Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas

Sélim Cornet, Christine Buisson, François Ramond, Paul Bouvarel, Joaquin Rodriguez

▶ **To cite this version:**

Sélim Cornet, Christine Buisson, François Ramond, Paul Bouvarel, Joaquin Rodriguez. Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas. 2018. hal-01909708v1

HAL Id: hal-01909708

https://hal.science/hal-01909708v1

Preprint submitted on 31 Oct 2018 (v1), last revised 18 Mar 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas

S. Cornet *, C. Buisson, F. Ramond, P. Bouvarel and J. Rodriguez

Abstract

Railway operations in dense traffic areas are very sensitive even to small disturbances, and thus require careful planning and real-time management. Dwell times in stations are in particular subject to a high variability and are hard to predict; this is mostly due to the interactions between passengers and the system during the dwelling process. This paper proposes an approach for estimating the minimum dwell time knowing the numbers of alighting, boarding and on board passengers, using Automatic Vehicle Location and Automatic Passenger Counting data. Based on the knowledge of this value, a method for estimating the conditional distribution of dwell time given passenger flows is presented. Numerical experiments are carried out on two stations located inside the dense traffic area of Paris suburban network. The obtained results show a broad applicability of these methods, that hence seem very promising.

Highlights

- Passenger flows between train and platform can be described by a single variable
- Minimum dwell time is obtained as a function of this variable
- Actual dwell time is the deterministic minimum dwell time plus random components
- Conditional distribution of dwell time given the passenger flow is computed

1 Introduction

Big cities such as Paris have been experiencing over the last years a steady growth of demand for passenger transportation, including commuter train services. For example, 70% of railway trips on the French network are made inside Paris suburban area, which represents more than 8 million trips per working day. In order to keep providing a good quality of service to passengers despite the progressive saturation of the network, operating companies seek to design timetables that perform well in saturated conditions. Knowledge of the mechanisms that determine trains dwell times is hence of paramount

*Corresponding author. Address: Sélim Cornet, SNCF Réseau, 10 rue Camille Moke, 93210 Saint-Denis, France. Email: selim.cornet2@reseau.sncf.fr

importance, as they dimension the capacity of the network [Abril et al., 2008] and are partly responsible for operations stability. Indeed, over-estimated dwell times lead to a sub-optimal use of the network capacity, which is to be avoided when the transportation demand is high. In addition, a structural source of instability lies in the process of alighting and boarding of passengers [Van Breusegem et al., 1991]: a train that arrives at a station a long time after the previous train will have to wait for many passengers to board, and is likely to dwell longer than expected. Therefore the headway with the previous train will grow, leading to an even higher dwell time at the next station, this phenomenon being amplified at each station unless a corrective action is taken.

Being able to estimate accurately dwell times would therefore allow the design of timetables that are less sensitive to this phenomenon. Finally, accurate estimations of trains dwell times are also needed during the operational phase, to facilitate traffic controllers decisions and to provide passengers with reliable information about their waiting and journey times.

Trains dwell times in stations are determined by several factors, namely:

- Alighting and boarding of passengers. The time required for this process depends on the number of alighting and boarding passengers, the load of the train and the platform, the technical times required for opening and closing doors, as well as the train and station longitudinal configurations [Daamen et al., 2008]. Passengers behavior matters as well: it is not uncommon for passengers to block doors in order to get on board while doors are closing, or because of the poor distribution of passengers in vehicles leading to jammed doors. This results in a longer dwell time.

- Timetable. In most countries trains are not allowed to depart ahead of schedule. Consequently early trains may have to dwell at stations even if the passengers exchange process is complete, in order to meet this constraint.

- Signaling. Operations safety is ensured in most countries by a fixed signaling system. In particular, a signal is located at the end of each platform and indicates whether the track ahead is free of any vehicle or not. Thus when trains are operated with short headways, it is not unlikely for a train to dwell a long time in station waiting for the track to be cleared and the signal to open.

- Driver behavior. Some drivers are likely to close doors as soon as the boarding and alighting process is complete and the timetable allows it, whereas others might wait longer for late passengers.

The time required for the alighting and boarding of passengers is hard to estimate accurately. Indeed, in addition to the factors listed above, it strongly depends on parameters that are not easily accessible: passenger distribution over the platform (which depends in particular on passengers' destinations [Kim et al., 2014], and on the weather for stations without a complete cover), the number of passengers carrying luggage, a bike or a stroller...

However, we verify in this paper the hypothesis that for a given passenger flow, a minimum dwell time is required for the passenger exchange process to be able to complete. A data-based method for computing it is presented. It requires, for each station under consideration, a dataset containing, for each train that served it, the amount of alighting and boarding passengers, as well as the train load on departure and the dwell time of the train at the station. Knowledge of the minimum dwell time (MDT) can indeed be useful for the capacity assessment of a network, or for implementing countdown systems in stations. All other disturbances (passengers with luggage or blocking doors, departure prevented by schedule, signaling) can only extend dwell time from this value. We also discuss on the probability to

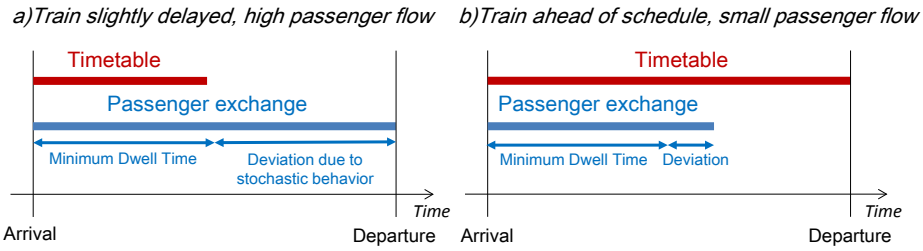

Figure 1: Examples of dwell time components

deviate from this minimum dwell time: the conditional distribution of dwell time given the passenger flow is estimated. Such distributions can be used for robust timetabling, for stochastic simulation as well as for providing traffic controllers the probability of various scenarios.

Figure 1 presents two examples of the possible components that determine dwell time. In both cases, passenger exchange requires a longer time than the MDT. However, in case a), the train cannot depart at the scheduled time as a long time is required for passenger exchange; conversely, in case b), the train is held in station after the passenger exchange process is complete, in order not to depart ahead of schedule.

The remainder of this paper is organized as follows. In the next section, we review some existing work about dwell time estimation in commuter train services, light rail transportation systems and bus. Section 3 presents the data used in this study and the preprocessing operations that were applied to it. The fourth section is dedicated to the method for estimating the minimum dwell time required for a given train. In the fifth section, these results are used for classifying the observations of the dataset, depending on which factor actually determined the dwell time. The sixth section is devoted to a discussion on the dependency of dwell times on passenger flows; we estimate the conditional distribution of dwell time given passenger volumes, thus highlighting the impact of extreme passenger flows. Section 7 concludes the paper.

2 State of the art

The boarding and alighting movement of passengers is one of the main determinants of dwell time in every public transport mode [Kraft, 1975]. However, not all existing dwell time models in the literature aim to describe this phenomenon explicitly. Therefore, previous studies on dwell time prediction can be classified in two categories: explicit models, which use as input passenger volume and behavior, and implicit models that use other inputs which do not require passenger counting.

Most explicit models resort to one of the following modeling approaches: data-based modeling and regression, or microscopic simulation of passenger movements. Some linear and non linear models are proposed by [Lin and Wilson, 1992]. They take into account the friction phenomenon that occurs when the number of standing passengers in vehicles is high. These models were applied to the MBTA metro green line in Boston for one-car and two-car trains. Several improvements on their work are made in [Puong, 2000], where a better understanding of the effects of crowding and number of doors is provided.

The development of Automatic Vehicle Location (AVL) and Automatic Passenger Counting (APC) systems in the 2000s opened new perspectives for dwell time estimation, as they provide larger and

more accurate amounts of data than those obtained by manual counting. The new availability of this type of data lead to a set of papers on the bus and metro cases. For example, [Dueker et al., 2004] provide descriptive statistics about dwell times and an estimation model for the Portland bus network. [Sun et al., 2014] make a deeper analysis of the bus boarding and lighting process, and exhibit a critical occupancy level beyond which the process is slowed down due to friction between passengers. The paper also highlights the effects of the rolling stock configuration on the dwelling process, by performing the same analyses on buses with different characteristics. [Buchmüller et al., 2008] adopted a different approach: instead of building a deterministic model, they divided the dwelling process into several subprocesses and computed the probability distributions of those subject to random variation. This allows to describe the inherent stochasticity of the phenomenon, as not all passengers behave in the same manner. A similar choice is made in [Longo and Medeossi, 2012] and [Larsen et al., 2014], where distributions of dwell times are used for stochastic simulation of train traffic. Recently, [D'Acierno et al., 2017] proposed an analytical model for dwell times on a metro line, and proved that traffic evolves toward an equilibrium in which dwell times can be computed as solutions of a fixed-point problem in finite dimension.

All parametric models, once calibrated, provide a closed form expression for dwell time that makes it easy to compute. On the other hand, microscopic simulation methods require a longer time for dwell time estimation but allow to get insights into passenger movements and behavior. Thus, [Zhang et al., 2008] designed a cellular automata-based simulation method able to reproduce congestion and negotiation between passengers at a microscopic level. This allowed them to estimate dwell times on Beijing metro network. [Yamamura et al., 2013] developed a multi-agent model for simulating passengers behavior, and used it for assessing the efficiency of some measures for reducing passenger congestion, such as using trains with larger doors or without seats. A similar study was conducted by [Schelenz et al., 2014] in order to determine the optimal bus layout from the passengers point of view. [Seriani and Fernandez, 2015] used a pedestrian traffic microsimulator for assessing the performance of pedestrian traffic management strategies in Metro de Santiago, and conducted experiments with volunteers in laboratory.

However, explicit models might not be the easiest to handle. Indeed, not all transportation systems are equipped with APC and passenger data can therefore be hard to obtain. In addition, systems equipped with APC do not always provide data in real-time, making difficult the use of such models for operational applications. For those reasons, some implicit models were developed ; these require input that are easier to collect, albeit influenced by passenger influx. [Hansen et al., 2010] use trains arrival delay as input and a robust regression method for estimating dwell times of intercity trains in the Netherlands. [Kecman and Goverde, 2015] applied Least Trimmed Squares and Random Forests regression methods on a data set containing trains type (local or intercity) and arrival delays, as well as information such as station size and peak hour. Local models on particular stations, where dwell times are predicted using a moving average on previous dwelling processes, are also presented and are found to perform better. [Li et al., 2016] predict the dwell time of a specific train at a given station using the information of dwell times of the same train at previous stations and dwell times of previous trains at the same stations. Improvements and a generality test of this approach are presented in [Li et al., 2018]. Similarly, [Xin and Chen, 2016] use k-nearest neighbors method for predicting bus dwell time at a given stop knowing its dwell times at the previous stops.

The concept of minimum dwell time is introduced by [Pedersen et al., 2018]; it is defined as the time required for boarding and alighting of passengers. The authors assume this minimum dwell time can be computed considering only delayed trains, and study the influence of temporal factors (such as hour in the day, day of the week, week of the year) on the minimum dwell time. However, the minimum dwell

time dependence on the passenger flow is not studied by the authors.

Some features of the models present in the literature are summed up in Table 1, in chronological order. Up to our knowledge, the intrinsic stochasticity in the dwelling process has drawn little research so far, [Buchmüller et al., 2008] and [Li et al., 2014] being some of the few attempts to model it. Our aim is to bridge this gap, by designing a method for computing the minimum dwell time required for boarding and alighting of a given volume of passengers, and estimating the probability of deviation from this minimum dwell time.

Source	Kind of model	Transportation mode	Modeling approach
[Lin and Wilson, 1992]	Explicit	Metro	PR
[Puong, 2000]	Explicit	Metro	PR
[Dueker et al., 2004]	Explicit	Bus	PR
[Buchmüller et al., 2008]	Explicit	Train	Di
[Zhang et al., 2008]	Explicit	Metro	Sim
[Hansen et al., 2010]	Implicit	Train	PR
[Yamamura et al., 2013]	Explicit	Train	Sim
[Sun et al., 2014]	Explicit	Bus	PR
[Li et al., 2014]	Implicit	Train	Di
[Seriani and Fernandez, 2015]	Explicit	Metro	Sim
[Kecman and Goverde, 2015]	Implicit	Train	PR and NPR
[Xin and Chen, 2016]	Implicit	Bus	NPR
[Li et al., 2016]	Implicit	Train	PR and NPR
[D'Acierno et al., 2017]	Explicit	Metro	An
[Pedersen et al., 2018]	Implicit	Train	Di
This paper	Explicit	Train	NPR and Di

Di: probability distributions, PR: parametric regression, NPR: non-parametric regression, An: analytical, Sim: microscopic simulation

Table 1: Some features of existing dwell time estimation models in public transport

3 Data collection and preprocessing

3.1 Experimental setup and available data

For this study, we designed an experimental setup for some stations of Paris suburban network, namely Bois-Colombes (BC) on line J and Houilles-Carrières (HC) on line L (these stations being selected because they are subject to both high and low passenger traffics depending on day and time). Results on other stations of these lines are provided in appendix. Services on these lines are operated by simple or double units of SNCF Class Z50000. Each unit has seven doorways of width 1.95 m, a length of 94.3 m, a seated capacity of 380 passengers (pax) and a total capacity of 760 pax. This rolling stock is equipped with an APC system using infrared lights above the doorways of vehicles. A train event recorder is also embedded, allowing to measure arrival and departure times at stations with higher accuracy than using data from track circuits.

The raw data archives of year 2017 were processed to build a dataset for each station under consideration and each direction (from suburbs to Paris - SP - or from Paris to suburbs - PS). Each observation of these datasets describes the passenger flow and the dwelling process of a given train, by:

- its theoretical timetable
- actual arrival and departure times (from which dwell time can be inferred)
- number of alighting and boarding passengers
- passenger load after departure
- number of doors

Data was filtered in order to keep only observations corresponding to dwell times lower than 180 seconds; indeed we assume that higher dwell times are due to incidents that fall out of the scope of this study. Furthermore, passenger data per unit has more interest for predicting dwell times than their absolute values; we therefore divided the number of boarding and alighting passengers as well as the passenger load by the number of units (one or two) of each train. All statistical analyses performed on these datasets required the use of Python's package Scikit-learn [Pedregosa et al., 2011].

In the sequel, we shall denote A the number of alighting passengers, B the number of boarding passengers, L the train load after departure and DT the dwell time (in seconds). Some descriptive statistics of these datasets are provided in Table 2. Note that all features A, B, L have a high standard deviation; this is due to the large variation of passenger flow between peak and off-peak hours.

Bois-Colombes - SP
23865 observations

	A	B	L	DT
Median	6.0	30.5	136.5	46.0
Mean	7.4	36.0	155.7	55.8
Std	6.8	31.6	120.7	29.4

Bois-Colombes - PS
24230 observations

	A	B	L	DT
Median	28.0	6.0	107.0	46.0
Mean	37.9	7.7	135.5	47.6
Std	31.9	7.0	99.0	15.8

Houilles-Carrières - SP
7989 observations

	A	B	L	DT
Median	5.0	18.0	61.0	34.0
Mean	9.0	29.7	111.6	36.8
Std	10.1	32.1	132.1	14.0

Houilles-Carrières - PS
7812 observations

	A	B	L	DT
Median	18.0	6.5	56.0	32.0
Mean	26.2	10.1	84.5	36.9
Std	24.5	11.0	84.0	16.6

SP : Suburbs to Paris - PS : Paris to suburbs

Table 2: Descriptive statistics of datasets

3.2 A single latent variable to describe passenger flows

Correlation matrices between parameters A, B, L are provided in Table 3. Unsurprisingly, these variables describing passenger flows are strongly correlated. Indeed, all three variables are representations of the same phenomenon (peak and off-peak hours). Namely, in the suburbs to Paris direction, there is a high

number of passengers boarding and a high passenger load during the morning peak hour and lower values the rest of the time. Similarly, in the Paris to suburbs direction, there is a high number of passengers alighting and a high passenger load in the evening peak hour and lower values otherwise. A consequence of this multicollinearity in the data is that no regression method can separate the influence of alighting, boarding and passenger load on dwell times. Moreover, predictions made using methods presented in the subsequent parts are only valid for cases within the range of the available data, and assuming the correlation patterns remain the same [Dormann et al., 2012].

Bois-Colombes - SP

	A	B	L
A	1	0.86	0.87
B	0.86	1	0.96
L	0.87	0.96	1

Bois-Colombes - PS

	A	B	L
A	1	0.74	0.93
B	0.74	1	0.78
L	0.93	0.78	1

Houilles-Carrières - SP

	A	B	L
A	1	0.44	0.55
B	0.44	1	0.87
L	0.55	0.87	1

Houilles-Carrières - PS

	A	B	L
A	1	0.49	0.86
B	0.49	1	0.68
L	0.86	0.68	1

Table 3: Correlation matrices of passenger data

To deal with this collinearity, we introduce a latent variable by performing a principal component analysis (PCA) on explanatory variables. The values of the three variables were previously scaled, as their orders of magnitude differ (usually between 0 and 100 for alighting and boarding passengers, between 100 and 1000 for passenger load). The ratio variance explained by the principal components p_1, p_2, p_3 is presented in Table 4.

	Bois-Colombes - SP	Bois-Colombes - PS	Houilles-Carrières - SP	Houilles-Carrières - PS
p_1	0.93	0.88	0.75	0.79
p_2	0.06	0.10	0.21	0.17
p_3	0.01	0.02	0.04	0.04

Table 4: Ratio of explained variance for each principal component

Note that a large part of the variance is explained by the first principal component p_1; the same occurred on most stations where we performed this operation (a table indicating the ratio of explained variance of the first component on other datasets is provided in appendix). From now on, we shall call it the principal component and denote it p. It will be used as the only representation of passenger volume in the sequel, the other components p_2, p_3 are dropped from the model. For a better understanding, its value was scaled to the interval $[0, 1]$, thus $p = 0$ corresponds to a train running almost empty with no passenger exchange at the platform and $p = 1$ to a train almost full with many alighting and/or boarding passengers. The linear regression lines of A, B, L against p for each dataset are provided on figure 2. For example, at the station of Houilles-Carrières in the Paris to suburb direction, a value of $p = 0.4$ corresponds approximately to $A = 100$ alighting passengers, $B = 30$ boarding passengers and $L = 350$ passengers on board after departure.

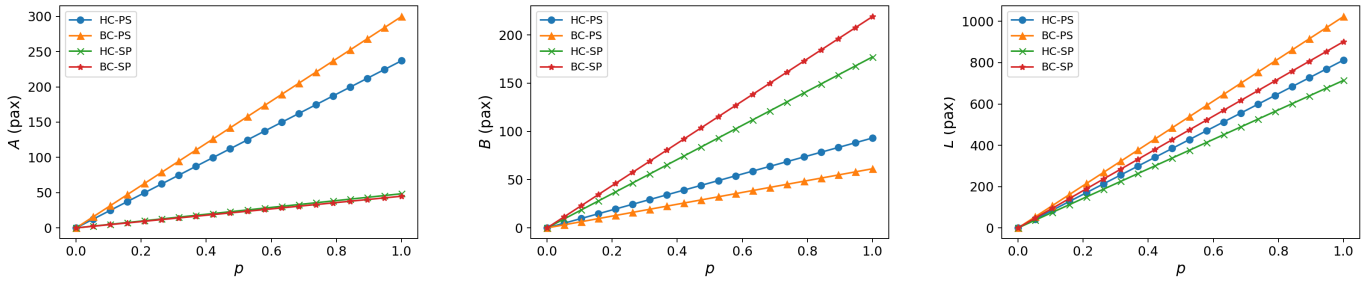

Figure 2: Correspondence between values of P and values of A, B, L

Finally, note that the data density is inhomogeneous along the p axis; many observations correspond to cases with small passenger influx (hence p is close to 0), whereas observations corresponding to extremely high values of p (close to 1) are scarce.

3.3 Towards the Minimum Dwell Time

A scatter plot of observations on the BC-PS and BC-SP datasets is provided on Figure 3, where the passenger flow p is plotted on the x-axis and the dwell time DT on the y-axis. The intrinsic stochasticity of the dwelling process duration can be visualized on that figure. Indeed, it does not only depend on timetable and passenger volumes but also on factors that are most likely unknown: passenger distribution over the platform and inside the train, presence of cumbersome luggage, driver behavior... We hence believe that dwell times at short stops in dense traffic areas cannot be predicted accurately using available data. As a matter of fact, after splitting each dataset into a training set and a test set, all our attempts to train a regression model on the training set (with explanatory variables A, B, L) and use it to predict the dwell time values of the test set were unsuccessful.

However, we observe that the lower bound of DT increases with the value of p; this seems to indicate, for a given value of the passenger flow p, the existence of a minimum value of dwell time for the passenger exchange process to be able to complete. We shall henceforth focus on this notion of Minimum Dwell Time (MDT); the next section is dedicated to a definition and an algorithm for computing it.

4 Estimating minimum dwell time

4.1 Definition

In this section, we refer to minimum dwell time (MDT) for a given volume of passengers (alighting, boarding and on board, described by the single variable p) as the shortest amount of time a train is required to dwell in station for the alighting and boarding process to be able to complete. It depends on the volume of passengers, rolling stock layout (number and width of doorways, capacity...) and station configuration (curvature, height of the step between the platform and the train...).

Up to our knowledge, so far this concept has only been studied by [Pedersen et al., 2018]. Their work is based on the hypothesis that a delayed train will dwell in stations only the minimum amount of time required for the exchange of passengers, and therefore conduct analyses by considering only delayed trains. In our opinion, this approach has several drawbacks and can be improved. First, we believe that

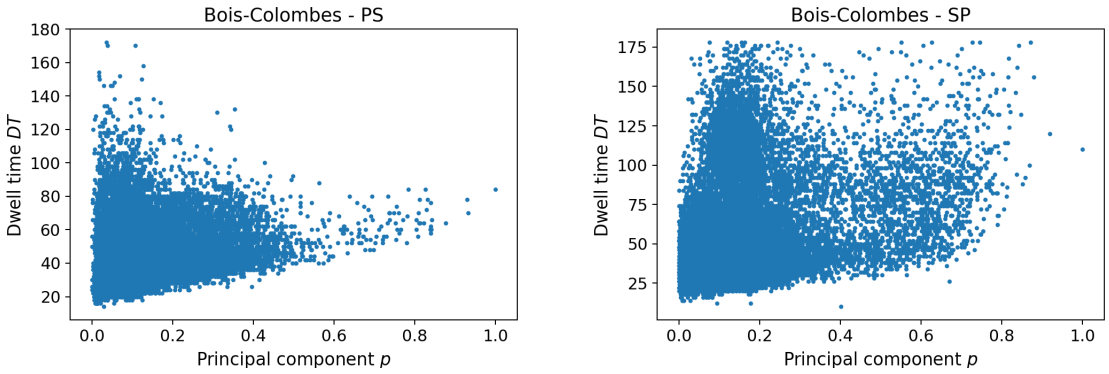

Figure 3: Dwell times for different values of p

not all delayed trains depart from stations as soon as alighting and boarding is complete; some can be prevented or discouraged from doing so by a red or yellow signal, extra seconds can be lost by passengers blocking doors or by the driver waiting for late passengers. In addition, such a definition makes the influence of passenger flows on the MDT difficult to study, although the variation of passenger volumes is the actual cause of the observed variation of the MDT according to temporal factors.

We propose instead to split the actual dwell time into a deterministic part, that is the MDT, and a stochastic part that is the difference between the minimum and actual dwell time. Thus the dwell time DT observed in the datasets can be written

$$DT = MDT(p) + Dev + \epsilon \tag{1}$$

where the minimum dwell time $MDT(p)$ is a deterministic function of p, the deviation Dev is a non-negative random variable, and ϵ is a random centered noise. Let us recall that the relationship used for computing p, as well as the function MDT and the random variable Dev, also depend on the station and direction under consideration. Note that a similar modeling approach was used in [Medeossi, 2008] for performing stochastic simulation of traffic; the minimal dwell time being however a constant value determined by the rolling stock features, the dependence on the passenger flow is not evaluated.

4.2 Computing the minimum dwell time

4.2.1 Algorithm

On figure 3, we can observe that the minimum dwell time is an increasing and sometimes approximately linear function of the passenger volume p. We propose to estimate it using a two-step process, whose principle can be presented the following way.

1. Selection of the data
We select points in each dataset according to the following procedure:

1. Choose a window width Δp.
2. Select all observations that satisfy $p \in [0, \Delta p)$

3. Among these observations, select the one with the lowest value of DT. Discard the other observations.

4. Repeat steps 2 and 3 on intervals $[\Delta p, 2\Delta p), \cdots, [n\Delta p, 1]$ (until the whole dataset has been browsed through).

Width Δp should be chosen in such a way that every window is likely to contain several points, but not too large so that every region is represented. The result of this step is a set of observations covering the whole range of values of p, and for which the observed dwell time is close to the MDT.

2. Regression

We then estimated the function $MDT(p)$ by applying a k-nearest neighbor regression method to the selected data, with a slight modification. Indeed, the function $MDT(p)$ can reasonably be assumed non-decreasing; we therefore enforced the same constraint on the estimation, by replacing any decreasing section by a piecewise constant one.

4.2.2 Calibration and validation on fictitious test datasets

We first verify the ability of this algorithm to accurately estimate the function $MDT(p)$ by applying it to fictitious datasets where the function is known. We built two datasets Test1 and Test2 where MDT was assumed to be given respectively by

$$MDT_1(p) = 20 + 60p \quad (2.1) \qquad \text{and} \qquad MDT_2(p) = 20 + 60p^2 \quad (2.2)$$

We built these datasets by selecting the values of p in the BC-SP dataset, and randomly generated the values of DT based on equation (1). We assumed Dev to follow an exponential distribution of parameter 20 and ϵ to follow a normal distribution $\mathcal{N}(0, 1)$. Some outliers were added following a uniform distribution in the space (p, DT). The corresponding points are plotted on Figure 4.

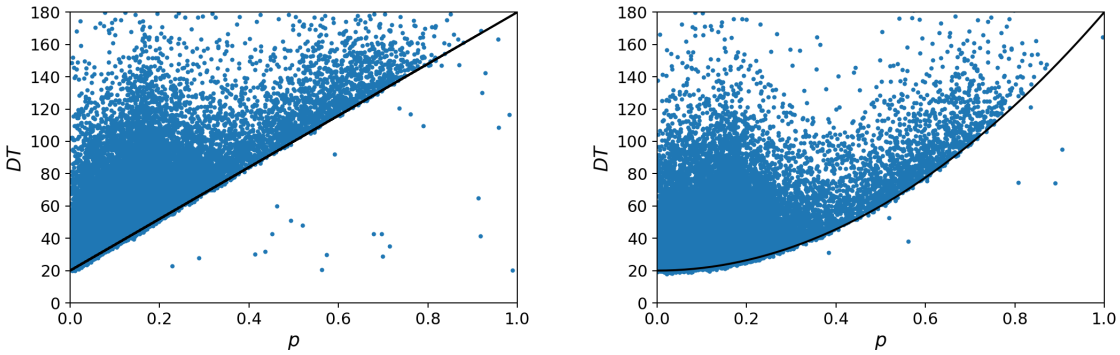

Figure 4: Fictitious test datasets. Left: eq (2.1), right: eq. (2.2)

We then applied the previous method to estimate the function $MDT(p)$, taking window width $\Delta p = 0.005$ and number of neighbors $k = 5$ (the choice of these parameters is discussed below). The selected points and obtained estimated function are plotted on Figure 5. We define the estimation error by the difference between the predicted value and the actual value (given by functions MDT_1 and MDT_2). The error evolution with p is plotted on Figure 6.

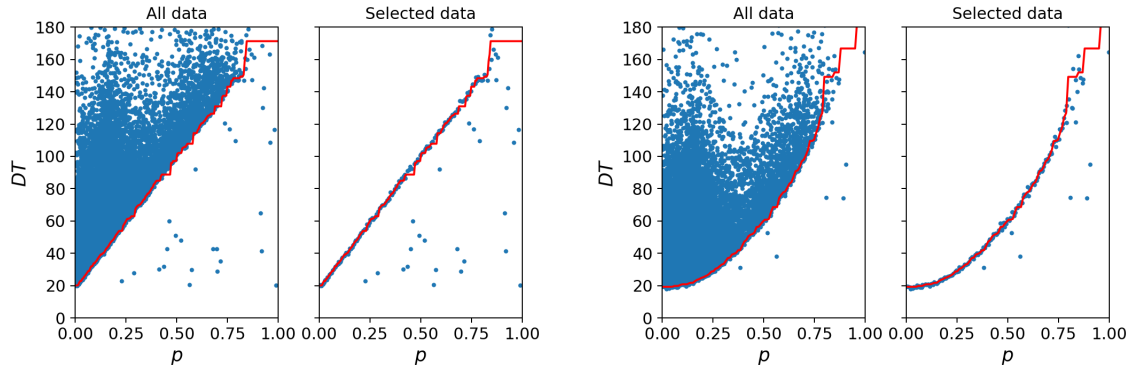

Figure 5: Selected data and estimated MDT function on test datasets. Left: eq (2.1), right: eq. (2.2)

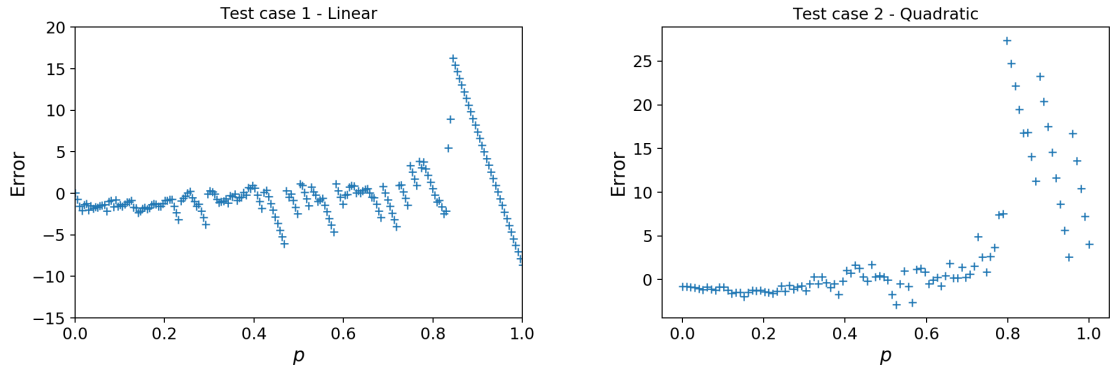

Figure 6: Error on estimated MDT function on test datasets. Left: eq (2.1), right: eq. (2.2)

We observe that in both cases, the value of MDT is estimated with good accuracy when the passenger flow p is not too high ($p < 0.8$). Beyond this level, the lack of points makes the prediction inaccurate; indeed, 99.9% of points fall within the range $p \in [0, 0.8]$.

The following method was used for tuning parameters k and Δp. 100 observations $p_1, \cdots, p_n$ were generated in the range $p \in [0, 0.8]$ (the range $p \in [0.8, 1]$ not being considered because, as previously mentioned, it contains too few points for the results to be meaningful). The model was then trained on the Test2 dataset and the predictions $\widehat{MDT}$ made on these 100 observations compared with the actual value of MDT_2. The Mean Absolute Error (MAE) was used as criterion:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |MDT_2(p_i) - \widehat{MDT}(p_i)| \tag{2}$$

The values of MAE for different sets of parameters $(\Delta p, k)$ are provided on the colormap of Figure 7. Although the lowest value (MAE=1.54 s) is attained for $\Delta p = 0.001$ and $k = 10$, in the sequel we used $\Delta p = 0.005$ and $k = 5$; indeed, this choice offers satisfactory performances (MAE=2.04 s) and guarantees some robustness, as the neighboring sets of parameters also provide good results.

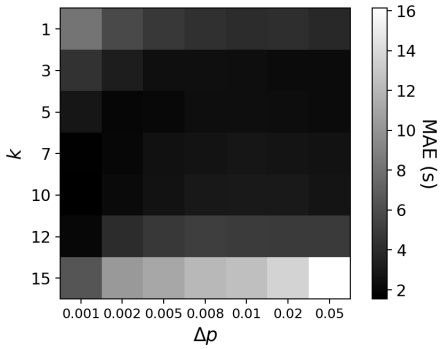

Figure 7: Colormap of MAE for various values of k and Δp

4.2.3 Results on real-world data

The same method was subsequently applied to the BC-SP, BC-PS, HC-SP and HC-PC datasets. The selected data and obtained regression curve are given on Figure 8. In all cases, we observe that the function $MDT(p)$ is approximately linear for small values of p, then increases with greater slope; this illustrates the phenomena of congestion that appear when the number of passengers is high.

We also note that the function seems to be station and direction-dependent. However, this assertion cannot be verified by simply comparing curves, as the way p is computed is itself station-dependent. Instead, we built two samples of realistic passenger volumes by randomly selecting 200 observations in the PS direction and 200 in the SP direction. We subsequently computed the associated values of p for each model, used the previous regression method to predict the MDT for these passenger volumes and compared the obtained results. Figure 9 represents the distribution along the principal component p of the difference between, respectively, the values of MDT predicted by BC-PS and HC-PS on the PS observations, and the values predicted by BC-SP and HC-SP on the SP observations.

This figure shows a small difference of predicted values for small passenger volumes, however it gets higher when the number of passenger exchanged increases. It seems therefore that the minimum dwell time indeed depends on the station under consideration. In addition, we observe that the difference is always negative; this suggests that for the same values of A, B, L, the minimum dwell time is always smaller at the station of Houilles-Carrières than at Bois-Colombes. This could be the result of different station layout (height of the platform, number and position of platform access for example).

5 Identifying determinants of dwell time

Let us recall that dwell time of trains in stations is not only determined by passenger volumes but also by their schedule (as trains are not allowed to depart ahead of schedule) and exterior events such as signaling. The knowledge of the minimum dwell time required for the process of alighting and boarding of passengers can help classifying available observations, depending on which of the previous factors actually determined the dwell time. For each observation, we define the minimum dwell time imposed by the timetable, denoted MDT-TT, by MDT-TT$= \max(\text{DT-TT} - \text{Delay}, 0)$ where DT-TT is the scheduled dwell time as defined by the timetable and Delay is the delay of the train upon arrival at the station. We also compute the minimum dwell time for the exchange of passengers (henceforth

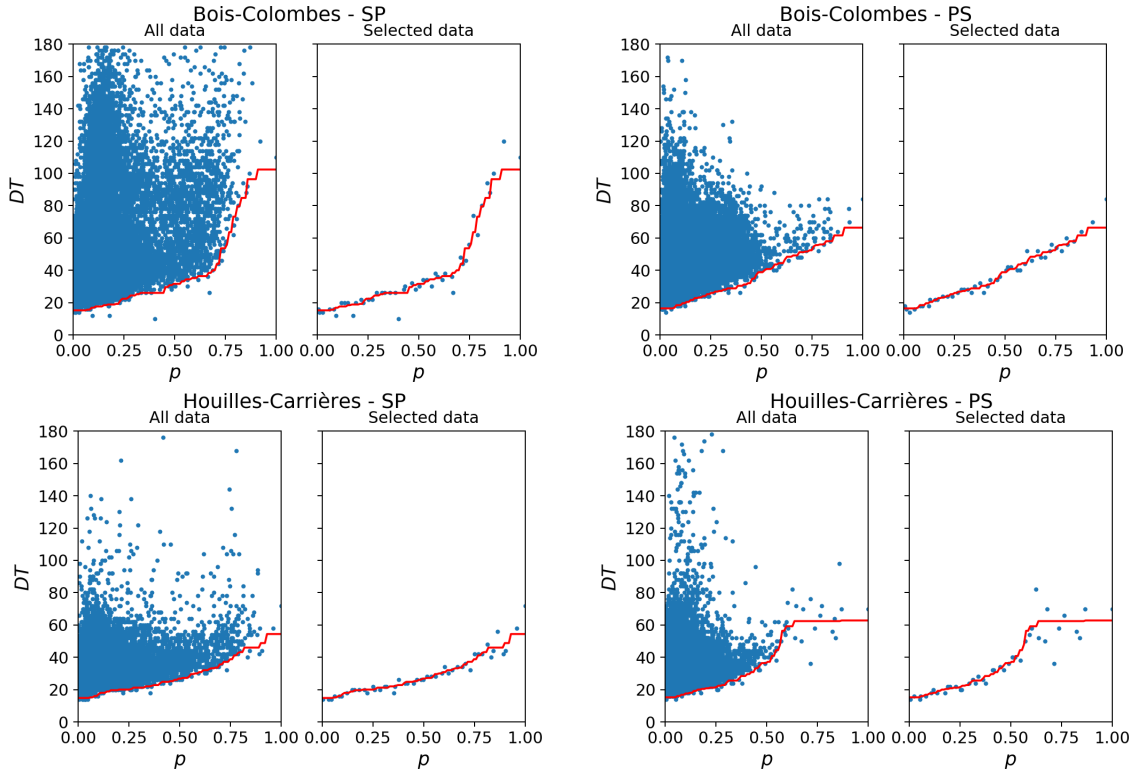

Figure 8: Selected data of each dataset and kNN regression

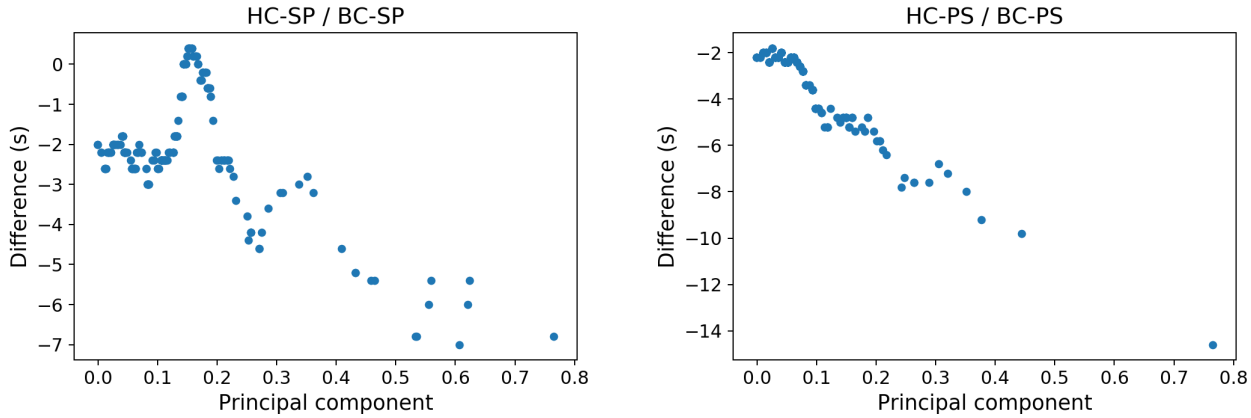

Figure 9: Difference of predictions

denoted MDT-P) with the method described in the previous section.

Observations are then classified in two sets: those where MDT-P>MDT-TT correspond to events were passengers volumes justified a longer dwell time in station than planned in the timetable. Conversely, those were MDT-P≤MDT-TT are events where the dwell time was most likely determined by the no-early-departure rule. Classified observations on BC-SP dataset are plotted on Figure 10. There are 16306 observations in the MDT-P≤MDT-TT case, and 7558 observations in the MDT-P>MDT-TT (similar proportions are observed on the other datasets).

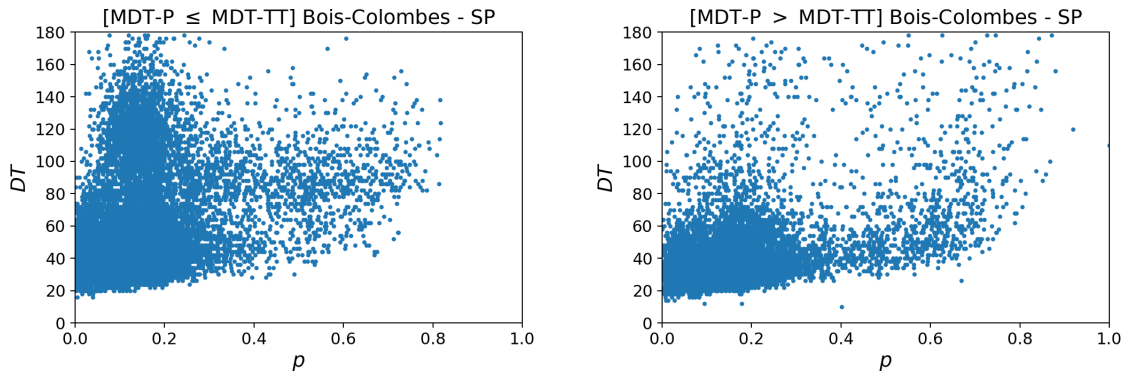

Figure 10: Classification according to the main determinant of dwell time

It is worth noting that most observations with high values of DT correspond to trains that arrived in station ahead of schedule; these long dwell times were due to the timetable constraint and not to a long passenger exchange process. We consequently discarded these observations from future examinations.

6 Estimating the conditional distribution of dwell time

The previous steps allowed us to build datasets (corresponding to the right part of Figure 10) where we can assume passenger exchange to be the main determinant of dwell time. We seek to use them for estimating the conditional distribution of dwell time given the passenger volume. From now on, we assume passenger volume P and dwell time DT to be random variables with densities f_P, f_{DT} and the couple (P, DT) to have density $f_{P,DT}$. Then, for a given passenger volume p, the conditional distribution of DT knowing $P = p$ can be obtained by

$$\mathbb{P}(DT \leq t | P = p) = \int_0^t \frac{f_{P,DT}(p, \tau)}{f_P(p)} \mathrm{d}\tau \tag{3}$$

We resorted to Kernel Density Estimation (KDE) [Silvermann, 1986] with Gaussian kernel for estimating the joint distribution $f_{P,DT}$. Consider the dataset as the realization of n independent and identically distributed random variables $(X_1, \cdots, X_n)$ with density $f_{P,DT}$, and $x = (p, t)$ a given point. The chosen kernel density estimator is defined by

$$\hat{f}(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{4}$$

where $K(u) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} u^T u\right)$ is the Gaussian kernel, and $h > 0$ is a smoothing parameter. The estimator $\hat{f}$ depends on $X_1, \cdots, X_n$ and can therefore be considered as a random variable. The quality of the estimation is measured by the Mean Integrated Squared Error (MISE)

$$\mathrm{MISE}(\hat{f}) = \mathbb{E}\left(\int (\hat{f}(x) - f_{P,DT}(x))^2 \mathrm{d}x\right) \tag{5}$$

[Silvermann, 1986] proves that $h = 0.96 \frac{1}{n^{1/6}}$ is an optimal choice if the density to estimate is the normal distribution, we chose this value of h even if the normal distribution hypothesis is not satisfied *a priori*.

Density $f_{P,DT}$ was then estimated on the BC-SP dataset after scaling the data. Marginal density f_P can subsequently be computed, as

$$f_P(p) = \int_0^{+\infty} f_{P,DT}(p,t)\mathrm{d}t \tag{6}$$

The integral being approximated with e.g. the trapezoidal rule (see [Stoer and Bulirsch, 1993]). The conditional cumulative distribution function (cdf) of equation (3) can then be computed for given values of p. This function is plotted for several values of p on Figure 11; similar curves were obtained on other datasets and are provided for illustrative purposes in appendix, on Figure 12. As the passenger flow increases, the slope of the cdf plot gets lower, implying a higher probability to deviate from the minimum dwell time. The plots corresponding to high passenger flows ($p > 0.5$) are often quite remote from each other: this highlights the phenomena of congestion at the doorway level that are more likely to happen in these situations.

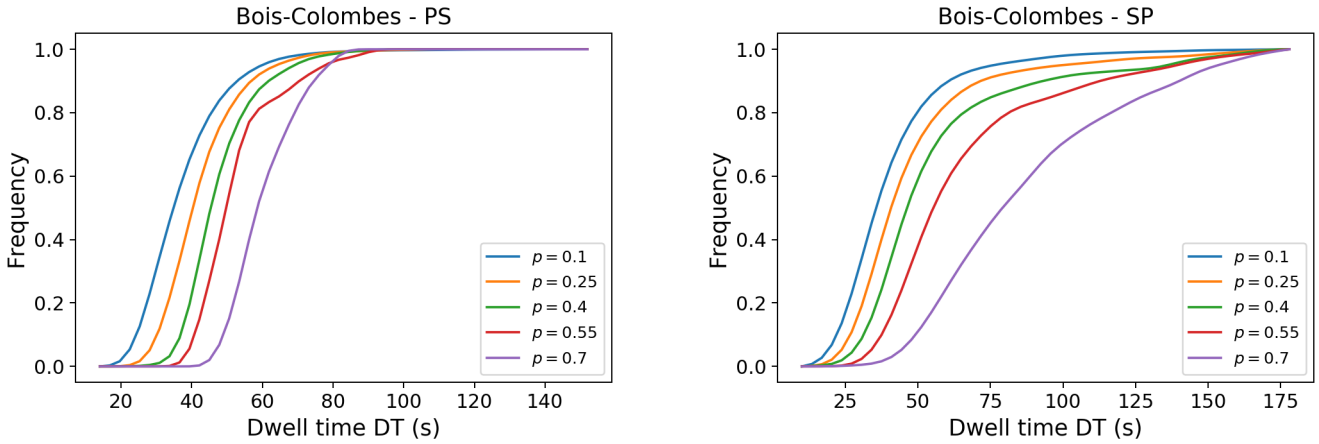

Figure 11: Conditional cdf of DT for different values of p

7 Conclusion and perspectives

This paper proposes a novel approach for assessing the dependency of dwell time on passenger flows in public transport. First, it is shown that in most cases, passenger flows can be described by a single (latent) variable. This considerably facilitates data processing as well as visualization. The concept of minimum dwell time is then investigated. A definition and a method for computing it from available data are proposed. Knowledge of the minimum dwell time for a given passenger volume has several practical uses: it allows a more accurate estimation of network capacity, as well as opportunity studies of implementing countdown systems for dwell time in stations. In addition, it helps classifying data according to the main determinant of dwell time, thus making the estimation of the conditional distribution of dwell time given passenger flows possible. Coupled with a passenger flow prediction model, these distributions can be used for performing quantitative assessments of timetable robustness, forecasting delay propagation in real time and providing reliable information to travelers, as well as for giving insights to traffic controllers into the potential consequences of their decisions.

The implementation of this dwell time model into a simulation tool should yield simulation results with higher accuracy, and will be part of our future work. The obtained results also seem to indicate that different patterns can be observed depending on the station under consideration, leading to contemplate a classification of stations with regards to this phenomenon. Finally, our method could also be improved by integrating data about phenomena that were neglected; using meteorological or signaling data could lead to more accurate predictions of dwell time.

Acknowledgments

This work has been partially financed by the ANRT (Association Nationale de la Recherche et de la Technologie) through the PhD number 2017/1065 with CIFRE funds and a cooperation contract between SNCF and IFSTTAR.

References

[Abril et al., 2008] Abril, M., Barber, F., Ingolotti, L., Salido, M., Tormos, P., and Lova, A. (2008). An assessment of railway capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(5):774–806.

[Buchmüller et al., 2008] Buchmüller, S., Weidmann, U., and Nash, A. (2008). Development of a dwell time calculation model for timetable planning. In *Computers in Railways*, pages 525–534, Toledo. WIT Press.

[Daamen et al., 2008] Daamen, W., Lee, Y.-c., and Wiggenraad, P. (2008). Boarding and Alighting Experiments: Overview of Setup and Performance and Some Preliminary Results. *Transportation Research Record: Journal of the Transportation Research Board*, 2042:71–81.

[D'Acierno et al., 2017] D'Acierno, L., Botte, M., Placido, A., Caropreso, C., and Montella, B. (2017). Methodology for Determining Dwell Times Consistent with Passenger Flows in the Case of Metro Services. *Urban Rail Transit*, 3(2):73–89.

[Dormann et al., 2012] Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

[Dueker et al., 2004] Dueker, K. J., Kimpel, T. J., Strathman, J. G., and Callas, S. (2004). Determinants of bus dwell time. *Journal of Public Transportation*, 7(1):21–40.

[Hansen et al., 2010] Hansen, I. A., Goverde, R. M., and van der Meer, D. J. (2010). Online train delay recognition and running time prediction. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 1783–1788.

[Kecman and Goverde, 2015] Kecman, P. and Goverde, R. M. P. (2015). Predictive modelling of running and dwell times in railway traffic. *Public Transport*, 7(3):295–319.

[Kim et al., 2014] Kim, H., Kwon, S., Wu, S. K., and Sohn, K. (2014). Why do passengers choose a specific car of a metro train during the morning peak hours? *Transportation Research Part A: Policy and Practice*, 61:249–258.

[Kraft, 1975] Kraft, W. H. (1975). *An Analysis of the Passenger Vehicle Interface of Street Transit Systems with Applications to Design Optimization*. PhD thesis, New Jersey Institute of Technology.

[Larsen et al., 2014] Larsen, R., Pranzo, M., D'Ariano, A., Corman, F., and Pacciarelli, D. (2014). Susceptibility of optimal train schedules to stochastic disturbances of process times. *Flexible Services and Manufacturing Journal*, 26(4):466–489.

[Li et al., 2016] Li, D., Daamen, W., and Goverde, R. M. P. (2016). 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*, 50(5):877–896.

[Li et al., 2014] Li, D., Goverde, R. M. P., Daamen, W., and He, H. (2014). Train dwell time distributions at short stop stations. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2410–2415.

[Li et al., 2018] Li, D., Yin, Y., and He, H. (2018). Testing the Generality of a Passenger Disregarded Train Dwell Time Estimation Model at Short Stops: Both Comparison and Theoretical Approaches. *Journal of Advanced Transportation*.

[Lin and Wilson, 1992] Lin, T.-M. and Wilson, N. (1992). Dwell time relationships for light railway systems. *Transportation Research Record*, 1361:287–295.

[Longo and Medeossi, 2012] Longo, G. and Medeossi, G. (2012). Enhancing Timetable Planning With Stochastic Dwell Time Modelling. In *Computers in Railways*, pages 461 – 471, New Forest. WIT Press.

[Medeossi, 2008] Medeossi, G. (2008). *Capacity and reliability in railway networks*. PhD thesis, Università degli studi di Trieste.

[Pedersen et al., 2018] Pedersen, T., Nygreen, T., and Lindfeldt, A. (2018). Analysis of temporal factors influencing minimum dwell time distributions. In *Computers in Railways*, pages 87 – 98, Lisbon. WIT Press.

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

[Puong, 2000] Puong, A. (2000). Dwell time model and analysis for the MBTA red line. Technical report, Cambridge.

[Schelenz et al., 2014] Schelenz, T., Suescun, n., Wikström, L., and Karlsson, M. (2014). Application of agent based simulation for evaluating a bus layout design from passengers' perspective. *Special Issue with Selected Papers from Transport Research Arena*, 43:222–229.

[Seriani and Fernandez, 2015] Seriani, S. and Fernandez, R. (2015). Pedestrian traffic management of boarding and alighting in metro stations. *Transportation Research Part C: Emerging Technologies*, 53:76–92.

[Silvermann, 1986] Silvermann, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

[Stoer and Bulirsch, 1993] Stoer, J. and Bulirsch, R. (1993). *Introduction to Numerical Analysis*. Springer.

[Sun et al., 2014] Sun, L., Tirachini, A., Axhausen, K. W., Erath, A., and Lee, D.-H. (2014). Models of bus boarding and alighting dynamics. *Transportation Research Part A: Policy and Practice*, 69:447–460.

[Van Breusegem et al., 1991] Van Breusegem, V., Campion, G., and Bastin, G. (1991). Traffic modeling and state feedback control for metro lines. *IEEE Transactions on Automatic Control*, 36(7):770–784.

[Xin and Chen, 2016] Xin, J. and Chen, S. (2016). Bus Dwell Time Prediction Based on KNN. *Green Intelligent Transportation System and Safety*, 137:283–288.

[Yamamura et al., 2013] Yamamura, A., Koresawa, M., Aadchi, S., Inagi, T., and Tomii, N. (2013). Dwell time analysis in urbain railway lines using multi-agent simulation. In *13th World Conference on Transportation Research (WCTR13)*, Rio de Janeiro.

[Zhang et al., 2008] Zhang, Q., Han, B., and Li, D. (2008). Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations. *Transportation Research Part C: Emerging Technologies*, 16(5):635–649.

Appendix

Station	NO - PS	EV - PS	NO - SP	EV - SP
Bécon-les-Bruyères	19111	0.68	19319	0.74
Colombes	24233	0.83	24199	0.91
Nanterre-Université	7592	0.66	7933	0.76
Sartrouville	7535	0.77	7723	0.83

NO: Number of observations, EV: ratio of explained variance

Table 5: Explained variance of the first principal component on other datasets

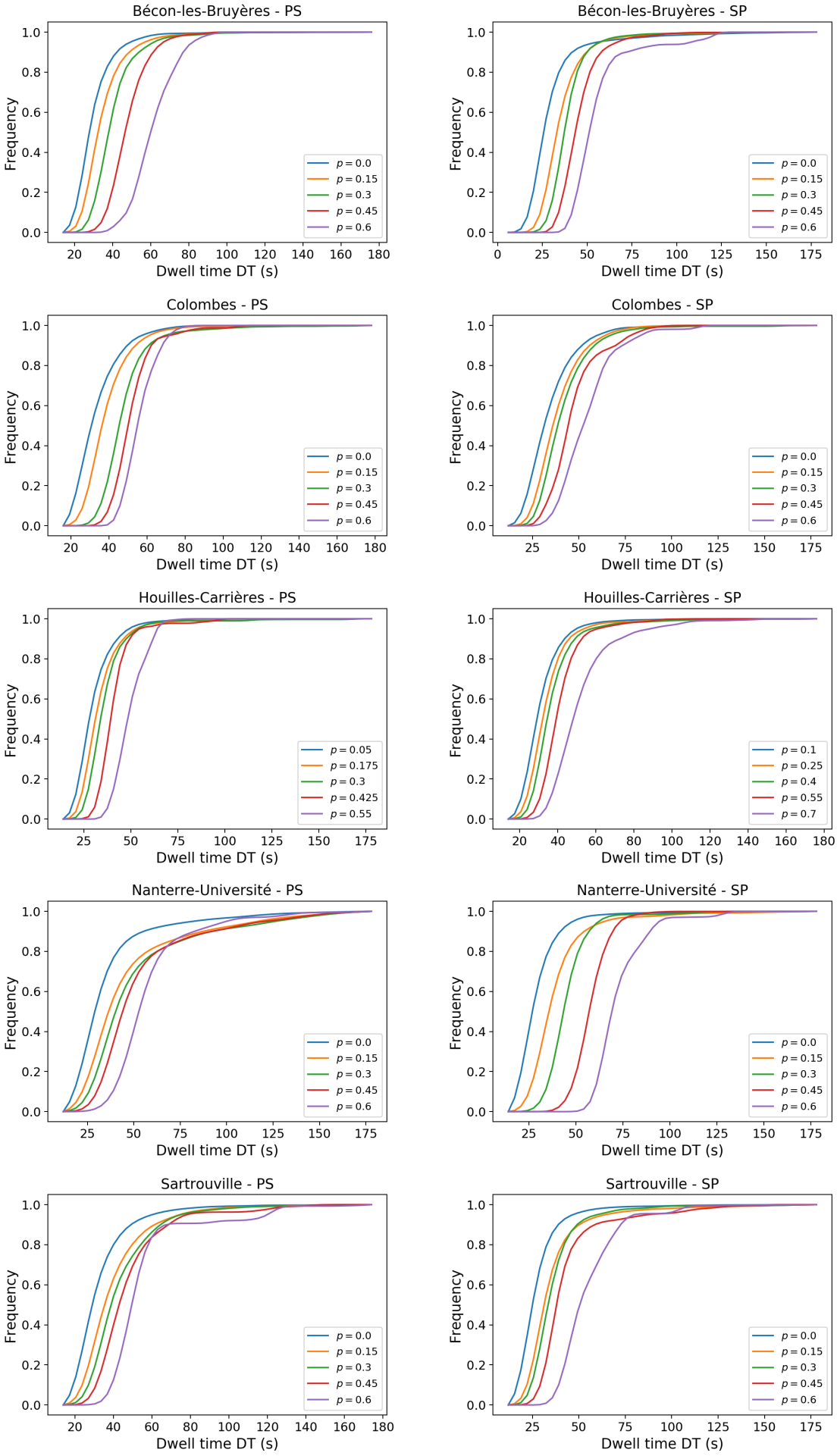

Figure 12: Conditional cdf of dwell time given passenger flow