Modeling passenger positioning on platforms

Justine Lebrun*[1], Etienne Côme[2], Rémi Coulaud[3], and Christine Keribin[4]

[1]PhD Student, Laboratoire de mathématiques d'Orsay & Transilien SNCF Voyageurs, Université Paris-Saclay, France
[2]Researcher, Université Gustave Eiffel, Laboratoire GRETTIA, France
[3]Head of Datalab' Mass Transit, Transilien SNCF Voyageurs, France
[4]Professor, Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, France

Short summary

Passenger flow management is a major issue for Transilien SNCF Voyageurs, which has to synchronise millions of passengers with thousands of trains daily. However, on the platforms, passengers often tend to congregate in specific areas, creating congestion. These imbalances not only disrupt passenger comfort but also affect their safety and train punctuality. Passenger positioning strategies depend mainly on the configuration of the platforms at the departure and arrival stations. In this study, we rely on these elements to propose a model that can be generalized to other stations. Our approach is based on discrete choice models, which are particularly well-suited to model this type of decision-making problem. This also allows us to quantify the impact of each factor on positioning choices.
Keywords: Discrete choice model, passenger counting data, passenger positioning, positioning strategy.

1 Introduction

In the Île-de-France region, Transilien SNCF Voyageurs operates in a densely populated area, transporting 3.4 million passengers daily. In this context, the goal is to ensure smooth traffic flow and avoid delays while ensuring passenger comfort. Transilien SNCF Voyageurs has therefore installed screens on the platforms to display, in real time, the number of passengers in each zone of the train Baietto et al. (2026). A screen example is shown in Figure 1. This system allows travelers to choose a less crowded car to improve their comfort and the fluidity of exchanges (Peftitsi et al. (2022), Zhang et al. (2017)).

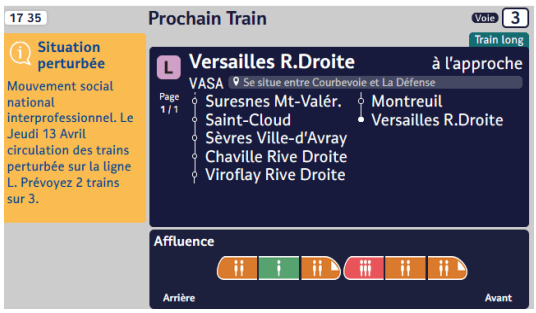


Figure 1: Example of a passenger information screen showing crowding levels by zone on Line L of the Transilien network.

Passengers do not choose their position on the platform at random but adopt various strategies (Kim et al., 2014) : minimizing the distance to the exit at the arrival station, minimizing the distance to the entrance at the departure station, and maximizing their comfort. Thus, the layout of the platforms, and in particular the location of the access points, appears to be a key indicator of passenger positioning ((Krstanoski, 2014), (Peftitsi et al., 2020)). As a result, some areas of the platforms are more attractive and can become overcrowded, which affects both passenger comfort

and safety, as well as production, with longer exchange times that can cause delays.

For example, Figure 2 shows the average proportion of passengers boarding at each door at Asnières-sur-Seine station for trains traveling from the suburbs to Paris. The most used areas are close to the platform accesses, confirming their influence on passenger behavior. We also observe that the front of the train is the most attractive zone. This can be explained by the fact that most passengers alight at Paris Saint-Lazare station, where the main exit is located at the front of the train. This highlights the influence of the destination station on passenger distribution.

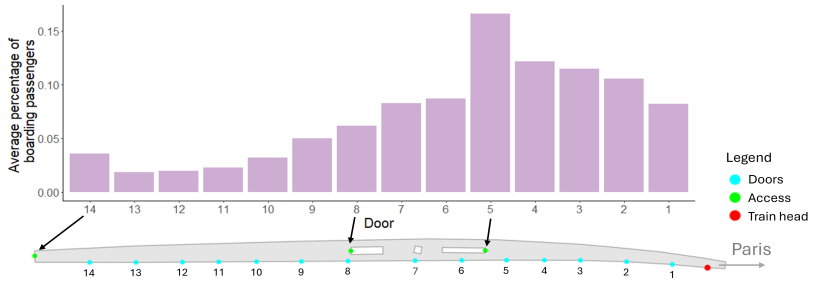


Figure 2: Average proportions of the number of passengers boarding per door at Asnières-sur-Seine station for trains traveling from the suburbs to Paris during the morning peak hours (top) represented in parallel with the platform (bottom).

Thus, the distribution of passengers on platforms is not always uniform and is not yet fully studied. The objective of this study is to model the positioning of passengers on platforms, based on the influencing factors identified in the literature. This involves, in particular, assessing and quantifying the impact of these factors on the choices made by passengers.
Since the exact distribution of passengers along the platform cannot be directly observed, we focus on the number of passengers boarding at each door of trains. We aim to develop a generalizable model that can be adapted to each station, knowing only its configuration. This approach can help Transilien Voyageurs SNCF optimize station layout and design, contributing to smoother passenger flow while improving safety and passenger comfort.

We start by outlining the scope of the study and the data used, before discussing the methodology employed. Next, we analyze and interpret our results while providing an assessment of the performance of our model. Finally, we conclude by summarizing the main contributions of this work and outlining our future research perspective.

2 Methodology

Perimeter of the study and data

We study 10 stations on a branch of the L Transilien line between Versailles Rive Droite and Paris Saint-Lazare (Figure 3), for the period from January 2022 to June 2024. We focus our study on contexts where we observed significant passenger traffic. We are therefore studying the morning peak period (between 6 a.m. and 9 a.m.) on weekdays and outside school holidays for trains traveling from the suburbs to Paris. Finally, we exclude terminal stations from our study due to their specific operating characteristics. Boarding volumes are high at these stations compared to intermediate stations due to longer train dwell times. In the end, we have a total of approximately 21000 stops.

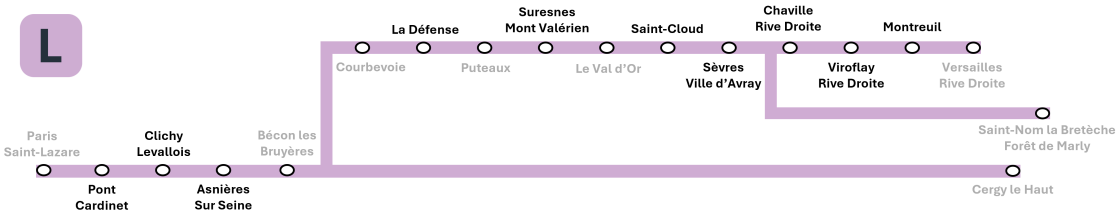


Figure 3: Scheme of the Transilien line L. Studied stations are indicated in **bold**.

In this work, we use two main sources of data:

- **Automatic passenger counting data.** This data is collected at each train door using infrared sensors. It provides information on the number of passengers boarding and alighting at each stop and for each door of a train. In our case, the trains studied have 14 doors. It should be noted that approximately half of the Transilien lines are equipped with counting equipment.
- **Platform geography data.** This data is available on *OpenStreetMap*, but is sometimes of low quality. We therefore carried out field visits to verify it. In our study, we use in particular, the geometry of the platforms, the location of access points, and the position of train stop signals. These signals correspond to the theoretical locations where the train is supposed to stop at the station. In reality, trains do not always stop at the exact same place, but we assume that this variability is negligible.

Method

We model the number of passengers boarding at stop i and door $p \in \{1, \dots, P\}$ by the random variable $M_{i,p}$. Its realization will be indicated by $m_{i,p}$, while the total number of passengers boarding at a stop will be denoted by m_i. The stop i is uniquely defined by a train $k \in \{1, \dots, K\}$, a station $s \in \{1, \dots, S\}$, and a date $t \in \{1, \dots, T\}$. The set of train stops at station s will be denoted by I_s.
Our modeling of $M_{i,p}$ is based on the geographical characteristics of the stations. More specifically, we use the distances $d^s_{p,a}$ between the doors p of the train and the access points a of station $s \in A_s$. The minimum distance between door p and the nearest access point is then denoted by $d^s_{p,min} = \min_a d^s_{p,a}$. In concrete terms, these distances refer to the walking distance passengers must complete to reach their boarding gate.

We use a discrete choice model (Ben-Akiva & Bierlaire, 1999) because it is appropriate for modelling the behaviour of an individual faced with multiple options. In our context, these options correspond to the different train doors. With a multinomial logit discrete choice model, we can define the probability that a passenger chooses the door p given the access (a_{dep}, a_{arr}) he used at his departure station s_{dep} and arrival station s_{arr} as:

$$p(porte = p \mid s_{dep}, s_{arr}, a_{dep}, a_{arr}) = \frac{\exp\left(\beta_1 d^{s_{dep}}_{p,a_{dep}} + \beta_2 d^{s_{arr}}_{p,a_{arr}}\right)}{\sum_{p'=1}^{P} \exp\left(\beta_1 d^{s_{dep}}_{p',a_{dep}} + \beta_2 d^{s_{arr}}_{p',a_{arr}}\right)}. \tag{1}$$

In reality, we do not know which access points were used or the arrival station of the passengers.

With regard to the arrival station, the study of passenger origins and destinations on the studied line identifies two major destination stations: La Défense, where access points are distributed along the platform, which, in principle, has little impact on passenger positioning; and Paris Saint-Lazare, whose main entrance is located at the front of the platform. Given these elements, we make an initial assumption that seems reasonable: we assume that all passengers alight at the terminus, that is, at Paris Saint-Lazare, and that they use the main entrance located at the front of the platform. We will then denote d^{PSL}_p the distance between a door p and the main entrance of Paris Saint-Lazare station.
We study trains with two non-connected carriages, but it is possible to move around inside each carriage. Thus, to reach the front of the train, passengers must be in the first carriage and therefore

board through one of the doors numbered 1 through 7. In the data we are analyzing, we often observe peaks in passenger boarding at the first accessible door of the first carriage. Therefore, we add a new explanatory variable to represent this phenomenon: $f^{g_{dep}}_{p,a_{dep}}$, which is equal to 1 for the door of carriage 1 closest to entrance a_{dep} and 0 otherwise.
Finally, regarding the departure station, we estimate the proportions $\pi_a^{s_{dep}}$ of use of each entrance a at station s, which become new parameters.
Thus, the observed boarding volumes are modeled by a mixture of multinomial distributions of the form:

$$M_{i,.} \sim \sum_{a=1}^{A_s} \pi_a^s \mathcal{M}(\boldsymbol{\alpha}_a^s, m_i), \text{ avec } \boldsymbol{\alpha}_a^s = \left[\frac{\exp(\beta_1 d^s_{p,a} + \beta_2 d_p^{PSL} + \beta_3 f_p^s))}{\sum_{p'=1}^{P} \exp(\beta_1 d^s_{p,a} + \beta_2 d_{p'}^{PSL} + \beta_3 f_{p'}^s))} \right]_{p \in \{1,\dots,P\}} . \quad (2)$$

The model parameters are therefore: a vector of proportions per station ($\boldsymbol{\pi}^s = [\pi_a^s]_{a \in \{1,\dots,A_s\}}$) and the parameters of the discrete choice model ($\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$). These parameters are estimated by maximum likelihood using an EM algorithm:

We start by building an initial discrete choice model based on stations with only one access point (in this case, we know that $\pi_a^s = 1$). This applies to Montreuil, Chaville Rive Droite, and Pont Cardinet stations. We then perform an iterative optimization that alternates between the following two steps:
Step 1: Estimated utilization rates for entrances at stations with more than one entrance.
Step 2: Re-estimation of the discrete choice model using stations with a single entrance as well as stations with multiple entrances, weighted by the estimated entrance weights ($\pi_a^{s_{dep}}$).

3 Results and discussion

We recall that an observation corresponds to a train stop at a station, and that we have a total of 21,000 observations in our dataset. We train our model on a training set, which accounts for 80% of the total dataset, and then test its performance on the remaining 20% of the data. The data is split randomly and uniformly across the entire study period.

Analysis of traveler positioning

We begin by explaining how we construct our final prediction using the forecasts of each access point.

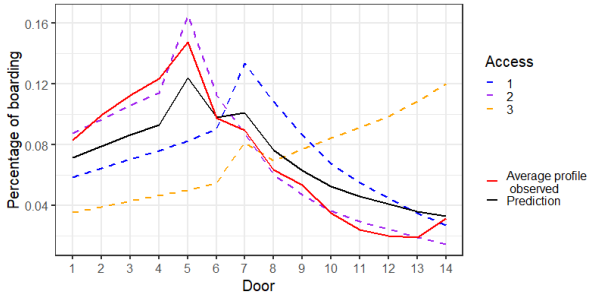


Figure 4: Percentage of boarding per door. Prediction at Asnières-sur-Seine station is shown in black; predictions for each access 1, 2 and 3 are shown in dashed lines in blue, purple and orange, respectively; the observed average profile is shown in red.

Figure 4 shows an example illustrating our method for the Asnières-sur-Seine train station.
Our model's predictions for the three station entrances (indicated by dashed lines) vary significantly from one entrance to another, showing higher usage for the doors closest to each respective entrance. The final prediction aligns closely with the prediction for access point 2, suggesting its dominant usage. In fact, the final prediction is a weighted average derived from the model's estimated usage proportions: $\pi^s_{a_1} = 0.29$, $\pi^s_{a_2} = 0.59$, and $\pi^s_{a_3} = 0.12$. These estimations confirm that access 2 is the most frequently used.

Now that we have outlined this process, we are applying the same methodology to all stations, presenting the final predictions in 5.

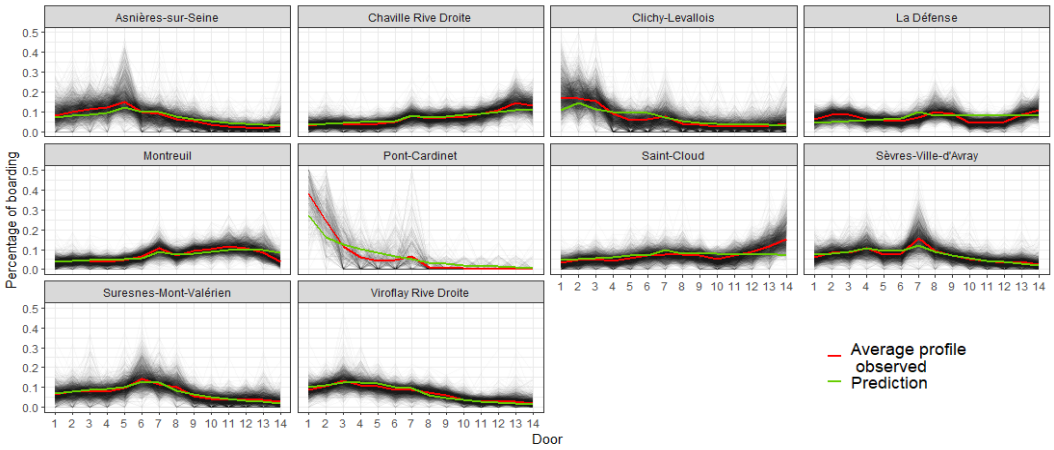


Figure 5: Distribution of passenger boardings at train doors at each station. The distributions at each stop are shown in black, the average profile is shown in red, and the prediction is shown in green.

Overall, the predictions (green in Figure 5) are fairly satisfactory and are close to the average distribution of passengers (red in Figure 5). However, we can see a lack of information at certain stations, such as La Défense or at the back of the platform in Saint-Cloud.
The analysis of station patterns reveals that Pont-Cardinet has the most atypical behavior, with the highest proportion of passengers using a single door. In fact, at this station, door 1 accounts for an average of 38% of passengers. This station also has the lowest total number of passengers boarding the train, which explains both why passengers do not spread out along the entire length of the platform and the greater variability observed in the distribution of boardings. At other stations, although there are some variations, these are usually moderate and remain close to the average.
Beyond this particular station, a similar trend is emerging at several other stations. At some stations, certain doors are more heavily used (for example, door 5 at Asnières-sur-Seine, door 7 at Sèvres-Ville-d'Avray, and door 6 at Suresnes-Mont-Valérien). This phenomenon occurs exclusively on doors of the first train car and most of the time on the first door through which passengers can board. This is consistent with our previous observations regarding passengers' preference for the front of the train due to their destination.

Model performance

We use two error metrics: root mean square error (RMSE) and mean absolute error (MAE).
We compare our model with a saturated Poisson model, which estimates a coefficient for each door-station pair. It is defined as follows:

$$M_{i,p} \sim P(\lambda_{i,p}), \quad \log(\lambda_{i,p}) = \beta_p^s + \log(m_i). \tag{3}$$

We also compare the model to a basic model, which assumes a uniform distribution of passengers along the platforms. The results are presented in Table 1.
As we might expect, the uniform model performs the worst, but it provides us with a lower bound for evaluating our results. The saturated model, on the other hand, provides the best performance, but it requires the estimation of a large number of parameters and cannot be generalized to new stations. Indeed, since we must estimate a coefficient for each door-station pair, we must re-estimate the model in order to apply it to a new station.
Our model lies between the other two in terms of performance, but it is the most flexible because it can be applied to any station while having a limited number of parameters. In addition, it is the only one able to evaluate the factors influencing passenger positioning.

Table 1: Models performance in terms of RMSE, MAE, and number of parameters.

Model	**RMSE train**	**RMSE test**	**MAE train**	**MAE test**	**Number of parameters**
Uniform model	4.18	4.24	3.13	3.17	0
Saturated model	2.67	2.69	1.95	1.96	140
Our model	3.17	3.20	2.30	2.31	20

Model interpretation

The coefficients obtained using the discrete choice model are significant and are reported in Table 2.

Table 2: Coefficient's values and their respective standard errors.

Coefficient	**Value**	**Standard error**
β_1	-0.013	3.65×10^{-5}
β_2	-0.06	2.95×10^{-5}
β_3	0.29	3.58×10^{-3}

We can see that the three coefficients associated with the distances have negative signs. This indicates that as these distances increase, the door becomes less attractive, which is consistent with what we expected. Since these two distances are in the same units, we can compare them: the distance from the departure station appears to have a greater influence on passengers' choice of boarding location than the distance to the destination. The coefficient associated with the term indicating the first door of the first carriage is positive and provides a bonus for those doors.

Discussion

In this study, we focused on modeling passenger boarding based on the departure station and assuming a fixed destination for all passengers. Passenger journeys are more complex, but we can estimate them by analyzing data from origin-destination surveys conducted among Transilien passengers.
On the other hand, boarding and alighting at stations along the same line are linked and influence each other. Therefore, it seems appropriate to study them simultaneously.
Our goal is to expand our model to capture the entire journey of passengers, from their departure station to their arrival station. We will thus aim to model both boarding and alighting at the same time. More specifically, we will model the joint probability of the pair of boarding and alighting doors, conditionally on the stations and access points at departure and arrival. In this case, given $d_{p_{dep},p_{arr}}$, the distance between door p_{dep} at the departure station and door p_{arr} at the arrival station, the probability is written as :

$$P(p_{dep}, p_{arr} | g_{dep}, a_{dep}, g_{arr}, a_{arr}) = \frac{\exp(\beta_1 d^{g_{dep}}_{p_{dep}a_{dep}} + \beta_2 d^{g_{arr}}_{p_{arr}a_{arr}} + \beta_3 d_{p_{dep},p_{arr}})}{\sum_{p'_{dep}=1,p'_{arr}=1}^{P} \exp(\beta_1 d^{g_{dep}}_{p'_{dep}a_{dep}} + \beta_2 d^{g_{arr}}_{p'_{arr}a_{arr}} + \beta_3 d_{p'_{dep},p'_{arr}})}. \tag{4}$$

We cannot directly calculate this probability because we do not observe the pairs of boarding and alighting doors.
However, we can use the data on arrivals and departures at the doors, which we observe separately. Furthermore, by estimating the usage rates of each access point at the departure and arrival stations and using origin-destination surveys, we can estimate marginal probabilities to reconstruct the desired joint probability. These results will be presented during the oral presentation.

4 Conclusions

In this work, we have succeeded in developing a model that accounts for passengers' positioning strategies based on walking distances within the station. We were able to quantify the impact of these distances and found that the distance from the departure station has the greatest influence on passengers' choices. We have obtained satisfactory results, but it is clear that there is still information we need to capture.

The next step is to develop a joint model for boarding and alighting in order to better account for positioning strategies at the starting station and the destination station. This will allow us to overcome the limitation associated with a fixed destination station.

Acknowledgements

This work is supported by the French National Association for Research and Technology (ANRT) under the CIFRE program (agreement no. 2024 / 1281) in partnership with SNCF, the Laboratoire de Mathématiques d'Orsay and the Laboratoire GRETTIA.

References

Baietto, M., Coulaud, R., Keribin, C., & Stoltz, G. (2026). Improved real-time crowding information through the modeling of passenger movements in trains with communicating coaches. *Data Science for Transportation*, *8*(1), 3.

Ben-Akiva, M., & Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science* (pp. 5–33). Springer.

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).

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.

Peftitsi, S., Jenelius, E., & Cats, O. (2022). Modeling the effect of real-time crowding information (rtci) on passenger distribution in trains. *Transportation Research Part A: Policy and Practice*, *166*, 354–368.

Zhang, Y., Jenelius, E., & Kottenhoff, K. (2017). Impact of real-time crowding information: a stockholm metro pilot study. *Public Transport*, *9*(3), 483–499.