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

Wang, Z.; Pel, A.J.; Verma, T.; Krishnakumari, P.K.; van Brakel, Peter; van Oort, N.

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the 15th International Conference on Advanced Systems in Public Transport (CASPT2022)

CASPT 2021

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

Full Paper

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

Abstract Predictions on public transport ridership are beneficial as they allow for sufficient and cost-efficient deployment of vehicles. At 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, in contrast, trip planner data is often available in (near) real-time. This paper analyzes 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. The trip planner data is studied using correlation analysis to select informative variables, that are then used to develop 4 supervised machine learning models (linear, k-nearest neighbors, random forest, and gradient boosting decision tree). The best performing model relies on random forest regression and reduces the error by approximately half compared to a baseline model based on the weekly trend. We show that this model performance is maintained even for prediction lead times up to 30 minutes ahead, and for different periods of the day.

Keywords Public Transport · Trip Planner · Bus Ridership Prediction · Machine Learning

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

Ziyulong Wang · Adam J. Pel · Panchamy Krishnakumari · Niels van Oort
Department of Transport and Planning, Delft University of Technology, 2600 GA Delft, The Netherlands
Trivik Verma
Department of Multi-Actor Systems, Delft University of Technology, 2600 GA Delft, The Netherlands
E-mail: {z.wang-19, a.j.pel, t.verma, P.K.Krishnakumari, N.vanOort}@tudelft.nl
Peter van Brakel
REISinformatiegroep B.V. (9292), 3511 MJ Utrecht, The Netherlands
E-mail: pvanbrakel@9292.nl

improves passenger satisfaction and leads to a higher level of PT service (Pel et al., 2014; Ohler et al., 2017). At an operational level, this prediction needs to be realized in the short-term with less than an hour.

Until now, such short-term passenger demand predictions have typically used Automatic Fare Collection (AFC) or Global System for Mobile Communications (GSM) data. Scholars have widely shown that AFC data is useful in predicting short-term passenger demand (see, for instance, Van Oort et al. 2015; Xue et al. 2015; Ding et al. 2016; Zhou et al. 2016; Wang et al. 2018). Such datasets, however, are collected over days and do not depict the variability in short-term (from real-time up to 30 minutes) ridership patterns (Pelletier et al., 2011; Van Oort et al., 2015). On the other hand, transit information can also be collected in real-time using mobile phone data. This data 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 AFC system) to reveal the spatial and temporal pattern and to offer insightful mobility patterns from strategical and tactical level. The same methodology is also seen in the passenger flow measurement at Paris metro (Aguiléra et al., 2014).

AFC and GSM (regardless of their lag time in data availability) inherently show realized ridership, as it occurs. But for ridership prediction, it is valuable to have data on travel intention, before the trip is executed. The latter is captured in data from trip planner apps. Therefore, trip planner data (especially when available in near real-time) provides a source for ridership prediction. 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 provides the same granular level of spatial and temporal information about possible trips, as smart card data (Ferreira et al., 2017), implying the possibility of its use for predicting ridership. Since users do not have to realize their trips for data to be aggregated, the user intent of a trip lodged in real-time and collected through digitized apps (e.g. 9292) [1] can prove very useful in predicting short-term PT ridership. 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 at an operational level.

In this paper, we investigate how trip planner data can be utilized for predicting short-term ridership. We use trip planner data provided by 9292 to predict the short-term ridership on two case study lines in the provinces of Groningen and Drenthe (in the Netherlands) during October 2019. We use smart card data provided by OV-bureau Groningen Drenthe (regional PT authority) to derive historical ridership patterns and validate our methods of prediction. We design two baseline models and four supervised machine learning (ML) models to predict short-term bus ridership and compare the performance of the models. 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 lead times (real-time to 10, 15, 30 min ahead), variability across a day and in space, day type, and line characteristic. We find that trip planner data contributes to the best per-

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

forming model with a feature/variable importance up to 50%, and this model can reduce the errors by half, compared to the baseline model based on the weekly trend. Moreover, this model performance is maintained for prediction lead times up to 30 minutes ahead, and different periods of the day.

The remainder of this paper is organized as follows: We present the data and methods in the following Section 2. In Section 3, we analyze the results of the models for ridership prediction. Lastly, in Section 4, we present our reflections and provide avenues for future research. Section 5 draws the main conclusions. Wang (2020) presents more details of the methodology and additional cases, which can be found in the supplementary materials.

2 Data and method

This section first presents the description of the data for a better understanding of the rest of the paper, including context, data description, and data analysis (Section 2.1). Then, it explains the different components of the method: the baseline models, the correlation analysis for variable selection, the ML models for ridership prediction with trip planner data, and the evaluation criteria, along with feature importance analysis (Section 2.2).

2.1 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, and with the largest market share of approximately 46 % [2] so that it is a representative set against the other competitors such as NS (the biggest railway operator), Google Maps, and ANWB.

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 rails, 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 position of the predicted arrival time of a transit vehicle at a stop or station 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 the methodology in two bus lines - Qliner 300 (inter-city, fast service) and Q-link 1 (connecting multiple important locations, including a

[2] The market share is estimated by Newcom: 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/

hospital and campus) 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 5,640 km^2. It is suitable for this study as bus is the only mode in that region, 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. More than half of the population of Groningen and Drenthe are below 45 years old, which is appropriate for this study.

Data description The study utilizes the smart card and trip planner data in October (2019). During this period, there was an official holiday for schools from 19th to 27th in the case study region, which influences 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 IP or location tracking is not available. Examples of trip planner question and answer data are presented in Tab. 1 and 2, respectively.

Table 1 Sample of fictitious 9292 question data

Unique ID	Request Date	Travel Date	Request Time	Desired Travel Time	Origin	Destination
{3A9EA79A}	2019-10-01	2019-10-01	17:36	17:36	1000187	1210130
{2C1F1461C}	2019-10-05	2019-10-06	01:44	06:30	1010777	1010750
{4B3558F8F}	2019-10-23	2019-10-23	08:27	08:30	zh*	1000145

*: Due to privacy concerns, an exact location typed in by the user will be hashed to a random string.

Table 2 Sample of fictitious 9292 answer data

Unique ID	Journey Sequence	Line Number	Modality	Estimated Travel Time
{3A9EA79A}	1	5	172	25
{3A9EA79A}	2	6	172	28
{2C1F1461C}	1	10	1	12
Unique ID	**Transport Company**	**Vehicle Departure Time**	**Stop (Origin)**	**Stop (Destination)**
{3A9EA79A}	219	17:40	1000187	1210130
{3A9EA79A}	219	18:10	1010777	1010750
{2C1F1461C}	15	06:35	1000188	1000145

In the Netherlands, a nationwide smart card system is in operation, for all modes, using tap-in and tap-out technology (see Van Oort et al. (2016) for a full description). The smart card data is 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 type as we have no information on the subscription type.

Data cleaning is conducted before the analysis steps to handle inaccurate recordings, duplicates, and special arrangements of the trips. Both the trip planner and smart card data do not have the information on 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 AVL data and stop names. This leads to around 5% loss of trip planner data and a 3% loss of smart card data, which is not significant. 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 can use 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 minutes before their trip. Most of the requests are sent within a prediction lead time of an hour. If we are interested in a larger prediction horizon, the drop in the percentage of requests is considerable.

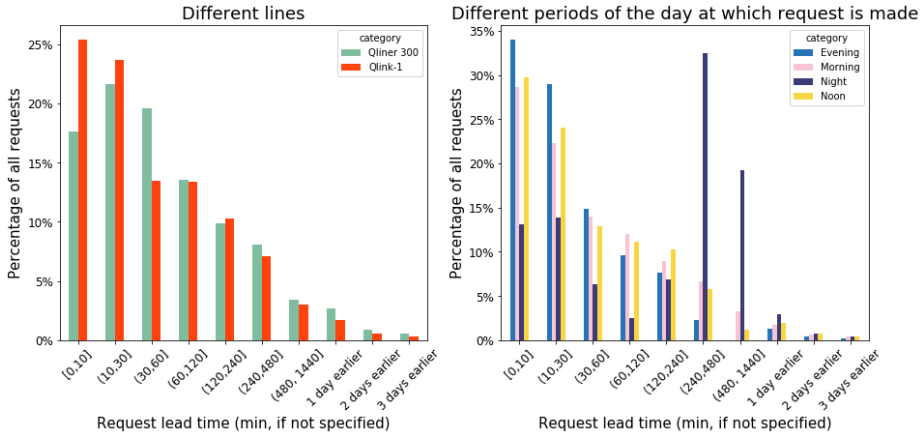

Fig. 1 Comparison of prediction lead time per line and per period

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

There is a significant negative influence on the ridership during weekend and holiday, which matches the findings in the literature (Chiang et al., 2011; Karnberger and Antoniou, 2020). However, the difference in ridership pattern over the day of the week is not apparent. The spatial characteristic is considerable, which is again in line with literature (Chakour and Eluru, 2016; Ding et al., 2016). The busy corridors with respect to ridership are usually a railway station, a Park and Ride (P+R) stop, a business area, or a city center. The average ridership of trips on each case study line is roughly below the seating capacity. However, it is the busy trips during peak hours that lower the level of service. It indicates the potential benefits of enhancing those crowded trips to improve comfort.

The joint distribution of trip planner requests and ridership is derived at stop-level, shown in Fig. 2. Although both distributions follow the same trend, there are more dots 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. Therefore, a linear relationship between them is hard to find.

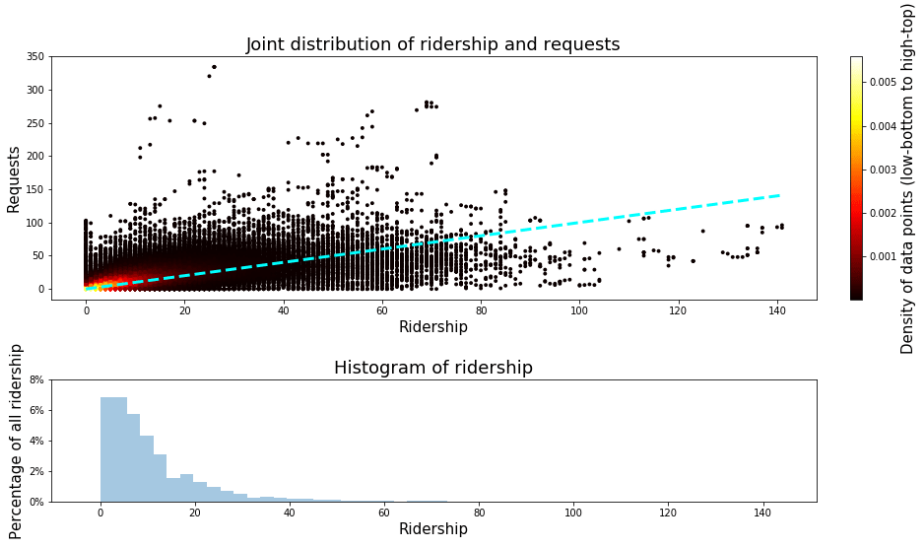

Fig. 2 Joint distribution of ridership and requests at stop-level

Moreover, the distribution of ridership is 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.

2.2 Methodology

Two baseline models are established in this paper. The first one is currently being used by PT operators. In such a model, the ridership of this week is estimated by the ridership of last week. The second baseline method 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.

For any ML model, the input variables are as important as the model itself. The selection of variables would always be an iterative process to reach a higher

performance of the models. We first choose the variables based on the literature and data analysis. Then, a variance-covariance analysis is performed to test the correlation between a specific variable and ridership (target). The insignificant variables will be kept out to avoid redundancy.

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 (Russell and Norvig, 2009). For other paradigms, readers are referred to the work of Dey (2016). Both regression and classification have been used for ridership prediction problem (Chiang et al., 2011; Xue et al., 2015; Ding et al., 2016; Zhou et al., 2016; Ohler et al., 2017; Wang et al., 2018; 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 (Raschka, 2015). Among the supervised models, we turn to more interpretable models suggested by Molnar (2019) 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: linear regression (LR), k-nearest neighbors regression (k-NNR), random forest regression(RFR), and gradient boosting decision tree regression (GBDTR).

We split the data into two instances, namely training data and test data with the 80/20 rule as a rule-of-thumb, also known as the Pareto Principle. Moreover, we assess the model at stop-level using the following metrics: mean absolute error (MAE), rooted mean square error ($RMSE$), coefficient of determination (R^2), and R^2 from cross-validation (Handelman et al., 2019). By stop level, we mean to measure the ridership between two consecutive stops during one bus trip.

We also explore the importance of the chosen features for the ridership prediction 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 et al., 2013). In this work, we use this technique to discover and quantify the usefulness of trip planner data.

For k-NNR, GBDTR, and RFR, we use *permutation feature importance*. 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 the interpretation. For tree-based models, *mean decrease in impurity* (MDI) is used to investigate the feature importance. Albeit this method is purely based on the training dataset and tends to inflate cardinality features, it is easily understandable and computationally light (Louppe, 2014).

3 ML models for ridership prediction with trip planner data

To predict the ridership at stop-level, we explore the ridership of a section between stops. To begin with, we illustrate the selection of variables and the variance-covariance matrix calculated from them in Section 3.1. Next, Section 3.2 discusses the model calibration, including sampling design and model tuning. After that,

we compare the model performance by the metrics in Section 3.3. Finally, Section 3.4 analyzes the best performing model further.

3.1 Covariance analysis and variable selection

Based on the data analysis, we list the variables considered in this study in Tab. 3. Other than the request-related variables, the other variables have been exhaustively studied in the literature.

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 2.1). 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 data and when the prediction is performed for further ahead in time.

With the considered input variables, we present the variance-covariance matrix in Fig. 3. The variance-covariance matrix shows the spread and deviation by the variance on the diagonal and the dependency between two variables by the covariance in other cells. Results, shown here, are derived from the case study lines during October. The top left corner represents the variance of ridership, which is the target we want to predict. This variance of ridership has a large span as high as 125.42, which means the prediction is valuable. The variance of requests also has a large value of 240.12. As a predictor variable, a high variation leads to a lower variation in the regression model and hence, a better prediction.

The most influencing factors are ridership-related and requested-related. The largest three positive covariances are seen in the request, request with at least 10 minutes ahead, and ridership last week. It confirms that the relationship between trip planner requests and ridership has a strong positive correlation, which implies the potential of the trip planner data to predict the ridership. Moreover, all covariances of trip planner variables with prediction lead time are strongly positive, which indicates that the travel purpose and behavior do not change when we consider further ahead in time. Plus, this covariance changes slightly when we focus on a specific period. However, we also notice that both line characteristics and temporal variables have unexpectedly small covariances. This implies that temporal and spatial influences on ridership are marginal. But we keep those in the prediction model as extensive literature prove their effectual 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).

3.2 Model calibration

The data analysis above shows that ridership is imbalanced. This imbalance is not inherently a problem. However, it is the conjunction between preference (crowded cases) and imbalance (more observations on the less crowded conditions) that causes a degradation of the performance of the most desirable instances (Fernández et al., 2018). The learning methods that we choose in this study (k-NNR, GBDTR,

Table 3 List of variables in ridership prediction with trip planner data

Variable	Explanation	Category	Unit/Coding
Ridership	The passengers on board	Numerical	Person
Ridership_mean	The historical average of ridership	Numerical	Person
Holiday	The autumn holiday	Categorical	One-hot encoding
Request†	The trip requests	Numerical	Record
Request_var†	The variance of trip requests, compared to the historical average	Numerical	Record
Request_mean	The historical average of trip requests	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 the trip	Categorical	One-hot encoding
Period	Peak or off-peak hour	Categorical	One-hot encoding
Ridership_last_week	The passengers on board of the same trip last week	Numerical	Person
Request_10†	The trip requests that are sent 10 minutes ahead	Numerical	Record
Request_var_10†	The variance of trip requests that are sent 10 minutes ahead	Numerical	Record
Request_15†	The trip requests that are sent 15 minutes ahead	Numerical	Record
Request_var_15†	The variance of trip requests that are sent 15 minutes ahead	Numerical	Record
Request_30†	The trip requests that are sent 30 minutes ahead	Numerical	Record
Request_var_30†	The variance of trip requests that are sent 30 minutes ahead	Numerical	Record

†: Variables with this symbol will be put into models in pairs, e.g. request and request_var.

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 *synthetic minority oversampling technique* (SMOTE) proposed by Chawla et al. (2002). SMOTE creates new instances of a minority class by using a K-Nearest-Neighbor approach. A random number of original observations are chosen

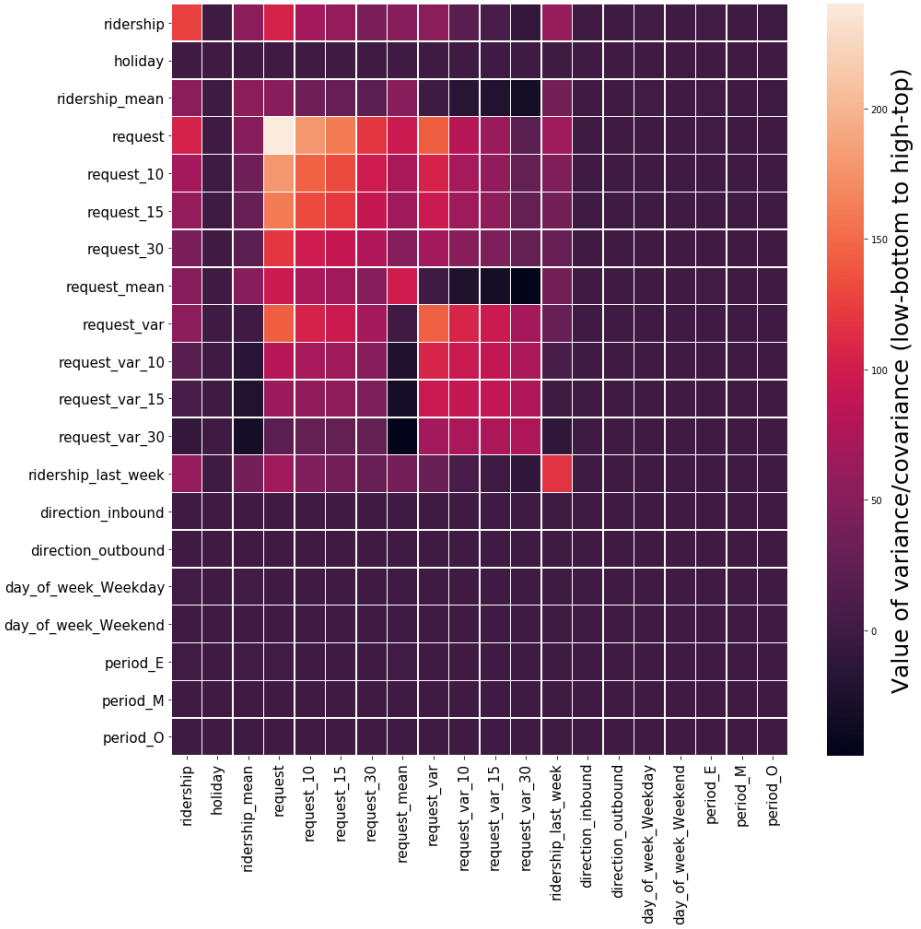

Fig. 3 Variance-covariance matrix at stop-level

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. (2018) indicate that a combination of SMOTE and undersampling performs the best.

A pair-wise study on random forest regressor with 4 undersampling and 6 oversampling strategies is developed to reach the optimal combination. The random forest regressor 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). Figure 4 presents the min-max scaled R^2 results of sampling design of each case study line through 5-fold cross-validation. Normally, the dataset is recommended to split up into k-partitions - 5 or 10 partitions as a rule of thumb (James et al., 2014). Both case study lines show a similar pattern through a 10-time experiment since the train/test split is random. Therefore, we resample the

data without undersampling the majorities but oversampling the minorities by 50%.

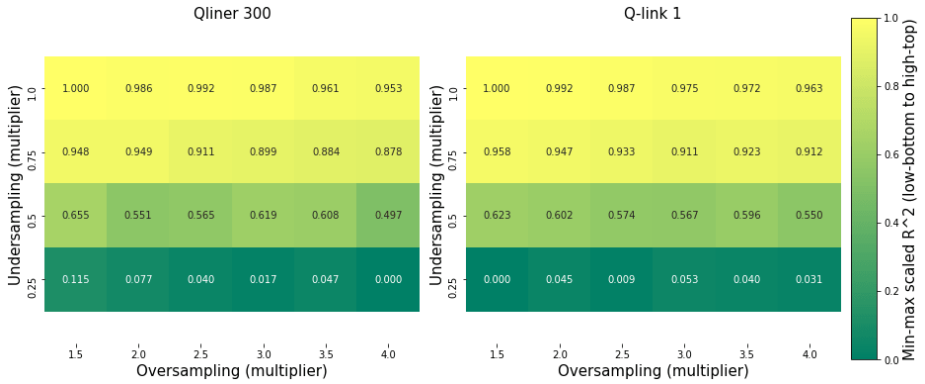

Fig. 4 Evaluation of different sampling designs per line

Given the optimal sampling design, 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). In this nested cross-validation, *stratified 5-fold cross-validation* is applied 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 et al., 2013). Tuning hyperparameters of non-parametric regression algorithms (such as k-NNR, GBDTR, and RFR) is of importance as they do not rely on the assumed shape or parameters of the underlying population distribution (Hopkins et al., 2018). The optimal hyper-parameters are presented in Tab. 4 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 and therefore GBDTR and RFR in this study (Probst et al., 2019). 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.

3.3 Model performance: comparison of models

With the tuned hyper-parameters of the models and the optimal method of sampling, we proceed to analyze the results of the four ML models for the two bus lines. We present the performance of the prediction models of Qliner 300 and Q-link 1 in Tab. 5 and 6, respectively.

In both cases, RFR outperforms the other models as shown by the R^2 from repeated random 5-fold cross-validation while for Qliner 300, GBDTR has the same score as that of RFR. In such a case, both RFR and GBDTR beat other models with a R^2 from cross-validation of 0.726. However, RFR outperforms GBDTR by

Table 4 Optimal hyper-parameters of the non-parametric models

	Qliner 300	Q-link 1
GBDTR	learning_rate = 0.01	learning_rate = 0.02
	n_estimators[1] = 13000	n_estimators = 25000
	max_depth[2] = 4	max_depth = 4
	min_samples_split[3] = 2	min_samples_split = 2
	min_samples_leaf[4] = 6	min_samples_leaf = 15
	subsample[5] = 1	subsample = 1
	max_features[6] = 11	max_features = 15
k-NNR	n_neighbors[7] = 14	n_neighbors = 10
RFR	bootstrap[8] = False	bootstrap = False
	n_estimators = 700	n_estimators = 400
	max_depth = 20	max_depth = 25
	min_samples_split = 2	min_samples_split = 2
	min_samples_leaf = 4	min_samples_leaf = 2
	max_features = 10	max_features = 20

[1]: Number of trees.
[2]: The maximum depth of a tree.
[3]: The minimal number of samples in a node for the node to be split.
[4]: The minimum number of samples in a leaf node.
[5]: The fraction of samples to be used for fitting the individual base learners.
[6]: The number of features randomly chosen as candidates for a split.
[7]: Number of neighbors to use.
[8]: Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.

Table 5 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	-
LR	5.573	9.870	0.276	0.595
GBDTR	4.664	7.114	0.624	**0.726**
k-NNR	4.699	7.277	0.607	0.666
RFR	**4.123**	**6.357**	**0.700**	**0.726**

the other three metrics. For Q-link 1, RFR significantly improves the prediction accuracy compared to other models. For the baseline model with weekly trend and linear regression model of Q-link 1, the performance was shockingly low with negative R^2 values. It means that a simple mean would work better than these

Table 6 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	-
LR	9.095	68.056	-17.879	0.536
GBDTR	9.462	13.667	0.239	0.796
k-NNR	5.622	9.235	0.652	0.698
RFR	**4.329**	**7.045**	**0.798**	**0.826**

two models, which indicates the failure of these models to find any meaningful relationship between the input and output.

Figure. 5 shows a specific instance of the predicted and actual values of Qliner 300. ML models mostly can capture quiet trips as well as beat the baseline models for busy trips. However, all models have higher average error and bias when the actual value of ridership increases, which implies the existence of heteroscedasticity. When the value of ridership is low, all models can function efficiently, where GBDTR tends to overestimate the prediction, k-NNR tends to underestimate while RFR is relatively neutral. However, the linear regression performs worse than even the baseline models, indicating an absence of a linear relationship between ridership and the selected variables.

3.4 Model performance: further analysis of RFR model

Residual analysis The best performing model -RFR- 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 differences in the PT operations. 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 Tab. 7.

Table 7 Overestimation and underestimation based on residual analysis of RFR

	Percentage		95th Percentile Absolute Error(Person)		Average of top 5 Percentile Absolute Error(Person)	
	Overestimation	Underestimation	Overestimation	Underestimation	Overestimation	Underestimation
Qliner 300	46.77%	53.23%	10.125	14.492	16.486	22.013
Q-link 1	50.51%	49.49%	12.079	18.285	18.353	25.833

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

The residuals of prediction vary with period and day type, shown in Fig. 6. For Q-link 1, it is the evening peak that has the highest variance of residuals

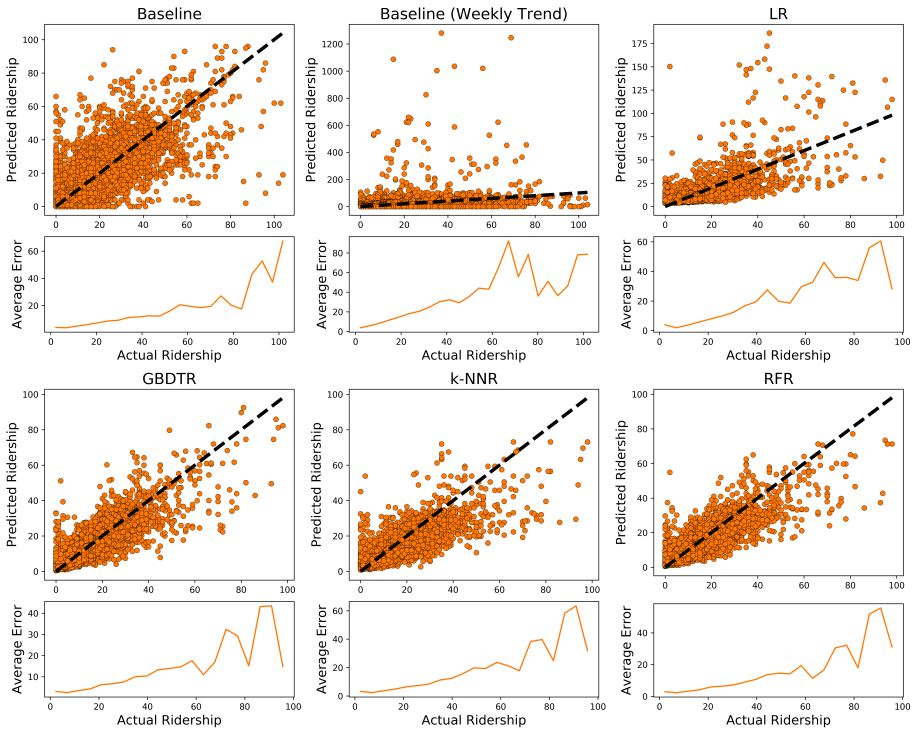

Fig. 5 Prediction vs. actuality plot of Qliner 300

because it is the second-highest commuting time. Besides, commuters have a non-uniform off-duty time so that the ridership is easy to fluctuate, and therefore a higher variance of prediction error. In contrast, Qliner 300 has a higher variance of residuals during the morning peak because it is a fast-service less-stopping line, and passengers flow into this line during the morning peak. This results in higher variance and predictive difficulty. Concerning day type, the variance of the weekday is higher than the weekends for both lines.

Feature importance We report the feature importance of RFR on Qliner 300 as an example in Fig. 7. Since it is a tree-based model, we can apply both permutation feature importance and MDI to calculate the feature importance and compare the results from both approaches to gain a complete understanding of the importance of features. We carry out MDI feature importance on the training set while we apply the permutation feature importance on the test set. In this case, we use the scenario with all trip planner requests without any prediction lead time.

No matter the feature importance measurement method, the first five contributing features are the average number of ridership, the ridership of last week, the number of requests, the average number of requests, and the variance of requests. The most influencing variable is usually the average number of ridership, which shows its relevance in ridership prediction. The MDI importance of Q-link 1 demonstrates that the number of requests supports the prediction the most with

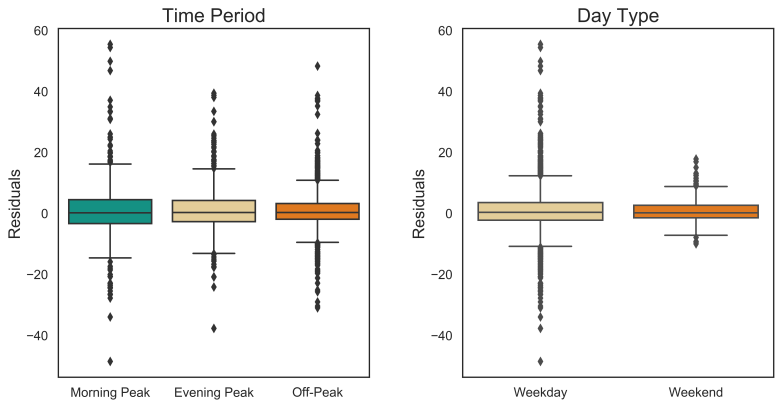

Fig. 6 Residuals of Qliner 300 per scenario (RFR)

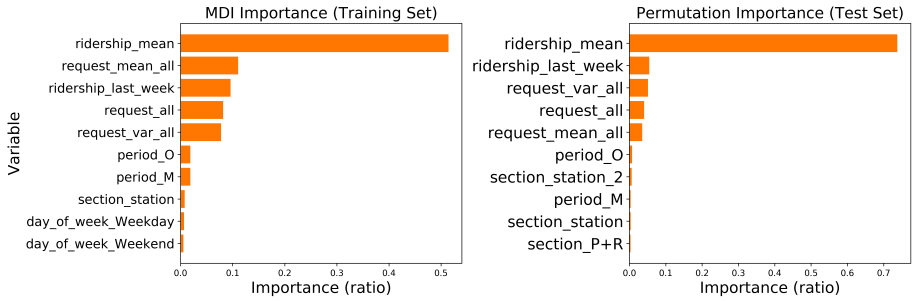

Fig. 7 Feature importance of Qliner 300 (RFR)

almost 35% of the importance score. Regardless of the approach of measurement, Q-link 1 regards requests-related variables are more important with almost 50% of the feature importance on average. In contrast, they play about 20% importance on average for Qliner 300. Still, at least one of the request-related variable often rank within the first three substantial variables, such as the average request in MDI importance and variance of request in permutation importance. Unexpectedly, both temporal and spatial variables exhibit minor importance for the predictions. 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).

Performance with prediction lead time We further analyze the performance of RFR with the same configuration and the same sampling design but with different prediction lead times. In other words, how the model performs with trip planner data that is further ahead of time, such as 10 minutes, 15 minutes, and 30 minutes before the vehicle start time. Figure. 8 displays the number of requests per prediction lead time per scenario (inset), the model performance of R^2 per scenario

(solid line with circle marker), and the feature importance of the request-related variables per scenario (dashed line with star marker). Request-related variables in each scenario contain the number of requests, the historical average of requests, and the variance of the requests.

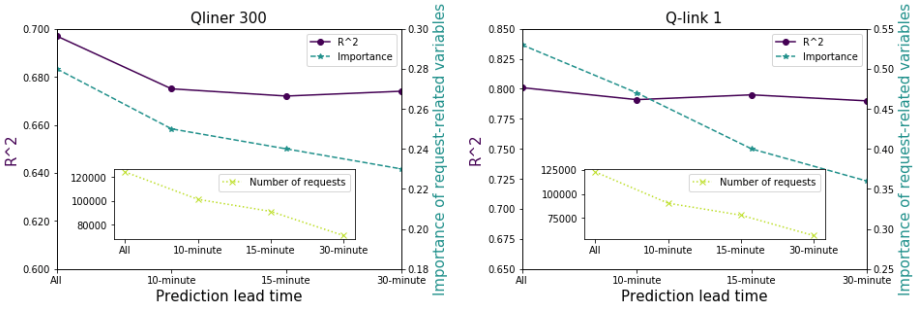

Fig. 8 Model performance with different prediction lead times in the short-term

By using trip planner data with different prediction lead times, the performance of RFR remains almost stable. Although the number of data and the correlation between requests and ridership drop, the model functions essentially the same, reflected by the R^2 value. Among all scenarios of Qliner 300, the model with only 10-minute in advance is the best as it is the closest time to the vehicle start time, containing most of the information. However, unlike Qliner 300, Q-link 1 has 15-minute ahead of time as the best-performing. It shows that people probably change their travel behavior for such a line when it moves from 15-minute to 10-minute ahead. 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 the number of requests and the average number of requests decrease. However, we see an increase in the importance of the variance of the requests. This leads to the sum of feature importance to remain the same.

Performance with requests sent from different periods In data analysis (Section 2.1), we conclude that people behave differently during different times while using such a trip planner. 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 hours, e.g. morning is from 4:00 to 10:00. Figure. 9 exhibits the results with the same layout as Fig. 8.

Essentially, we see the model performance, and the summed request-related importance has the same trend with the number of requests. The best performances are seen from 10:00 to 16:00 when the number of requests is the largest, compared to including all the requests. 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

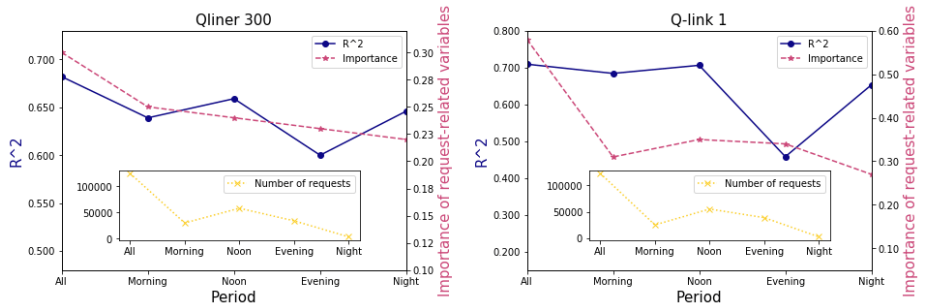

Fig. 9 Model performance with requests sent from different periods

predicts worse for both cases. It means that during the evening, the relationship between the variables and ridership is 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 requests becomes more important, while the number of requests tends to be less influencing.

4 Discussion

In this study, we investigated how trip planner data can contribute to the short-term prediction of bus ridership and built 2 baseline and 4 supervised machine learning models with a selection of variables, based on correlation analysis. We found that trip planner data can have feature importance up to 50% in the best performing model (RFR), which can reduce the 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 privacy concerns of the 9292 trip planner can be minimized. 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, the underestimation of the trip planner's importance can be avoided. 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.

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 here in this paper, where the sparse data in the high-value domain is oversampled. This optimal sampling design was based on a pair-wise study, which implies that better methods for sampling can be explored.

Enhancement of baseline models Several failed models can be substantially improved, including the baseline model and the baseline model with weekly trend. Missing recordings were notable, which results in a deterioration of the baseline model (currently used in practice) performance. The baseline model with weekly trend was strongly biased due to the high week multiplier factor on several sections of last week. Smoothing can be added to elevate the model by considering the trip of yesterday or so forth.

Inclusion of more significant variables The inclusion of new significant variables should be considered for improving the predictions. During the residual analysis, we found out 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 capture the relationship.

Findings on feature importance In this study, the contribution of the historical ridership variable was significant, 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. Unexpectedly, the day of the week was also found to be not significant. Only morning peak and off-peak variables marginally influenced the model. Some high-traffic locations were also insignificant.

5 Conclusion

In this paper, we proposed a method to analyze the effectiveness of trip planner data in predicting short-term bus ridership. The case studies showed that the best performing model relied on the RFR can reduce the errors by almost half, compared to a baseline model based on the weekly trend. This model 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 prediction lead times up to 30 minutes ahead, and for different periods of the day. Regardless of the approach of measurement, the trip planner data can roughly have a 35% feature importance on average. The presented method explained the use of real-time transit information followed by apps at 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. Discussions are 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 able to capture the temporal and spatial influence such that it can enhance the operational performance of transit operators.

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. 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 large data source with the knowledge of user behavior. 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 operation of the PT system and improve its level of service.

Acknowledgements 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. Transportation Research Part C: Emerging Technologies 43:198 – 211, special Issue with Selected Papers from Transport Research Arena

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

Branco P, Torgo L, Ribeiro RP (2017) Smogn: a pre-processing approach for imbalanced regression. In: First International Workshop on Learning with Imbalanced Domains: Theory and Applications, pp 36–50

Breiman L (2001) Random forests. Machine learning 45(1):5–32

Cawley GC, Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research 11:2079–2107

Chakour V, Eluru N (2016) Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal. Journal of Transport Geography 51:205–217

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:321–357

Chiang WC, Russell RA, Urban TL (2011) Forecasting ridership for a metropolitan transit authority. Transportation Research Part A: Policy and Practice 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. Transportation Research Record 2652(1):50–58

Dey A (2016) Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies 7(3):1174–1179

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)

Elias D, Nadler F, Stehno J, Krösche J, Lindorfer M (2016) Somobil – improving public transport planning through mobile phone data analysis. Transportation Research Procedia 14:4478 – 4485, transport Research Arena TRA2016

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets. Springer

Fernández A, Garcia S, Herrera F, Chawla N (2018) SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research 61:863–905

Ferreira MC, Fontesz T, Costa V, Dias TG, Borges JL, e Cunha JF (2017) Evaluation of an integrated mobile payment, route planner and social network solution for public transport. Transportation Research Procedia 24:189–196

Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, Lee MJ, Asadi H (2019) Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. American Journal of Roentgenology 212(1):38–43

Hao S, Lee DH, Zhao D (2019) Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system. Transportation Research Part C: Emerging Technologies 107:287 – 300

He H, Ma Y (2013) Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons

Hopkins S, Dettori JR, Chapman JR (2018) Parametric and nonparametric tests in spine research: Why do they matter? Global spine journal 8(6):652–654

James G, Witten D, Hastie T, Tibshirani R (2014) An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated

Karnberger S, Antoniou C (2020) Network–wide prediction of public transportation ridership using spatio–temporal link–level information. Journal of Transport Geography 82(May 2019):102,549

Khoshgoftaar TM, 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 (ICTAI 2007), IEEE, vol 2, pp 310–317

Kuhn M, Johnson K, et al. (2013) Applied predictive modeling, vol 26. Springer

Louppe G (2014) Understanding random forests: From theory to practice. 1407.7502

Molnar C (2019) Interpretable Machine Learning. Lulu

Mulley C, Clifton GT, Balbontin C, Ma L (2017) Information for travelling: Awareness and usage of the various sources of information available to public transport users in nsw. Transportation Research Part A: Policy and Practice 101:111–132

Noursalehi P, Koutsopoulos HN, Zhao J (2018) Real time transit demand prediction capturing station interactions and impact of special events. Transportation Research Part C: Emerging Technologies 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 - Volume 1: VEHITS,, SciTePress, pp 300–311

Pel AJ, Bel NH, Pieters M (2014) Including passengers ' response to crowding in the Dutch national train passenger assignment model. Transportation Research Part A 66:111–126

Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19(4):557 – 568

Probst P, Wright MN, Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(3):e1301

Raschka S (2015) Python machine learning. Packt Publishing Ltd

Russell S, Norvig P (2009) Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall Press, USA

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8(1):25

Van Oort N, Brands T, De Romph E (2015) Short-term prediction of ridership on public transport with smart card data. Transportation research record 2535:105–111

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, CRC Press Boca Raton, FL, chap 11, pp 197 – 224

Wang X, Zhang N, Zhang Y, Shi Z (2018) Forecasting of short-term metro ridership with support vector machine online model. Journal of Advanced Transportation 2018:3189,238

Wang Z (2020) Predicting short-term bus ridership with trip planner data: A machine learning approach. Master's thesis, Delft University of Technology, URL http://resolver.tudelft.nl/uuid:f1e4b495-d2ad-4a1e-803e-13e6c9b39f4a

Xue R, Sun DJ, 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

Zhou C, Dai P, Wang F, Zhang Z (2016) Predicting the passenger demand on bus services for mobile users. Pervasive and Mobile Computing 25(2013):48–66

6 Supplementary materials

In the project with wider scope (Wang, 2020), we also test two more case study lines, including Line 50 Groningen-Assen (city-city bus line) and Line 35 Groningen-Oldehove (city-village bus line). The performances are presented in Tab. 8 and Tab. 9.

Table 8 Performance of short-term prediction models (line 50)

Line 50	**MAE (Person)**	**RMSE (Person)**	**R^2**	**Repeated Random 5-Fold Cross-Validation**
Baseline	4.026	5.938	0.494	-
Baseline with weekly trend	6.799	12.022	-1.423	-
LR	4.491	10.959	-0.780	0.515
GBDTR	11.819	13.693	-1.778	0.657
k-NNR	**3.248**	**4.864**	**0.639**	0.686
RFR	3.578	5.179	0.603	**0.770**

Table 9 Performance of short-term prediction models (line 35)

Line 35	MAE (Person)	RMSE (Person)	R^2	Repeated Random 5-Fold Cross-Validation
Baseline	3.423	6.158	0.341	-
Baseline with weekly trend	4.225	10.338	-0.858	-
LR	3.732	12.577	-1.656	0.530
GBDTR	2.243	3.922	0.742	0.811
k-NNR	2.458	4.204	0.703	0.747
RFR	**1.938**	**3.493**	**0.795**	**0.831**

For line 50, GBDTR performs the worst due to the different distribution of data (sparse in the high-value domain) and a relatively large oversampling strategy (350%). GBDTR is sensitive to "noisy" data (in our case, it is the underlying difference in the distribution of the training set and test set). We have a trade-off between k-NNR and RFR. The margin between k-NNR and RFR is small, with respect to MAE, RMSE, andR^2. However, we have a better R^2 from the cross-validation of RFR. In this specific case, the sampling design varies considerably every time, and therefore less sensitivity to training data is of importance. Hence, we regard RFR is the best performing model.

The prediction vs. actuality plot of Q-link 1 is shown in Fig. 10.

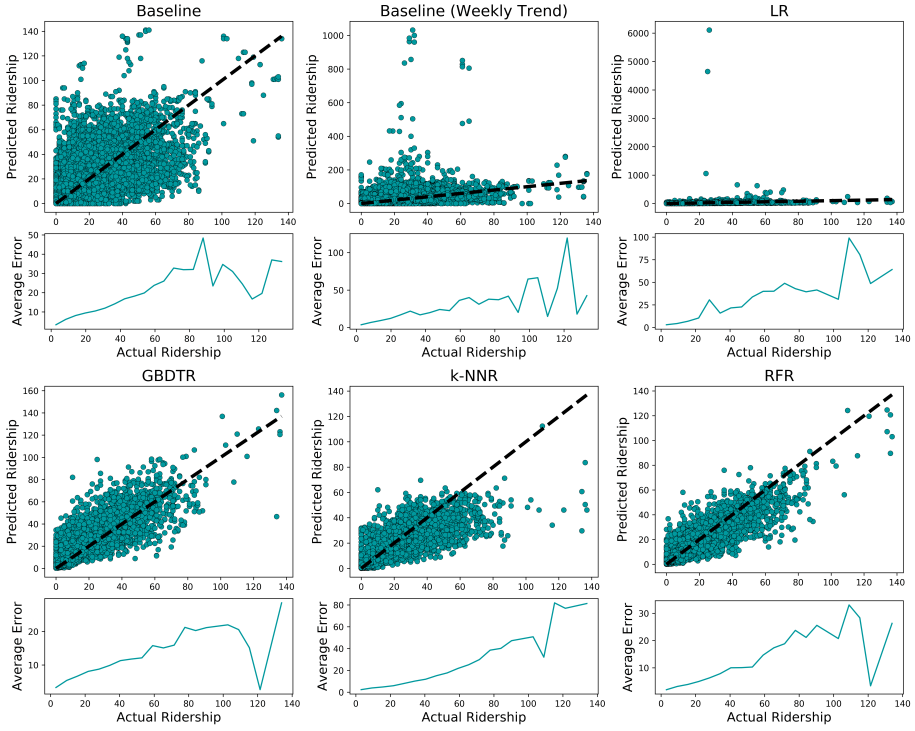

Fig. 10 Prediction vs. actuality plot of Q-link 1

The feature importance plot of Q-link 1 is shown in Fig. 11.

Full information of four cases on model performance with prediction lead times is shown from Tab. 10 to Tab. 13.

Table 10 Performance of RFR with timing advance of trip planner requests (Qliner 300)

Qliner 300	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	124350	0.397	4.131	6.385	0.697	0.090	0.080	0.110
10-minute	101313	0.328	**4.199**	**6.612**	**0.675**	**0.070**	**0.070**	**0.110**
15-minute	91176	0.304	4.269	6.647	0.672	**0.070**	0.060	**0.110**
30-minute	70824	0.274	4.215	6.623	0.674	0.060	0.060	**0.110**

Table 11 Performance of RFR with timing advance of trip planner requests (Q-link 1)

Q-link 1	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	122913	0.618	4.308	6.992	0.801	0.340	0.090	0.100
10-minute	90408	0.541	4.375	7.156	0.791	**0.300**	**0.070**	**0.100**
15-minute	77920	0.501	**4.317**	**7.099**	**0.795**	0.230	**0.070**	**0.100**
30-minute	56664	0.458	4.387	7.181	0.790	0.200	0.060	**0.100**

Table 12 Performance of RFR with timing advance of trip planner requests (line 50)

Line 50	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	91719	0.560	3.609	5.241	0.593	0.210	0.160	0.060
10-minute	72666	0.411	**3.897**	**5.656**	**0.526**	**0.130**	0.100	0.070
15-minute	63489	0.411	4.408	6.337	0.405	**0.130**	**0.110**	0.070
30-minute	44667	0.313	5.853	7.472	0.173	0.090	0.080	**0.090**

Full information of four cases on model performance with requests sent during different periods is shown from Tab. 14 to Tab. 17.

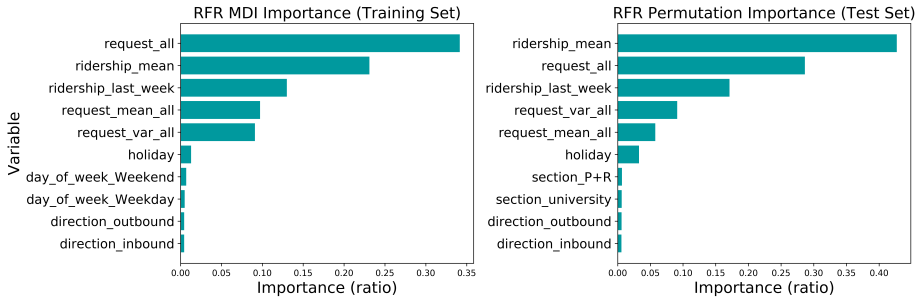

Fig. 11 Feature importance of Q-link 1 (RFR)

Table 13 Performance of RFR with timing advance of trip planner requests (line 35)

Line 35	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	25822	0.707	1.839	3.298	0.817	0.310	0.180	0.060
10-minute	18231	0.630	1.910	3.460	0.800	**0.300**	**0.120**	0.060
15-minute	15572	0.576	1.943	3.543	0.789	0.280	0.110	**0.070**
30-minute	10908	0.472	**1.862**	**3.409**	**0.805**	0.210	0.100	**0.070**

Table 14 Performance of RFR with requests at different times (Qliner 300)

Qliner 300	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	124350	0.397	4.279	6.538	0.682	0.100	0.110	0.090
Morning	29644	0.307	4.546	6.967	0.639	**0.050**	**0.120**	**0.080**
Noon	57783	0.230	**4.401**	**6.774**	**0.659**	**0.050**	**0.120**	0.070
Evening	33792	0.072	4.711	7.337	0.600	0.030	**0.120**	0.080
Night	3131	0.123	4.509	6.907	0.646	0.010	**0.120**	0.090

Table 15 Performance of RFR with requests at different times (Q-link 1)

Q-link 1	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	122913	0.618	5.098	8.443	0.709	0.360	0.110	0.110
Morning	25423	0.308	5.366	8.803	0.684	0.060	**0.150**	**0.100**
Noon	55479	0.436	**5.191**	**8.499**	**0.706**	**0.130**	**0.150**	0.070
Evening	39173	0.247	7.112	11.527	0.458	0.090	**0.150**	**0.100**
Night	2838	0.145	5.567	9.227	0.653	0.020	**0.150**	**0.100**

Table 16 Performance of RFR with requests at different times (line 50)

Line 50	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	91719	0.560	4.202	6.096	0.449	0.230	0.070	0.170
Morning	22890	0.376	5.707	8.022	0.046	**0.110**	**0.100**	**0.090**
Noon	42251	0.350	**4.194**	**5.802**	**0.501**	0.100	0.090	**0.090**
Evening	24615	0.096	4.489	6.305	0.411	0.040	**0.100**	**0.090**
Night	1963	0.134	8.126	11.346	-0.907	0.020	0.090	0.060

Table 17 Performance of RFR with requests at different times (line 35)

Line 35	Number	Correlation	MAE (Person)	RMSE (Person)	R^2	Request	Request_mean	Request_var
All	25822	0.707	2.257	3.882	0.747	0.340	0.070	0.200
Morning	7191	0.479	3.293	5.677	0.459	0.130	**0.130**	0.100
Noon	12568	0.489	**2.373**	4.343	0.683	**0.140**	0.110	**0.140**
Evening	5488	0.050	3.601	6.500	0.291	0.050	0.110	0.090
Night	575	0.187	2.451	**4.307**	**0.688**	0.040	0.110	0.080