-value < 0.05) are in bold.

Table 2: Model results

Time int.	λ_{11}	$p_{11,69,1}$	$p_{11,32,1}$	$p_{11,72,1}$	$p_{11,83,1}$	$p_{11,69,2}$	$p_{11,32,2}$	$p_{11,72,2}$	$p_{11,83,2}$	mult. R^2	p-value
7-9	**0.039**	1.7	**2.8**	**4.5**	**2.8**	-0.23	-1.2	2.1	1.2	0.92	9.6e-138
16-18	**0.044**	1.7	**2.8**	**2.6**	**2.6**	**-3.8**	0.5	0.19	**3.0**	0.93	4.8e-146

Considering the morning peak, all the variables related to the vehicles arriving at the stop from North-East result to be not significant (Tab. 2). Regarding lines 69, 72 and 83 (which arrive at the stop of Zürich, Bucheggplatz from the stop of Zürich, Milchbuck), we can explain this by noting that passengers coming from Zürich, Milchbuck do not have convenience in transferring from these lines to line 11 at Zürich, Bucheggplatz, because they have other direct connections (not shown in the figure) from Zürich, Milchbuck towards North. Regarding line 32, we can explain its associated low significance by noting that it reaches Zürich, Bucheggplatz from North, so passengers are less likely to continue their journey towards the same direction from which they came. In addition, we observe that even line 69 from West is not significant: we can explain this by noting that in the morning line 69 is mainly used in the direction towards West, since it connects two major transfer points of the network (Zürich, Milchbuck and Zürich, Bucheggplatz) with an hospital and the university campus). Similar considerations can be done for the evening peak.
Focusing on the significant values, according to the results we can see that the arrival rate of the passengers from the external is slightly higher in the evening peak, where a passenger arrives, on average, every 22.7s. Lines 32, 72, and 83 arriving from South-West are the most significant, in particular line 72 which, according to the model, takes 4.5 passengers (for each arrival) who transfer to line 11. We can explain this by noting that these lines reach the stop of Zürich, Bucheggplatz from opposite directions with respect to line 11, and after crossing large portions of the city.
The last step of this work is using the model in Equ. (1) for predicting the boardings at the stop. In particular, we are interested in looking whether modeling the passenger transfers allows to enhance the predictions made, with respect to modeling only the passenger arrivals at the considered stop. To this end, we compare three prediction models, all based on Equ. (1), differing in the independent variables used. The first model (only line headway) uses only the real headway (H_{11}) to predict the boardings. In other words, it assumes passengers arrive at the stop randomly, without distinguishing between those who arrive from the external on their own (e.g., by walking or by bicycle) and those who arrive on board another PT vehicle. In this model, λ_{11} represents the arrival rate of all the passengers arriving at the stop (without considering how they arrived there). The second model (significant lines) uses both the real headway and the passenger transfers from the lines that resulted significant in the significance test (see Tab. 2). The third model (all lines) uses both the real headway and the passenger transfers from all the lines at the stop. In addition, we consider as baseline model the average boardings in the considered time interval. Two thirds of the dataset was used for training the model, and the remaining for testing.
Tab. 3 shows the results in terms of Root Mean Square Error (RMSE, pax).
Firstly, the three proposed models perform significantly better than the baseline one. Secondly, the best performance is achieved by the significant lines model, whose RMSE is 5% lower than the one of the only line headway model. We can then see that, in the considered case study, modeling both passenger transfers and passenger arrivals slightly improve the quality of the predictions made.

Table 3: Prediction results (RMSE)

Model	**7.00-9.00**	**16.00-18.00**
Average	14.0	11.6
Only line headway	11.9	8.0
Significant lines	11.3	7.6
All lines	11.5	7.9

4 Conclusions

This work presents a methodology to model the passenger boardings at transfer stops, which considers both the passengers who reach the stop on their own from the external (e.g., by foot or bicycle) and passengers who reach the stop on board a public transport vehicle. The former category of passengers is assumed to arrive at the stop randomly, whereas the latter one is assumed to arrive at the same time of the vehicle they are on. Differently from the passenger arrival times, the PT vehicle arrivals can be more easily predicted (Buchel & Corman (2022)).
The proposed methodology is then applied to the stop of Zürich, Bucheggplatz, which is a major transfer stop of the PT network of the Swiss city of Zurich. Results from the predictions made show an improvement in the RMSE of 5% in the case the passenger transfers are considered, with respect to the case in which predictions are made by considering only random passenger arrivals. Even if the correlation between the independent variables is low, it can partly explain the obtained improvement. Therefore, in the future we plan to apply the proposed method in less regular networks, where the effect of delayed arrivals could be more relevant. In addition, results could be used to understand and predict the passenger transfers at the stops.

Acknowledgements

The authors would like to thank Verkehrsbetriebe Zürich (VBZ) for providing the passenger count data used in this work.

References

Buchel, B., & Corman, F. (2022). What Do We Know When? Modeling Predictability of Transit Operations. *IEEE Transactions on Intelligent Transportation Systems*, 1–12.

Ding, C., Wang, D., Ma, X., & Li, H. (2016). Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. *Sustainability*, *8*(11), 1100.

Drabicki, A., Kucharski, R., & Cats, O. (2022). Mitigating bus bunching with real-time crowding information. *Transportation*, *50*(3), 1003–1030.

Drabicki, A., Kucharski, R., Cats, O., & Szarata, A. (2021). Modelling the effects of real-time crowding information in urban public transport systems. *Transportmetrica A: Transport Science*, *17*(4), 675-713.

Gallo, F., Sacco, N., & Corman, F. (2023). Network-wide public transport occupancy prediction framework with multiple line interactions. *IEEE Open Journal of Intelligent Transportation Systems*, *4*, 815–832.

Gallo, F., Spanninger, T., & Corman, F. (2023). Weather effects on the public transport ridership of the city of Zurich at a stop-to-stop level. In *2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Nice, France, June 14-16, 2023* (p. 1-7).

Hoppe, J., Schwinger, F., Haeger, H., Wernz, J., & Jarke, M. (2023). Improving the Prediction of Passenger Numbers in Public Transit Networks by Combining Short-term Forecasts with Real-time Occupancy Data. *IEEE Open Journal of Intelligent Transportation Systems*, *4*, 153–174.

Jenelius, E. (2020). Personalized predictive public transport crowding information with automated data sources. *Transportation Research Part C: Emerging Technologies*, *117*, 102647.

Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2015). Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios. *Journal of Intelligent Transportation Systems*, *19*(3), 273–288.

Rodrigues, F., Borysov, S. S., Ribeiro, B., & Pereira, F. C. (2017). A Bayesian Additive Model for Understanding Public Transport Usage in Special Events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2113–2126.

Roncoli, C., Chandakas, E., & Kaparias, I. (2023). Estimating on-board passenger comfort in public transport vehicles using incomplete automatic passenger counting data. *Transportation Research Part C: Emerging Technologies*, *146*, 103963.

Shelat, S., van de Wiel, T., Molin, E., van Lint, J. W. C., & Cats, O. (2022). Analysing the impact of COVID-19 risk perceptions on route choice behaviour in train networks. *PLOS ONE*, *17*(3), e0264805.

Wood, J., Yu, Z., & Gayah, V. V. (2023). Development and evaluation of frameworks for real-time bus passenger occupancy prediction. *International Journal of Transportation Science and Technology*, *12*(2), 399–413.

Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., ... Li, Z. (2017). A Real-Time Passenger Flow Estimation and Prediction Method for Urban Bus Transit Systems. *IEEE Transactions on Intelligent Transportation Systems*, *18*(11), 3168–3178.