One-Station-Ahead Forecasting of Dwell Time, Arrival Delay and Passenger Flows on Trains Equipped with Automatic Passenger Counting (APC) Device

Rémi Coulaud, Christine Keribin, Gilles Stoltz

One-Station-Ahead Forecasting of Dwell Time, Arrival Delay and Passenger Flows on Trains Equipped with Automatic Passenger Counting (APC) Device

Rémi COULAUD[1,2], Christine KERIBIN[2], Gilles STOLTZ[2]

[1]SNCF Voyageurs – Transilien, 10 rue Camille Moke, 93220, Saint-Denis, France
[2]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
Corresponding Author: Rémi Coulaud (remi.coulaud@sncf.fr)

Abstract

We consider a suburban railway network line in the greater Paris area, with sub-branches. Trains of the line are equipped both with automatic vehicle localization (AVL) and automatic passenger counting (APC) devices, leading to a rich data set with simultaneous measurements of variables related to railway operations (arrival delay, dwell time) and passenger flows (numbers of passengers alighting and boarding, total load at departure). We aim for one-station-ahead forecasting of each of these five variables independently from each other. To do so, we build a bi-auto-regressive approach consisting of using the past values of the variable of interest along a first dimension, given by past stations along the train ride, and along a second dimension, given by past trains at the station. A building block of this approach is a train-station representation that accommodates different types of train services. We identify repeated patterns in this representation and exploit this fact. Indeed, the proposed bi-auto-regressive models are based on linear regressions whose coefficients depend on the stations and possibly only on the location of the train ride within a repeated pattern. This results in models that have a smaller complexity than extremely local models tailored to the timetables, with no significant decrease in accuracy.

Keywords: dwell time, arrival delay, passenger flow, forecasting, bi-auto-regressive models

1. Introduction

We consider a suburban railway network line, namely, line H of the SNCF railway network of the greater Paris area, in the direction from suburbs (from the origin station "Pontoise" and from the intermediate station "Montsoult-Mafflier") to Paris (terminus station "Gare du Nord"), see Figure 1.

We are interested in the short-term (next station or next train) forecasting of some variables related to train stops in a given station: numbers A and B of passengers alighting and boarding, load L at departure, as well as dwell time T and arrival delay ΔA. We use X to refer to any of these quantities in a generic way. At a high-level, our approach consists of predicting the value $X_{k,s}$ of a quantity for the k-th train in the s-th station based on recent observations of the same quantity at the same station for earlier trains and at earlier stations for the same train, i.e., we consider some auto-regressive modelling, where "time" is measured through pairs (k, s) of trains and stations.

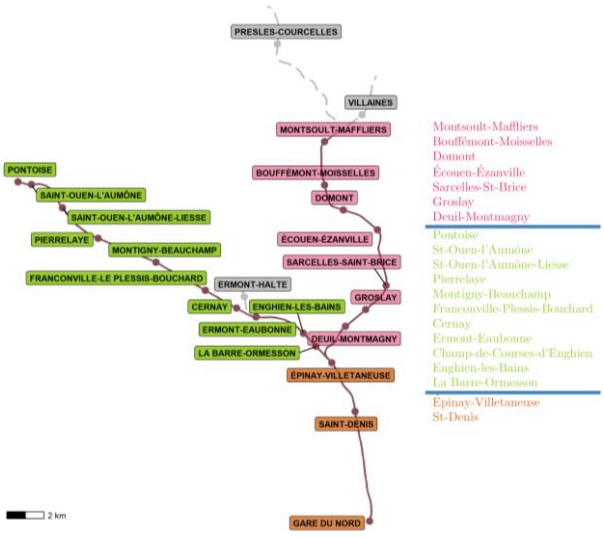


Figure 1*:* Left picture: The subset of the railway line considered, consisting of two branches (in green and in pink) merging for the final three stations (in brown). Stations not considered are in grey. Right column: List of the corresponding stations except for the terminus station, with the same colour code.

We do this in a novel way: we introduce our methodology in Sections 2.1 and 2.2 and then compare it to existing ones in Section 2.3.

2. Methodology

Our simplified representation of line H is composed of two branches that merge for the final three stations. As there is no scheduled train overtake, we may reindex the scheduled timetable from time in x-axis (left part of Figure 2) to train number in x-axis (right part). Trains are ranked according to their stops at the second station of the common part of line H. We discard the terminus station in the timetable and in our predictions, as the forecasting of most quantities of interest (dwell time, number of passengers alighting and boarding, load after departure) is of no interest or not applicable.

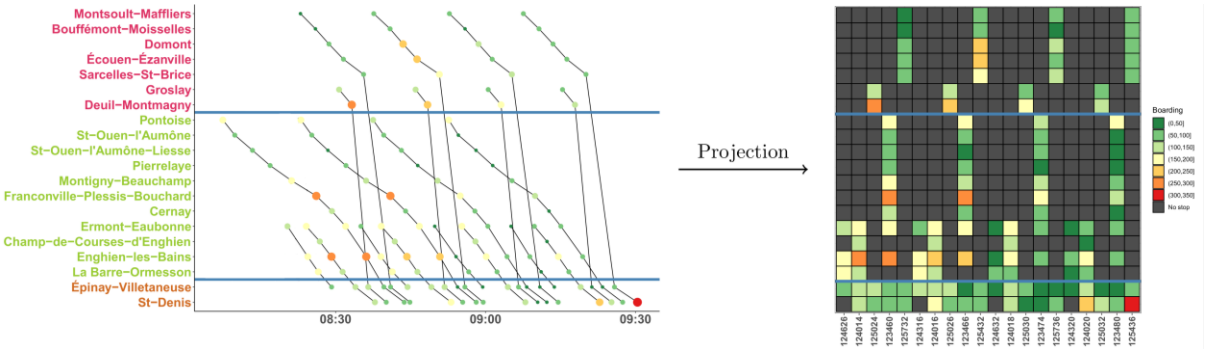


Figure 2: Projection of time-station timetable (left graph) into a train-station timetable (right graph). Colours and sizes of circles indicate the level of the boarding. Trains do not stop at stations marked in dark grey on the right graph.

2.1. L-shaped Neighbourhoods for Prediction

To forecast some information (for example, for train 29 at station Épinay-Villetaneuse: purple cell on Figure 3), we may use past information at the station of interest (for earlier trains: pink cells) and along the past stations of the given train ride (blue cells). Of course, we may restrict our attention to a shorter memory range ($P = 2$ past trains in dark pink and $Q = 3$ earlier stations in dark blue). Strikethrough cells contain future information and cannot be used for prediction. All in all, the information to be used for prediction is shaped like an inverse-L: we will refer to it as an L-shaped neighbourhood with sizes P and Q.

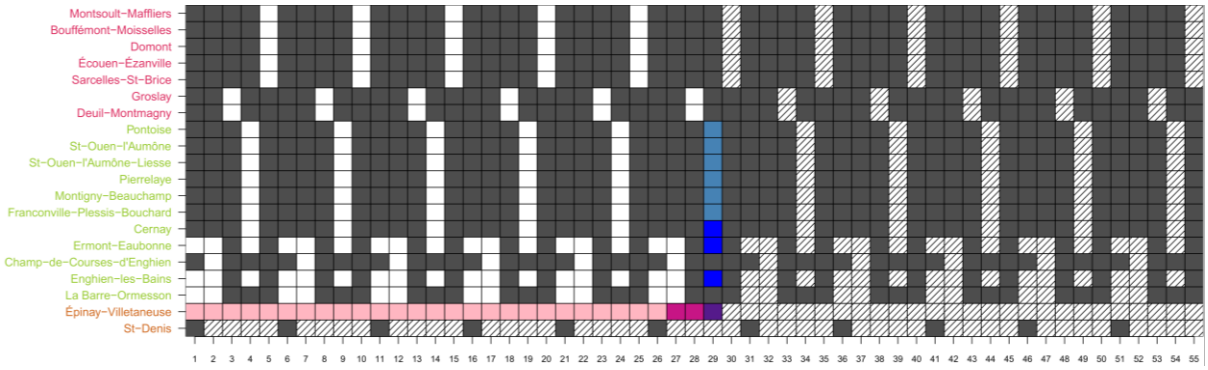


Figure 3: Identification of the underlying structure in the train-station graph of the morning peak hour. Dark grey cells mean no stop for corresponding train-station pairs. Information available at earlier stations and for earlier trains are in blue and pink, respectively. Dark blue and dark pink denote the most recent information. Strikethrough cells contain future information and cannot be used for prediction.

In the right graph of Figure 2, we read an underlying structure that consists of 4 repetitions of a given pattern. structure, see Figure 3. We denote by M the periodicity of the repetitions: here, the same sub-structures arise every $M = 5$ trains. On top of these intra-day repetitions, we also consider inter-days repetitions, i.e., each day

may be seen as a realization of a given stochastic process.

For the sake of clarity, Figure 4 depicts the repetitions of the L-shaped neighbourhood introduced in Figure 3, during the morning peak hour of a given day.

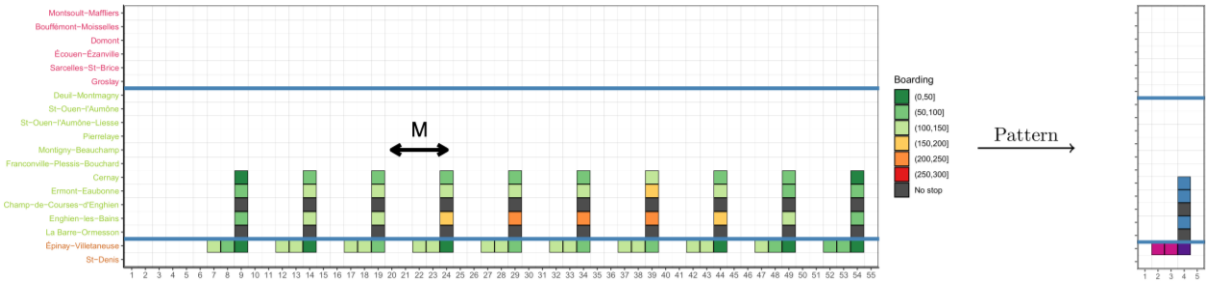
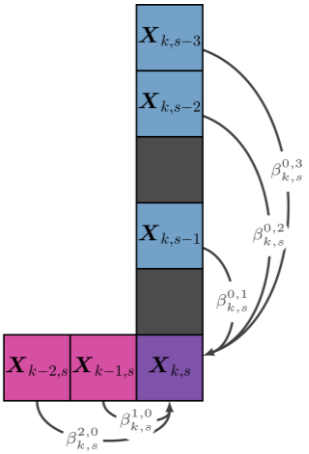


Figure 4: Illustration of the repetitions of a given L-shaped neighbourhood in the train-station graph of the morning peak hour.

2.2. L-shaped Regression Models, with Different Degrees of Stationarity

We model a quantity $X_{k,s}$ of interest (dwell time *T*, number *B* of passengers boarding , etc.) at a station *s* for train *k* as an affine function of the same quantities of interest $X_{k-p,s}$ and $X_{k,s-q}$ in the L-shaped neighbourhood considered – hence the name of L-shaped regression models. The three models introduced differ by the flexibility allowed for the coefficients. In the "stationary" model, all instances of the L-shaped neighbourhoods within the peak hours and among days are considered repetitions of the same stochastic process, while in the "non-stationary" model, days are considered repetitions of the same stochastic process, but no specific structure is assumed within the peak hour. The respective linear regression models have coefficients that only depend on the station *s* and the location of *k* within a pattern (which is given by the value of *k* modulo *M*, denoted by *k*[*M*]), and, on the contrary, that may fully depend on *k* and *s*. A model lying between these two extremes is referred to as "semi-stationary", where the intercept coefficient may fully depend on *k* and *s* (which models some variation of the level within the morning peak hour) but the coefficients for explanatory variables only depend on *k*[*M*] and *s* (which models some intrinsic relationship with neighbouring values). Formally, the modelling equations read as follows, where $\varepsilon_{k,s}$ denote the error terms:



Non-stationary: $$X_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k,s} X_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k,s} X_{k,s-q} + \varepsilon_{k,s}$$

Semi-stationary: $$X_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} X_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} X_{k,s-q} + \varepsilon_{k,s}$$

Stationary: $$X_{k,s} = \beta^{0,0}_{k[M],s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} X_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} X_{k,s-q} + \varepsilon_{k,s}$$

2.3. Literature Review and Main Contributions

Main contribution #1: Assessing forecasting methods on 5 variables. In the public transportation literature, short-term forecasting methods are typically built for one specific variable at a time: e.g., dwell time in Kecman

& Goverde [4], Li et al. [5], arrival delay in Corman & Kecman [2], passenger load at departure in Bapaume et al. [1], Jenelius [3], Pasini et al. [6]. Although the underlying methods are often generic, they were each tested only for a specific variable. The richness of our data set allows for the assessment of each model considered on 5 different variables. We chose to evaluate performance one-single-step ahead as Li et al. [5] did but keep in mind that this is merely a first step towards operational solutions, which require rather multiple-step-ahead forecasts as noted by Bapaume et al. [1].

Main contribution #2: Train-station representation despite sub-branches. Conversions of time-station representations into train-station ones were already performed in Bapaume et al. [1], Jenelius [3] on simpler networks with a unique branch and simpler train rides, with no overtake. We extend such conversions to a case where there are sub-branches and several train services types. We note however that the more complex approach by Corman & Kecman [2] allows for such multiple sub-branches, but leads to models with significantly more parameters (see below), while we aim for frugal models.

Main contribution #3: Bi-auto-regressive modelling. To the best of our knowledge, a systematic exploration of the power of auto-regressive modelling using both the recent past along the train ride and the recent past at the station considered was not offered by the literature so far. Instead, Jenelius [3] built auto-regressive models along the train ride using levels of the variable of interest at the station considered, formed by averages over some possibly short- and long-term past.

Main contribution #4: A balance between frugality and complexity. Some models of the literature are possibly too frugal and hold independently of the stations and train rides considered (i.e., do not depend on s and k, with our notation), as in Li et al. [5]. Some other approaches rely on possibly too many parameters, as the one in Corman & Kecman [2] which proposes a modelling even more complex than what we termed the "non-stationary approach" above, where all coefficients depended on s and k. Thanks to the identification of repeated patterns, possibly combined with the existence of a trend taken into account by the non-stationary approach, we propose an intermediate modelling, where coefficients depend on s and on $k[M]$, the location of the train ride within a repeated pattern. However, our approach does not accommodate well deviations to the scheduled timetable so far, just as the one by Corman & Kecman [2]; it is thus somehow less flexible to these deviations than the approach of Li et al. [5], for instance, which however ignores a significant part of information available and rely on a local view given by a single train ride.

Note. Some ad hoc, possibly very complex models (e.g., deep-learning based treatment of images, see Bapaume et al. [1], Pasini et al. [6]), were also proposed for some of the variables considered, but they are out of the scope of this contribution, which targets generality and simplicity of the models constructed.

3. Presentation of the Data Set

Our approach crucially relies on having identical timetables from day to day: we therefore have to restrict our attention to working days. We do so for the morning peak hours (55 trains daily during the 6h33 - 9h28 time range) and for the 106-day-long period ranging from January 7, 2019 to July 5, 2019. We recall that we consider 20 stations. All in all, the data set consists of about 34,000 observed stops. Passenger flow variables (numbers A and B of passengers alighting and boarding, load at departure L) are measured by automatic passenger counting (APC) sensors. Railway operations variables (dwell time T and arrival delay ΔA) are measured by automatic vehicle localization (AVL) and track data. There are few missing values (10% for passenger flow variables and 6.5% for railway operations variables).

As explained in Section 2, abidance by the timetable is key to run our methodology. When trains take over or are suppressed, we clear locally the corresponding data (e.g., both train rides in case of a takeover). Doing so, 78% of the 34,000 potential observations are available and used.

We split the dataset into two data sets: a train set (January 7 - May 20) and a test set (May 21 - July 5), accounting for 70% and 30% of the observations, respectively. We estimate the parameters of the L-shaped regression models on the train set, by ordinary least squares, and compute their associated performance on the test set, which we report next. Our main indicator of performance is the mean absolute error (MAE).

4. Results

For the sake of space, we only report results for symmetric L-shaped neighbourhoods, i.e., with $P = Q$. The case $P = Q = 0$ corresponds to predictions given by average values on the train set, per cell (k, s). "Real" predictions based on the local context use $P = Q \geq 1$. Table 1 reports the global MAE (i.e., the MAE averaged over all possible cells of Figure 2) of the methods introduced in Section 2.2, for various values of $P = Q$. The reference method consists of the non-stationary L-shaped regression models with $P = Q = 1$, which is the closest to what the literature considered so far (see Section 2.3). We however note that for two variables at least (dwell time T and number of passengers alighting A), reporting average values (i.e., using $P = Q = 0$) results in decent predictions.

One issue of the reference method is the large number of coefficients to be estimated – around 900. We now address the wish to reduce the complexity of the method while preserving performance. There is a general balance in statistical models between their intrinsic ability to model the phenomenon at stake, which usually requires more coefficients, and the need to properly estimate these coefficients, which requires not having to estimate too many of them given a data set of fixed size.

Models			**Railway operations**		**Passenger flow**		
Name	*L-Shape*	Number of coefficients	T [s]	ΔA [s]	A [count]	B [count]	L [count]
Non-stationary	P = Q = 0	327	9,7	35,8	10	21	70
	P = Q = 1	915	9,5	16,1	9	18	22
Semi-stationary	P = Q = 1	403	9,2	16,1	10	18	23
	P = Q = 2	440	9,2	15,8	9	18	23
	P = Q = 3	466	9,1	15,8	9	18	23
Stationary	P = Q = 1	76	9,3	16,2	10	21	30
	P = Q = 2	113	9,2	15,8	9	20	29
	P = Q = 3	139	9,2	15,9	9	20	29

Table 1: Global MAE (mean absolute error) of some of the forecasting methods considered (indicated in the first two columns). For each method, we report the number of coefficients it uses (column 3), as well as the global MAE achieved for each of the five variables to be predicted (columns 4-8): dwell time T, arrival delay ΔA, numbers A and B of passengers alighting and boarding, load at departure L. The reference method is in blue and good alternative methods are in green.

For railway operations variables (dwell time T and arrival delay ΔA) stationary L-shaped regression models with $P = Q = 2$ rely on only about 100 coefficients while obtaining slightly better performance than the reference model. For passenger flow variables, a good alternative model consists of semi-stationary L-shaped regression

models with $P = Q = 1$: it requires about twice fewer coefficients than the reference method while obtaining an only slightly worse performance.

We move to a more local study of the performance, and report the difference of MAE between the reference method (non-stationary model with $P = Q = 1$) and the alternative methods (green cells in Table 1). We observe that for a majority of train-station pairs, the MAE (over repetitions within peak hours and over days) are almost equivalent, i.e., differ by at most 1 s or 1 passenger; see the white cells on Figure 5. This is especially remarkable for the dwell time T. The arrival delay ΔA and the number B of passengers boarding are locally better predicted by the alternative models. On the contrary, the alternative models are less accurate for the local prediction of the load at departure L (especially in one of the sub-branches) and the number A of passengers alighting (especially in the common part of the line).

References

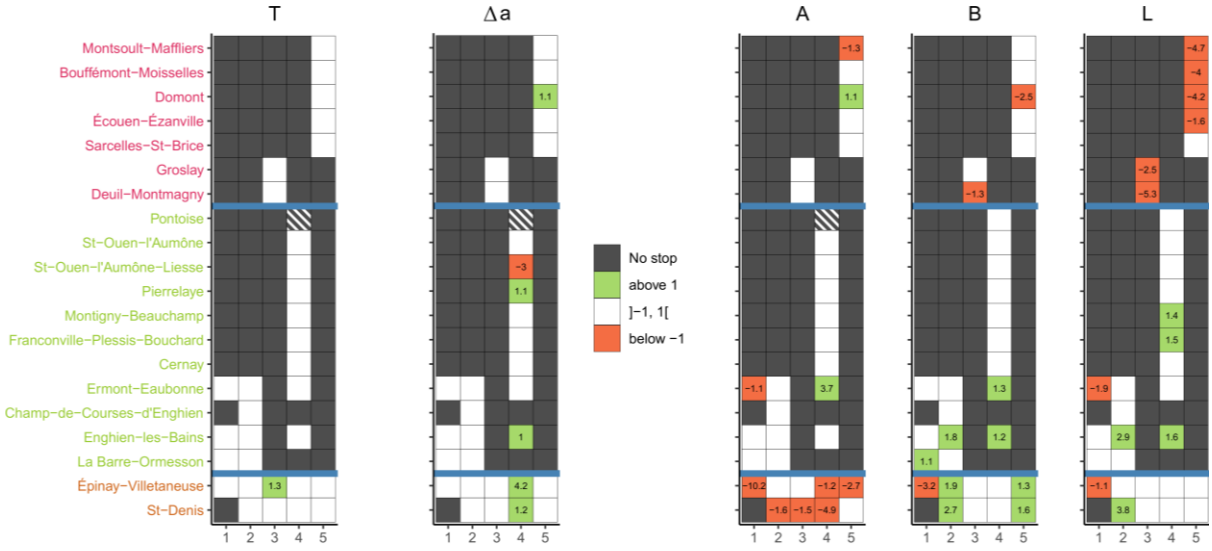


Figure 5: Average difference of performance between the alternative methods introduced in Table 1 and the reference method. Each column corresponds to a variable of interest. Each cell reports the average over corresponding train-station pairs during the peak hour and over the days. Significant improvements over the reference method are in green, deteriorations are in orange, while white denotes equivalence. The numbers indicate the average difference in MAE.

[1] Bapaume, T., Côme, E., Roos, J., Ameli, M., & Oukhellou, L., "Image Inpainting and Deep Learning to Forecast Short-Term Train Loads", *IEEE Access*, vol. 9, pp. 98506-98522, 2021.

[2] Corman, F., & Kecman, P., "Stochastic prediction of train delays in real-time using Bayesian networks", *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 599-615, 2018.

[3] Jenelius, E., "Data-driven metro train crowding prediction based on real-time load data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 21(6), pp. 2254-2265, 2019.

[4] Kecman, P., & Goverde, R. M., "Predictive modelling of running and dwell times in railway traffic". *Public Transport*, vol. 7(3), pp. 295-319, 2015.

[5] Li, D., Daamen, W., & Goverde, R. M., "Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station", *Journal of Advanced Transportation*, vol. 50(5), pp. 877-896, 2016.

[6] Pasini, K., Khouadjia, M., Same, A., Ganansia, F., & Oukhellou, L., "LSTM encoder-predictor for short-term train load forecasting", In *Proceedings of the ECML PKDD conference*, volume III, pp. 535-551, 2019.