# Real-time occupancy predictions of public transport vehicles

**Federico Gallo**
**Francesco Corman**
**Nicola Sacco**

**Conference Paper STRC 2022**

STRC

**22nd Swiss Transport Research Conference**

Monte Verità / Ascona, May 18 – 20, 2022

# Real-time occupancy predictions of public transport vehicles

Federico Gallo, Francesco Corman

Institute for transport planning and systems
ETH Zurich, Switzerland

E: fgallo@ethz.ch
francesco.corman@ivt.baug.ethz.ch

Federico Gallo, Nicola Sacco

Department of Mechanical, Energy, Management and Transportation Engineering
University of Genova, Italy

E: nicola.sacco@unige.it

May 2022

## Abstract

In this paper we address the problem of predicting the crowding of urban public transport vehicles with a network-wide approach where more transit lines and their features are jointly considered. In particular, the aim is to investigate whether the real-time availability of the on-board passenger countings can help to improve the quality of the short-term predictions made for the sections in advance, with respect to the predictions made only with historical data and current delays. To do so, a dataset related to the Zurich public transport network is considered, with a wide range of features including historical passenger countings, scheduled and real departure times, weather, scheduled and real connections and overlappings with other lines, day typology and holidays. We then add to this dataset features regarding the occupancy levels observed at the previous stops, which are then used in the performed experiments. To handle the large dimension of the dataset, predictions are made using an efficient gradient-boosted decision tree algorithm called LightGBM.

Results show that real-time occupancies can help to make more accurate predictions up to 8/9 stops in advance, even if a dependency on the stop position is observed. Results also show a negative effect of the most central stops, in particular the ones coinciding with important transfer points, on making predictions with data related to the sections located previously such stops. Moreover, the performed experiments can help understanding better the role some stops have in the public transport network.

## Keywords

Public Transport, Occupancy Level Prediction, Machine Learning

# 1. Introduction

With an ever-growing population and mobility demand in urban areas, public transport is more and more recognized as essential to limit traffic-related negative effects like the increase of emissions, noise, and traffic congestion. To guarantee its attractiveness, in addition to criteria like frequency of services, reliability, speed and number of transfers it is essential also the comfort perceived by the users, which is strictly related to the crowding of vehicles.

Knowing in advance when and where a public transport network will be crowded can help passengers in planning and making their trip, allowing them to opt for a less crowded service if they wish. This is particularly feasible in complex urban public transport networks, where passengers have usually different alternatives in terms of lines and services (e.g., wait for the next service or for another line if it is less crowded), and will became more and more relevant in the future thanks to the expected increase and spread of flexible and remote working strategies (Yu et al., 2019), which give workers more freedom in deciding the commuting times, the commuting days and also where to work.

Therefore, giving crowding information to the passengers is expected to result in a better distribution between them and the available lines and services, contributing in particular to a decrease in the occurrences of high crowding levels. These latter are indeed a major source of discomfort for the passengers, and can arise reliability problems like bus bunching (Drabicki et al., 2022) as well as sanitary problems like an increased spreading of Covid-19 during the recent pandemic outbreak (Thomas et al., 2022). Related to this latter aspect, recent studies have revealed a willingness of passengers to choose less crowded trips even if this implies longer travel times, and even to pay more for travelling in less crowded services (Marra et al., 2022; Shelat et al., 2022).

Finally, knowing in advance the crowding levels can help the public transport provider to better manage the available resources and to introduce additional services when and if these are needed.

## 1.1 Literature review

Different studies have addressed the topic of predicting the number of passengers on public transport vehicles, showing the importance and the consideration for this research topic; they mainly differ in the type of data used and the prediction model employed.

Many studies in literature make use of Smart Card data, which is available in a variety of large cities with metro lines (Xue et al., 2015; Ding et al., 2016; Zhang et al., 2017; Pereira et al., 2015; Rodrigues et al., 2017). This type of data has the advantage of covering all trips, allowing to understand the movements of individual passengers such as transfers between lines and

origin-destination patterns. With Smart Cards it is also possible to consider the effect of different fares on the passenger travel choices.

In addition to Smart Card data, Pereira et al., 2015 and Rodrigues et al., 2017 used data mining techniques to extract from the web context information about local events to forecast the number of passengers depending on the characteristics of an event. Similarly to Rodrigues et al., 2017, Ni et al., 2016 adopted a similar approach to make predictions. In such a paper authors extract data of nearby events by scraping Twitter for hashtags corresponding to certain events.

Historical data, Automatic Vehicle Location (AVL) and Automatic Passenger Counting (APC) data are instead used in Samaras et al., 2015 and in Jenelius, 2019 to predict passenger demand.

The increasing large availability of mobile data has also provided new possibilities to explain and predict passenger numbers: Vandewiele et al., 2017 collected data from mobile phone surveys where passengers could indicate the perceived level of crowding during their trip. Due to the data source, the considered problem deals with classification and not with regression.

As for the prediction framework, due to the cyclical component of commuting concerning the time of the day and the day of the week, most classical statistical prediction approaches have relied on time-series forecasting methods such as ARIMA and Kalman Filters (Ni et al., 2016; Xue et al., 2015; Zhang et al., 2017).

More recently, research has shifted more towards Machine Learning and Deep-Learning techniques that can model the non-linearity and the feature interactions. While Rodrigues et al., 2017, and Baek and Sohn, 2016, have successfully applied Neural Networks, Samaras et al., 2015, Ding et al., 2016, and Vandewiele et al., 2017, have shown that Neural Networks were outperformed in their cases by tree-based algorithms like Random Forests and Gradient Boosted Decision Trees.

Also Markov Chains have been proven to provide good results, since they can capture memoryless dependencies between the crowding of close (in time) public transport services (Więcek et al., 2019).

In summary, the state of the art demonstrates the efficacy of different typologies of data and methods in predicting on-board passenger numbers, but they mainly focus on individual lines or routes. However, an extensive urban public transport network is a complex system, with a large variety of phenomena and events that can influence the on-board crowding: passengers often can choose between different lines, routes and modes of transport to reach their destination. In addition, the high frequency of services makes passengers flexible with their departure times. To cope with these issues, Jamar et al., 2020 proposed to model the possibility for passengers of taking multiple lines by adopting a prediction framework at a network level,

where both the line overlaps over a path and the different line arrivals at the stops are considered. In this paper we further develop this approach, by explicitly investigating, at a network level, the role real-time crowding information of vehicles has in influencing the crowding of such vehicles in the next sections of a given line.

## 1.2 Making crowding predictions at a network level

In Jamar et al., 2020, authors took as input a dataset related to the Zurich public transport network, which consisted of 10'785'072 rows each representing the number of passengers aboard a public transport vehicle (tram or bus) between two stops of the Zurich network from 10/12/2017 to 09/12/2018.

To investigate the usefulness of working at a network level, from the dataset authors created a number of features related to the overlapping between different lines (these latter defined as segments of a network served by more than one line, where users usually do not distinguish between the various lines); these features included the length of the overlaps, the lines and the stops involved in them, and the headway on the overlaps (resulting from all the lines serving the overlap). It is worth pointing out here that in this paper for working at a network level we do not mean simply making predictions for many lines of a network, but considering (as described above) the data related to the other lines of a network when making the predictions for a given line.

To investigate the performance of predictions made several days in advance before the departures (long-term predictions), authors then developed two prediction models: the first one was built using only scheduled timetable data and historical passenger countings; the second one was instead built considering also real-time data like real departure times, delays and weather.

From the comparison between a benchmark model (built from the historical average of the occupancy grouped by time, line and stop), the scheduled-time model and the real-time model, different observations where made: firstly, that making predictions at a network level allows to get more accurate results with respect to considering each line alone; secondly, that considering real-time data like current delays, real headways and real vehicle arrivals at the stops leads to an improvement of the predictions made, with respect to the ones made many days in advance, when this kind of information is not available.

An important open issue left by the above described work is how to effectively use the real-time passenger countings to improve the accuracy of the predictions made. Indeed, with the real-time measured passenger countings it is possible to compute, for a given line, time and section, the error between the real value and the value predicted by the proposed model. This

error can then be used to improve the predictions along the next sections of the line. However, it has been shown in such work that the error does not remain constant along the line; in other words, if the predicted crowding at a certain section differs from the observed one of a certain value, it is not possible to directly correct the predictions in the next sections with the found value, because after few stops this would worsen the predictions. Thus, a further investigation of the error behavior as well as its spatial and temporal propagation is needed.

The goal of this paper is to start from the above mentioned results and present a strategy to implement the real-time passenger counting data coming from automated on-board measurements, and to determine whether this kind of data can be useful to improve the quality of short-term predictions. In other words, once a certain number of passengers is observed on a section at a certain time and for a certain line, the goal is the one of using it to improve the quality of the predictions made with the real-time model described above, which, in the following, will be addressed as base model. The related base scenario is therefore the one where no information on real-time crowding at previous sections is supposed to be available and predictions are made only with historical data and current delays.

# 2. Methodology

In this section we describe the dataset used, the followed procedure to add the real-time crowding information to it, and the Machine Learning model chosen to make the prediction of the occupancies.

Since the aim is the one of having the lowest root mean square error (RMSE) possible, this metric will be used to address the quality of the occupancy predictions (number of passengers on board in each section of a public transport service). Furthermore, to simulate the application of the considered methodology in a real case where information must be provided to the passengers, we consider three occupancy levels for which computing the discrete accuracy (percentage of correct classifications): low, when the occupancy is less than the half of the seating capacity of the vehicle (each passenger has one seat for himself and another one for his belongings); medium, when the occupancy is less than the seating capacity (each passenger has at least one seat for himself); high, when the occupancy is greater that the seating capacity (some passengers have to stand).

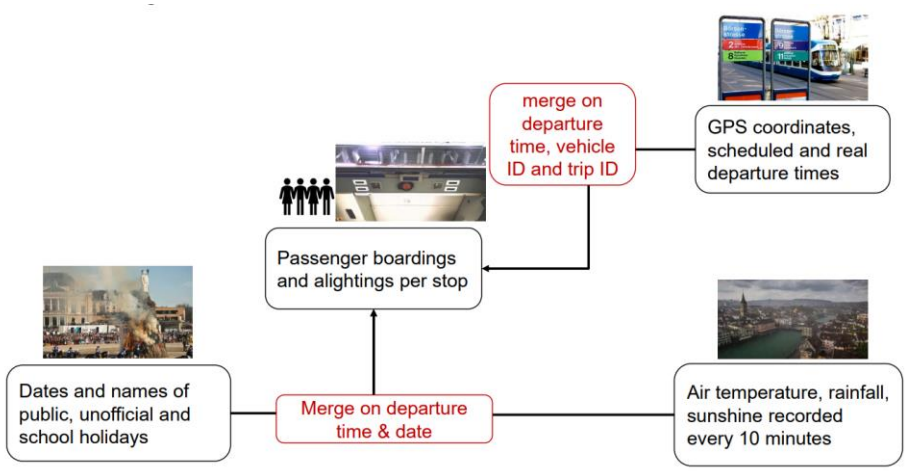
## 2.1 Data

The considered dataset consists of a fraction of the one used by Jamar et al., 2020, where only the tram lines of the Zurich network are considered. The dataset is the result of the merging of four different types of data:

- Passenger boardings and alightings per stop;
- Scheduled and real departure times;
- Weather;
- Holidays.

Fig.1 shows the merging steps performed to build the dataset.

Figure 1: Data aggregation performed to build the considered dataset



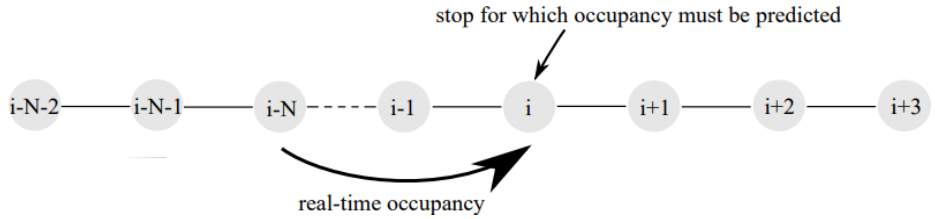

To simulate the conditions for the base scenario, several network features were then added to the merged dataset: the most important ones are the overlaps between lines, the scheduled and the real headway on a corridor served by more lines, and the last lines arrived at the stop.

## 2.2 Modeling real-time crowding information at previous stops

To simulate the availability of the real-time occupancy in the sections located prior a public transport stop for which occupancy predictions must be made, we performed different experiments; in each experiment we added to each row of the original dataset described in Sec. 2.1 one or more features representing the measured real-time occupancy on the vehicle in the sections located a certain number N of stops away from the stop *i*, which is the one for which predictions must be made. In this way, by comparing the results of the different experiments, it is possible to investigate the efficacy of this kind of data in enhancing the quality of the predictions. In particular, by comparing the quality of the predictions made with the real occupancy measured at different stops away, it is possible to see how far the real-time occupancy can influence the occupancy in the following sections, and thus how far this kind of data can be used to improve the quality of the predictions made.

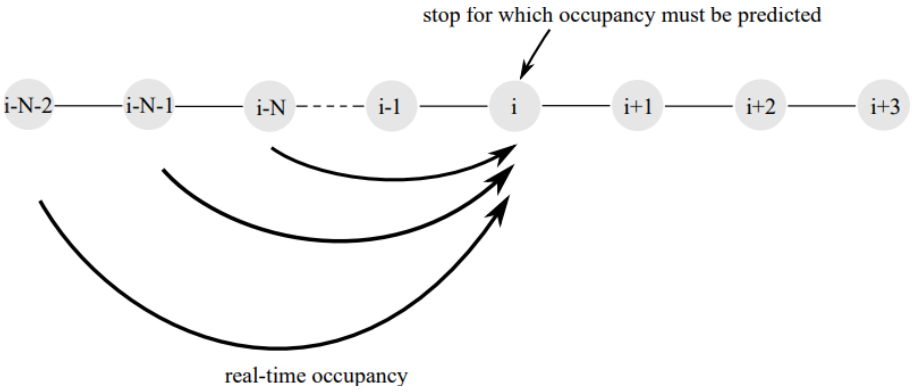
Fig. 2 shows graphically the considered framework: to make predictions at the stop *i*, the real-time occupancy at the N-th stop away is added as an additional feature to the dataset. This operation is done for each stop in the dataset apart from the first ones, for which this kind of data is not yet available.

Figure 2: Scheme of the considered framework to make predictions with the real-time occupancy observed N stops before.




To investigate whether knowing the real-time occupancy in more than one consecutive section can improve the quality of the predictions, we performed some experiments where we added to each row of the original dataset more than one measured occupancy in the past consecutive sections (see Fig. 3).

Figure 3: Scheme of the considered framework to make predictions with the real-time occupancy observed in some consecutive segments, the last one of them located N stops before.




The main aim of such experiments is to investigate whether the crowding on the future sections of a line depends only on the last observed occupancy or also on the past observations; in other words, whether the Markovian memoryless property holds.

## 2.3 Prediction model

To make the occupancy predictions for each of the considered experiments, we make use of the LightGBM algorithm, which is an improved version with high efficiency of the Gradient Boosted Decision Tree algorithm (Ke et al., 2017). The reason of such a choice lies into the fact that this model already provided good results for a similar dataset, compared with other popular

Machine Learning methods like Generalized Linear Models, Support Vector Machines, and Neural Networks (Jamar et al., 2020). Compared to Support Vector Machines and Neural Networks, tree-based algorithms like LightGBM do not require any preprocessing such as scaling, and the use of bagging or boosting procedures make these algorithms robust when handling co-linear features and features with low importance.

The tuning of the model parameters, like the number of leaves and the fraction of features used at each boosting round, was done by means of an exact grid search. The final considered parameter values for the base scenario are: 17282 as the number of iterations; 0.95 as feature fraction; 25 as number of leaves; 0.2 as learning rate.

# 3. Results

All the experiments have been performed with the software R using the library LightGBM for the predictions.

In this section we present and discuss the main results obtained for the lines 7 and 11 (Fig. 4), two of the most crowded tram lines of the Zurich public transport network. All the other tram lines were considered when building the overlap features as described in Sec. 1.2. Both these two lines start their trip in the outskirts of the city and cross the city center for ending their trip in another outskirt area. During the central portion of the trip these lines cross important streets and interchange stations between other lines. They also cross the main railway station of the city (Zurich HB). All the 30 stops of line 7 and the relevant 29 sections between them are shown in Fig. 5, while all the 34 stops of the line 11 and the relevant 33 sections are show in Fig. 6.

Figure 4: Path of the lines 7 (black) and 11 (green) of the Zurich public transport network with indication of the direction of travel and of the four stops considered in the experiments described in Sec. 3.2.

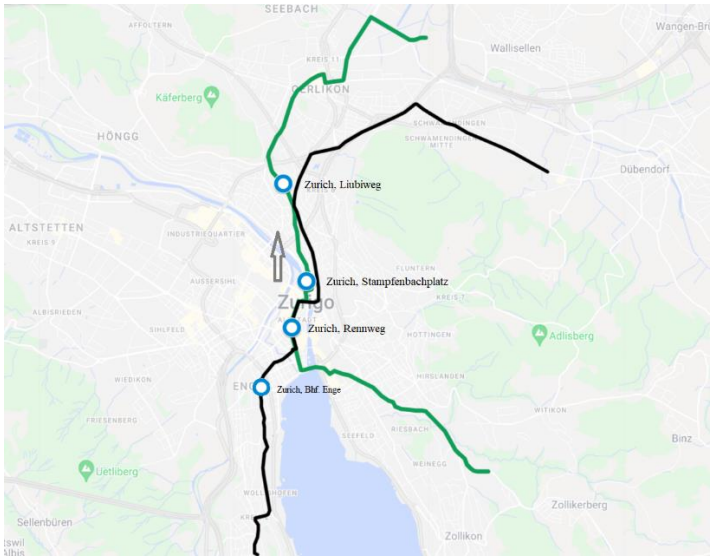


Figure 5: Stops of the line 7 of the Zurich public transport network in the considered direction of travel.

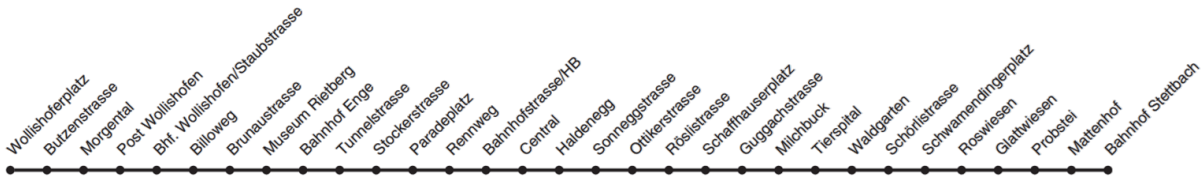
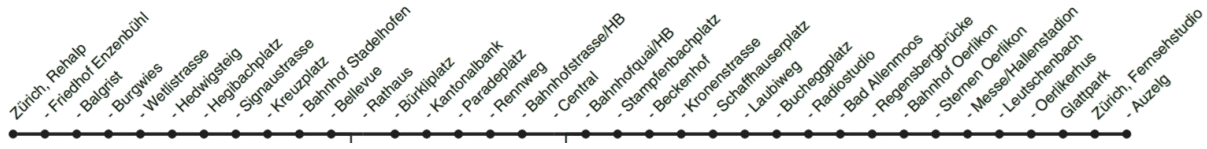



Figure 6: Stops of the line 11 of the Zurich public transport network in the considered direction of travel.




As for line 11, Fig. 7 shows the histogram of the occupancy values. The minimum measured value is 0, the maximum is 231 and the average is 36. The most frequent vehicle used (Cobra tram) has 90 seats and a capacity of 220 passengers.

Figure 7: Histogram of the measured occupancy values for line 11. The colored rectangles indicate the low (green), medium (orange) and high (red) occupancy level ranges considering the capacity of the most common vehicle used on the line.

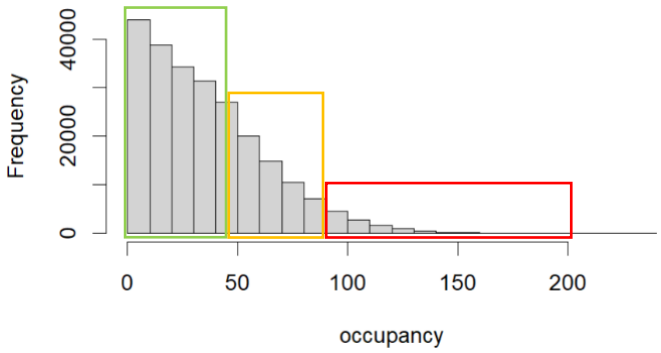
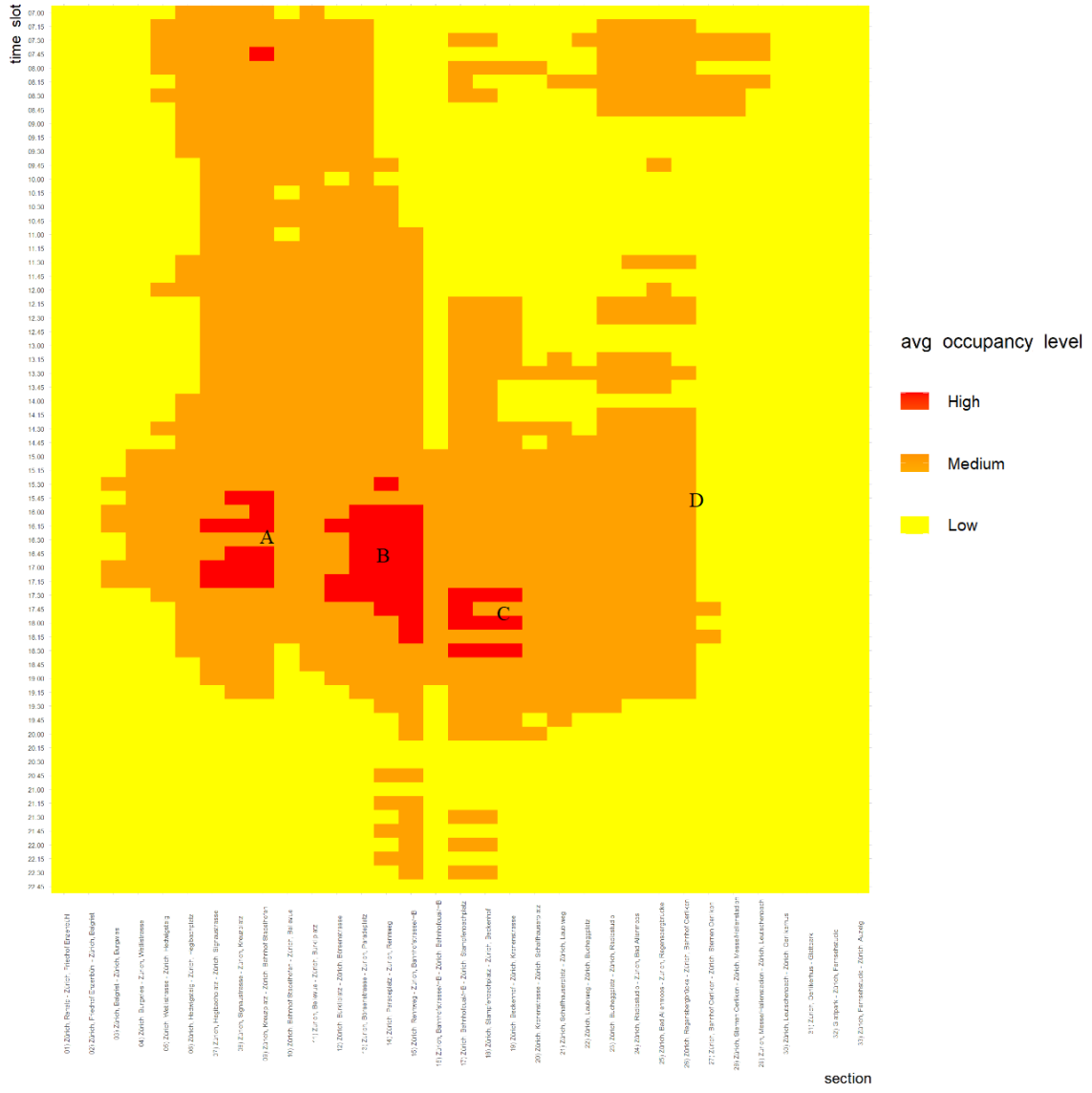



Fig. 8 shows instead the heatmap of the average occupancy level in a normal working day as a function of the section (x axis) and the time (y axis) of the day (15-minutes intervals): it emerges that the average occupancy is always low in the peripheral sections of the line, when it travels in the outskirts, while the occupancy is usually higher in the central sections. In particular there is a peak in the occupancy level in the afternoon, indicatively from 16 pm to 18 pm in some central sections; their intermittency can be explained by pointing out that the first two peaks of high occupancy (A and B) spatially end at two stops which are important transfer points with other lines and, in particular, with other train lines (Zurich, Bahnhof Stadelhofen and Zurich, Bahnhofstr./HB, respectively), so it can be assumed that many passengers get off the tram to

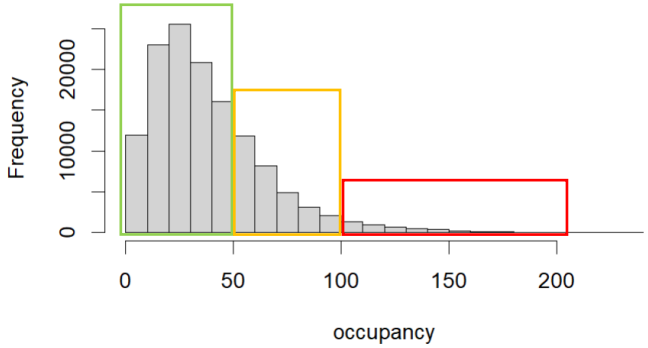
take other lines and trains. A similar conclusion can be done to explain the transition from medium to low occupancy (D), which almost always takes place after the stop of Zurich, Bahnhof Oerlikon, other important transfer point with trains. The third peak (C) of high occupancy ends instead when the line has already crossed the center and the number of passengers onboard starts, globally, to decrease.

Figure 8: heatmap of the average occupancy level per section and time of line 11 during a working day.



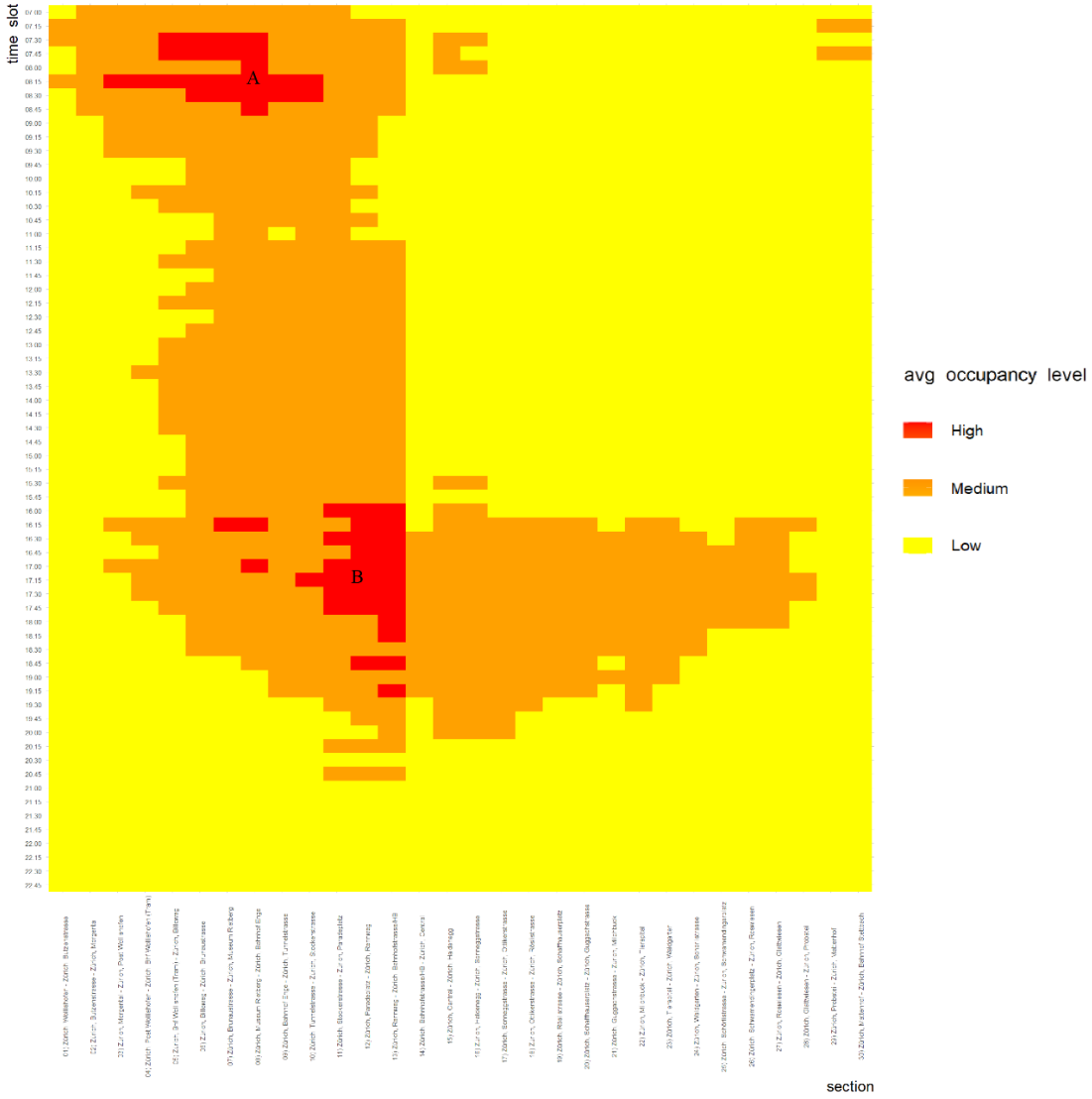

As for line 7, Fig. 9 shows the histogram of the occupancy values. The minimum measured value is 0, the maximum is 238 and the average is 39. The most frequent vehicle used (Tram 2000 Sanfte with pony) has 101 seats and a capacity of 253 passengers.

Figure 9: Histogram of the measured occupancy values for line 7. The colored rectangles indicate the low (green), medium (orange) and high (red) occupancy level ranges considering the capacity of the most common vehicle used on the line.




As for the occupancy level, Fig. 10 shows that, differently from line 11, line 7 has a peak in the occupancy level also in the morning (A) in the sections directed towards the city center, while

Figure 10: heatmap of the average occupancy level per section and time of line 7 during a working day.



the afternoon peak (B) is spatially more concentrated than the one observed for line 11, and it is located in the most central sections of the line. In particular, it starts from the stops Zurich, Tunnelstrasse and Zurich, Stockerstrasse (the first central stops of the line) and ends at the stop of Zurich, Bahnofstrasse/HB, the stop located in front of the main railway station.

To understand the potential of using real-time occupancies at previous stops in improving the quality of predictions, we first present results obtained for all the stops of the two lines except from the first 8 (since, for them, no observations at previous sections can be available).

## 3.1 Line-level perspective

In this section we compare the results obtained in the different experiments we performed on line 7 and line 11, without distinguishing between the stops or their position along the line.

Since the base model, which is the one representing the scenario where the information about real-time crowding is not available at neither of the previous stops, has more difficulties in predicting the high occupancies rather than the low occupancies (as it can be seen in Tab. 1, which shows the confusion matrix for the predictions made for line 11 with the base model), we focus in this section on the afternoon times (16-19 pm) of a working day, because in such period the majority of high occupancy levels is observed.

Table 1: Confusion matrix of the predictions for line 11 made in the base scenario (no information on previous stops available)

| | **Predicted occupancy level** | | |
|---|---|---|---|
| **Real occupancy level** | Low | Medium | High |
| Low | 77% | 23% | 0% |
| Medium | 6% | 89% | 5% |
| High | 1% | 50% | 49% |

Tab.2 shows the RMSE of the predicted occupancy (continuous values) and the accuracy (percentage of correct classifications) of the predicted occupancy level (discrete values) for different experiments. The first one, with no additional features created, represents the base scenario with no information available regarding the occupancy at previous stops.

By looking at the other rows of the table, it is possible to see that the knowledge of the real-time occupancy observed in the previous stops improves the quality of the predictions. In particular, and as expected, the closer the observed occupancy is to the stop at which predictions are made, the more accurate the predictions are, both in terms of RMSE and accuracy. For

instance, knowing the real-time occupancy on the vehicle can improve the accuracy of the predictions made 5 stops in advance up to 4%, and reduce the RMSE up to 10%.

Table 2: Accuracies of predictions for line 7 and 11

| Added features | RMSE (line 7) | Accuracy [%] (line 7) | RMSE (line 11) | Accuracy [%] (line 11) |
|---|---|---|---|---|
| None | 14.78 | 0.79 | 15.34 | 0.78 |
| Measured occupancy 8 stops before | 14.33 | 0.80 | 14.36 | 0.80 |
| Measured occupancy 7 stops before | 14.18 | 0.80 | 14.23 | 0.80 |
| Measured occupancy 6 stops before | 14.01 | 0.81 | 13.97 | 0.81 |
| Measured occupancy 5 stops before | 13.85 | 0.82 | 13.86 | 0.82 |
| Measured occupancy 4 stops before | 13.37 | 0.83 | 13.25 | 0.82 |
| Measured occupancy 3 stops before | 12.58 | 0.84 | 12.79 | 0.83 |
| Measured occupancy 2 stops before | 11.39 | 0.86 | 11.52 | 0.85 |
| Measured occupancy 1 stops before | 9.48 | 0.89 | 9.59 | 0.88 |

To investigate whether knowing the observations of the occupancy in many previous stops after the last observed one can provide better results than just knowing the occupancy observed at only a single previous stop, we then performed different experiments by adding more than one observed occupancy in the previous sections. Tabs. 3, 4 and 5 show the results.

Table 3: Accuracies of predictions for line 7 - multiple observations from 5 stops before and backward

| Added features | RMSE | Accuracy [%] |
|---|---|---|
| Measured occupancy 5 stops before | 13.85 | 0.82 |
| Measured occupancy 5 and 6 stops before | 13.82 | 0.82 |
| Measured occupancy 5, 6 and 7 stops before | 13.83 | 0.82 |
| Measured occupancy 5, 6, 7 and 8 stops before | 13.92 | 0.82 |

Table 4: Accuracies of predictions for line 11 – multiple observations from 2 stops before and backward

| Added features | RMSE | Accuracy [%] |
|---|---|---|
| Measured occupancy 2 stops before | 11.52 | 0.85 |
| Measured occupancy 2 and 3 stops before | 11.60 | 0.85 |
| Measured occupancy 2, 3 and 4 stops before | 11.58 | 0.85 |

Table 5: Accuracies of predictions for line 11 - multiple observations from 5 stops before and backward

| Added features | RMSE | Accuracy [%] |
|---|---|---|
| Measured occupancy 5 stops before | 13.26 | 0.82 |
| Measured occupancy 5 and 6 stops before | 13.07 | 0.82 |
| Measured occupancy 5, 6 and 7 stops before | 13.07 | 0.82 |
| Measured occupancy 5, 6, 7 and 8 stops before | 13.15 | 0.82 |

As it can be seen, the RMSE and the accuracy are not significantly improved by the additional features, and it seems the knowledge of only the last available measured occupancy is sufficient to improve the predictions, which seem to be independent on the past observations. As a result, in the following section we will not use these additional features anymore in the experiments.

## 3.2 Stop-level perspective

From Sec. 3.1 it emerges that real-time measurements can improve the quality of the predictions at a line level. However, the Zurich tram lines usually have many transfer points and cross several central sections of the Zurich Network. For this reason, the aim of this section is to present and discuss the results obtained by focusing on some specific stops, to investigate possible spatial differences that can be present in the results. In particular, we focus on three stops of the line 11 and on one stop of the line 7.

The first considered stop is Zurich, Rennweg. It is one of the most central stops, located in the Bahnhofstrasse, an important central street of Zurich with many demand attractors and overlaps between other tram lines. From Tab. 6 it is possible to see that at this stop the base model already performs better than the average (first line of Tab. 2). All the stops used for the observation of the real-time occupancy are located in the city center; for this reason it is not useful to make use of the real-time information on crowding for more than 5 stops away, due to the large variety of uncertain events that can happen in these zones of the city.

Table 6: Accuracies of predictions for line 11 at the stop Zurich, Rennweg

| Added features | RMSE | Accuracy [%] |
|---|---|---|
| None | 13.98 | 0.82 |
| Measured occupancy 5 stops before (Zurich, Bahnhof Stadelhofen) | 13.26 | 0.83 |
| Measured occupancy 4 stops before (Zurich, Bellevue) | 12.08 | 0.84 |
| Measured occupancy 3 stops before (Zurich, Burkliplatz) | 11.13 | 0.85 |
| Measured occupancy 2 stops before (Zurich, Kantonalbank) | 10.64 | 0.86 |
| Measured occupancy 1 stops before (Zurich, Paradeplatz) | 8.79 | 0.89 |

The second considered stop is Zurich, Stampfenbachplatz. This is the first stop after the stop located in front of the main railway station, and the first one after the most central sections. From Tab. 7 it is possible to see here a different behavior with respect to the one observed at Zurich Rennweg. Firstly, the performance of the base model is significantly worse. Secondly, there is a big increase in the accuracy (and a drop in the RMSE) for the last two experiments, and in particular for the last one. This can be explained by the large variety of events than can happen in the proximity of the main railway station (arrival of trains, transfer between other lines, …). In other words, the two stops near the main railway stations act as a 'barrier' that makes it difficult to predict what will happen after them using data related to sections before them. Once passed this barrier, the predictions suddenly get better. A spatial dependency of the effectiveness of real-time observations in improving the quality of the predictions has been observed also in the prediction of delays (Buchel and Corman, 2022).

Table 7: Accuracies of predictions for line 11 at the stop Zurich, Stampfenbachplatz

| Added features | RMSE | Accuracy [%] |
|---|---|---|
| None | 13.52 | 0.77 |
| Measured occupancy 5 stops before (Zurich, Kantonalbank) | 12.77 | 0.78 |
| Measured occupancy 4 stops before (Zurich, Paradeplatz) | 12.67 | 0.78 |
| Measured occupancy 3 stops before (Zurich, Rennweg ) | 12.43 | 0.79 |
| Measured occupancy 2 stops before (Zurich, Bahnhofstr./HB) | 10.12 | 0.83 |
| Measured occupancy 1 stops before (Zurich, Bahnhofquai/HB) | 5.86 | 0.90 |

The third considered stop is Zurich, Laubiweg. This stop was chosen as representative of the stops located sufficiently far from the central sections of the line. From the results shown in Tab. 8, it can be seen that for this stop the predictions of the base model can be enhanced using data up to 8 stops in advance. Also in this case it is observable the effect of the stops next to the main railway station, after whom there is the most important increase in the accuracy of predictions.

Table 8: Accuracies of predictions for line 11 at the stop Zurich, Laubiweg

| Added features | RMSE | Accuracy [%] |
|---|---|---|
| None | 12.59 | 0.80 |
| Measured occupancy 8 stops before (Zurich, Paradeplatz) | 11.67 | 0.81 |
| Measured occupancy 7 stops before (Zurich, Rennweg) | 11.70 | 0.82 |
| Measured occupancy 6 stops before (Zurich, Bahnhofstr./HB) | 10.13 | 0.82 |
| Measured occupancy 5 stops before (Zurich, Bahnhofquai/HB) | 8.37 | 0.85 |
| Measured occupancy 4 stops before (Zurich, Stampfenbachplatz) | 8.03 | 0.86 |

| | | |
|---|---|---|
| Measured occupancy 3 stops before (Zurich, Beckenhof) | 7.85 | 0.87 |
| Measured occupancy 2 stops before (Zurich, Kronenstr.) | 6.97 | 0.89 |
| Measured occupancy 1 stops before (Zurich, Schaffhauserplatz) | 4.91 | 0.91 |

The last considered stop is Zurich, Bahnhof Enge, which is one of the first central stop of line 7 in the direction of the city center. From Tab. 9 it can be seen in this case that the gain in the accuracy and the decrease of the RMSE is more constant, due to the lack of important transfer points in this portion of the line. In addition, the real-time crowding information can be exploited up to 8 stops backward to increase the quality of the predictions.

Table 9: Accuracies of predictions for line 7 at the stop Zurich, Bahnhof Enge

| **Added features** | **RMSE** | **Accuracy [%]** |
|---|---|---|
| None | 15.60 | 0.79 |
| Measured occupancy 8 stops before (Zurich, Wollishoferplatz) | 14.24 | 0.80 |
| Measured occupancy 7 stops before (Zurich, Butzenstr.) | 13.86 | 0.82 |
| Measured occupancy 6 stops before (Zurich, Morgental) | 13.01 | 0.82 |
| Measured occupancy 5 stops before (Zurich, Post Wollishofen) | 12.35 | 0.84 |
| Measured occupancy 4 stops before (Zurich, Bhf. Wollishofen) | 12.08 | 0.85 |
| Measured occupancy 3 stops before (Zurich, Billoweg) | 11.57 | 0.85 |
| Measured occupancy 2 stops before (Zurich, Brunaustr.) | 10.58 | 0.85 |
| Measured occupancy 1 stops before (Zurich, Museum Rietberg) | 9.66 | 0.86 |

# 4. Conclusions and future research directions

In this work we implement real-time crowding information in a network-level prediction framework, and we analyze the effectiveness of using such kind of data in improving the accuracy of the predictions. Results allow to conclude that real-time observations can be used to significantly improve the predictions made for a vehicle on the sections in advance, but this improvement progressively decays when increasing the distance between the observed stop and the stop at which the prediction is made. Moreover, spatial dependencies are observed, since the position on the line of the observed stop and of the predicted stop has an influence on the quality of the results. In general, real-time information is less useful when there are important transfer points or demand attractors between the observed stop and the predicted one; as a matter of fact, these points introduce more sources of uncertainties related to all the possible events which may happen in the proximity of them.

We also find that predictions are not influenced by the past observations once the crowding level at a certain stop is observed. This memoryless property could be exploited in the future paving the way for including Markovian Chains or Markovian Fields in the prediction framework.

Further directions of the proposed approach include a specific investigation of the spatial relationships between neighbor stops, an investigation of the role crowding information related to past services can have on the quality of the predictions, a deeper modeling of the overlaps between lines and the study of other prediction models.

# 5. References

Baek, J., Sohn, K., 2016. Deep-Learning Architectures to Forecast Bus Ridership at the Stop and Stop-To-Stop Levels for Dense and Crowded Bus Networks. Applied Artificial Intelligence 30, 861–885. https://doi.org/10.1080/08839514.2016.1277291

Buchel, B., Corman, F., 2022. What Do We Know When? Modeling Predictability of Transit Operations. IEEE Trans. Intell. Transport. Syst. 1–12. https://doi.org/10.1109/TITS.2022.3145243

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, 1100. https://doi.org/10.3390/su8111100

Drabicki, A., Kucharski, R., Cats, O., 2022. Mitigating bus bunching with real-time crowding information. Transportation. https://doi.org/10.1007/s11116-022-10270-3

Jamar, L., Buchel, B., Corman, F., 2020. A Network-Wide Approach to Predicting Urban Pub lic Transport Passenger Numbers at a Stop-to-Stop Level.

Jenelius, E., 2019. Data-Driven Bus Crowding Prediction Based on Real-Time Passenger Counts and Vehicle Locations.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Curran Associates, Inc. 30.

Marra, A.D., Sun, L., Corman, F., 2022. The impact of COVID-19 pandemic on public transport usage and route choice: Evidences from a long-term tracking study in urban area. Transport Policy 116, 258–268. https://doi.org/10.1016/j.tranpol.2021.12.009

Ni, M., He, Q., Gao, J., 2016. Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. IEEE Trans. Intell. Transport. Syst. 1–10. https://doi.org/10.1109/TITS.2016.2611644

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, 273–288. https://doi.org/10.1080/15472450.2013.868284

Rodrigues, F., Borysov, S.S., Ribeiro, B., Pereira, F.C., 2017. A Bayesian Additive Model for Understanding Public Transport Usage in Special Events. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2113–2126. https://doi.org/10.1109/TPAMI.2016.2635136

Samaras, P., Fachantidis, A., Tsoumakas, G., Vlahavas, I., 2015. A prediction model of passenger demand using AVL and APC data from a bus fleet, in: Proceedings of the 19th Panhellenic Conference on Informatics. Presented at the PCI '15: 19th

Panhellenic Conference on Informatics, ACM, Athens Greece, pp. 129–134. https://doi.org/10.1145/2801948.2801984

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, e0264805. https://doi.org/10.1371/journal.pone.0264805

Thomas, M.M., Mohammadi, N., Taylor, J.E., 2022. Investigating the association between mass transit adoption and COVID-19 infections in US metropolitan areas. Science of The Total Environment 811, 152284. https://doi.org/10.1016/j.scitotenv.2021.152284

Vandewiele, G., Colpaert, P., Janssens, O., Van Herwegen, J., Verborgh, R., Mannens, E., Ongenae, F., De Turck, F., 2017. Predicting Train Occupancies based on Query Logs and External Data Sources, in: Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion. Presented at the the 26th International Conference, ACM Press, Perth, Australia, pp. 1469–1474. https://doi.org/10.1145/3041021.3051699

Więcek, P., Kubek, D., Aleksandrowicz, J., Stróżek, A., 2019. Framework for Onboard Bus Comfort Level Predictions Using the Markov Chain Concept. Symmetry 11, 755. https://doi.org/10.3390/sym11060755

Xue, R., Sun, D. (Jian), Chen, S., 2015. Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach. Discrete Dynamics in Nature and Society 2015, 1–11. https://doi.org/10.1155/2015/682390

Yu, R., Burke, M., Raad, N., 2019. Exploring impact of future flexible working model evolution on urban environment, economy and planning. Journal of Urban Management 8, 447–457. https://doi.org/10.1016/j.jum.2019.05.002

Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., Tian, C., Li, X., Huang, B., Li, Z., 2017. A Real-Time Passenger Flow Estimation and Prediction Method for Urban Bus Transit Systems. IEEE Trans. Intell. Transport. Syst. 18, 3168–3178. https://doi.org/10.1109/TITS.2017.2686877