Passenger flow forecasting on transportation network: sensitivity analysis of the spatiotemporal features

Johanna Baro, Mostepha Khouadjia

Passenger flow forecasting on transportation network: sensitivity analysis of the spatiotemporal features

Johanna Baro *IRT SystemX*
Saclay, France
johanna.baro@irt-systemx.fr

Mostepha Khouadjia *IRT SystemX*
Saclay, France
mostepha.khouadjia@irt-systemx.fr

February 1, 2022

Abstract

Predicting the crowding level in train stations or the passenger load in trains can be useful to enrich the information available to passengers and improve train regulation processes or service quality levels. The main issue to handle when forecasting passenger flows is the structural variability of the related time series induced by the irregularity of train schedule and the influence of several contextual factors, such as calendar information and the characteristics of the served station. Forecasts depend on different contextual variables that generally have a spatial component, a temporal component, or both. We study the sensitivity of the spatiotemporal features of machine learning forecast models. Our main goal is to understand how the spatiotemporal features affect the performance of the models. First, we propose to study the impact of spatial and temporal inputs such as the served station, the train route or direction, and the type of day on the forecasting results to set up the best way to build a set of machine learning models to predict the passenger load of trains. Second, we address the effect of the temporal aggregation level on model performances for the forecasting task. The proposed models are based on ensemble machine learning approaches and have been deployed on a line of the Paris greater area railway network. A fine-grained evaluation is conducted as a support of the model's sensitivity analysis.

Sensitivity analysis, forecast, time series, Random Forest, passenger load, transportation network

1 Introduction

In recent years, the larger availability of information systems collecting data on public transport networks such as automatic fare collection (AFC), automatic passenger counts and loads in trains and bus (APC), or automatic vehicle locations (AVL) has allowed providing more detailed information and analysis on the state of transport systems. Long-term and short-term prediction of passenger flows is an active topic as it is an essential task to efficiently plan the transport offer, as well as to provide real-time information to passengers. Indeed, research studies have analyzed the effects of crowding on passengers' behavior and operation management, often highlighting impacts on waiting and travel time, route choice, and passengers' wellbeing [1].

In transport applications related to roads network, traffic prediction has been an active research subject for more than 40 years [2]. Various methods have been proposed to perform long-term and short-term predictions in this field. Studies focusing on flow prediction in public transport are more recent but rely on the large state of the art of this related problem. Methods can be split into three main categories: naïve, parametric, and nonparametric approaches. Naïve approaches such as historical averages are often used as a baseline to compare the performance of newly proposed models. Among parametric approaches are notably ARIMA-based models [3], Kalman filter-based models [4], or Bayesian networks [5]. Among the nonparametric are notably tree ensemble learning approaches [6–8] or neural network approaches [8–10].

Most of the studies which forecast passenger crowds based on ridership obtained from AFC provide results at the metro station level [5, 7, 9–11]. Their datasets have different characteristics but most of the studies focus on subway or light rail transit systems and they use aggregated data on grids with 1, 2, 10, 15, or 30 minutes intervals. In this situation, the studied dataset corresponds to regular time series. Only two studies [12,13] are performing forecast per vehicle (tram or metro train), using a time unit corresponding to the vehicle timestamp at a station. But still, the problem in [12] is a regular time series problem because metro trains are circulating every 5 minutes. It is also worth noting that the majority of the studies are performing short-term forecasts whereas [7] considered long-term forecasts.

Globally, most of the studies are experimental and applied to a small portion of the considered public transport system. Models are trained on few stations or a single line of the network and almost no study, except [9] which addresses the complete network, discusses the scalability of the proposed approach and studies the suitable right spatial and temporal granularity to obtain the most relevant forecasts. Those questions are related to the question of the impact of the spatiotemporal scale at which the data are observed on the regularity and the variability of the mobility pattern. [14–16] perform analytical studies of the spatiotemporal patterns presented in AFC data to gain knowledge on the regularity or variability of those patterns. Although no predictive model is developed in those studies, they are informative on the impact of the data granulometry on variability and hence on the possibility to obtain meaningful

predictions.

To deploy valuable forecasts on a public transport network it is crucial to take into account the following points:

- It is unrealistic to develop and maintain a model per station for a large public transport network containing hundreds of stations. Therefore, it is important to understand to which extent a single model can provide accurate forecasts for multiple stations. Knowing that stations can have different profiles, the quality of the forecast can be impacted according to their characteristics. To this end, [8] compared different models trained for each station to a model trained for all the stations at once, while [9] compared models trained for each station to models trained for stations regrouped based on their similar mobility pattern.
- The time interval on which the passenger flow is aggregated has an impact on the variability of the time series and thus on its predictability. Most of the studies use a fixed time step, often around 15 minutes which has been identified by [14] as a change point for the variability patterns for 3 different transportation networks. But being able to obtain prediction results on different time steps can be beneficial for planning as well as for real-time monitoring of trains capacity. [11] tries different time steps to validate the performance of the proposed approach, while [9] proposed an adaptative time step chosen according to the stations' profiles.

To contribute to this existing work, we propose to study and compare the results of prediction models of the passenger load trained on subsets with different spatiotemporal targets, and on datasets with different levels of spatiotemporal aggregation. Building light, transferable, and easily deployable models is a prerequisite in many industrial applications. Having this goal in mind, we analyze the spatiotemporal sensitivity of forecasting models based on the Random Forest algorithm [17]. The rest of this paper is organized as followed: Section 2 presents the available dataset, Section 3 presents the forecasting method and the different models that will be compared, then section 4 presents the sensitivity analysis of the different spatiotemporal features.

2 Data description

The study focuses on the French suburban railway Transilien line H operated by SNCF and located in Paris greater area. The railway line carries approximately 250,000 passengers daily with trains circulating during service hours from 5 am to 2 am the next day. The dataset covers 41 stations, organized in 4 branches connecting Paris city center to different suburban towns, located in the northern area of Paris from January 2016 to December 2017. The dataset consists of records of the passenger load collected via sensors in vehicles and calculated at each time a train makes a stop in a station.

Figure 1 shows examples of the passenger load plotted according to the observed station, the direction of the trains, and the type of day. The load is

influenced by many contextual spatiotemporal variables, among which are: the time of the day (morning and evening rush hours, etc.), the day of the week (weekday or weekend), or the day of the year (holidays), the station location (city center, suburb), the direction of the train and its mission, that is the list of served stations knowing that trains can be omnibus or direct. The effect of this last variable is visible on the figure, for Paris Nord or Saint-Denis stations, through oscillations in passengers load of two consecutive trains. Figure 1 also shows that train stops are not evenly distributed across stations: there are more records i.e. train stops in the inner city and dense suburbs stations than in suburbs stations located far away from Paris. We have to process an unbalanced spatiotemporal dataset with more observations on some stations, and trains passing at different frequencies according to the time of the day and the location (branch, station). In addition to those contextual factors, public transport demand and therefore passenger load can also be impacted by external events (social, cultural, sport, etc.). Hence, the model should integrate all the temporal, spatial, and exogenous factors listed above to provide accurate forecasts.

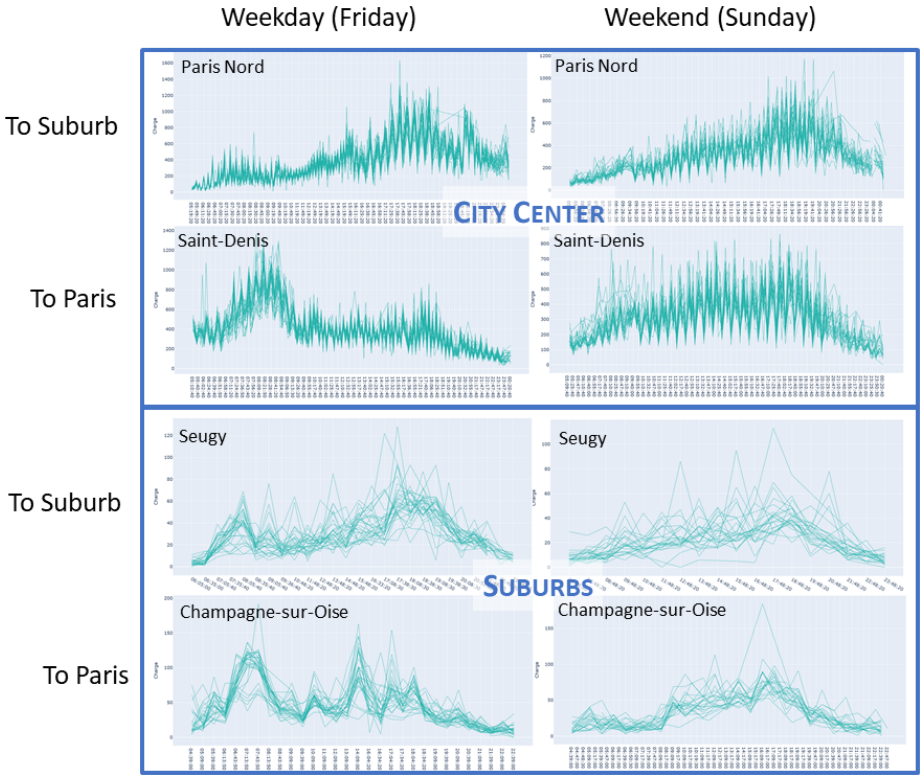

Figure 1: Variation of the passenger load profiles per station and direction (row), and type of day (column)

3 Methodology

3.1 Forecasting using Random Forest Regression

In previous works [18] on the same dataset, we compared passenger load predictions obtained with different machine learning models such as XGBoost [19], Random Forest [17], and a neural network based on an LSTM architecture. Due to their close performances, we chose in this study to work with the Random Forest tree ensemble learning approach for its low computational cost, its interpretability, and its good performance even compared to more complex approaches based on deep neural networks. Random Forests are well suited in our application where the processed dataset contains complex spatial and temporal dependencies that can be represented through feature engineering.

Tree ensemble methods are machine learning algorithms used to solve regression and classification problems. Those methods rely on the combination of weak estimators, here binary decision trees, to obtain a more robust estimator. Tree ensemble methods can be split into two categories: the averaging methods relying on independent estimators to build the robust estimator [17,20], and the boosting methods relying on estimators built sequentially to obtain the final one [19,21,22]. Random Forest is a popular algorithm of the averaging methods category which relies on a bagging method that proposes to combine and average several trees built independently to obtain a single and more performant estimator with reduced variance. The introduction of randomization processes with replacements on the records and the features (bootstrap samples) allows building complex models while limiting overfitting.

3.2 The different forecasting models

We study the results of passenger load prediction models trained on different subsets and on datasets with different levels of spatiotemporal aggregation and different features.

3.2.1 Dataset splitting and the different sub-models

Training subsets are defined using spatiotemporal variables that supposedly create patterns in the passenger load of the trains: the station at which a train stops, the direction of the train, and the type of days of service. The type of day is given by 6 categories (used by the transport authority) composed by weekdays, Saturdays and Sundays in and outside holidays periods, while the direction of trains is given by the 2 categories from and to Paris city center. The patterns created by those variables are observable in many public transport networks and most of the studies reviewed in section 1 treat them by deploying specific predictive models per station or per type of day for example. If we were to apply the same approach to our dataset, we would have to train up to 492 models by considering the different stations, types of day, and direction of the trains (see Table 2). From an operational point of view, it seems very difficult to keep track

of such a large number of models, especially when considering that line H is only one of the 15 railway lines of the Transilien network operated by SNCF, and one of the 31 lines metro and railway lines covering Paris metropolitan area, which serves only 41 stations on more than 700 stations. Having this difficulty in mind, one of our goals is to assess the performance of the predictive models trained on different subsets. The goal is indeed to diminish the number of models to maintain to ease the operational transfer of such a predictive approach.

3.2.2 Data aggregation and the different spatiotemporal scales

As stated by previous studies such as [14–16], the spatiotemporal aggregation of a dataset determines its variability, hence its predictability. To better understand the variability of the passenger load and its impact on the forecast, we train different models built on two criteria: the different spatiotemporal aggregation scales. As shown in Table 1, the spatial unit varies from train to station, and the temporal unit ranges from a minute to an hour for both predictive temporal horizons (long-term and short-term).

Table 1: Predictive models based on the forecasting horizons and the dataset spatiotemporal scales

Spatiotemporal scale		Forecast horizon	
Spatial unit	**Temporal unit**	**Long-term Features**	**Short-term Features**
(Train id, Station id)	Timestamp (hh:mm:ss)	Calendar Contextual Theoretical schedule	Calendar Contextual Theoretical schedule Realized schedule Train lag Station lag
Station id	15 minutes interval 30 minutes interval 60 minutes interval	Calendar Contextual Agg. theoretical schedule	Calendar Contextual Agg. theoretical schedule Agg. realized schedule Station lag

Predictions are computed for each train stop at a specific time (first row of Table 1). In this situation, the location of the forecast, that is the train stop, and the time of the forecast is the theoretical schedule at which the train is supposed to arrive at the station. Predictions are also computed on different aggregated datasets (other rows of Table 1). The aggregation consists of regrouping records in time intervals of fixed size: the passenger load is not predicted for a single train, but for all the trains passing at the station in the given interval. Here the target variable is the sum of the passenger load of the trains contained in the time interval, which are trains that are scheduled to stop at the target station during the interval.

3.2.3 Models features

To build the predictive models, the following features are considered:

- Contextual features: 1) the train stop id (i.e the station) encoded in one-hot, 2) the planned route of the train, i.e. list of served stations summarized as a mission code encoded in one-hot, 3) a boolean indicating if the train stop is the departure station of the train.
- Schedule features: Theoretical (resp. Realized) arrival and departure time at each train stop, expressed in time passed since the beginning of the daily service starting at 4 am.
- Calendar information: 1) the day of the year represented with a cyclical encoding using sine and cosine functions, 2) one-hot encodings of the weekdays and the holidays

Single-step predictions for the next timestamp or time interval t are computed. But we studied two forecast horizons: long and short terms predictions (see Table 1). Long-term predictions rely only on the contextual and calendar available a long time in advance. They are supposed to provide results that can be interpreted as mean results over detailed historical traffic context, that remain valid in a yearly horizon, and can be used for planning purposes. Short-term predictions rely on the same features and on lagged features which can capture the short-term dynamics useful for real-time monitoring and information. Those features are obtained by applying a lag on the passenger load or the realized schedule features sorted by:

- *Train lag*, that is the state of the train at its previous stops;
- *Station lag*, that is the state of the previous trains passing at the same station as the train stop.

The features considered in the aggregated models are globally the same as in the disaggregated model. Schedule features are average values over the aggregated trains and only lagged features on stations are considered as short-term feature candidates. The definition of lagged features on trains is more complex on the aggregated datasets due to the configuration of the network. The trains passing at a target station during a given time interval are not necessarily coming from the same previous stations because of the difference in missions (branches, direct vs. omnibus). To avoid mixing information coming from different locations of the network, we decide to exclude those lag features in the study.

4 Sensitivity analysis

In this section, we discuss the performances of the different models trained in the various settings described in the previous section. First, we study the

patterns associated with specificities of the spatiotemporal units and compare the results of the trained models on different subsets split using spatiotemporal characteristics. Second, we study the non-explainable variability and the effect of the aggregation of the dataset on the known variability.

4.1 Experiments settings

Experiments have been carried out with the Random Forest implementation of the scikit-learn library [23]. Models have been trained on the data of the year 2016, while the year 2017 is used as an independent test set. The experiments presented in section 4.2 have been performed at the disaggregated train stop scale using long-Term features on all the subset split summarized in table 2. And the experiments presented in section 4.3 have been performed on all the scales and types of features identified in Table 1.

Models are evaluated using the Root Mean Square Error (RMSE) metric (1), and mainly Weight Mean Absolute Percentage Error (WMAPE) metric (2) to take into account the difference of volume of passengers:

$$RMSE = \sqrt{\frac{\sum_t \sum_j (F_{tj} - A_{tj})^2}{n}} \tag{1}$$

$$WMAPE = \frac{\sum_t \sum_j |F_{tj} - A_{tj}|}{\sum_{t=1}^{n} |A_{tj}|} \tag{2}$$

where A_{tj} (resp. F_{tj}) is the observed (resp. forecasted) passenger load at time t and location j (train stop or station) and n is the total number of samples.

An iterative grid search with a cross-validation procedure has been used to fine-tune the hyperparameters of the models. This procedure has been applied on different training sets split using the type of day variable where each training set is composed of long-term and short-term features but without any lag features. We choose to split the dataset by type of days to have a manageable number of training sets, but with different characteristics in terms of size and pattern of passenger load. It allowed us to evaluate if the characteristics of the subset have an impact on the parameters and if so, on which parameters. In the end, a unique set of hyperparameters has been selected to be used to train all the models on the split dataset. Among the most significant parameters, we chose to train the Random Forest with deep trees ($max\_depth = 50$) combined with a minimal impurity decrease of 0.01. It allows the creation of large and complex trees capable of representing complex patterns using a large amount of information extracted from the features of the largest datasets. But the second parameter allows adapting to the smaller datasets by stopping the training early and creating smaller trees as shown by Table 2. The same set of parameters has been used for the models trained at the different spatiotemporal scales, with only an adaptation on the maximum numbers of samples and features as the training set in those experiments is bigger.

Table 2: Parameters of the independent models training

Models Split	Number of models	Training Time (mm:ss)*	Mean depth of forests
Global	x 1	06:06	48,9
Direction	x 2	04:05	45,0
Type of day	x 6	04:54	43,7
Type of day x Direction	x 12	03:35	36,3
Station	x 41	05:12	29,4
Station x Direction	x 82	06:40	26,3
Station x Type of day	x 246	07:35	18,5
Station x Type of day x Direction	x 492	09:46	15,8

* Evaluated on a virtual machine with 32 vCPU, 115GB of RAM.

4.2 How many independent models to train?

As presented in section 1, most of the studies on the forecasting of mobility flows consider that spatiotemporal characteristics such as the considered station or the type of days create such a great heterogeneity in the dataset, that it is best to train several models according to those characteristics. The question here is to what extent the forecasting models can discover those patterns and associate them to the different spatiotemporal units. Table 3 presents the global performances of the models trained on the different subsets split by spatiotemporal characteristics.

Table 3: Global performances of the independent models

Models		TRAIN (2016)		TEST (2017)	
Split	Number	RMSE	WMAPE	RMSE	WMAPE
Global	x 1	43,44	12,43	54,29	14,37
Direction	x 2	42,14	11,92	53,62	14,06
Type of day	x 6	42,60	11,90	49,37	13,87
Type of day x Direction	x 12	41,52	11,48	49,07	**13,71**
Station	x 41	40,15	10,87	53,12	13,74
Station x Direction	x 82	**39,97**	10,78	53,00	13,72
Station x Type of day	x 246	40,49	10,87	49,41	13,74
Station x Type of day x Direction	x 492	40,35	**10,82**	49,35	13,72

The main result is that no matter the split of the dataset, with the set of chosen hyperparameters, the performances seem to remain constant from the global model to the detailed 492 independent models, even if the global model performs a little bit worst than the smaller models. To asset the significance of the differences between the independent models, figure 2 presents the heatmap of the residuals difference (i.e. the difference between the residuals of a model A and the residuals of a model B) and highlight with black boxes pairs of models for which the null hypothesis of the Wilcoxon signed-rank test [24] cannot be rejected ($p - value > 0.05$). This test which allows comparing the distributions of pairs of forecast residuals indicates that most of the independent models still provide statistically different results. But the p-values should be used with

caution in our experiments which use a very large number of samples [25]. As displayed in the center of each cell, the median of the residuals differences is still very small for most of the models. With medians smaller than one passenger and weighted error metrics varying between 13 and 15 passengers (WMAPE), we can conclude that the different models provide mostly the same results. Indeed the passenger countings differences between the forecast models are too small to impact the quality of the information provided to the end-users.

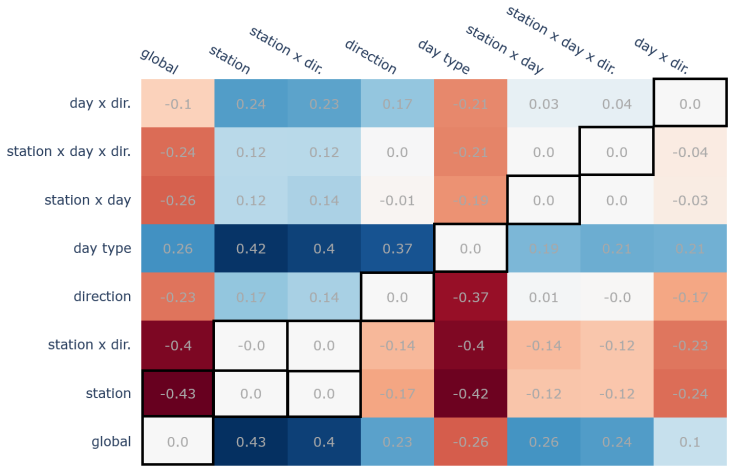

Figure 2: Models split comparison: median of the residuals differences & Wilcoxon test results

Looking at the results detailed according to the spatiotemporal variables, the performances remain similar regardless of the model. Figure 3 provides results according to the direction of the trains on the network. The prediction performances for trains from and to Paris city center are globally similar even if direction determines the passenger load according to the time of the day.

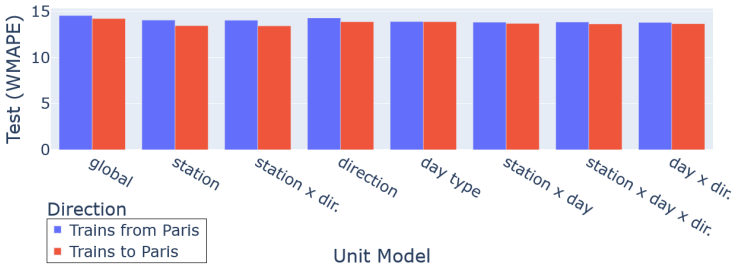

Figure 3: Models split comparison: details per directions

Figure 4 presents results according to the type of day. It shows that the performances of the models are globally similar no matter what the model is.

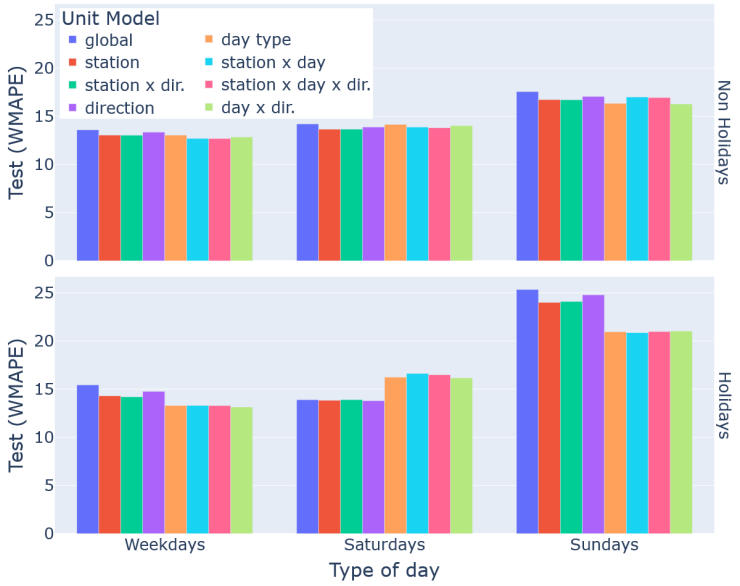

Figure 4: Models split comparison: details per type of day

But it confirms that the quality of the forecasts differs according to the type of day, even with the most 492-unit models. Holidays are harder to predict, especially Sundays and Saturdays, while non-holidays and especially weekdays (larger dataset, and supposedly with the most regular patterns such as home-to-work mobility patterns) exhibit better performances.

Figure 5 provides results per station and type of day. Here again, it appears that all the models have similar performances on all stations except for some stations located specifically on the Paris – Pontoise branch of the line. Differences between models appear for some trains (in direction to Paris) circulating during the weekdays and for some trains circulating (from Paris) the Saturdays and Sundays, stopping at the stations of Montigny-Beauchamp and Saint-Ouen-l'Aumône. For those stations, the models trained on the subsets split by type of day and direction gave worst performances than the other models on the test set. The location of those stations on the same branch led to think that a spatial pattern could not be discovered for this specific branch, that has the specificity to share most of its stations with another suburban railway line (RER C) on which we do not have any data. Figure 5 also shows that the performances differ per station for all the models. Looking at the WMAPE metric that normalizes the error with the mean passenger load, making the error per station comparable, it appears that performances deteriorate in small suburban stations such as Seugy, Luzarches, or Champagne-sur-Oise. The passenger load of those stations is always smaller, but also it exhibits a higher variability than in inner-city stations making the forecast problem harder.

Overall, it appears that training a global model to forecast the passenger load

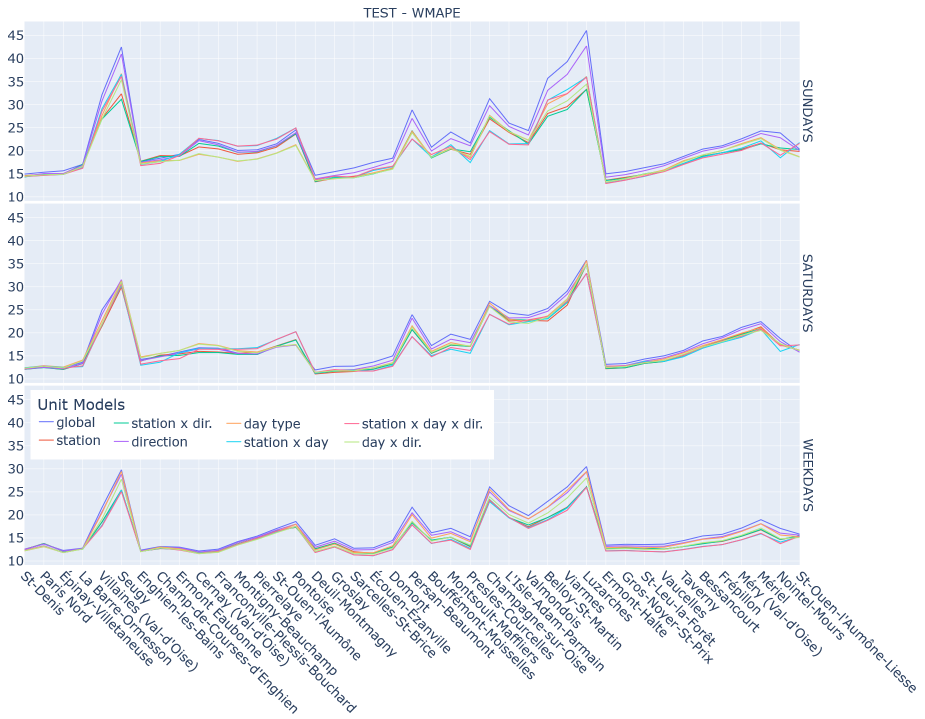

Figure 5: Models split comparison : details per stations and types of day

of all the trains circulating at any station and any type of day on railway line H is similar to training a large number of independent models, even if spatiotemporal variables indeed condition the quality of the forecast. Of course, those results are determined by the set of hyperparameters chosen to train the Random Forests. We favor here having a unique model to maintain, but the compromise between the size of the model and the number of models could lead to another choice depending on the application. Notably, when considering the training time showed in Table 2, it appears that the overall training time increases with the number of models to train, but the global model takes more time to train than the models split by direction and type of day. If a short training time is a requirement, the split according to those criteria offers a compromise between the unique training on a large and complex dataset and numerous trainings on simpler datasets that are costly from a time point of view.

4.3 How the spatiotemporal aggregation affects the performances?

In this section, we investigate the impact of the aggregation of the spatiotemporal units on the variability of the dataset. The question here is to what extent

the aggregation helps to obtain better forecasts of the passenger load. To answer this question, we will compare the results of models trained on different spatiotemporal units. Table 4 gives an overview of the performances of the long-term and short-term models trained on 6 spatiotemporal units: the passenger load at each train stop on a given timestamp, and the passenger load per station on time steps of 1, 5, 15, 30, and 60 minutes.

Table 4: Performances of the LT and ST models trained on different spatiotemporal scales

		Long-Term		Short-Term	
Spatial Unit	**Temporal Unit**	**TRAIN (2016)**	**TEST (2017)**	**TRAIN (2016)**	**TEST (2017)**
Station	60 minutes	**9,46**	**11,96**	*9,41*	*13,04*
	30 minutes	10,48	12,96	10,61	14,21
	15 minutes	10,83	13,39	10,85	14,48
	5 minutes	10,77	13,34	10,68	14,37
	1 minutes	10,67	13,24	10,68	14,37
Train	Disagregated	10,83	13,88	**4,07**	**5,43**

The short-term 'Train' model using the train and station lag features outperforms all the other models. Notably, this model outperforms all the other short-term models where the load is aggregated per station, knowing that those models only use station lag features. The train lag features are determinants to obtain accurate predictions of the passenger load. Figure 6 presents a heatmap of the medians of the residuals differences used to compute the Wilcoxon signed-rank test. For all the pairs of models, the null hypothesis can be rejected but here again the large sample size is prone to caution in the interpretation of the p-values. Looking at the median, the used features (short or long term) as well as the spatial unit condition the results. However, the temporal aggregation seems to have a smaller influence on the results which translates in medians closer to zero. Considering the WMAPE metrics and the median of the residuals, the temporal aggregation seems to slightly affect the performance as the error slightly decreases with the aggregation of the data per station.

Figure 7 presents a comparison of the forecasting results aggregated per station and type of day for all the models. Here again, the train_ST_LT model outperforms all the other models on the majority of the stations, except the departure stations where the train lag features are null by default. For the departure stations, all the other short-term and long-term models, no matter what the spatiotemporal scale, provide similar results. For those stations, the train_ST_LT performs worst than all the other models because from an analysis of the importance of the features this model relies mostly on the train lag features. It is not able to learn properly the passenger load behavior of those stations based on the other features.

When comparing the models of the other no-departure stations, it appears that the temporal aggregation plays in favor of the accuracy of the forecasts per station, the 60 minutes time steps being more accurate than 30, 15, 5, and 1 minute time steps, and the 'train_' model for long-term and short-term models. And interestingly when comparing the short-term and long-term models at the

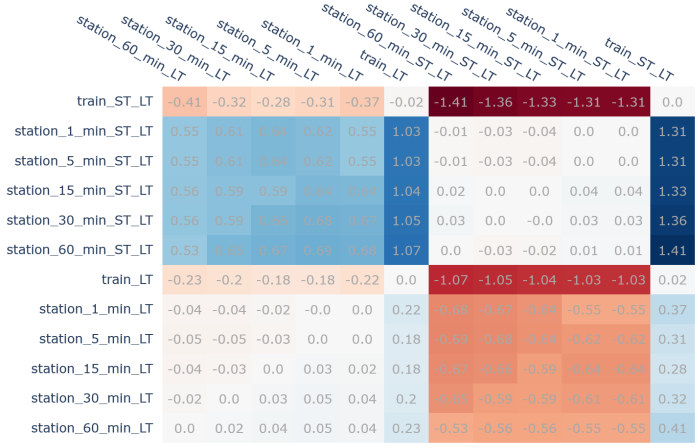

Figure 6: Models scales comparison: median of the residuals differences

same spatiotemporal scale for the 'station_' forecast, it seems that lag features are not beneficial no matter what the scale as long-term forecast models are better. Short-term features seem slightly beneficial only for stations where the frequency of trains is high.

Except for the 60 minutes aggregation, the performances of the other temporally aggregated models seem to depend more on the station profile than on the temporal aggregation. The frequency of the trains varies according to the location: in Paris city center it is around 3 minutes during peak hours, whereas in suburbs it is around 30 minutes during peak hours and 1 hour during off-peak hours. Hence, the temporal aggregation does not reduce the variability homogeneously. In some situations instead of helping it seems to lose some information or to enhance the variability compared to the other stations on which the aggregated models may focus more during the training. But as stated previously the impact of temporal aggregation on the performance is still overall minimal.

Figure8 compares models results according to the direction of the trains and the different peak hours of the day. The morning and evening peaks are identified according to the adaptation of the trains offer with more frequent trains passing at each station. The morning peak goes from 6 am to 10 am, while the evening peak goes from 4 pm to 8 pm. The off-peak period is defined as the complementary period of service outside those two peaks. Except for the train_ST_LT model, similar patterns in the results are observable through the scales depending only minimally on the type of forecast. When looking at trains circulating from Paris to the suburbs, the morning peak (red) and the off-peak hours (green) are the most difficult periods to predict. It is expected that the evening peak shows more regularity in this direction because from the urban organization it is the time of the day where workers are leaving Paris city center to go back to residential areas located in the suburbs. The off-peak period is

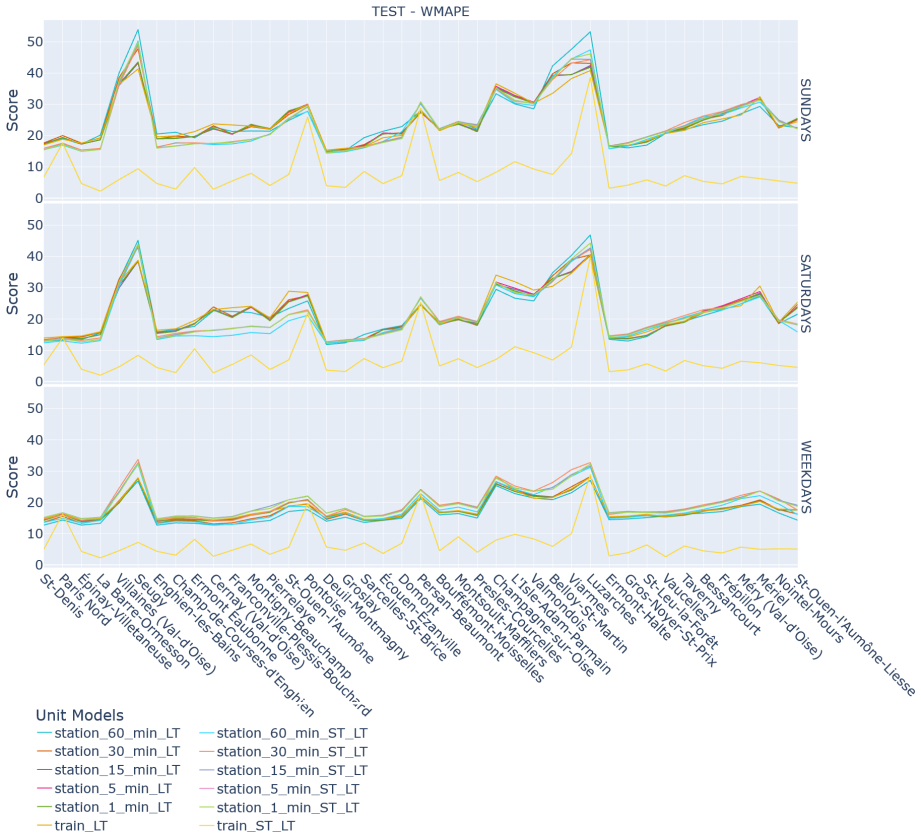

Figure 7: Models scales comparison: details per stations and types of day

easier to predict than the morning peak for stations close to Paris because it may contain more regular mobility patterns associated with work-related mobility but with staggered hours. When looking at trains circulating from the suburbs to Paris, the off-peak (green) is the most difficult period to predict, probably because it regroups all the non-work-related patterns associated for example to leisure activities, followed by the evening peak. The morning peak period in this direction presents more regularity, probably because it is the time of the day where workers are going to Paris to work.

From the comparison of the results of the different models, it appears that aggregating the dataset helps a little in the prediction but it is not determinant on the quality of the results. Results are consistent through scales, although using the disaggregated dataset provides better forecasting models because it helps to incorporate determinant lag features on the previous status of each train. Using similar train lag features has not been tested on the aggregated scales because of the network architecture of the studied railway line. Hence, training models on aggregated datasets could be preferable for the long-term

Figure 8: Models scales comparison: details per direction and peak hours

forecast where the train precision is not required, while training models on the disaggregated dataset could be preferable to obtain the most accurate and short-term forecast.

5 Conclusion

In this paper, we studied the sensitivity of spatiotemporal features on the prediction of the passenger load on board the train. This study offers valuable insights on the impact of the spatiotemporal features involved in the construction of the machine learning models which can help to easily deployed and maintained such models in real-world settings. First, it appeared that it was sufficient to train a single Random Forest model for all the stations of the studied railway line and that it was not necessary to train several models for each station, direction, or type of day. With the relevant set of hyperparameters, it was possible to train a forest with trees complex enough to represent various patterns depending on the location of the trains or on the temporal context of circulation of the trains. Second, it appeared that the forecast results were consistent through different scales of spatiotemporal aggregation. When data of the passenger load was aggregated according to the station on a regular temporal grid of 1, 5, 15, 30, or 60 minutes, the aggregation played only a little in favor of the forecast by bringing more regularity to the dataset. It could notably be beneficial to train long-term forecast models. But observing the passenger load of each train was the best way to obtain accurate short-term forecast models relying on the passenger load observed at the previous train stop. This model was the most accurate one of the study, even if performances on departure stations of the trains were similar to the other models. For those stations, the other contextual features related to the calendar, schedule, or location were still relevant enough to obtain valid forecasts.

Future research should investigate more extensively the evolution of model responses over time. It would be interesting to analyze the impact of different sizes of historical data to train the models to evaluate the necessary amount of information to obtain relevant forecasts. It could also help to better understand to which extent the predictive models remain valid through time. Moreover, future research should extend the sensitivity analysis to other predictive approaches such as the previously studied LSTM neural network model [18] compare the performances of the different models in those various spatiotemporal settings.

Acknowledgment

This research is a part of the IVA Project, which aims to enhance traveler information in the mobility context. It is carried out under the leadership of the Technological Research Institute SystemX, with the partnership and support of the transport organization authority Ile-De-France Mobilités (IDFM), the railway operator SNCF, and public funds under the scope of the French Program

"Investissements d'Avenir".

References

[1] A. Tirachini, D. A. Hensher, and J. M. Rose, "Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand," *Transportation Research Part A: Policy and Practice*, vol. 53, pp. 36–52, Jul. 2013.

[2] E. Vlahogianni, M. Karlaftis, and J. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transportation Research Part C: Emerging Technologies*, vol. 43, Jun. 2014.

[3] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, "Using an ARIMA-GARCH Modeling Approach to Improve Subway Short-Term Ridership Forecasting Accounting for Dynamic Volatility," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1054–1064, 2018.

[4] S. Liang, M. Ma, S. He, and H. Zhang, "Short-Term Passenger Flow Prediction in Urban Public Transport: Kalman Filtering Combined K-Nearest Neighbor Approach," *IEEE Access*, vol. 7, 2019.

[5] J. Roos, S. Bonnevay, and G. Gavin, "Short-Term Urban Rail Passenger Flow Forecasting: A Dynamic Bayesian Network Approach," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2016, pp. 1034–1039.

[6] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees," *Sustainability*, vol. 8, no. 11, p. 1100, Nov. 2016.

[7] F. Toqué, M. Khouadjia, E. Come, M. Trepanier, and L. Oukhellou, "Short long term forecasting of multimodal transport passenger flows with machine learning methods," in *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 560–566.

[8] F. Toqué, E. Come, L. Oukhellou, and M. Trepanier, "Short-Term Multi-Step Ahead Forecasting of Railway Passenger Flows During Special Events With Machine Learning Methods," in *Conference on Advanced Systems in Public Transport and TransitData (CASPT)*, 2018, p. 15p.

[9] J. Zhang, F. Chen, and Q. Shen, "Cluster-Based LSTM Network for Short-Term Passenger Flow Forecasting in Urban Rail Transit," *IEEE Access*, vol. 7, pp. 147 653–147 671, 2019.

[10] Y. Liu, Z. Liu, and R. Jia, "DeepPF: A deep learning based architecture for metro passenger flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 18–34, Apr. 2019.

[11] L. Tang, Y. Zhao, J. Cabrera, J. Ma, and K. L. Tsui, "Forecasting Short-Term Passenger Flow: An Empirical Study on Shenzhen Metro," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3613–3622, Oct. 2019.

[12] E. Jenelius, "Data-Driven Metro Train Crowding Prediction Based on Real-Time Load Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 6, pp. 2254–2265, Jun. 2020.

[13] L. Heydenrijk-Ottens, V. Degeler, D. Luo, N. van, and H. van Lint, "Supervised learning: Predicting passenger load in public transport," in *Conference on Advanced Systems in Public Transport and TransitData (CASPT)*, 2018, p. 9.

[14] C. Zhong, M. Batty, E. Manley, J. Wang, Z. Wang, F. Chen, and G. Schmitt, "Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data," *PLOS ONE*, vol. 11, no. 2, Feb. 2016, publisher: Public Library of Science.

[15] W. Geng and G. Yang, "Partial Correlation between Spatial and Temporal Regularities of Human Mobility," *Scientific Reports*, vol. 7, no. 1, p. 6249, Jul. 2017.

[16] E. Manley, C. Zhong, and M. Batty, "Spatiotemporal variation in travel regularity through transit user profiling," *Transportation*, vol. 45, no. 3, pp. 703–732, May 2018.

[17] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[18] K. Pasini, M. Khouadjia, F. Ganansia, and L. Oukhellou, "Forecasting passenger load in a transit network using data driven models," in *12th World Congress on Railway Research (WCRR)*, Tokyo, Japan, Oct. 2019.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Aug. 2016.

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[21] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] F. Wilcoxon, "Individual Comparisons by Ranking Methods," in *Breakthroughs in Statistics: Methodology and Distribution*, ser. Springer Series in Statistics, S. Kotz and N. L. Johnson, Eds. New York, NY: Springer, 1992, pp. 196–202.

[25] M. Lin, H. C. Lucas, and G. Shmueli, "Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem," *Information Systems Research*, vol. 24, no. 4, pp. 906–917, Dec. 2013.