Modeling intra-rail transfer volumes in a suburban train station using passengers counting data

Mehdi Baali*[1], Rémi Coulaud[2], and Christine Buisson[3]

[1]PhD Student, LICIT-ECO7 & SNCF Voyageurs, Université Gustave Eiffel, France
[2]Head of Datalab' Mass Transit, SNCF Voyageurs, France
[3]Senior researcher, LICIT-ECO7, Université Gustave Eiffel & ENTPE, France

Short summary

Transferring is unavoidable in urban traveling and transferring times are perceived as more uncomfortable than running times. Many studies developed transfer estimation methods. However, no work focused on the common situation of intra-rail transfer where no tap-out data is available. This work proposes a model based on entering volumes and automatic passenger counting data to estimate transfer proportions. The model is assessed in a station of the Paris suburban railway network. A satisfying fit of the data is achieved. Transfer proportions obtained with the model are compared with values inferred from the origin-destination survey. The model and survey agree for cases with no alternative to get to the desired destination from the transfer station. Other cases show heterogeneous agreement levels.
Keywords: Railway, transfer, modeling, passenger counting data.

1 Introduction

During urban travels, users commonly need to transfer to reach their desired destination. Transfer times are seen to be more repulsive than travel times (Nielsen et al., 2021; Yap et al., 2024). Therefore, optimizing transfers is an aim for many public transport companies.
To do so, one needs to estimate transfer volumes properly. A common practice is to use the origin-destination surveys. However, surveys are expensive and operators cannot afford to do them frequently. For example, such surveys in the Paris suburban railway network are done every four years. This makes the transfer volume estimate quite outdated. Surveys were widely discussed in the literature (Simon & Furth, 1985).
The more and more common availability of data helps to better estimate transfer volumes. Several approaches use automatic fare collection (AFC). Some authors based their estimates on tap-in and tap-out (Yap et al., 2017). If a tap-in was registered shortly after a tap-out for the same ticket or smartcard, it will be considered a transfer. Other studies deduce transfers using tap-in and successive trips (Hofmann & O'Mahony, 2005; Gordon et al., 2018). For example, if a passenger validated a ticket (or a smartcard) in one bus and then validated the same ticket (or smartcard) in a second bus that stopped at a station served by the first bus shortly before the second bus; this passenger was considered transferring.
However, these approaches do not hold for the estimate of intra-rail transfers where no tap-in or tap-out is required. Some studies tackled this issue using tap-in at the entrance of the network and tap-out at the exit of the network, see Yap et al. (2024) for an example on the London Underground. Transfer stations were inferred using automatic vehicle location (AVL) to deduce the most probable transfer station. To our knowledge, no study tackled the estimation of intra-rail transfer volumes where no tap-out is mandatory to exit the network. Yet, it is a common situation, the networks in Paris or New York are well-known examples.
Though intra-mode transfers are generally less unpleasant than inter-mode transfers, they are very common (14% of people who took a train in the Paris region in 2023, had taken another train before). The work presented here aims to estimate transfer volumes in intra-rail contexts where no tap-out is mandatory to exit the network.
To do so, a model will be proposed based on other available data sources: automatic passenger counting (APC), transport plans, and tap-in volumes from ticket barriers.
The manuscript is structured as follows. First, the useful data pieces are introduced along with the perimeter studied. Then, the model and assumptions are presented. The model is assessed

using statistical criteria and the transfer patterns obtained are compared to the ones inferred from the origin-destination survey. A conclusion ends this manuscript.

2 Methodology

Data presentation

In the Paris suburban railway network, it is not always[1] mandatory to validate a ticket or a smartcard to exit a station. Then, the use of tap-out data cannot be generalized at the line level. Instead of tap-out data, other kinds of data will be used in this study.

Aggregated entering data For any station on the network, it is mandatory to validate a ticket or a smartcard to enter the station. Such data is aggregated in time at the hour level.

Theoretical and realized transport plan The transport plan is the set of arrival and departure times of the trains (both theoretical and realized). From a rail tag system, the actual arrival and departure times of any train at any station are measured. The theoretical arrival and departure times are fixed at the beginning of the year. Except for small adjustments and delays, a given train number k should run every day at the prescribed theoretical arrival and departure times.

Passenger counting New rolling stocks in several lines in the Paris suburban network are equipped with infrared sensors. These sensors perform automatic passenger counting (APC) for each door, for alighting people and boarding people. These APC data are then aggregated across the train (for any train number k, at any station s, and at any date d). A study from the data provider shows an absolute deviation of less than 5 people on average compared to manual counting. Notice that failures during the on-board processing lead to data losses. As a result, data from around 30% of the trains are missing, and are uniformly distributed.

Origin-destination survey Every four years, an origin-destination survey is performed across the Paris suburban network. People are approached at stations and questioned about their whole journey. The latest survey is from the spring of 2022. Here, we use the survey data to estimate the transfer proportions to be compared with our model. Survey data contains biases and inaccuracies (Simon & Furth, 1985) due to: medium response rate (around 600 people for the case study), low survey frequency, varying seasonality (in particular weather conditions during the survey)...

Case study

The station *Argenteuil* is our case study. Figure 1 illustrates the railway node of *Argenteuil* and the two lines stopping there. This station was chosen for an important daily volume of passengers (around 26,000 people boarding every day on average) and the availability of the above-mentioned data sources: entering data is representative of the volume as the station is equipped with ticket barriers, all trains stopping at *Argenteuil* are equipped with sensors that provide APC data.
This station is a transfer station between the J4 line and the J6 line in the Paris suburban network. Line J4 is omnibus between Paris and *Argenteuil* and has a unique suburban terminus. Line J6 is direct between Paris and *Argenteuil* and has multiple suburban termini.
The possible transfers in this station can be qualitatively discussed:

- Transfers inside the J4 line are rather unlikely. People who transfer inside J4 are either alighting to wait for someone and then board again in the same direction, or they have missed their stop so they board in the opposite direction. Surveys show less than 1% of people doing so.
- For transfers inside the J6 line, if the train is toward the suburbs, it is very unlikely to see people backtracking to Paris as *Argenteuil* is the first stop. If the train is toward Paris, transfers may exist between the various services, but such a transfer should appear before arriving at *Argenteuil*, making them unlikely to appear (less than 1% in the survey).

[1] It is mandatory when there are ticket barriers, but only 48% of stations have such equipment.

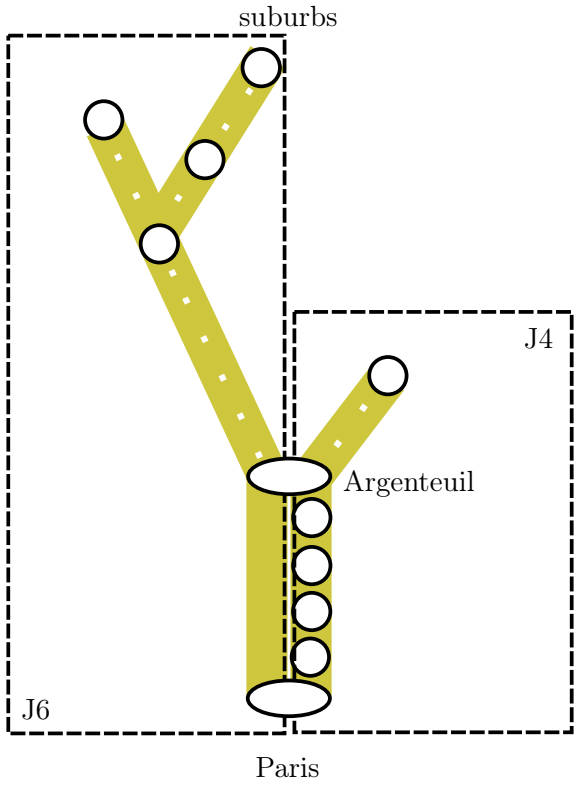

Figure 1: Scheme of the lines at Argenteuil.

From \ To	J6 toward suburbs	J6 toward Paris	J4 toward suburbs	J4 toward Paris
J6 toward suburbs	Unlikely	Very unlikely	Unlikely	Unlikely
J6 toward Paris	Unlikely	Very unlikely	Likely	Likely
J4 toward suburbs	Likely	Unlikely	Very unlikely	Unlikely
J4 toward Paris	Likely	Very unlikely	Unlikely	Very unlikely

Table 1: Qualitative assessment of the likeliness of the possible transfers at *Argenteuil*.

- A transfer from "J6 toward suburbs" to "J4 toward Paris" corresponds to backtracking. It would be faster and more comfortable to directly take a J4 train toward the suburbs between Paris and *Argenteuil*. Such a transfer is unlikely (less than 1% in the survey).
- Transfers from "J6 toward suburbs" to "J4 toward suburbs" might be possible. It corresponds to people taking the first train toward the suburbs and then transferring to a connecting station. Yet, surveys show less than 1% of people doing so.
- The other transfers are the most observed (with more than 3% in the survey) and are of interest in this work.

Table 1 summarizes the likeliness of transfers at *Argenteuil*.
The data used in this work were collected between the 1st of September and the 11th of December in 2022. Only working days outside of school holidays were used.

Method

The principle of the model is to fit the boarding volumes ($B_{k,s,d}$) from the available data to infer transfers. k is the train number, s is the station, and d is the date.
People who board a train may be divided into two categories: people who board the first train of their travel (primo-boarding people), and people who transfer (transferring people).
Primo-boarding people ($P_{k,s,d}$) enter the railway network at the station s. We assume their amount to be linear with the volume of people entering the station s during the hour when the train k stops ($P_{k,s,d} = p_{k,s,d}E_{k,s,d}$). Primo-boarding volume may depend on the headway (h), the

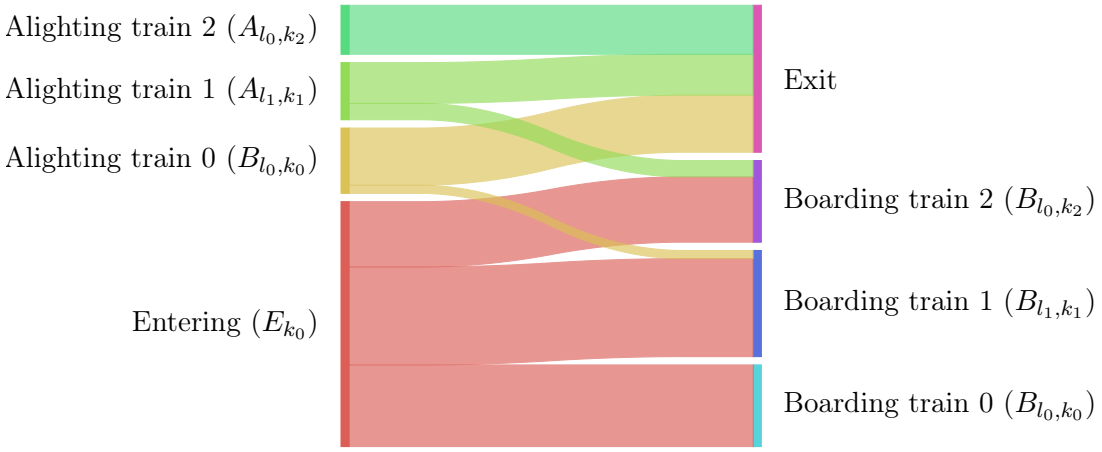

Figure 2: Sankey diagram of a simple example with three trains to illustrate the model. The exit volume is not known. Boarding volumes are assumed to be a mixture of alighting from another line and entering. The volumes represented here are only illustrative and do not correspond to any truth.

departure time (t) of the train, and other variables ($\{x^i\}$) (which may not be available). Formally, $P_{k,s,d} = p(h_{k,s,d}, t_{k,s,d}, \{x^i_{k,s,d}\})E_{k,s,d}$.
Transferring people $C_{k,s,d}$ to a given line l_0 are assumed to be a proportion of alighting people from other lines $l \neq l_0$ whose trains stopped in the station between train k and the previous train of the line l_0: $C_{k,s,d} = \sum_{l \neq l_0} \gamma_{l,s} A_{l,k,s,d}$.
In summary, we assume the following model:

$$B_{l_0,k,s,d} = p(h_{k,s,d}, t_{k,s,d}, \{x^i_{k,s,d}\})E_{k,s,d} + \sum_{l \neq l_0} \gamma_{l,s} A_{l,k,s,d}. \quad (1)$$

Figure 2 illustrates the model for a simple example with three trains.

Additional hypotheses and data selection

The total dataset is composed of 16,642 triplets (k, s, d). A line and a direction are chosen and we focus either on the morning peak hours (i.e. 6 a.m - 10 a.m) or the evening peak hours (i.e. 4 p.m - 8 p.m). For the sake of simplicity, we illustrate the data selection only for J4 trains toward the suburbs, during morning peak hours (941 triplets).
As a first approximation, we consider $p_{k,s,d}$ to have additive dependencies: $p_{k,s,d} = f_h(h_{k,s,d}) + f_t(t_{k,s,d}) + \sum_i f_i(x^i_{k,s,d})$.
Since no information is available on the x^is variables, they will be considered a noise term: $\sum_i f_i(x^i_{k,s,d})E_{k,s,d} = \epsilon$.
To simplify the dependency on departure times, we only consider on-time trains (907 triplets). Then, we can assume: $f_t(t_{k,s,d}) = \beta_{k,s}$, meaning that for a given train k at a station s, the departure time term is the same every day.
Besides, a significant correlation may be observed in some cases between the headway on one line and the alighting volumes from other lines. This would result in a mis-estimation of parameters in the fit. Thus, we only consider cases where the headway is equal to the theoretical headway. Then, we can assume: $f_h(h_{k,s,d}) = \alpha_{k,s}$ (838 triplets).
With these assumptions, the model can be written as:

$$B_{l_0,k,s,d} = \sum_j \delta^{j,k} \lambda_{j,s} E_{j,s,d} + \sum_{l \neq l_0} \gamma_{l,s} A_{l,k,s,d} + \epsilon, \quad (2)$$

where $\lambda = \alpha + \beta$, and $\delta^{j,k}$ is the Kronecker symbol (its value is 1 if $j = k$, 0 instead).
Non-negligible correlations exist between $\delta^{j,k} E_{k,s,d}$ and $A_{l,k,s,d}$ due to the effect of the hour of the day (more people overall at 7 a.m than at 9 a.m.) and the day of the week (more people overall on Tuesdays than on Fridays). Then, we divide all the variables by their average over hours and

days. It leads to :

$$B'_{l_0,k,s,d} = \sum_j \delta^{j,k} \lambda'_{k,s} E'_{k,s,d} + \sum_{l \neq l_0} \gamma_{l,s} A'_{l,k,s,d} + \epsilon', \tag{3}$$

where for any variable X, $X' = X/\widehat{X}$, $\widehat{X}$ being the average of X over hours and days.
According to the qualitative assessment proposed in Table 1, transfer terms corresponding to "Unlikely" or "Very unlikely" transfers are neglected.
Moreover, any train number appearing less than 5 times is suppressed from the dataset. The final dataset contains 413 triplets (k,s,d).
The parameters λ' and γ' are then estimated from a linear fit. We are particularly interested in the values of γ'. Indeed, $\gamma'_{l,s}$ can be interpreted as the proportion of alighting people from l line that transfer to l_0 line at station s.

Model fitting and assessment

The above-presented model is fitted using the function *OLS* from the module *statsmodels.api* in Python.
A benchmark model is set to show the relevance of adding transfer terms. This benchmark is constructed similarly to the model but without the transfer terms:

$$B^{benchmark}_{l_0,k,s,d} = \sum_j \delta^{j,k} \mu'_{k,s} E'_{k,s,d} + \epsilon'. \tag{4}$$

The approach is then assessed by comparing the model with the benchmark using AIC and BIC indicators.

$$AIC = 2k - 2\ln(L), \tag{5}$$

where k is the number of parameters and L the likelihood.

$$BIC = k\ln(n) - 2\ln(L), \tag{6}$$

where n is the number of values in the dataset.
Both of these criteria help to compare two models by penalizing the addition of variables. A lower AIC or BIC means a more appropriate model. One may interpret the AIC as saying that the model is $\exp\left(\frac{AIC_{model} - AIC_{benchmark}}{2}\right)$ times as probable as the benchmark to be appropriate. The same interpretation is possible for the BIC: the model is $\exp\left(\frac{BIC_{model} - BIC_{benchmark}}{2}\right)$ times as probable as the benchmark to be appropriate. Thus, at a 5% threshold, one needs to have a difference of more than 6 in AIC and BIC to state a significant difference.
R^2 are also computed to give an insight into the fit quality.
Another assessment is done by comparing the γ' values with the proportion of transfer estimated from the surveys.

3 Results and discussion

In this section, the relevance of the transfer terms in the model is justified using AIC and BIC criteria. Then, the parameters are discussed through the transfer proportions obtained from the survey.

Model assessment

Table 2 shows the performances for the fitting on "J4 toward suburbs". In both cases, AIC and BIC show that the model is more appropriate than the benchmark at a 5% threshold. In addition, R^2 coefficients indicate a high overall performance of the fit.
Concerning the fitting performances for "J4 toward Paris" (see Table 3) AIC and BIC show no significant difference between model and benchmark. Nevertheless, R^2 coefficients remain high, which means a good fit. This might be interpreted as either a low contribution of transfers in boarding volumes or a bad estimate due to primo-boarding people to Paris. For this second potential interpretation, primo-boarding people going toward Paris may take a J4 or J6 train according to their destination (Paris or another station) and arrival time at *Argenteuil*. As none of this information is available, the model assumes the same distribution every day between J6 and J4. This might be an oversimplification.

Model	Day-period	AIC	BIC	R^2
Model	Morning peak	3377	3457	0.905
Benchmark	Morning peak	3387	3464	0.908
Model	Evening peak	3653	3746	0.907
Benchmark	Evening peak	3666	3754	0.904

Table 2: Performances for the fit on "J4 toward suburbs". In both cases for both indicators, the model is significantly more appropriate than the benchmark at a 5% threshold.

Model	Day-period	AIC	BIC	R^2
Model	Morning peak	3864	3951	0.893
Benchmark	Morning peak	3863	3946	0.893
Model	Evening peak	3961	4060	0.877
Benchmark	Evening peak	3964	4059	0.876

Table 3: Performances for the fit on "J4 toward Paris". In both cases for both indicators, no significant difference can be observed at a 5% threshold.

A similar assumption has been made for boarding people on "J6 toward suburbs" where multiple termini exist. Boarding people will take the first train going to their destination. As we do not know the destination for any passenger, the model assumes the same distribution every day for primo-boarding and a constant proportion of transfer among alighting people from other lines. Yet, AIC and BIC show that the model is significantly more appropriate than the benchmark, R^2 still being high, see Table 4.

Transfer flow

In this section, γ' parameters are discussed by comparison with the survey data. As discussed earlier, survey data may contain biases and inaccuracies. Thus, rather than considering it as the ground truth, the order of magnitude should be the same as our model's parameters.
Figure 3 shows the proportions of transferring alighting people, estimated with the model and the survey, in each case. For transfers from "J6 toward Paris" to "J4 toward suburbs", the orders of magnitude of the proportions obtained with the model are coherent with those of the survey.
By contrast, the proportions of alighting people who transfer from "J6 toward Paris" to "J4 toward Paris" are distinct, especially in the morning peak hours. This highlights a bad estimation of this proportion by the model (already mentioned earlier). This high difference in the morning is correlated to a particularly important volume of primo-boarding people. Then, as already discussed, primo-boarding people may take any train toward Paris (either J4 or J6) in a non-trivial way, which may disturb the estimation of transfers.
Concerning alighting people who transfer to "J6 toward suburbs", model estimates also differ from survey estimates. People boarding on "J6 toward suburbs" will take a train stopping at their destination station. Then, the number of boarding varies due to unequal demand for any possible destination. This phenomenon disturbs the estimation of the model's parameters (including transfer proportions).

Model	Day-period	AIC	BIC	R^2
Model	Morning peak	1670	1724	0.933
Benchmark	Morning peak	1685	1732	0.930
Model	Evening peak	2649	2738	0.927
Benchmark	Evening peak	2687	2768	0.920

Table 4: Performances for the fit on "J6 toward suburbs". In both cases for both indicators, the model is significantly more appropriate than the benchmark at a 5% threshold.

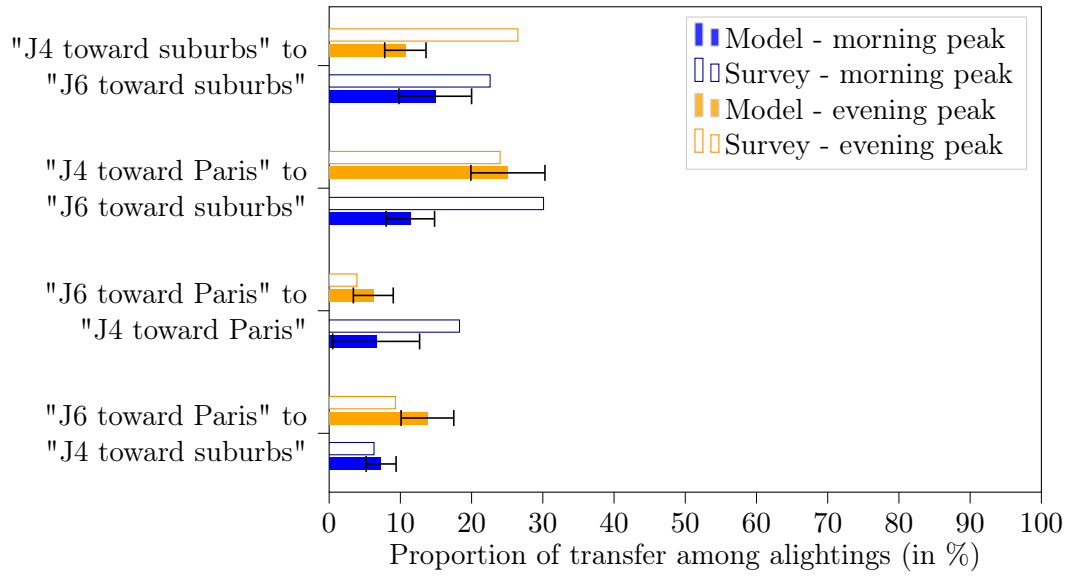

Figure 3: Summary of the obtained proportion (in %) of alighting that transfer. Proportions are presented for morning peak hours and evening peak hours, from the model and the survey. Transfer proportions from "J6 toward Paris" to "J4 toward suburbs" agree the most between the model and the survey. Survey data was collected in the spring of 2022, and data used in the model was collected in the autumn of 2022.

Limitations

Other limitations in the approach shall be highlighted. Demand dynamics may fluctuate from one day to another for various reasons. A noticeable one is inter-mode transfers: people coming from the bus and transferring in *Argenteuil* will arrive later if the bus has some delay. This limitation could be partially solved with a higher granularity of entering data (which is available at the hour level so far). For example, with a granularity at the minute, one might see the additional effect of bus arrival on the number of boarding. Using such data will help future improvement.
Another limitation is the sample size. More than half of the triplet (k, s, d) are lost during the preprocessing and not all train numbers are conserved. Thus, the estimated transfer proportion may be disturbed by a non-exhaustive representation of the trains. Extending the temporal perimeter would increase the size of the dataset but may hide seasonal effects (due to weather, for example). Finally, more stations should be assessed to prove the quality of the model. Unfortunately, rolling stocks equipped with infrared counting sensors do not run on lines with higher demands (RER A, RER B, RER D). Thus, most potential case studies do not see enough passenger volumes for such a study. For example, new rolling stock for RER B and RER D will be deployed by 2030, it might be interesting to discuss the proposed model from these new data.

4 Conclusions

In this work, we presented a model to estimate intra-rail transfers. The results showed an important contribution of transfers to model boarding volumes. The transfer proportions estimated from the model were discussed based on the transfer proportions derived from the origin-destination survey. A satisfying coherence was observed for J4 trains toward the suburbs. Notice that it was the only case studied with no alternative. In other cases, the coherence was mitigated. Future work will help confirm the relevance of the approach.
The model developed in this work can be useful to complement origin-destination surveys in estimating transfer proportion. Knowing transfer proportions is crucial to optimize transfer comfort (transfer duration, stopping platform choice, train waiting policies). Another implication of this model is modeling and predicting boarding volumes in connecting stations in a more general framework: considering delay and disruption. Future work will aim to generalize the model for this purpose.

Acknowledgements

We want to thank Arnaud Marville who participated in the early stages of this study during his internship.

References

Gordon, J. B., Koutsopoulos, H. N., & Wilson, N. H. (2018). Estimation of population origin–interchange–destination flows on multimodal transit networks. *Transportation Research Part C: Emerging Technologies*, *90*, 350–365.

Hofmann, M., & O'Mahony, M. (2005). Transfer journey identification and analyses from electronic fare collection data. In *Proceedings. 2005 ieee intelligent transportation systems, 2005.* (pp. 34–39).

Nielsen, O. A., Eltved, M., Anderson, M. K., & Prato, C. G. (2021). Relevance of detailed transfer attributes in large-scale multimodal route choice models for metropolitan public transport passengers. *Transportation Research Part A: Policy and Practice*, *147*, 76–92.

Simon, J., & Furth, P. G. (1985). Generating a bus route od matrix from on-off data. *Journal of Transportation Engineering*, *111*(6), 583–593.

Yap, M., Cats, O., van Oort, N., & Hoogendoorn, S. (2017). A robust transfer inference algorithm for public transport journeys during disruptions. *Transportation research procedia*, *27*, 1042–1049.

Yap, M., Wong, H., & Cats, O. (2024). Passenger valuation of interchanges in urban public transport. *Journal of Public Transportation*, *26*, 100089.