Forecasting passenger load in a transit network using data driven models

Kevin Pasini, Mostepha Khouadjia, Fabrice Ganansia, Latifa Oukhellou

Forecasting passenger load in a transit network using data driven models

Kevin PASINI[1,2], Mostepha KHOUADJIA[2], Fabrice GANANSIA[3], Latifa OUKHELLOU[1]

[1] Université Paris-Est, IFSTTAR, Cosys-Grettia, Champs-sur-Marne, France
[2] IRT SystemX, Saclay, France.
[3] SNCF- Innovation & Recherche, St Denis, France
Corresponding Author: Kevin PASINI (kevin.pasini@irt-systemx.fr)

Abstract

Passenger load forecasting can be valuable in transportation planning, operation management and for enriching the information available to passengers, particularly in high-density megacities. This paper investigates the long and short term forecasting of passenger loads in a transit network by using multiple sources of data (on-board headcount data and train timetables). With each passenger load being treated as a time series, one of the main challenges of this study is related to the dependence of the temporal dynamics of the time series to be predicted on the railway timetable. Machine learning models are proposed to predict the passenger load on each train passing each station. We will compare different models, including a random forest, and a gradient boosting tree. Different types of features (calendar, hour, last passenger load, train delay, and train route) will be considered to measure their contributions to the prediction task. The experiments are conducted on a real historical dataset covering the period from 2015 to 2016. The dataset was collected on a railway transit network line operated by SNCF in suburban Paris.

Keywords: Machine learning, forecasting, train load, passenger information.

1. Introduction

Passenger load forecasting can be valuable in transportation planning, operation management and for enriching the information available to passengers, particularly in high-density megacities [1]. For example, the greater Paris region saw approximately 10 million daily passengers in 2017. Providing passengers with train load forecasts in addition to expected arrival times of future trains can be useful to enable better planning of their journeys, which can improve global comfort and avoid overcrowding on trains. Such predictive indicators related to flow management can also be used by public transport authorities and transport operators either to enrich public transport route planning or to better estimate transport demand, which could improve synchronization of train traffic with passenger flow.

The basic idea that we investigate in this paper is the forecasting of passenger loads in a transit network by using on-board headcount data and train timetables. With each passenger load being treated as a time series, one of the main challenges of this study is related to the dependence of the temporal dynamics of the time series to be predicted on the railway timetable. The experiments are conducted on a real dataset covering the period from 2015 to 2016. This set of predictive tools could be integrated into mobility service platforms with the aim of providing citizens or transport operators with real-time information about the fluidity of transport networks and traffic conditions. Moreover, this kind of predictive information could also be useful in the context of Mobility as a Service (MaaS).

The forecasting is achieved by using machine learning models including a random forest, a gradient boosting tree and a fully connected network. A considerable amount of research has been conducted on passenger flow forecasting by applying data mining and machine learning approaches [2] [3] [4]. Most of the studies use smart card data and focus on the forecasting of passenger affluence at an

aggregated level (per 15 minutes or 30 minutes time horizon). Here, we focus on the forecasting of passenger load taking account in addition to calendar information and train operation, real-time train schedules. This impacts the time step of the time series that we should predict.

The paper is organized as follows. Section 2 details the dataset and the objectives of the study. Section 3 describes the methodology proposed to solve the forecasting task. Section 4 summarizes the obtained results. Section 5 concludes the paper.

2. Data and objectives

Our study focuses on data collected from a railway line operated by SNCF. This line serves the northern area of suburban Paris and carries approximately 250,000 passengers daily. As shown in Figure 1, this line can be divided into four branches serving different terminals. This transport network structure induces complex railway operations. The dataset includes both counting data of passengers boarding and alighting at each station and real-time timetable information. These heterogeneous sources of data enable us to reconstruct the passenger load on each train on each inter-station link, as shown in Figure1.

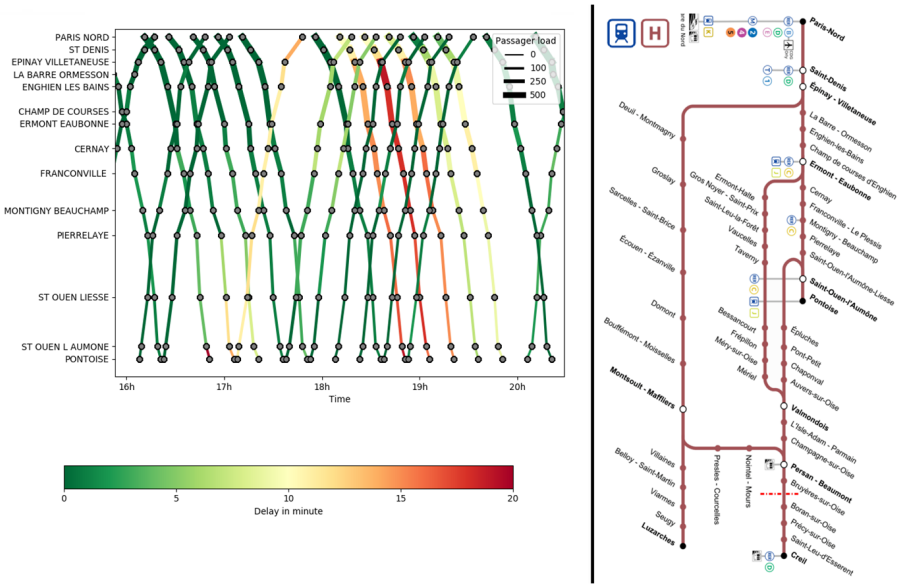

Fig. 1: Train table of one branch of a SNCF railway line and the map of the line

Figure 2 shows an example of weekday and weekend daily train passenger loads collected from two stations. It can be seen that passenger load is strongly impacted by several factors including calendar factors (hour, type de the day), train operation, location of the station. This induces a high variability of the time series to be predicted.

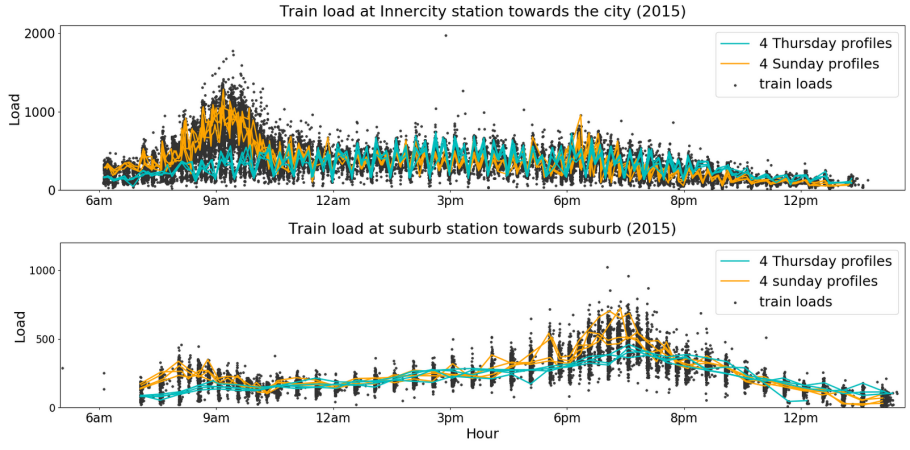

Fig. 2: Examples of passenger loads collected on two different stations.

3. Forecasting methodology

3.1 Forecasting issues

The forecasting of train passenger load at each station is a non-trivial and challenging task, mainly due to the influence of several factors [5]. These factors are related to transport demand and supply and can be summarized as follows:

- Calendar factors including the type of the day (weekday, weekend, day-off, school break), the concerning month and the hour.
- Spatial factors such as the information about the area where the station is situated (residential, employment, commercial, leisure, etc.).
- Transport network factors such as the planning and schedule of the offer which can has a high variability according to the service at the station that could lead to irregular temporal structure of train passage time series. Additional factors related to the impact in case of schedule modification due to unexpected events (disturbance, incident,...) or even plannified ones (strike, public demonstration) and the incurred delay as a result of these events have to be considered.
- Exogenous factors such as the climatic conditions.
- Sensors that collect the flows of passengers boarding or alighting the trains which could provide the data subject to different interpretations depending on the context. For example, null value has different meanings, such as no passengers, missing value, or sensor malfunction.

In light of these factors, we address the forecasting passenger load task within the framework of machine learning..

3.2 Feature engineering

The purpose of the features is to easier understand the context of the forecasting problem and then might help to solve it. For our case, these features summarize the influencing factors mentioned above in section 3.1 and are categorized according to their nature into two categories:

Long-term features: Mainly related to calendar and train service attributes and regroup:

- **Calendar features (CA):**
    - Day of the year: Position of the day in the year encoded by cosine and sine of (2 x 4) frequencies.
    - Day type: Position of the day in the week encoded from 1 to 7.
    - Holiday type: True if the day falls on an extra day off, school or public holiday.
    - Minutes: Minutes of the day encoded by cosine and sine of (2x4) frequencies.
- **Train service features (TS):**
    - Train routes: Feature related to the train routes that serve the considered station, which is the result of applying PCA dimensionality reduction on the train routes.

Short-term features (ST): In addition, we also consider short-term features by considering a lag windows that ranges between 1 to 6 past observations:

- Delay: Difference between the real and the theoretical schedule of the train at the station.
- Load: Number of passengers on the train for each of the last 6 passages at the station.

We propose to evaluate the importance of these features according to the performance of the forecasting of passenger load. For that purpose, we note CAL when only calendar long-term features are used in the training of the model, LT when both long-term calendar and train service features are used as inputs of the model, and ALL when long and short term features are used to build the model.

3.3 Forecasting models

We evaluate five models for train load forecasting on both suburb and inner-city stations. These models

range from classical baseline model to advanced machine learning models:

1. **Last Value (LV)**: It is the simplest forecasting that consists of forwarding the last observed load on the train to the next one at the same station.
2. **Contextual Average (CA)**: It consists of using the average load of trains that are committed on the same day type and time slice.
3. **Gradient Boosting (XGB)**: A regressor model that produces a prediction model in the form of an ensemble of weak decision trees and by weighting average prediction by boosting.
4. **Random FOREST (RF)**: A regressor model that produces a prediction model in the form of an ensemble of weak decision trees where the prediction is given by bagging.
5. **Long-Short Term Memory (LSTM)**: A recurrent neural network able to capture long-range temporal dependencies in the past due to its ability to avoid the vanishing gradient problem.

4 Experimental results and discussion:

For the experimental part, we propose to evaluate the contribution of the different features on the forecasting performances. For this purpose, we train XGB with the three sets of features (CAL, LT and ST). The obtained XGB models are compared with each other and with the chosen baseline models namely LV and CA. Once the best set of features is determined, we evaluate the performance of the advanced machine learning models which are XGB, RF and LSTM.

The training and the evaluation of the models are carried out on datasets related to suburb and inner-city stations. These datasets concern the period from January 2015 to June 2016, and they are split into training (year 2015 ≈ 66%) and test sets (year 2016≈33%). The best parameters of the models XGB and RF have been selected by a random search in conjunction with cross-validation on the main parameters of the algorithms (deep of trees, boosting, and bagging parameters), while for the LSTM the parameters have been fixed empirically. The assessment is performed based on each time step by measuring the root mean square error (RMSE) and a weighted absolute percentage error (WAPE) indicators. RMSE is the standard forecasting metric and WAPE is a normalized MAE that can be interpreted as the percentage of the overall error compared to the average value of the dataset. The error obtained for both training and test sets are given in Table 1.

Table 1: Baselines and XBG model performances on suburb and inner-city stations.

Model	*Suburb*		*Inner-city*	
	WAPE	*RMSE*	*WAPE*	*RMSE*
			Train score	
LV	18.2	36.5	42.0	187.0
CA	13.9	29.2	26.3	121.8
XGB - CAL	12.55	25.8	14.0	77.8
XGB - CAL+TS	11.9	24.6	9.6	54.7
XGB - ST	8.4	28.0	8.4	59.7
XGB - ALL	**7.34**	**15.6**	**5.9**	**32.4**
			Test score	
LV	23.9	46.8	46.0	205.0
CA	19.1	39.8	29.4	125.0
XGB - CAL	17.4	37.34	16.4	92.4
XGB - CAL+TS	16.8	36.3	12.6	74.0
XGB - ST	20.52	40.3	21.23	110.4
XGB - ALL	**16.0**	**33.8**	**12.1**	**71.2**

The results show that more we extend the feature set, better are the performances. The XGB-ALL with both long and short term features outperforms baselines and advanced models which are built based on one of the subset of features. These features help in the generalization of the XGB and avoid overfitting on the training data. We can observe a gain in performances between XGB-CAL and XGB-CAL+TS and by adding short term (ST) features to this later to obtain XGB-ALL. This gain is more

pronounced on the inner-city than the suburb station due to the train service variability. In the other hand, we can see that short-term features do not provide enough information on influencing factors to allow good predictions. The combination of short and long term features allows to capture the contextual information as well as the underlying dynamic of the train service at the station in order to provide the most accurate forecasts. Furthermore, the difference between the train and test errors can partly be explained by the lack of generalization of the models, but also by the evolution of the mobility demand on public transport from one year to another and that covers train and test set periods.

Table 2: Models performances with both short and long term features (ALL).

Model	*Suburb*		*Innercity*	
	WAPE	*RMSE*	*WAPE*	*RMSE*
		Train score		
XGB - ALL	7.3	15.6	5.9	32.4
RF - ALL	**5.1**	**11.4**	**5.1**	**27.8**
LSTM - ALL	13.3	26.5	13.1	69.3
		Test score		
XGB - ALL	16	33.8	12.1	71.2
RF - ALL	**15.9**	**33.5**	**12.0**	**70.8**
LSTM - ALL	16.9	36.5	18.5	85.0

Considering the best configuration of the set of features as the combination of both short and long term features, we propose to compare the performances of the advanced machine learning models trained on this set of features. As shown in the Table 2, the RF outperforms XGB and LSTM on suburb and inner-city datasets. LSTM provides poor performances on the test set in comparison with the ensemble learning models. This can be explained by the fact that LSTM has trouble with handling the irregular temporal structure of time series related to the train service and which is caused by its strong variability. While the ensemble learning models RF and XGB are able to identify contextual situations as snapshots independently of the underlying service variability.

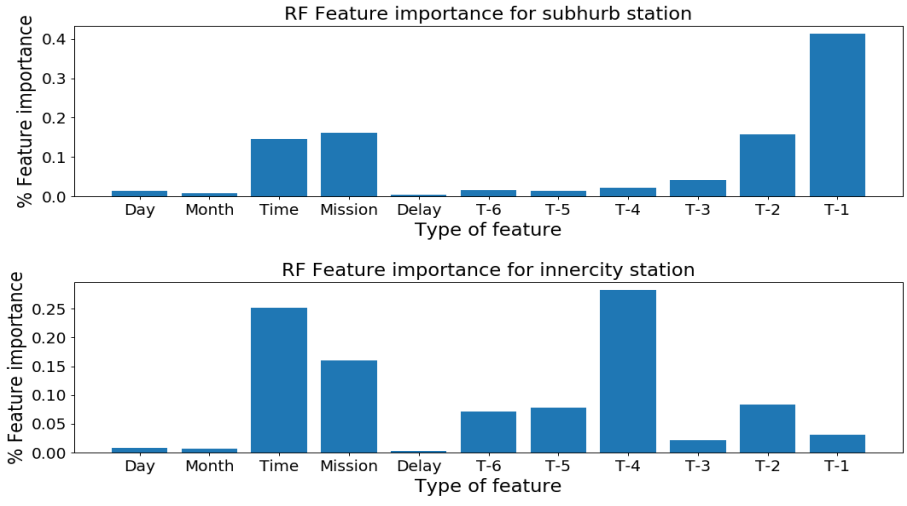

Fig. 3: Feature importance for the model RF-ALL on both stations.

When we analyze the feature importance of RF-ALL model on both stations, we can observe (see Figure 3) that the last load values are the most relevant features for the model, followed by the time and train service features. Against all expectations, the calendar features do not seem to be enough informative for the model in determining the short-term forecast, while, the delay feature is not reliable due to errors and noise in data transmission of train transit times at the station. A difference appears

between the suburb and the inner-city stations when we focus on the load features: indeed, for the suburb station, the most informative load feature corresponds to the last load at the passage of the train at the station. For the inner-city the most informative feature is the load on board of the fourth train in the lag window of the passages at the station. This can be explained by the fact that in the suburb station there is a single train service with unique route, where, in inner-city station there are several train services which serve different routes. The fourth train in the lag window is the one that probably served the same route than the train passing by the station at the prediction time.

5. Conclusion

In this paper, we investigated train load forecasting with advanced machine learning models and the contribution of the features that are part of the construction of these models. The obtained results have shown that machine learning models, and more particularly ensemble learning approaches such as XGB or RF can address the train load forecasting by using a combination between short and long term features that translate influencing factors. Furthermore, these models have proven their ability to deal with temporal irregularity of train service unlike other approach such as LSTM. Future works will carry out to enhance LSTM performances through redesigning its architecture.

Acknowledgment

This study is undertaken as part of the IVA Project coordinated by IRT SystemX and involves several partners, such as IFSTTAR, SNCF and the public transport authority of Ile-de-France.

References

[1] Ceapa, I., Smith, C., and Capra, L., “Avoiding the crowds: Understanding tube station congestion patterns from trip data,”ACM SIGKDD Int. Workshop on Urban Computing, pp. 134–141, 2012.

[2] Toqué, F., Côme, E., El Mahrsi M. K., and Oukhellou L., “Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks,”IEEE 19th Int. Conference on Intelligent Transportation Systems (ITSC), pp. 1071–1076, 2016.

[3] L. Heydenrijk-Ottens, V. Degeler, D. Luo, N. van Oort, and J. van Lint, “Supervised learning: Predicting passenger load in public transport,” in Conference on Advanced Systems in Public Transport, Brisbane, Australia, CASPT 2018.

[4] Wang, C. Wu, and Gao, X., “Research on subway passenger flow combination prediction model based on rbf neural networks and lssvm,” in Control and Decision Conference (CCDC), 2016 Chinese.IEEE, 2016, pp. 6064–6068.

[5] Zhang, J., Zheng, Y., and Qi, D., “Deep spatio-temporal residual networks for citywide crowd flows prediction.” in AAAI, 2017, pp. 1655–1661.