Effectiveness of trip planner data in predicting short-term bus ridership

Wang, Ziyulong; Pel, Adam J.; Verma, Trivik; Krishnakumari, Panchamy; van Brakel, Peter; van Oort, Niels

DOI
10.1016/j.trc.2022.103790

Publication date
2022

Document Version
Final published version

Published in
Transportation Research Part C: Emerging Technologies

Transportation Research Part C 142 (2022) 103790

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Effectiveness of trip planner data in predicting short-term bus ridership

Ziyulong Wang [a,*], Adam J. Pel [a], Trivik Verma [a], Panchamy Krishnakumari [b], Peter van Brakel [c], Niels van Oort [a]

[a] Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands
[b] Department of Multi-Actor Systems, Delft University of Technology, P.O. Box 5015, 2600 GA Delft, The Netherlands
[c] REISinformatiegroep B.V. (9292), P.O. Box 19319, 3501 DH Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Public transport
Trip planner
Bus ridership prediction
Machine learning

ABSTRACT

Predictions on Public Transport (PT) ridership are beneficial as they allow for sufficient and cost-efficient deployment of vehicles. On an operational level, this relates to short-term predictions with lead times of less than an hour. Where conventional data sources on ridership, such as Automatic Fare Collection (AFC) data, may have longer lag times and contain no travel intentions, in contrast, trip planner data are often available in (near) real-time and are used before traveling. In this paper, we investigate how such data from a trip planner app can be utilized for short-term bus ridership predictions. This is combined with AFC data (in this case smart card data) to construct a ground truth on actual ridership. Using informative variables from the trip planner dataset through correlation analysis, we develop 3 supervised Machine Learning (ML) models, including k-nearest neighbors, random forest, and gradient boosting. The best-performing model relies on random forest regression with trip planner requests. Compared with the baseline model that depends on the weekly trend, it reduces the mean absolute error by approximately half. Moreover, using the same model with and without trip planner data, we prove the usefulness of trip planner data by an improved mean absolute error of 8.9% and 21.7% and an increased coefficient of determination from a 5-fold cross-validation of 7.8% and 18.5% for two case study lines, respectively. Lastly, we show that this model performance is maintained even for the trip planner requests with prediction lead times up to 30 min ahead, and for different periods of the day. We expect our methodology to be useful for PT operators to elevate their daily operations and level of service as well as for trip planner companies to facilitate passenger replanning, in particular during peak hours.

1. Introduction

Predicting public transport (PT) ridership is vital to address the increasing passenger demand (Van Oort et al., 2016; Noursalehi et al., 2018; Hao et al., 2019). It allows operators for allocating vehicles sufficiently and cost-efficiently, which improves passenger satisfaction and leads to a higher level of PT service (Pel et al., 2014; Ohler et al., 2017). On an operational level, this prediction needs to be realized in the short term with less than an hour.

* Corresponding author.
E-mail addresses: Z.Wang-19@tudelft.nl (Z. Wang), A.J.Pel@tudelft.nl (A.J. Pel), T.Verma@tudelft.nl (T. Verma), P.K.Krishnakumari@tudelft.nl (P. Krishnakumari), pvanbrakel@9292.nl (P. van Brakel), N.vanOort@tudelft.nl (N. van Oort).

https://doi.org/10.1016/j.trc.2022.103790
Received 10 July 2021; Received in revised form 17 May 2022; Accepted 2 July 2022
Available online 18 July 2022

Until now, such short-term passenger demand predictions have typically used Automatic Fare Collection (AFC) or Global System for Mobile Communications (GSM) data. Research has widely shown that AFC data are useful in predicting short-term passenger demand with different techniques, such as elasticity model (Van Oort et al., 2015), interactive multiple model (Xue et al., 2015), and Machine Learning (ML) models (Ding et al., 2016; Zhou et al., 2016; Liu and Chen, 2017; Ma et al., 2019). Such datasets, however, are collected over days and do not depict the variability in short-term (i.e., from real-time up to 30 min) ridership patterns (Pelletier et al., 2011; Van Oort et al., 2015). Instead, transit information can also be collected in real-time using mobile phone data, which is essential for representing, analyzing, and planning the PT system (Elias et al., 2016). De Regt et al. (2017) fused GSM data with smart card data (retrieved from an AFC system) to reveal the spatial and temporal pattern and to offer insightful mobility patterns from strategical and tactical levels. The same methodology was also seen in the passenger flow measurement of the Paris metro (Aguiléra et al., 2014). Similar to the short-term ridership prediction, some studies expounded upon short-term Origin–Destination (OD) prediction with AFC data (Liu et al., 2019; Zhang et al., 2021) or short-term bus arrival prediction with Automatic Vehicle Location (AVL) data (Pang et al., 2019; Liu et al., 2020).

Nevertheless, these data sources (AFC, AVL, and GSM) are all realized data, which means that they are always logging data or by definition historical data such that they do not contain the intention of traveling. It is valuable to have data on travel intention in ridership prediction before the trip is executed, which AFC and GSM (regardless of their lag time in data availability) fail to accomplish by themselves. This intention can be captured in data from trip planner apps, especially since the data can be available in near real-time. Applications that make the collection of planner data available, provide integrated travel information to its users, which helps users realize their travel needs and brings in convenience and flexibility (Ferreira et al., 2017). Thus, as a proxy, trip planner data provide the same granular level of spatial and temporal information about possible trips equal to the smart card data (Ferreira et al., 2017), implying the possibility of their combination for predicting ridership. Since users do not have to realize their trips for data to be aggregated, their intents of a trip are lodged in real-time and collected through digitized apps (e.g., 9292.[1]) The proliferation of this kind of trip planner app offers a unique opportunity to combine trip planner data and smart card data, which could potentially cater to the substantial interest of operators in matching the vehicle supply and passenger flow demand on an operational level.

In this paper, we investigate how trip planner data in combination with smart card data can be utilized for predicting short-term ridership. In particular, whether the travel intents inherently stored in the trip planner data can improve the effectiveness of short-term ridership prediction. The trip planner data are provided by 9292 in conjunction with the smart card data and AVL data offered by OV-bureau Groningen Drenthe (regional PT authority) to predict the short-term ridership on two case study lines in the provinces of Groningen and Drenthe (in the Netherlands) during October 2019. Moreover, we design three baseline models without trip planner data and three supervised ML models with trip planner data to predict short-term bus ridership and compare the performance of the models. By comparing with the baseline models, we prove that the ML models can help PT operators renew their current models and the incorporation of trip planner data can improve the prediction on top of the smart card data. Using the model performance scores, we further infer the role of trip planner data in the short-term prediction of ridership patterns using variables such as trip request with lead times (real-time to 10, 15, 30 min ahead), variability across a day and in space, day type, and line characteristic. Trip planner data contribute to the best-performing model with a feature importance up to 50%, and this model can reduce the mean absolute error by approximately half, compared to the baseline model based on the weekly trend. Furthermore, this model performance is maintained for prediction lead times up to 30 min ahead in the trip planner requests, and for different periods of the day. The contribution of this paper lies in analyzing the added value of including a novel data source, namely travel planner requests, in a short-term bus ridership prediction method. Hence, we aim to be complementary to other studies where the many state-of-the-art ML methods themselves are reviewed and analyzed (see e.g., Bhavsar et al., 2017; Veres and Moussa, 2020; Xie et al., 2020).

The remainder of this paper is organized as follows: We describe our problem in the following Section 2. Then, the data and methods used in this study are presented in Section 3. Next, we analyze the results of the models for ridership prediction (Section 4). In Section 5, we present our reflections and provide avenues for future research. Lastly, Section 6 draws the main conclusions. More details of the methodology and additional cases can be found in Wang (2020).

2. Problem description

This study seeks to test the usefulness and the effectiveness of trip planner data in short-term ridership prediction in combination with smart card data so that PT operators can cope with the sudden variation of passenger influx by properly maneuvering the vehicle supply with the potential passenger demand. Herein, we define the temporal prediction horizon from real-time to 30 min as short term, and in the remainder of the paper, we refer to the short term for any prediction up to 30 min. This time span is feasible for the trip planner undertakings to update the information and notify the potential crowdedness level for PT operators to dispatch rolling stocks whilst facilitating the route choice of passengers.

The spatial scope of our research is specified as the on-board passenger per trip per section, i.e., from stop to stop. Therefore, the trip planner and the smart card data are collected on an OD level. Namely, both smart card transactions and trip planner requests are presented with a single trip and with a certain traveler. However, no information entailing the subscription type and the background of the traveler is presented, nor it is possible to link the trip planner user with the smart card owner due to privacy

[1] A PT travel information company based in the Netherlands, covering all PT modes - https://9292.nl/.

concerns. Normally, the collection of AFC data takes days, and thus we assume the ridership last week and the historical average of the trip are readily available, except for one baseline model where we take the ridership of yesterday into account. In contrast, we envisage that trip planner data retrieving has no delay as fetching data from a real-time Application Programming Interface (API) is comparably fast and straightforward.

Summarizing the objective, the contributions of this paper are fivefold:

- For the first time, the emerging and novel data source – trip planner requests – has been explored and investigated in a short-term ridership prediction model.
- We extensively analyzed the dimensions in the trip planner data and its correlation with observed ridership from the AFC data in the short term.
- We develop several interpretable, machine-learning prediction models with and without the trip planner data, and investigate their performance on two different bus lines in the Netherlands with different characteristics.
- Results highlight that the best-performing model with the trip planner data feature importance of up to 50% can reduce the mean absolute error by approximately half, compared to the baseline model based on the weekly trend, which is the current prediction model used by the public transport authority. Compared with the same model without trip planner data, the inclusion of trip planner requests decreases the mean absolute prediction error by 8.9% and 21.7% and an increased coefficient of determination from a 5-fold cross-validation of 7.8% and 18.5% for two case study lines, respectively.
- The prediction model performance with trip planner data is maintained for prediction lead times up to 30 min head in the trip planner requests and for different periods of the day, in particular during peak hours.

3. Data and method

This section first explains the data for a better understanding of the rest of the paper, including context, data description, and data analysis (Section 3.1). Then, it presents the different components of the method: the baseline models without trip planner data, the correlation analysis for variable selection, the ML models for ridership prediction with trip planner data, and the evaluation criteria, along with the feature importance analysis (Section 3.2).

3.1. Data

Trip planner. In this study, we use the trip planner data from 9292. 9292 is an interactive trip planner, established in 1992, the Netherlands. It is notably the biggest one in the country and with the largest market share of approximately 46%[2] so that it is a representative set against the other competitors such as NS (Nederlandse Spoorwegen, the main railway operator), Google Maps, and ANWB (Royal Dutch Touring Club).

Every day, it has 600,000 active devices with 4 to 5 requests per device on average, resulting in around 3 million requests per day.[3] It provides local information of the Netherlands and includes PT information of all modes such as bus, metro, train and light rail, which matches the interest of this study. The users can access such a trip planner either through a mobile app, tablet app or a web browser. With the filled-in information of origin, destination, and preference, the planner searches the database for the transport supply and provides the most suitable and possibly multi-modal trip alternatives with the corresponding temporal and spatial details. It can also provide the predicted arrival time of a transit vehicle at a stop as real-time transit information. Hence, it could benefit passengers by reducing waiting time and correspondingly increase the ridership of transit as a result of elevated transit service and perceived personal security (Brakewood and Watkins, 2019).

Case study. We apply our methods on two bus lines — Qliner 300 (inter-city, fewer stops, 54 to 56 min scheduled journey time from terminal to terminal, scheduled shortest headway 12 min,[4]) Q-link 1 (inner-city, connecting multiple important locations, including a hospital and campus, 23 min (off-peak hours) or 27 min (peak hours) scheduled journey time from terminal to terminal, 30-min headway) in region Groningen and Drenthe, two provinces in the northeast of the Netherlands, covering a total population of 1,076,157 and an area of 5640 km^2. For two more case study lines (inter-city service with multiple stops and city-village shuttle), readers are referred to the study by Wang (2020).

The selected region is suitable for this study as bus is the only local mode there, and bus users are much more inclined to be mobile app users, compared to train users as the timetable is not frequently adjusted (Mulley et al., 2017). Besides, Mulley et al. (2017) concluded that age has a strong negative impact on the usage of the trip planner and more than half of the population of Groningen and Drenthe are below 45 years old, which is appropriate for this study.

[2] 9292 hires Newcom to estimate the market share in 2019: https://www.newcom.nl/.

[3] 9292 hires Flurry to measure the number of unique devices per day on which the app is used at least once: https://www.flurry.com/.

[4] The headway varies, depending on the day type and the times of the day, for details, see Wang, 2020.

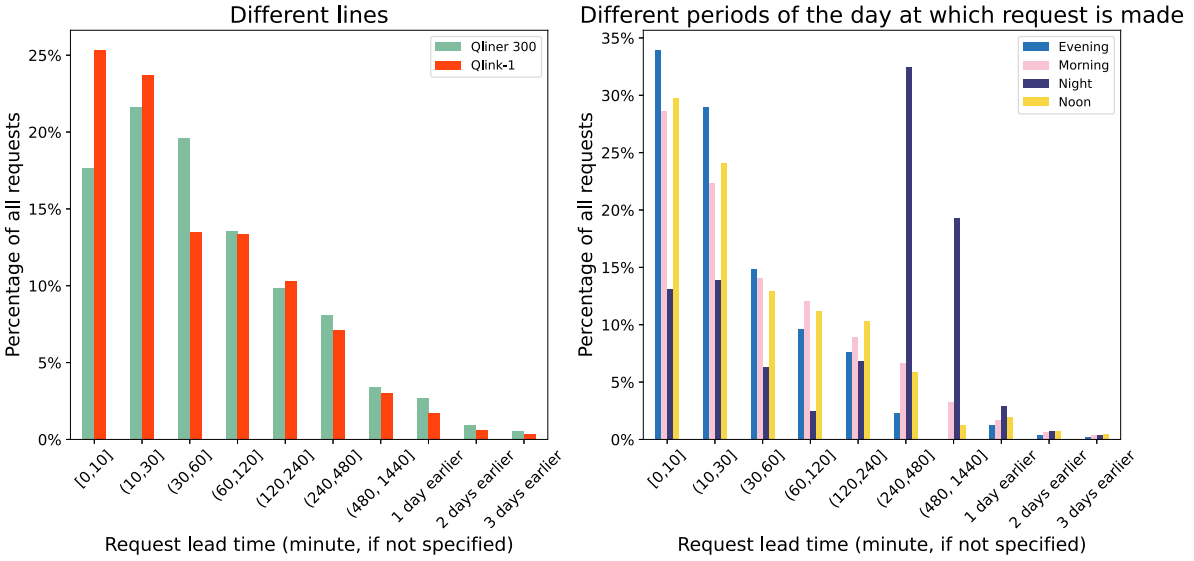

Fig. 1. Comparison of the request number with different prediction lead times per line and per period.

Data description. This study utilizes the smart card and trip planner data in October 2019. During this period, there was a public holiday for schools from 19th to 27th in the case study region, which influenced the travel of students, teachers, and other school-related jobs.

The trip planner data contain four parts, namely stops, modality, answer, and question. A question is the recording of a trip request (desired point-to-point travel information) while an answer is the trip advice accordingly. Only the answer with the least travel time is recorded in the trip planner dataset, and user ID, IP, and location tracking are not available.

The smart card data are retrieved from the nationwide AFC system, using tap-in and tap-out technology (see Van Oort et al., 2016 for a full description). The smart card data are split into trips, namely a tap-in and tap-out of a single leg of a journey with the corresponding spatio-temporal details. However, it is not possible to distinguish the user ID nor the user type as we have no information on the card number and subscription type.

Data cleaning is conducted before the analysis to handle inaccurate recordings, duplicates, and special arrangements of the trips. Both the trip planner and smart card data do not entail the trip number (vehicle recording), and they do not have standardized systems for the stop numbers or names. Thus, we have to map the ridership and requests onto the vehicles based on the AVL data and stop names. This leads to around a 5% loss of trip planner data and a 3% loss of smart card data, which is insignificant. Consequently, there are 26,544 rows of data for Qliner 300 and 39,258 for Q-link 1. For further details on the data cleaning process and results, see Wang (2020).

Data analysis. In order to unveil the usefulness of trip planner data with a certain prediction lead time, we measure the difference between vehicle start time and requested travel time of passengers. The number of requests generally drops with the increase of the prediction lead time in every case study line as shown in the left part of Fig. 1. People prefer asking for route advice 10 to 30 min before their trip. Most requests are sent within a prediction lead time of an hour, and the number of requests drops considerably when the prediction horizon is longer.

It is intuitive that during different periods of the day, people behave differently while using such a trip planner. We differentiate these periods and cluster them into four groups with the same time horizon of 6 h. The right part of Fig. 1 testifies that people plan their trip at least 8 h before the trip during the night but remain roughly the same behavior for the other three periods.

As Fig. 2 indicates, the average number of ridership per line is mostly stable during weekdays, whereas a significant negative influence on the ridership and trip planner requests can be seen on weekends and holidays. Particularly, we observe a high standard deviation of the ridership in every case study line from the density plot of Fig. 2 due to the holidays and the Fridays before the weekends. This phenomenon means that the ridership numbers are spread out within our temporal scope and match the findings in the literature, such as Chiang et al. (2011), Karnberger and Antoniou (2020). However, there is no accident/breakdown, nor special/social events in the case study that could affect ridership as the literature suggests (Pereira et al., 2015; Li et al., 2017; Tao et al., 2018). Additionally, the difference in ridership pattern over the day of the week is not dramatic. On the contrary, spatial characteristics are considerable, which is again in line with the literature (Chakour and Eluru, 2016; Ding et al., 2016). The busy corridors with respect to ridership are usually railway stations, Park and Ride (P+R) stops, business areas, or city centers. The average ridership of each trip in the case study is mostly below the seating capacity. However, it is the busy trips during peak hours that lower the level of service. This implies the potential benefits of improving comfort for those crowded trips.

The scatter plot of trip planner requests and ridership and the histogram of ridership are derived on a section level per line shown in Fig. 3. For the two case study lines, the distributions of request and observed ridership follow the same trend whereas

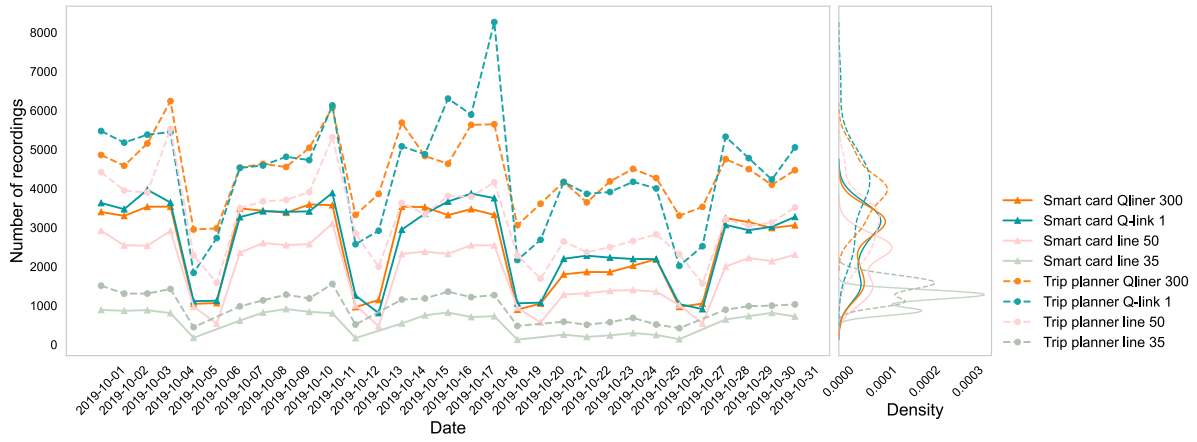

Fig. 2. Distribution of ridership and requests over days in October.

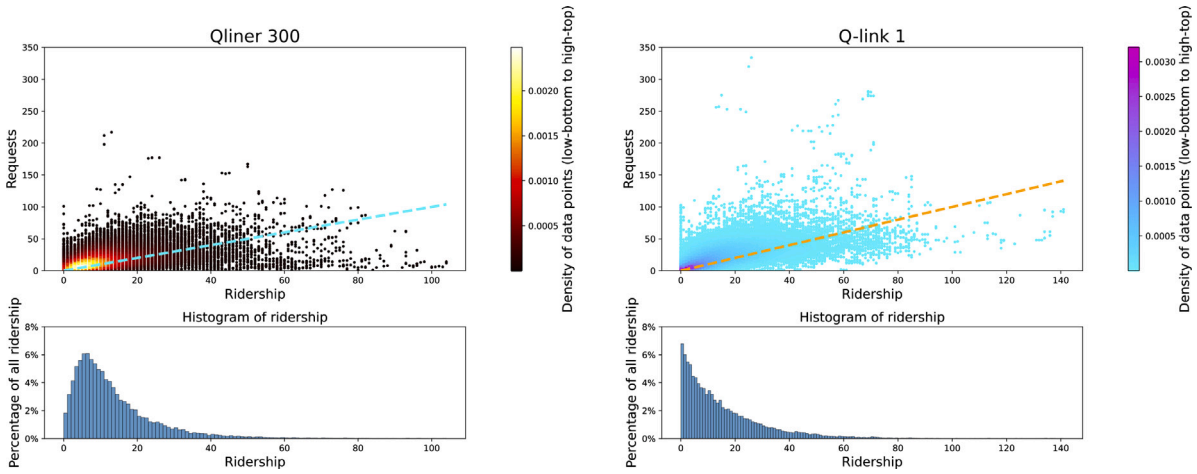

Fig. 3. Scatter plot of ridership and requests and histogram of ridership on a section level per line.

there are more data points that lie above the line, which means that the number of requests is generally larger than the ridership for a given trip, i.e., the realized trips. Besides, several outliers lie beyond the line remarkably, in particular for Q-link 1. Therefore, a linear relationship between them is hard to find.

There could be multiple reasons that lead the number of requests per person/device to be higher than the actual number of passengers on-board and we summarize the generic ones as follows: (1) searching trip advice multiple times; (2) checking trip advice that were looked up before; (3) investigating the PT timetables or disruptions; (4) seeking new trip advice due to an unrealizable current one; (5) maintaining or testing by the trip planner company. It also could be other causes, depending on the situation.

Moreover, the histograms of ridership for both case studies are right-skewed, which means that we have an imbalanced distribution on the target that we want to predict. If we calibrate the ML model by randomly sampling from these observations of ridership, the minimization of errors under less crowded conditions will naturally outweigh that of crowded conditions. This is not necessarily optimal if PT operators may wish to prioritize the predictions for crowded situations. This kind of issue is prevalent in many domains within predictive tasks (Branco et al., 2017). In Section 3.2, we propose an approach to capture the rarest and relevant cases equally as the majority.

3.2. Methodology

Trip planner data emerge as a new type of big data that could be leveraged to help transport operators forecast the passengers on-board. A thorough investigation into the prediction model with trip planner data is essential to deepen the understanding of how and to what extent it could help ridership prediction, not just a “black box” predictive modeling. Our model formulation is based on a five-step approach using four types of supervised learning (including one baseline model) and two state-of-the-practice baseline models without trip planner data that emphasizes its effectiveness. A flow chart with a brief description of each step is presented in Fig. 4.

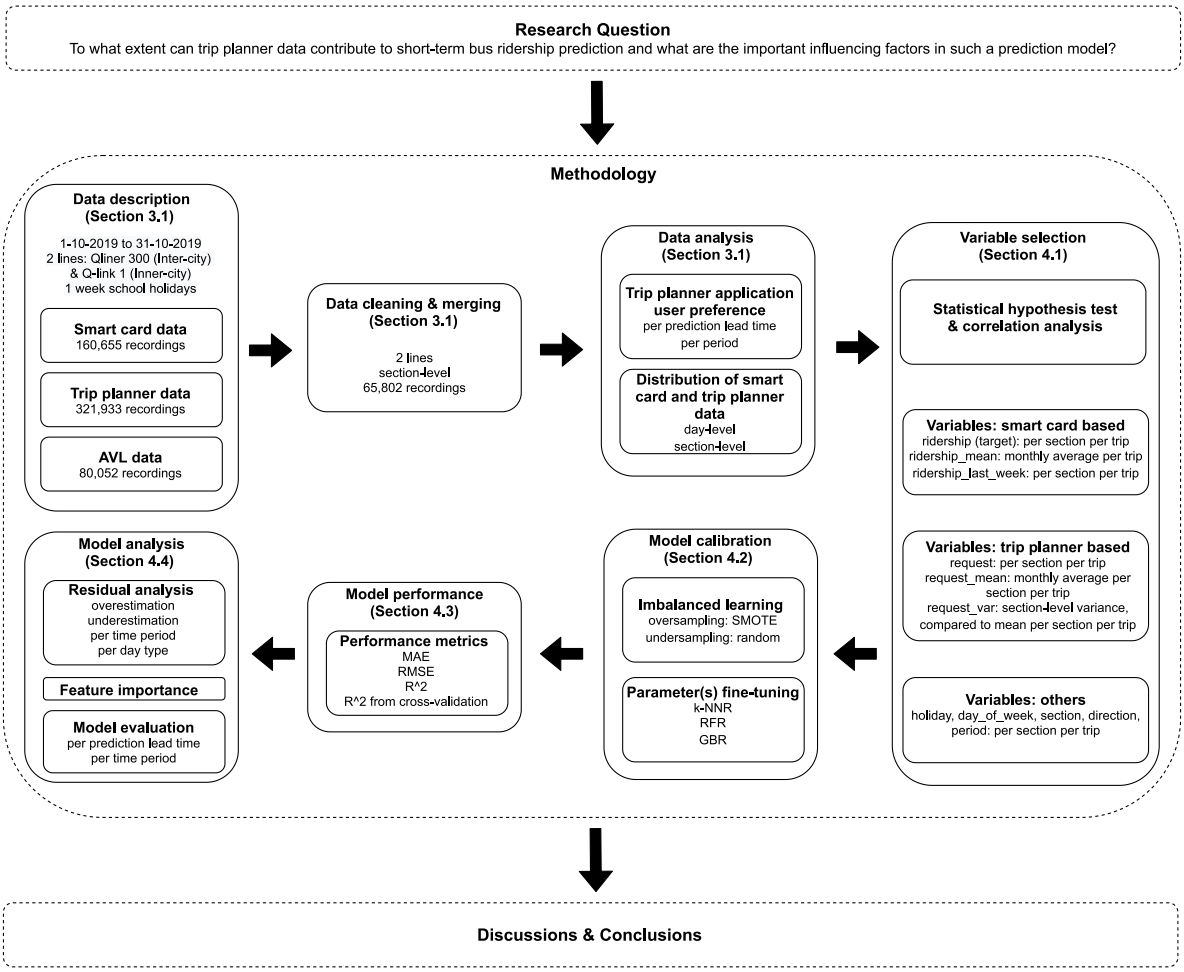

Fig. 4. Overview of the main research question and the proposed method.

First of all, three types of data have been introduced, cleaned, merged, and analyzed as above-mentioned. We unveil the user behavior and preference based on the data analysis to understand the effective prediction lead time. Moreover, we also observe the spatio-temporal influence on the ridership and the trip planner requests. By deriving the scatter plot of ridership and planner requests, there is a positive correlation between the actual passenger number on-board and the trip request asked in advance on a section level.

Second, the input variables are as important as the model itself for any ML model. The selection of variables would always be an iterative process to reach a higher model performance. We start with the variables based on the literature and data analysis. Then, a hypothesis test followed by a correlation analysis is performed to test the correlation between a specific variable and ridership (target). The insignificant variables are kept out to avoid redundancy. The target variable – ridership – is obtained from the AFC system, while we get the other input variables from datasets, including the AFC system, AVL system and trip planner request database.

Third, we establish three baseline models and three supervised ML models with trip planner data. The first baseline model is currently being adopted by PT operators. In such a model, the ridership of this week is estimated by the ridership of last week. The second one uses a multiplier to ridership to capture the weekly trend. This multiplier is calculated by the ridership of the day before divided by the ridership of the day before from last week. The last baseline model is the best-performing ML model among the chosen ML models without trip planner data. In particular, it only uses the AFC and AVL data, which a PT operator possesses.

There are four learning paradigms in ML, namely *supervised*, *unsupervised*, *semi-supervised* and *reinforcement learning*. We choose supervised learning in this study as it uses labeled training datasets to build the model and maps an input to the desired output based on example input–output pairs. For other paradigms, readers are referred to the work of Sarker (2021). Both regression and classification have been extensively applied for ridership prediction problems (Chiang et al., 2011; Xue et al., 2015; Ding et al., 2016; Zhou et al., 2016; Ohler et al., 2017; Karnberger and Antoniou, 2020). In this paper, we aim to forecast the number of passengers on-board in a specific section. It is a continuous quantity and is, therefore, a regression problem.

There are numerous models for regression, and there is no single model that can be the best for every scenario. Among the supervised models, we turn to more interpretable models suggested by Molnar (2022) because they can explain themselves so that we can know the importance of trip planner data in such a model. We decide to include the following interpretable ML models: *k-nearest neighbors regression (k-NNR)*, *random forest regression (RFR)*, and *gradient boosting regression (GBR)*. Furthermore, we first randomly split the data into two instances (i.e., training data and test data) with the 80/20 rule as a rule-of-thumb, also known as the Pareto Principle. The dependent variable (i.e., ridership) and independent variables (i.e., other input variables) have thus been assigned into two independent subsets, one for training and the other for testing. Then, we also split the data into different ratios, namely 90/10, 70/30, and 60/40. Different training/test strategies result in a stable outcome. Moreover, we report the results from the cross-validation to provide findings on the unbiased estimate of the model performance. For this reason, we mainly present the results from the 80/20 train-test split in the following section if not specified. After the split, 21,235 and 5309 rows of Qliner 300 data, as well as 31,406 and 7852 rows of Q-link 1 data, are for the ML models training and testing, respectively.

Fourth, we calibrate the model with a combination of oversampling and undersampling strategies. For quality and control reasons, a PT operator is typically more interested in crowded trips than quiet ones. Nonetheless, those trips are poorly represented by the data as they are rare in the observations. The conjunction of preference (crowded cases) and imbalance (more observations on the less crowded conditions) causes a degradation of the performance of the most desirable instances (Fernández et al., 2018a). The learning methods that we have chosen for this study (k-NNR, GBR, and RFR) do not give equal importance to the minority class as the majority class. Therefore, we resample the training data to tackle this issue (He and Ma, 2013). We randomly undersample the majorities and oversample the minorities by applying the *Synthetic Minority Oversampling Technique* (SMOTE) proposed by Chawla et al. (2002). SMOTE creates new instances of a minority class by using a k-NN approach. A random number of original observations are chosen and for each of their K neighbors, a new sample is created as a linear combination of the initial observation and its neighbor. Chawla et al. (2002) and Fernández et al. (2018b) indicate that a combination of SMOTE and undersampling performs the best.

Tuning hyperparameters of non-parametric regression algorithms (such as k-NNR, GBR, and RFR) is of importance because they do not rely on the assumed shape or parameters of the underlying population distribution. To this end, we perform *nested k-fold cross-validation* to calibrate the model and investigate the robustness of the model as this technique is able to avoid information leakage and significant bias, caused by applying *k-fold cross-validation* twice (Cawley and Talbot, 2010). Normally, the dataset is recommended to split up into k-partitions — 5 or 10 partitions as a rule of thumb (James et al., 2021). In the nested cross-validation, we implement *stratified 5-fold cross-validation* in the inner loop to equally capture the class while *random permutation 5-fold cross-validation* is adopted in the outer loop to approximate the reality (Kuhn and Johnson, 2013).

Lastly, we assess the model on a section level using the following metrics: *mean absolute error* (MAE), *root mean square error* ($RMSE$), *coefficient of determination* (R^2), and R^2 from cross-validation (Handelman et al., 2019). We also explore the importance of the chosen features in the best-performing model. Feature importance measures the relative importance of each feature when making a prediction by assigning scores to the input features (Kuhn and Johnson, 2013). In this study, we practice this technique to discover and quantify the usefulness of trip planner data. For k-NNR, GBR, and RFR, *permutation feature importance* can be utilized. This is computed by measuring the decrease in a model score when a single feature value is randomly shuffled (Breiman, 2001). Although it considers the interactions among features, it is computationally efficient and facilitates interpretation. For tree-based models, we also could use *mean decrease in impurity (MDI)* to investigate the feature importance. This method is easily understandable and computationally light, even though it is purely based on the training dataset and tends to inflate cardinality features (Louppe, 2014). The performance evaluation continues with comprehensive residual analysis from the aspects of over- and under-estimation and per temporal attribute. We wrap up the analysis with an investigation into the model performance with various prediction lead times in trip planner requests and during different periods.

4. ML models for ridership prediction with trip planner data

To predict the ridership on the section level, we begin with illustrating the selection of variables and present the correlation matrix calculated from the variables in Section 4.1. Next, Section 4.2 discusses the model calibration, including sampling design and model tuning. After that, we compare the model performance by the metrics in Section 4.3. Due to the potential interest of peak hours by any PT operators, we also present the model performance during congested periods in Section 4.4. Finally, Section 4.5 analyzes the best performing model further.

4.1. Correlation analysis and variable selection

Based on the data analysis and the literature, we list the variables considered on the section level in this study in Table 1. Other than the request-related variables, other variables have been exhaustively studied and proven their profoundness in the literature.

We derive the historical average of ridership and requests per trip based on the one-month smart card and trip planner data as this average number does not change substantially unless there is a service change or a huge and long-lasting incident. Furthermore, there are 50 sections coded as dummy variables for the two case study bus lines. We include the day of the week but are not considering each day of the week since it is not significant as aforementioned (Section 3.1). For the baseline ML model, we exclude all request-related variables from the list. Additionally, we put variables with prediction lead time into models in pairs. For instance, request and request_var would be a pair to see how the model performs when we have all the trip planner data available, while other variables with a specific prediction lead time would test how the model performs with fewer trip planner request data and

Table 1
List of variables in ML ridership prediction models with trip planner data on the section level.

Variable	Explanation	Category	Unit/Coding
Ridership (target)	The passengers on-board per section per trip	Numerical	Person
Ridership_mean	The monthly average of ridership per trip	Numerical	Person
Holiday	The autumn holiday	Categorical	One-hot encoding
Request[a]	The trip requests per section per trip	Numerical	Record
Request_mean	The monthly average of requests per section per trip	Numerical	Record
Request_var[a]	The section-level variance of requests, compared to the request_mean per section per trip	Numerical	Record
Day_of_week	Weekday or weekend	Categorical	One-hot encoding
Section	The section that a vehicle traverses during a trip	Categorical	One-hot encoding
Direction	The direction of a trip	Categorical	One-hot encoding
Time period	Morning peak or evening peak or off-peak hours	Categorical	One-hot encoding
Ridership_last_week	The passengers on-board of the same trip last week per section per trip	Numerical	Person
Request_10[a]	The trip requests that are sent 10 min ahead per section per trip	Numerical	Record
Request_var_10[a]	The section-level variance of requests that are sent 10 min ahead, compared to the request_mean per section per trip	Numerical	Record
Request_15[a]	The trip requests that are sent 15 min ahead per section per trip	Numerical	Record
Request_var_15[a]	The section-level variance of requests that are sent 15 min ahead, compared to the request_mean per section per trip	Numerical	Record
Request_30[a]	The trip requests that are sent 30 min ahead per section per trip	Numerical	Record
Request_var_30[a]	The section-level variance of requests that are sent 30 min ahead, compared to the request_mean per section per trip	Numerical	Record

[a]Variables with this symbol will be put into models in pairs, e.g., request and request_var.

when the prediction is performed for further ahead in time. Note that the variable "time period" means the morning (weekdays, 6:30 to 9:00 AM) or the evening peak (weekdays, 4:00 to 6:30 PM), or the off-peak hours, according to the Dutch principle train operator NS.[5]

With the considered input variables, we first run a two-sample T-test to determine whether there is statistical evidence that shows the associated population means are significantly different. Since we are only interested in the relationship between ridership and other variables, we run the test in pairs with the ridership. The result shows that only ridership and the historical ridership come from the same population, which is rational. All other variables and the ridership are statistically different with a zero p-value.

After the hypothesis test, the correlation matrix based on case study lines is presented in Fig. 5. Note that we have excluded the visualization of the variable "section" because of its overwhelming number. It shows the normalized spread and deviation by the variance on the diagonal and the dependency between two variables by the normalized covariance in other cells. In this way, variables with different units and scales become comparable. Before normalizing, the variance of ridership has a large span as high as 125.42, which means the prediction is valuable. As a predictor variable, the variance of requests also shows a high value of 240.12, which indicates that a better regression model can be expected.

The most influencing factors are ridership-related and request-related. The largest three positive correlations are seen in the monthly average of ridership, the number of requests, and the ridership of last week. It confirms that the logged smart card data is

[5] https://www.ns.nl/binaries/_ht_1449754213072/content/assets/ns-en/22471_nsr_brochure_travelling_by_train_a5.pdf

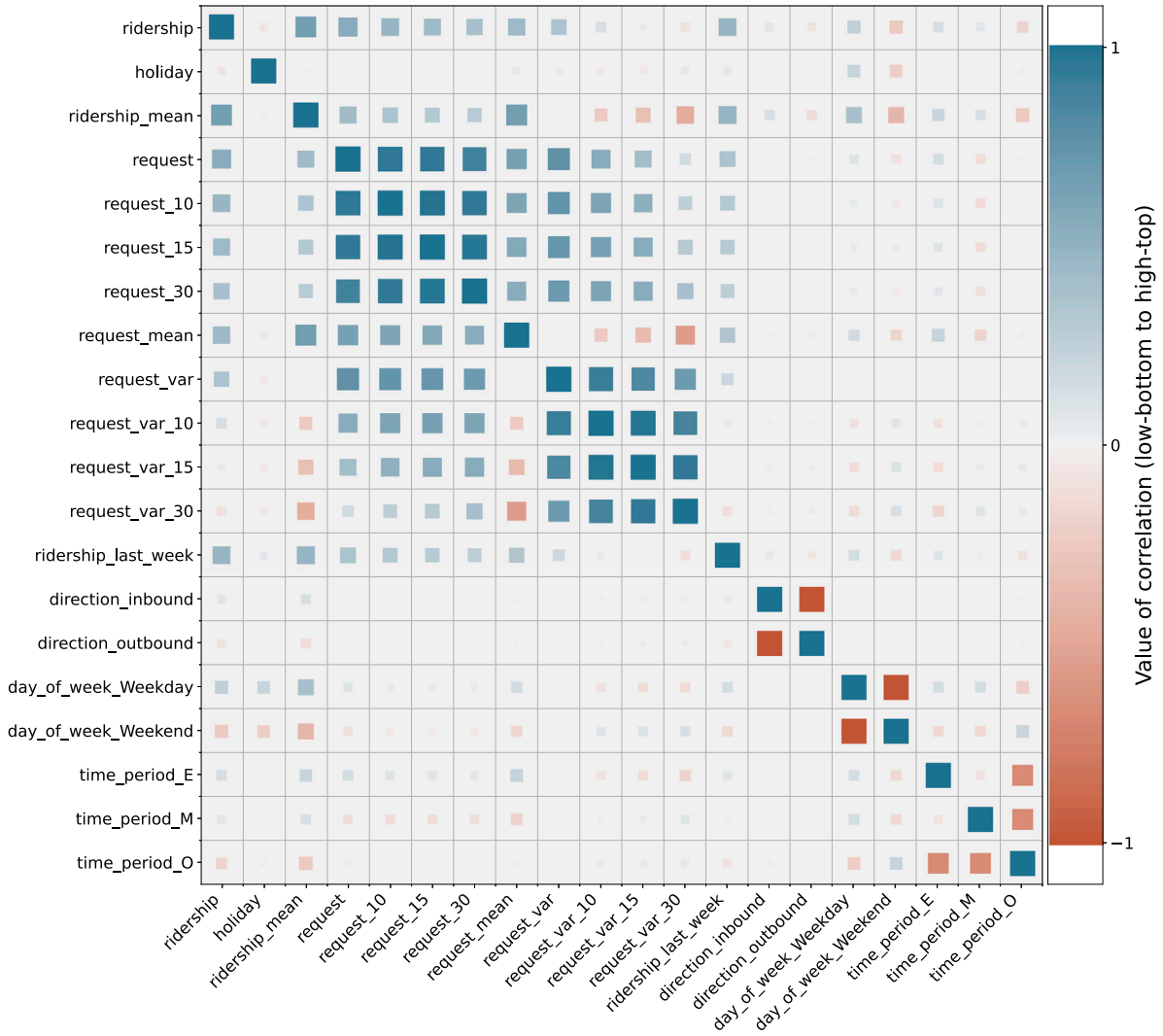

Fig. 5. Correlation matrix on the section level.

a sound basis for ridership prediction and the potential of using trip planner data for predicting ridership based on a strong positive correlation between the ridership and requests. Moreover, almost all covariances of trip planner variables with prediction lead times are strongly positive, except for the variance of request with a 30-min prediction lead time. This indicates that the travel purpose and behavior do not change drastically when we consider further ahead in time. It is worthwhile noticing that this correlation decreases continuously when we focus on further ahead in time. However, we also see that both line characteristics and temporal variables have unexpectedly small correlations. This implies that temporal and spatial influences on ridership are marginal and could already be inherently reflected by the historical smart card data or the request-related variables, such as the variance and the mean. But we keep those in the prediction model as extensive literature proves their influences (Chiang et al., 2011; Xue et al., 2015; Chakour and Eluru, 2016; Ding et al., 2016; Ohler et al., 2017; Karnberger and Antoniou, 2020).

4.2. Model calibration

We develop a pair-wise study with 4 undersampling and 6 oversampling strategies on Random Forest Regressors to seek the optimal combination. The Random Forest Regressor is chosen as it tends to focus more on the prediction accuracy of the majority class, which often results in low accuracy for the minority class (Khoshgoftaar et al., 2007). Fig. 6 presents the min–max scaled R^2 results of sampling design of each case study line through 5-fold cross-validation. From a maximizing model calibration perspective, an oversampling ratio in the interval 1–1.5 without undersampling is found to be optimal from both case study lines through a 10-time experiment. Since PT operators are more concerned with the congested cases, we therefore have a slight preference towards

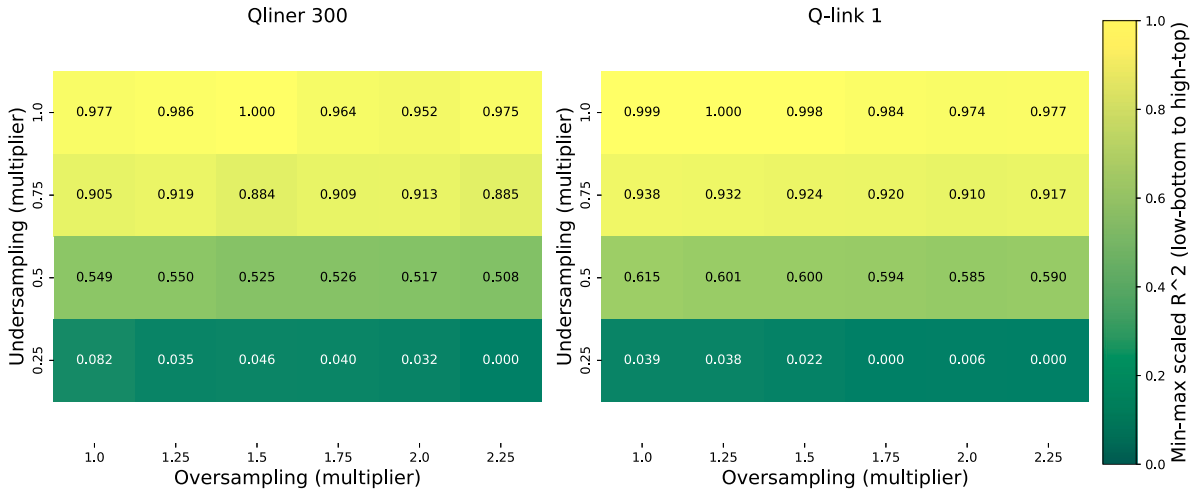

Fig. 6. Evaluation of different sampling designs per line.

the upper end of the interval. For that reason, we choose a value of 1.5 for oversampling to make the representative groups more profound.

Given the optimal sampling design, we present the optimal hyper-parameters in Table 2 for both bus lines. R^2 of the training dataset is optimized by cross-validated grid-search over a pre-defined parameter grid. For tree-based models, a process called regularization can help to use hyper-parameters to control the structure of the decision tree-based models. As for k-NNR, the only hyper-parameter we need to tune is the nearest K, which is calculated by conducting a sensitivity analysis of different K based on the Euclidean distance. Besides, we observe that the parameters obtained in each fold do not vary dramatically when calibrating. This means that this type of model does not require regular retraining, unless there is a change of travel or trip planner user behavior, e.g., PT service change or pandemic.

4.3. Model performance: comparison of models

With the tuned hyper-parameters of the models and the optimal resampling strategies, we present the overall performance of the prediction models of Qliner 300 and Q-link 1 in Tables 3 and 4, respectively. Note that the resampling strategies are only executed on the training data, whereas the model evaluation is carried out against unsampled test data.

Compared to the baseline model that PT operators currently adopt, ML models have effectively elevated the prediction performance. In particular, the results from the RFR without trip planner data have already improved the model by half on average. For the baseline model with weekly trend, the performance was shockingly low with negative R^2 values. It means that a simple mean would work better than these two models, which indicates the failure of these models to find any meaningful relationship between the input and output.

Regarding the ML models with trip planner data, RFR outperforms the others, as shown by every dimension in the metrics in both cases. The R^2 from repeated random 5-fold cross-validation of Qliner 300 tells that the GBR has almost the same score as RFR. However, RFR outperforms GBR by the other three metrics. For Q-link 1, RFR significantly improves the prediction accuracy compared to other models. Most importantly, both GBR and RFR with trip planner data outrun the baseline RFR that does not contain trip planner data, particularly in the Q-link 1 case. Compared with the RFR without trip planner data, the inclusion of trip planner data improves the prediction model performance by around 7.8% for Qliner 300 and 18.5% for Q-link 1, reflected by the R^2 from cross-validation. Given the baseline RFR model has already significantly enhanced the short-term prediction, the incorporation of trip planner data undeniably further improves it. Additionally, the prediction performance also increases with the abundance of data.

Figs. 7 and 8 show the scatter plots of the predicted and actual values of Qliner 300 and Q-link 1, subsequently. ML models can mostly capture quiet trips as well as beat the baseline models for busy trips. Notwithstanding, we observe the existence of heteroscedasticity as all models have higher average error and bias along with the rise of ridership. When the value of ridership is low, all models can function efficiently, where GBR tends to overestimate the prediction and k-NNR tends to underestimate while RFR is relatively neutral. Comparing the RFR model with trip planner data and the baseline RFR without trip planner data, the magnitude of the error is lower, indicating a better capture of the variation of the ridership in using such data. It is worth mentioning that by using trip planner data, a much more reliable prediction can be expected in the busy trips since the magnitude of average prediction error decreases by approximately 20 persons for Qliner 300 and roughly 30 persons for Q-link 1. This prediction area is also conjointly focused by PT operators.

Table 2
Optimal hyper-parameters of the non-parametric models.

	Qliner 300	Q-link 1
GBR	learning_rate = 0.01 n_estimators[a] = 12000 max_depth[b] = 4 min_samples_split[c] = 2 min_samples_leaf[d] = 3 subsample[e] = 1 max_features[f] = 10	learning_rate = 0.02 n_estimators = 23000 max_depth = 4 min_samples_split = 2 min_samples_leaf = 10 subsample = 1 max_features = 15
k-NNR	n_neighbors[g] = 14	n_neighbors = 10
RFR without trip planner data	bootstrap[h] = True n_estimators = 1000 max_depth = 18 min_samples_split = 2 min_samples_leaf = 8 max_features = 12	bootstrap = True n_estimators = 800 max_depth = 40 min_samples_split = 2 min_samples_leaf = 3 max_features = 19
RFR with trip planner data	bootstrap = False n_estimators = 800 max_depth = 18 min_samples_split = 2 min_samples_leaf = 4 max_features = 11	bootstrap = False n_estimators = 1000 max_depth = 28 min_samples_split = 2 min_samples_leaf = 2 max_features = 18

[a]Number of trees.
[b]The maximum depth of a tree.
[c]The minimal number of samples in a node for the node to be split.
[d]The minimum number of samples in a leaf node.
[e]The fraction of samples to be used for fitting the individual base learners.
[f]The number of features randomly chosen as candidates for a split.
[g]Number of neighbors to use.
[h]Whether bootstrap samples are used when building trees. If "False", the whole dataset is used to build each tree.

Table 3
Performance of short-term prediction models (Qliner 300).

Qliner 300	MAE (Person)	RMSE (Person)	R^2	R^2 from repeated random 5-fold cross-validation
Baseline	5.761	9.060	0.461	–
Baseline with weekly trend	10.721	30.408	−6.179	–
Baseline RFR without trip planner data	4.494	6.867	0.644	0.669
GBR	4.449	6.895	0.641	0.717
k-NNR	4.624	7.125	0.617	0.663
RFR with trip planner data	**4.093**	**6.237**	**0.707**	**0.721**

Table 4
Performance of short-term prediction models (Q-link 1).

Q-link 1	MAE (Person)	RMSE (Person)	R^2	R^2 from repeated random 5-fold cross-validation
Baseline	7.744	12.588	0.287	–
Baseline with weekly trend	9.920	26.071	−2.049	–
Baseline RFR without trip planner data	5.652	9.234	0.633	0.691
GBR	4.984	7.938	0.729	0.792
k-NNR	5.361	8.685	0.676	0.682
RFR with trip planner data	**4.235**	**6.901**	**0.806**	**0.819**

4.4. Model performance: comparison of models during peak hours

Due to the strong interest of PT operators, we also compare the selected models during peak hours as shown in Tables 5 and 6, sequentially. In spite of the type, the model performance drops during peak hours as the observations are fewer, and therefore the correlation between the selected variables and the ridership is hard to be captured. Among all the models, RFR with trip planner data still outperforms noticeably. Compared with the baseline RFR without trip planner data, we see an improvement in the prediction by 4.4% for Qliner 300 and 22.5% for Q-link 1 through measuring R^2 from cross-validation. The elevation in prediction during peak hours is more profound for Q-link 1 but less remarkable in Qliner 300. However, the inclusion of trip planner in GBR and k-NNR is less significant compared to the overall performance. Whereas the other baseline models worsen to a large extend, ML models

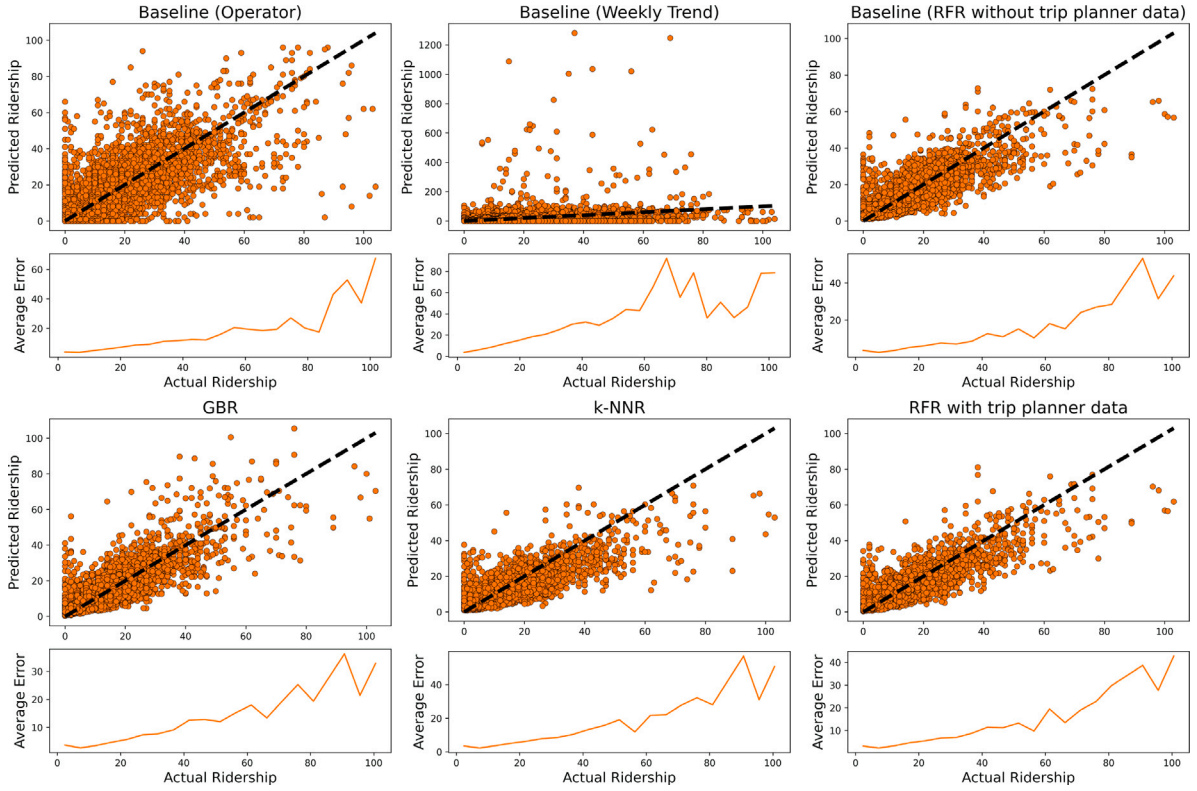

Fig. 7. Prediction vs. actuality plot of Qliner 300.

Table 5
Performance of short-term prediction models during peak hours (Qliner 300).

Qliner 300	MAE (Person)	RMSE (Person)	R^2	R^2 from repeated random 5-fold cross-validation
Baseline	7.744	12.215	0.389	–
Baseline with weekly trend	16.149	49.142	−11.286	–
Baseline RFR without trip planner data	5.550	8.528	0.650	0.662
GBR	5.574	8.819	0.626	0.680
k-NNR	5.916	9.271	0.587	0.636
RFR with trip planner data	**5.399**	**8.430**	**0.658**	**0.691**

Table 6
Performance of short-term prediction models during peak hours (Q-link 1).

Q-link 1	MAE (Person)	RMSE (Person)	R^2	R^2 from repeated random 5-fold cross-validation
Baseline	9.898	15.584	0.181	–
Baseline with weekly trend	14.553	47.181	−6.408	–
Baseline RFR without trip planner data	6.603	10.586	0.639	0.645
GBR	6.453	10.318	0.657	0.759
k-NNR	6.815	10.829	0.622	0.637
RFR with trip planner data	**5.445**	**8.774**	**0.752**	**0.790**

remain relatively stable. The results further support the usage of trip planner data in the RFR when the recorded historical data is fewer.

4.5. Model performance: further analysis of RFR model

Residual analysis — over- and under-estimation. The best performing model – RFR with trip planner data – is used to investigate the residuals, especially the over- and under-estimation of the prediction. Over- and under-estimation is of interest as they cause

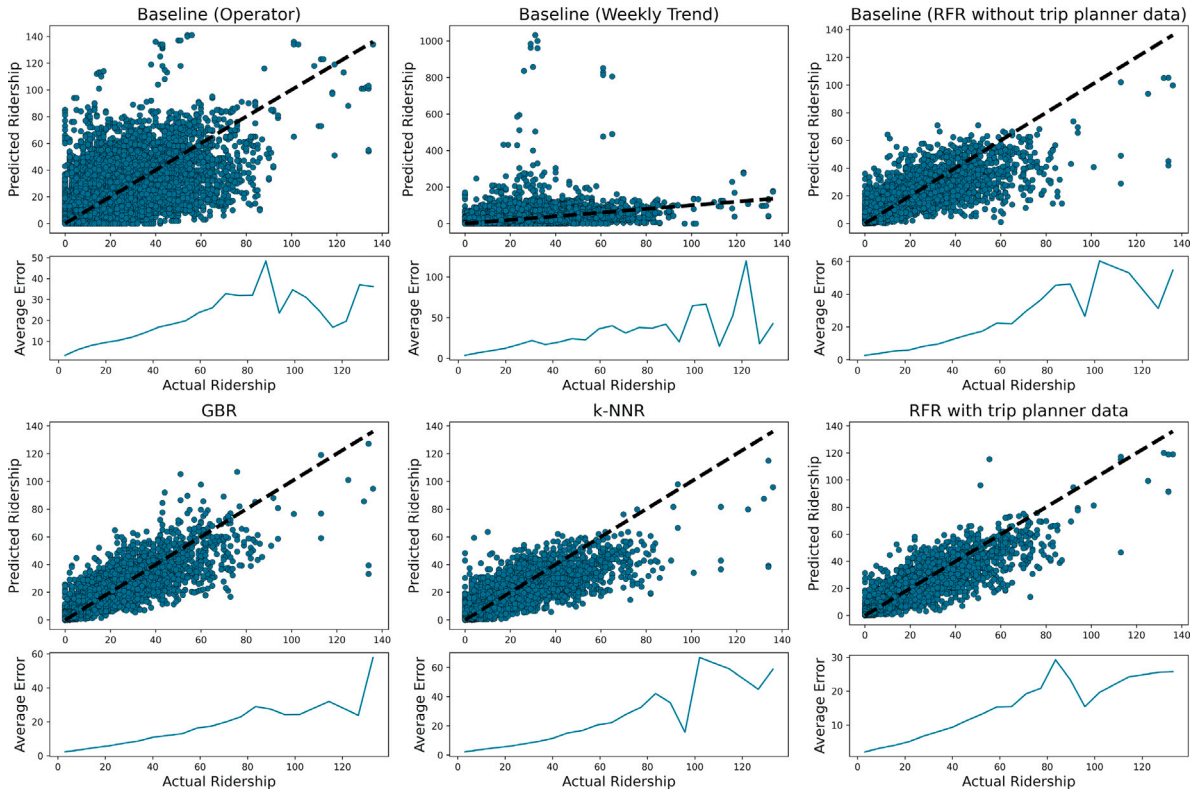

Fig. 8. Prediction vs. actuality plot of Q-link 1.

Table 7
Over- and under-estimation based on residual analysis of RFR with trip planner data.

	Percentage		95th percentile absolute error (person)		Average of top 5 percentile absolute error (person)	
	Overestimation	Underestimation	Overestimation	Underestimation	Overestimation	Underestimation
Qliner 300	46.450%	53.550%	11.557	13.730	18.022	20.186
Q-link 1	50.955%	49.045%	11.534	18.173	17.446	26.338

differences in the PT operations. Specifically, overestimation of ridership leads to wasted supply while underestimation results in a lower level of service. The analysis of the over- and under-estimation with residuals of RFR is presented in Table 7.

Qliner 300 has more underestimation than overestimation and the difference between them is around 7%. Q-link 1 is rather balanced with a similar percentage of over- and under-estimation of the ridership prediction. Both cases have a tendency to underestimate the actual values, and this tendency is much more profound when we see the average of the top 5 percentile error.

Residual analysis — Times of the day and day type. The residuals of prediction vary with period and day type, and therefore we present the correspondingly residual plots of Qliner 300 and Q-link 1 in Figs. 9 and 10. For Qliner 300, the model has a high variance of residuals during both the morning and evening peaks with evening slightly higher. This matches the nature of Qliner 300 as it is a fast-service limited-stop line, and passengers flow into this line during peak hours, especially evening with a dispersed passenger flow. This results in higher variance and predictive difficulty. The influence of non-uniform off-duty time during the evening peak is much more significant for Q-link 1, in which the highest residual variance can be observed. The off-peak category of both case study lines also presents a high variance in residuals because of the heteroscedasticity. This heteroscedasticity could be from missing important independent variables (which we will discuss in the following section) or the data type that has been applied in the study, for instance, the cross-sectional study that is applied in this paper. Concerning day type, the variance of the weekday is higher than the weekends for both lines, which is in line with the rational passenger behavior.

Feature importance. We report the feature importance of RFR on Qliner 300 and Q-link 1 in Figs. 11 and 12, separately. Since RFR is a tree-based model, both permutation feature importance and MDI can be applied to calculate the feature importance, and we are able to compare the results from both approaches to gain a complete understanding of the feature importance. Thus, we carry out the MDI feature importance on the training set and implement the permutation feature importance on the test set. Note that, the trip planner requests in this scenario are without any prediction lead time.

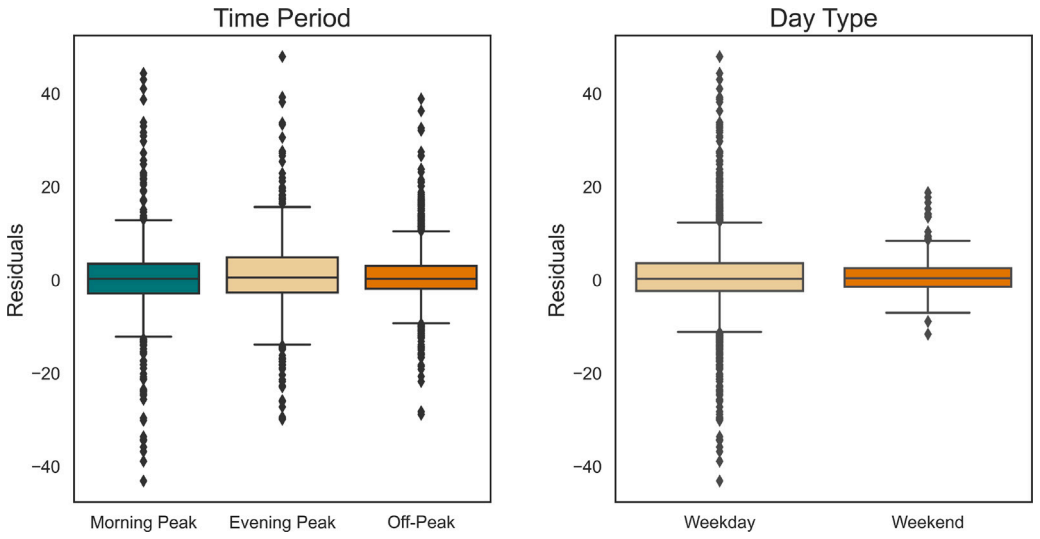

Fig. 9. Residuals of Qliner 300 per time of the day and per day type (RFR with trip planner data).

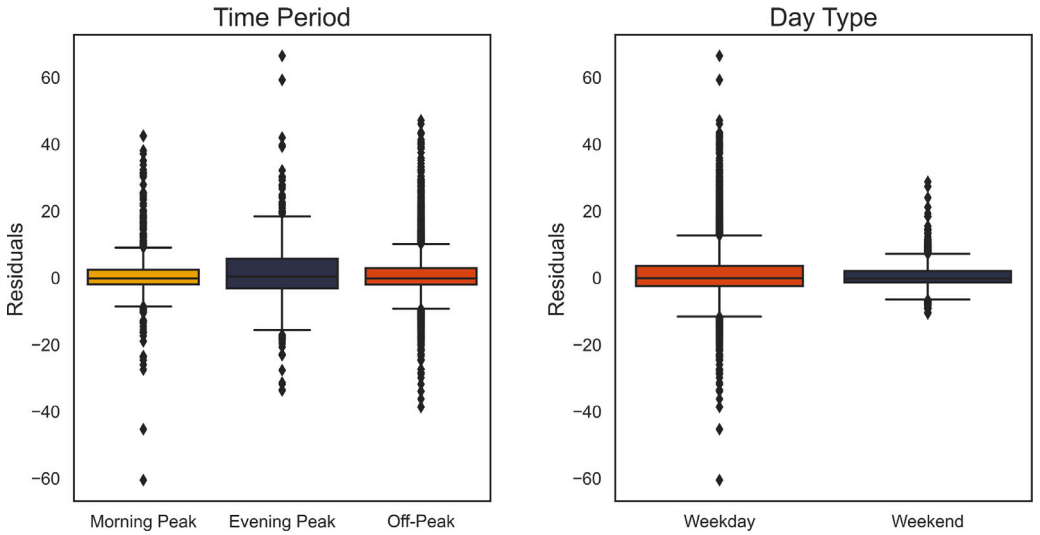

Fig. 10. Residuals of Q-link 1 per time of the day and per day type (RFR with trip planner data).

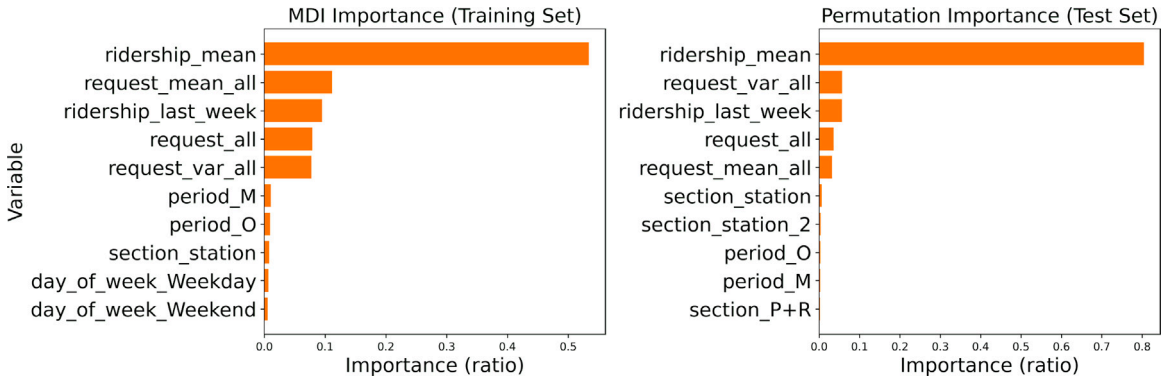

Fig. 11. Feature importance of Qliner 300 (RFR with trip planner data).

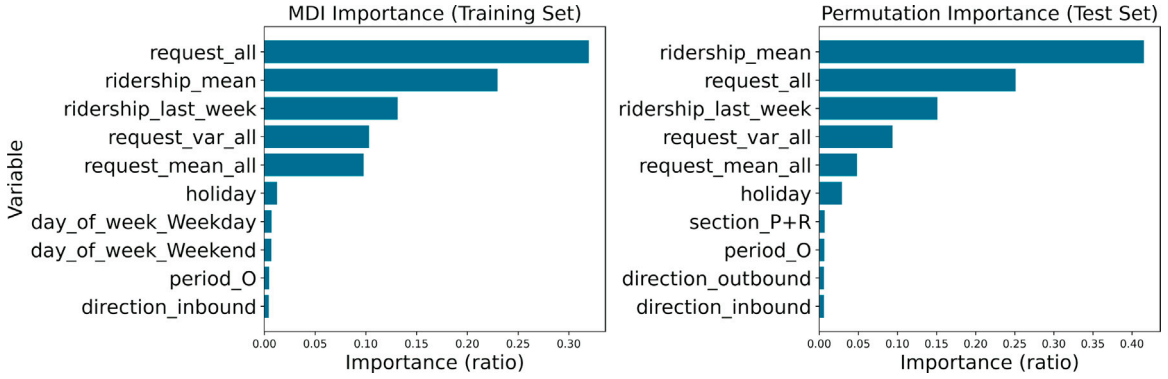

Fig. 12. Feature importance of Q-link 1 (RFR with trip planner data).

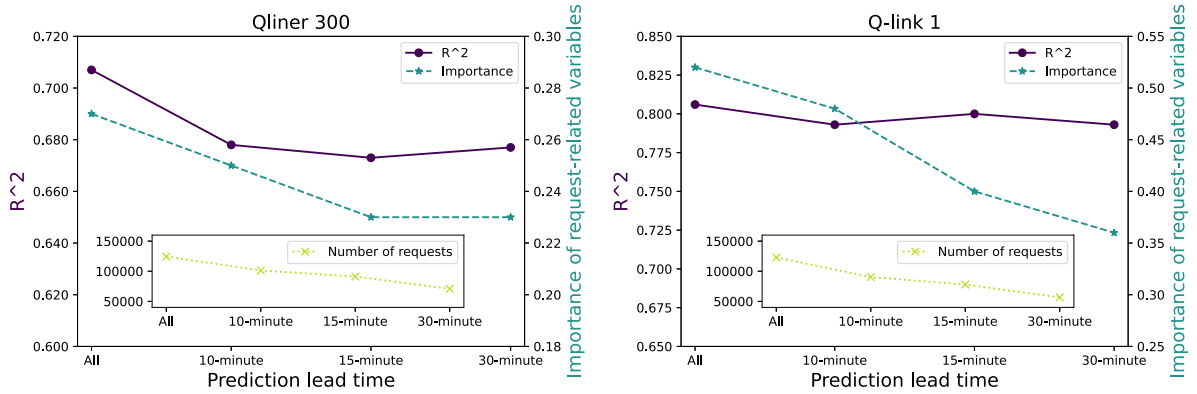

Fig. 13. RFR model performance with different prediction lead times stored in the trip planner requests.

Regardless of the feature importance measurement method, the first five contributing features are the historical average of ridership, the ridership of last week, the number of requests, the average number of requests, and the variance of requests. Consistent with the literature, smart card data is undoubtedly the most critical data source. Specifically, the monthly average of ridership is usually the most contributing variable. Notably, the request-related variables are essential in Q-link 1 where we see almost 50% importance on average. In particular, the MDI importance of Q-link 1 demonstrates that the number of requests supports the prediction the most with approximately 35% of the importance score. In contrast, the total average of request-related variables is only about 20% importance for Qliner 300. Still, at least one of the request-related variables often ranks within the first three substantial variables, such as the average number of requests in the MDI importance and the variance of requests in the permutation importance. In line with the correlation analysis, both temporal and spatial variables have low feature importance for the predictions. We envision that the change in trip planner requests and historical ridership in our model already reflects the impact of spatial–temporal factors. These factors strongly influence the number of requests and ridership as the data analysis indicates, but they contribute less once we have a much more direct source in the prediction model, such as smart card data or trip planner data. Our feature importance analysis also backs up the argument that the impurity-based feature importance can inflate the importance of numerical features (Strobl et al., 2007).

Model performance with prediction lead times in trip planner data. We further analyze the performance of RFR with trip planner data by using the same configuration and the same sampling design but with different prediction lead times in trip planner requests. In other words, how the model performs with trip planner data that is further ahead of time, such as 10 min, 15 min, and 30 min before the vehicle start time. Fig. 13 displays the number of requests per prediction lead time per scenario (inset), the model performance based on R^2 per scenario (solid line with circle marker), and the total feature importance of the request-related variables per scenario (dashed line with star marker). In each scenario, the request-related variables contain the number, the monthly average, and the variance of requests.

By using trip planner data with different prediction lead times, the performance of RFR remains practically stable. Although the amount of data and the correlation between the request and ridership drop, the model functions essentially the same, reflected by the R^2 value. Among all scenarios with a prediction lead time of Qliner 300, the model with requests sent only 10-min in advance is the best as it is the closest time to the vehicle start time, containing most of the up-to-date travel intentions. However, unlike Qliner 300, Q-link 1 has the scenario of 15-min ahead of time as the best-performing. It accordingly shows that people probably change

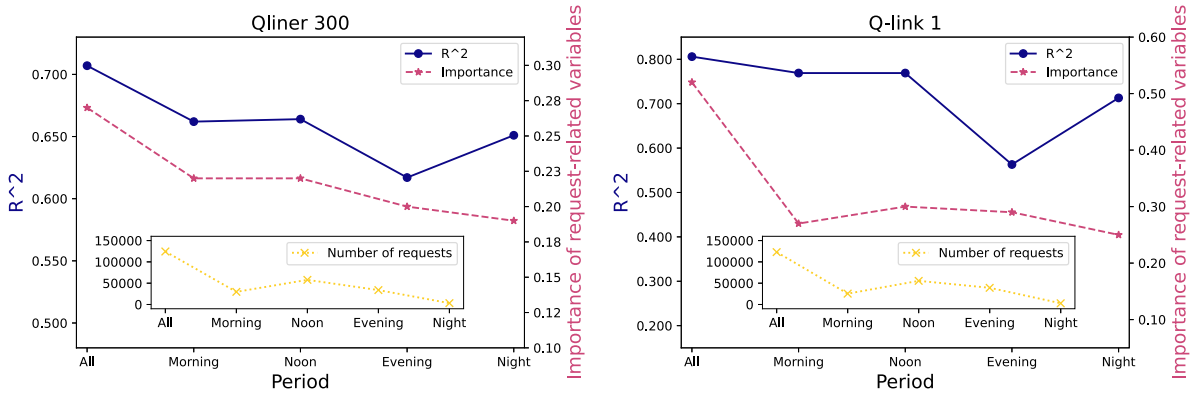

Fig. 14. RFR model performance with the trip planner requests sent from different periods.

their travel behavior when travel with such a line from 15-min to 10-min ahead. Thus, depending on the line characteristics, people opt for using such a trip planner differently. Sometimes, when it comes close to the vehicle start time, it has an adverse effect on the prediction model.

The further ahead in time, both the importance of request number and the average number of requests decrease. Nonetheless, there is an increase in the importance of the request variance, which leads to the total feature importance of request-related variables remaining the same.

Model performance with trip requests sent from different periods. The data analysis reveals that people behave differently during different times of the day when using a trip planner (Section 3.1). Thus, we execute the model again with the same optimal hyper-parameters and sampling design, but we investigate the performance of RFR by leveraging trip requests sent from different periods. Each period has the same horizon of 6 h, e.g., morning is from 4:00 to 10:00. Fig. 14 exhibits the results with the same layout as Fig. 13.

The model performance and the summed request-related importance have the same trend as the number of requests. The best performances are seen from 10:00 to 16:00 when the request number is the largest, compared to all the requests enclosed. Users send very few requests at night (from 22:00 to 4:00), which is around 5% of all requests. But the performance does not degrade dramatically. In contrast, it is during the evening (from 16:00 to 22:00) when RFR performance deteriorates for both cases. It means that during the evening, the relationship between the variables and ridership becomes more complicated. Concerning the feature importance of request-related variables, the request importance drops sharply when the number of requests is low. The role of average and variance of the request stands out more, while the number of requests tends to be less influencing.

5. Discussion

In this study, we investigated how trip planner data can contribute to the short-term prediction of bus ridership and built 3 baseline models without trip planner data and 3 supervised ML models with trip planner data. We unveiled that the incorporation of trip planner data can improve ridership prediction reliability, particularly in busy trips where observations are fewer. Trip planner data can provide feature importance up to 50% in the best-performing model (RFR), which can reduce the mean absolute error by approximately half, compared to the baseline model that is established by the weekly trend. However, in this section, we will discuss the limitations and further opportunities for this research domain.

Data availability. Broader studies can be carried out if more information about trip planner data is available, and if the privacy concerns about the trip planner application can be reduced. It is unknown whether it is a single journey with multiple legs or a group of people traveling with one trip planner request. Therefore, if the user ID is available, we can reduce the underestimation of trip planner's importance. Moreover, knowing the alternatives provided for a piece of trip advice can be meaningful so that we can study the user preference and behavior. Besides, distinguishing between the user type will be advantageous to understand the travel preference among different kinds of travelers. Also, with this additional information, we can gain insights into how often different types of people make requests. Lastly, two types of events could significantly influence the ridership in PT, i.e., an accident/breakdown (which cannot be predicted) and predictive events (such as concerts, sports events, markets, school holidays). In our data, we have a one-week school holiday, which lies in the scope of the latter, and 9292 has already been investigating extensively in this regard through correlation analysis. However, how could this type of data contribute to sudden events is not yet well understood.

Partial observability. The AFC system has a penetration rate of approximately 93% in the Netherlands with both tap-in and tap-out at the ticket machines. Only 5% of the passengers travel with a paper ticket, and 2% of the fares are evaded. This makes the smart card data a relatively comprehensive observation of the passengers on-board in our paper. On the other hand, the trip planner 9292 has a market share of roughly 46%. Although it is the most representative one in the Netherlands, there are also other competitive companies and passengers who travel without using such an app, which means that only some part of the trip planner data correlate with the smart card data and consequently becomes partially observable. The penetration rate of the AFC system and the usage of the trip planner are case-study specific. Hence, we recommend the PT operators to regularly and manually validate the ridership observation and the trip planner companies to be aware of the usage statistics for a more accurate ridership prediction.

Endogenous ridership data. We have derived the variable of the historical average of ridership based on the one-month smart card data, and we have used the same data source to acquire the prediction target in training and test data. In addition, the variable of ridership last week (i.e., ridership of the same trip at a specific section from last week) also comes from the same data, for which the time is sufficient for a PT operator to retrieve. We envisage that the monthly average number of ridership remains almost unchanged except if the PT operators offer different services or a drastic operational incident occurs. Incorporating endogeneity may have biases on the prediction results, and thus it is recommended to get the historical average based on other past months' data or the same month last year when possible.

Error weighting and data imbalance. Note that if we calibrate the model based on a random sample from all observations, then naturally the minimization of errors under regularly observed ridership conditions (e.g., less busy conditions) will outweigh the minimization of errors under rarely observed ridership conditions (e.g., very busy conditions). This is not necessarily optimal from the perspective of the PT operator who may wish to prioritize having good predictions specifically for irregular situations. This can be solved (directly) by using error-weighting in the calibration process or (indirectly) by using non-random sampling. The latter is applied herein, where the sparse data in the high-value domain is oversampled. The calibrated model is then evaluated with the unbiased sample. The optimal sampling design was based on a pair-wise study with several oversampling and undersampling strategies. From a maximum calibration perspective, the optimal interval lies in 1.0 to 1.5 oversampling without undersampling. In this paper, we have a slight preference towards 1.5 due to the preference of PT operators. Aside from the popular sampling method to deal with the imbalanced data that we have applied in this paper, future studies could also apply algorithm centered approach or even hybrid approach as explained in Kaur et al. (2019).

Enhancement of baseline models. Several failed models can be substantially improved, including the baseline model and the baseline model with the weekly trend (currently used in practice). First, missing recordings were notable, which results in a deterioration of the model performance. Second, the baseline model with weekly trend was strongly biased due to the high weekly multiplier on several sections of last week. Smoothing can be added to elevate the model by considering the trip of yesterday or other factors. Regardless, this advises PT operators to renew their current prediction models.

Inclusion of more significant variables. The inclusion of new significant variables should be considered for improving the predictions. During the residual analysis, we discovered that the existence of heteroscedasticity in the prediction. Since we have already transformed the variable, the other solution for improving it is to add more contributing variables so that the model can better capture the relationship between the target and the independent variables. For instance, the two case studies in this paper are less frequent, and therefore a variable of headway is kept out, but further studies on frequent lines with a headway shorter than 12 min should perhaps consider this (Van Oort and Van Nes, 2009). Such as incidents or special events, these kinds of variable can be incorporated into the model to better capture the variability of ridership during special times if applicable. From the desirable data perspective, user-related features could be added if allowed, such as the user IP address, the subscription type, OD spatial characteristics (unencrypted door-to-door travel). In terms of available data, data cleaning and merging could be more accurate if PT operators, smart card data management firms, and trip planner companies share the same naming system.

Findings on feature importance. In this study, the monthly average of ridership was contributing significantly, which supports the literature that it is a sound basis for the ridership prediction. However, the temporal and spatial feature importances were minor, which is contradictory to previous studies. We consider that this is because the influence of spatio-temporal features is partially covered by the fluctuation of the monthly ridership average and the trip request. Nevertheless, their impacts are still critical in any research that tries to understand the correlation between ridership and environmental factors as the data analysis revealed.

Retraining of ML models. Given the prediction improvement, ML models are significantly better than the baseline models that are currently practiced by the PT authorities. Nonetheless, ML models are complex to calibrate and implement, particularly when it relates to re-calibration. In this study, we found that the parameters did not vary considerably in the selected models when calibrating and thus suggest less frequent retraining when applying. Although the performance will be jeopardized when a change of travel behavior happens, the chosen interpretable and supervised ML models are comparably simpler to be retrained.

Deep Learning (DL) models. The primary goal of this paper is to explore and analyze whether the inclusion of a novel data source can add value to a short-term bus ridership prediction method, namely travel planner requests. Consequently, we have not opted for many state-of-the-art DL methods (such as Fully Connected Neural Network, Long Short Term Memory, Convolutional Neural Network), although ensemble models are also effective in prediction, e.g., Ahmad et al. (2017), Nawar and Mouazen (2017). However, future studies could apply these techniques to predict the short-term bus ridership with trip planner data since we have shed light on the usefulness of the trip planner requests, such as the method applied in Hao et al. (2019) or Zhang et al. (2020).

6. Conclusion

In this paper, we proposed a method to analyze the effectiveness of trip planner data in predicting short-term bus ridership. The presented method explained the use of real-time transit information followed by apps on an operational level. Such information could help PT operators cope with short-term passenger demand and could facilitate the trip planner to notify its users about the crowdedness level.

The case studies showed that the best-performing model relied on the RFR can reduce the mean absolute error by almost half, compared to a baseline model based on the weekly trend. The incorporation of trip planner data can considerably improve the prediction performance by comparing with the same model without trip planner data, in particular when the observations are fewer. Those busy trips are more interested by PT operators and therefore need such insights. The RFR model with trip planner data generally reached a balanced estimation, and the temporal variation of the prediction was in line with the temporal variation of the ridership with specific line characteristics. Moreover, the model performance was maintained even for trip planner requests with prediction lead times up to 30 min ahead, and for different periods of the day. Regardless of the measurement approach, the trip planner data could roughly have a 35% feature importance on average. Afterward, discussions were made to further explore the capability of trip planner data.

The paper has reported that, it is novel and useful to combine trip planner data and the historical ridership data to realize the short-term ridership prediction as trip planner data is often available in (near) real-time. In this way, we can substantially avoid the long collection time of smart card data and we are able to capture the temporal and spatial influence such that it can enhance the operational performance of transit operators. More importantly, the traditional data sources (e.g., AFC, AVL and GSM) do not contain the travel intentions since they are by definition logging data. Whereas the trip planner data can represent such purposes as the apps are normally used before the trip.

Although the study takes place in the Netherlands with the local trip planner, it is scalable and portable to many other case studies with a similar data provision. Nevertheless, the application and choice of prediction lead times would not only depend on the user behavior of the trip planner but also rely on the network scale and the transport mode. The trip planner is and will become more and more beneficial to PT users. Accordingly, millions of trip planner data provide a unique and real-time huge data source with the knowledge of user behavior. In particular, the travel intention stored in the trip planner requests could correlate to the latent demand and thus trigger the change of the PT operations on a tactical and strategical level. If we can minimize the privacy concerns and other technical limitations, PT operators and researchers will offer a better understanding of the role of the trip planner in ridership prediction and facilitate the operations of the PT system and improve its level of service.

CRediT authorship contribution statement

Ziyulong Wang: Conceptualization, Methodology, Data curation, Software, Visualization, Writing – original draft. **Adam J. Pel:** Methodology, Visualization, Supervision, Writing – review & editing. **Trivik Verma:** Methodology, Visualization, Writing – review & editing. **Panchamy Krishnakumari:** Methodology, Visualization, Writing – review & editing. **Peter van Brakel:** Resources, Project administration. **Niels van Oort:** Funding acquisition, Resources, Project administration, Writing review & editing.

Acknowledgments

The research leading to these results has received funding and data from REISinformatiegroep B.V. (9292). We also thank OV-bureau Groningen and Drenthe for providing the smart card data.

References

Aguiléra, V., Allio, S., Benezech, V., Combes, F., Milion, C., 2014. Using cell phone data to measure quality of service and passenger flows of Paris transit system. Transp. Res. C 43, 198–211.

Ahmad, M.W., Mourshed, M., Rezgui, Y., 2017. Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build. 147, 77–89.

Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., Dera, D., 2017. Machine learning in transportation data analytics. In: Chowdhury, M., Apon, A., Dey, K. (Eds.), Data Analytics for Intelligent Transportation Systems. Elsevier, pp. 283–307.

Brakewood, C., Watkins, K., 2019. A literature review of the passenger benefits of real-time transit information. Transp. Rev. 39 (3), 327–356.

Branco, P., Torgo, L., Ribeiro, R.P., 2017. SMOGN: A Pre-processing approach for imbalanced regression. In: Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, Vol. 74. PMLR, pp. 36–50.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11 (70), 2079–2107.

Chakour, V., Eluru, N., 2016. Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal. J. Transp. Geogr. 51, 205–217.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: SYnthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chiang, W.-C., Russell, R.A., Urban, T.L., 2011. Forecasting ridership for a metropolitan transit authority. Transp. Res. A 45 (7), 696–705.

De Regt, K., Cats, O., Van Oort, N., Van Lint, H., 2017. Investigating potential transit ridership by fusing smartcard and global system for mobile communications data. Transp. Res. Rec. 2652 (1), 50–58.

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.

Elias, D., Nadler, F., Stehno, J., Krösche, J., Lindorfer, M., 2016. SOMOBIL–Improving public transport planning through mobile phone data analysis. Transp. Res. Proc. 14, 4478–4485.
Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018a. Learning from Imbalanced Data Sets, first ed. Springer.
Fernández, A., Garcia, S., Herrera, F., Chawla, N.V., 2018b. SMOTE For learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J. Artificial Intelligence Res. 61, 863–905.
Ferreira, M.C., Fontesz, T., Costa, V., Dias, T.G., Borges, J.L., e Cunha, J., 2017. Evaluation of an integrated mobile payment, route planner and social network solution for public transport. Transp. Res. Proc. 24, 189–196.
Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Huang, S., Brooks, M., Lee, M.J., Asadi, H., 2019. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. Am. J. Roentgenol. 212 (1), 38–43.
Hao, S., Lee, D.-H., Zhao, D., 2019. Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system. Transp. Res. C 107, 287–300.
He, H., Ma, Y., 2013. Imbalanced Learning: Foundations, Algorithms, and Applications, first ed. Wiley-IEEE Press.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. Resampling methods. In: James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), An Introduction to Statistical Learning: with Applications in R, second ed. Springer New York, NY, pp. 197–223.
Karnberger, S., Antoniou, C., 2020. Network–wide prediction of public transportation ridership using spatio–temporal link–level information. J. Transp. Geogr. 82, 102549.
Kaur, H., Pannu, H.S., Malhi, A.K., 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput. Surv. 52 (4), 1–36.
Khoshgoftaar, T.M., Golawala, M., Van Hulse, J., 2007. An empirical study of learning from imbalanced data using random forest. In: 19th IEEE International Conference on Tools with Artificial Intelligence, Vol. 2. ICTAI 2007, IEEE, pp. 310–317.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, first ed. Springer.
Li, Y., Wang, X., Sun, S., Ma, X., Lu, G., 2017. Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. Transp. Res. C 77, 306–328.
Liu, L., Chen, R.-C., 2017. A novel passenger flow prediction model using deep learning methods. Transp. Res. C 84, 74–91.
Liu, Y., Liu, Z., Jia, R., 2019. DeepPF: A deep learning based architecture for metro passenger flow prediction. Transp. Res. C 101, 18–34.
Liu, H., Xu, H., Yan, Y., Cai, Z., Sun, T., Li, W., 2020. Bus arrival time prediction based on LSTM and spatial-temporal feature vector. IEEE Access 8, 11917–11929.
Louppe, G., 2014. Understanding Random Forests: From Theory to Practice (Ph.D. thesis). University of Liege, Belgium, arXiv:1407.7502.
Ma, X., Zhang, J., Du, B., Ding, C., Sun, L., 2019. Parallel architecture of convolutional bi-directional LSTM neural networks for network-wide metro ridership prediction. IEEE Trans. Intell. Transp. Syst. 20 (6), 2278–2288.
Molnar, C., 2022. Interpretable Machine Learning, second ed. URL: https://christophm.github.io/interpretable-ml-book.
Mulley, C., Clifton, G.T., Balbontin, C., Ma, L., 2017. Information for travelling: Awareness and usage of the various sources of information available to public transport users in NSW. Transp. Res. A 101, 111–132.
Nawar, S., Mouazen, A.M., 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. Sensors 17 (10), 2428.
Noursalehi, P., Koutsopoulos, H.N., Zhao, J., 2018. Real time transit demand prediction capturing station interactions and impact of special events. Transp. Res. C 97, 277–300.
Ohler, F., Krempels, K., Möbus, S., 2017. Forecasting public transportation capacity utilisation considering external factors. In: Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems. VEHITS, pp. 300–311.
Pang, J., Huang, J., Du, Y., Yu, H., Huang, Q., Yin, B., 2019. Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network. IEEE Trans. Intell. Transp. Syst. 20 (9), 3283–3293.
Pel, A.J., Bel, N.H., Pieters, M., 2014. Including passengers' response to crowding in the dutch national train passenger assignment model. Transp. Res. A 66, 111–126.
Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transp. Res. C 19 (4), 557–568.
Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2015. Using data from the web to predict public transport arrivals under special events scenarios. J. Intell. Transp. Syst. 19 (3), 273–288.
Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. SN Comput. Sci. 2 (3), 1–21.
Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25.
Tao, S., Corcoran, J., Rowe, F., Hickman, M., 2018. To travel or not to travel: 'weather' is the question. Modelling the effect of local weather conditions on bus ridership. Transp. Res. C 86, 147–167.
Van Oort, N., Brands, T., De Romph, E., 2015. Short-term prediction of ridership on public transport with smart card data. Transp. Res. Rec. 2535 (1), 105–111.
Van Oort, N., Brands, T., De Romph, E., Yap, M., 2016. Ridership evaluation and prediction in public transport by processing smart card data: A dutch approach and example. In: Kurauchi, F., Schmöcker, J.-D. (Eds.), Public Transport Planning with Smart Card Data, first ed. CRC Press Boca Raton, FL, pp. 197–224.
Van Oort, N., Van Nes, R., 2009. Regularity analysis for optimizing urban transit network design. Public Transp. 1 (2), 155–168.
Veres, M., Moussa, M., 2020. Deep learning for intelligent transportation systems: A survey of emerging trends. IEEE Trans. Intell. Transp. Syst. 21 (8), 3152–3168.
Wang, Z., 2020. Predicting short-term bus ridership with trip planner data: A machine learning approach. Delft University of Technology, URL: http://resolver.tudelft.nl/uuid:f1e4b495-d2ad-4a1e-803e-13e6c9b39f4a.
Xie, P., Li, T., Liu, J., Du, S., Yang, X., Zhang, J., 2020. Urban flow prediction from spatiotemporal data using machine learning: A survey. Inf. Fusion 59, 1–12.
Xue, R., Sun, D.J., Chen, S., 2015. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. Discrete Dyn. Nat. Soc. 2015, 1–11.
Zhang, J., Che, H., Chen, F., Ma, W., He, Z., 2021. Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method. Transp. Res. C 124, 102928.
Zhang, J., Chen, F., Cui, Z., Guo, Y., Zhu, Y., 2020. Deep learning architecture for short-term passenger flow forecasting in urban rail transit. IEEE Trans. Intell. Transp. Syst. 22 (11), 7004–7014.
Zhou, C., Dai, P., Wang, F., Zhang, Z., 2016. Predicting the passenger demand on bus services for mobile users. Pervasive Mob. Comput. 25, 48–66.