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

Selim 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

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 railway system during the dwelling process. This paper presents a data-driven approach for assessing the influence of the numbers of alighting, boarding and on board passengers on the dwell time. We propose to split the dwell time into a deterministic component depending on the passenger flow, called the Minimum Dwell Time, and a random component. A method for estimating the minimum dwell time is provided. Based on the knowledge of this value, observations can be classified according to the main determinant of dwell time, namely timetable constraints or passenger exchange. The latter observations are used for estimating the conditional distribution of dwell time given passenger flows. Numerical experiments are carried out on stations located inside the dense traffic area of Paris suburban network. The obtained results indicate that the presented method can be used for a variety of applications, such as capacity assessment or stochastic simulation.

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

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

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 even in saturated conditions. Knowledge of the mechanisms that determine trains dwell times is hence of paramount 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. It is therefore desirable to design 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:

- Technical features of the rolling stock, mainly the time required for door opening and closing.
- 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, 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 used 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. Some examples are the 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 make (and verify) the hypothesis that for a given passenger flow, a minimum dwell time is required for the passenger exchange process to be able to complete. All other disturbances (passengers with luggage or blocking doors, departure prevented by schedule, signaling) can only extend dwell time from this value. Figure 1 presents an example of the possible components that determine

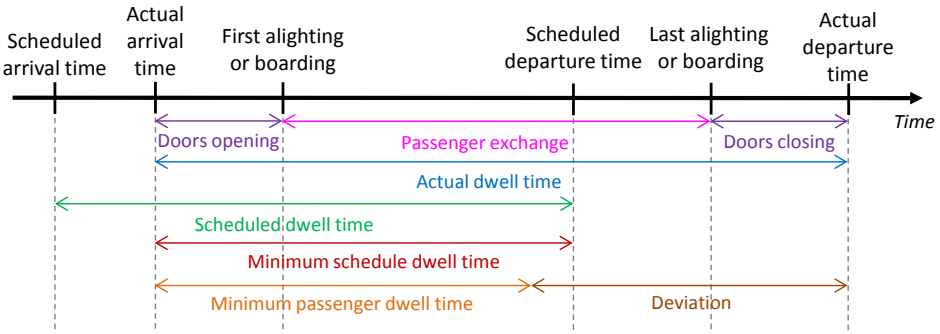

Figure 1: Example of dwell time components

dwell time, here in the case of a train arriving slightly late at the platform. In that case, the train could have departed on schedule if passenger alighting and boarding had been complete after the minimum passenger dwell time. Yet, here a deviation due to random behavior of passengers caused passenger exchange to extend and the train to depart with delay.

The aim of this paper is to present a new data-based method for assessing the impact of passenger flow on dwell times. We show that for a given passenger flow, a minimum dwell time is required and present a data-based algorithm for computing it. Knowledge of the minimum dwell time can indeed be useful for the capacity assessment of a network, or for implementing countdown systems in stations. It also allows to classify each observation depending on whether the dwell time was mainly determined by timetable constraints or passenger exchange. The algorithm requires a dataset containing, for each station under consideration and 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. We finally discuss on the probability to deviate from the minimum dwell time: the conditional distribution of dwell time given the passenger flow is estimated using the previously classified observations. Such distributions can be used with a long-term passenger demand forecast model (see for example [Ceder, 2007]) for robust timetabling and stochastic simulation; or with a short-term passenger flow prediction method ([Wei and Chen, 2012]) for providing traffic controllers with the probability of various scenarios.

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 station and passenger flow. 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 numbers, thus highlighting the impact of extreme passenger flows. Section 7 concludes the paper. Additional details and results are given in appendix.

2 State of the art

Boarding and alighting movements of passengers are some 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 number and behavior; and implicit models that use other inputs which do not require passenger counting, such as day, hour, delay or headway with the previous train. Note that all these features are however usually correlated with passenger flows.

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 alighting 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 correlated with 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 smallest dwell time (including door opening and closing) allowing boarding and alighting of passengers to complete. The authors assume that this minimum dwell time can be computed by 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 explicit dependence on the passenger flow is not studied by the authors.

Some features of the main models in the literature are summed up in Table 1, in chronological order. As said before, the dependence of the models on passenger flows can be either explicit or implicit, depending on whether the model includes passenger flows in its variables. This is indicated in the "Passenger flows" column.

Source	Passenger flows	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

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 number of passengers, and estimating the probability of deviation from this minimum dwell time.

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. A partial map of these lines, featuring the considered stations, the performed services and their frequencies in peak hours is given on figure 2. Lines L2, L3 and J4 each run on a dedicated double-track. 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; captors are located above each doorway and are able to detect separately alighting and boarding movements (from which the train load can be deduced). Note that when the train is extremely crowded, passengers may have to get off the train in order to letting other passengers alight, before boarding again. Those are counted both as alighting and boarding passengers. However, the train occupation rates observed in the datasets suggest that this phenomenon seldom occurs.

Comparisons with manual countings showed that the average error in the measure of alighting and boarding movements is below 5%, whereas the train load is known with less than 15% error. The accuracy of this system has therefore been judged sufficient to use it instead of manual countings by the transport organization authority of Paris suburban area. 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 passengers alighting and boarding the train (these numbers being only known at the train level, they cannot be detailed for each door),
- 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, 99.9% of observations correspond to lower dwell times. 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

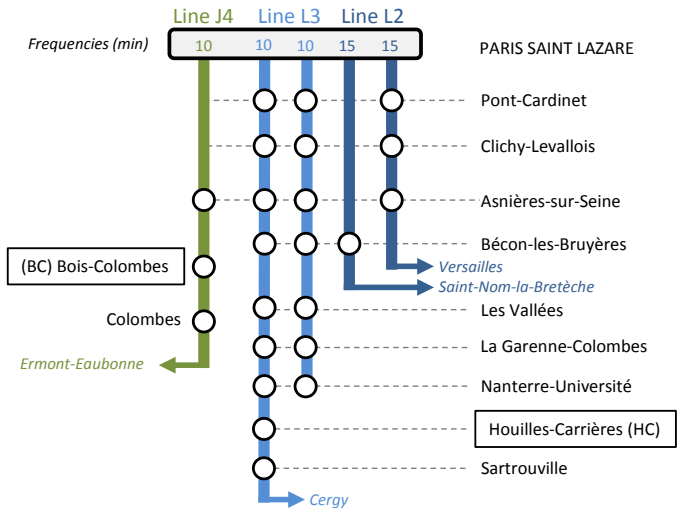

Figure 2: Partial map of lines J and L

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 passengers alighting the train, B the number of passengers boarding the train, 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 (std); 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].

To deal with this collinearity, we introduce a latent variable by performing a principal component analysis (see e.g. [Jolliffe, 2011]) on explanatory variables A, B, L. This technique consists in building new variables p_1, p_2, p_3 (called the principal components) as a linear combination of the old ones, in such a way that the new variables are uncorrelated and have maximum variance for the data. The values of the three variables A, B, L 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 of variance explained by the principal components p_1, p_2, p_3 is presented in Table 4.

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

	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, or the reduced passenger flow, and denote it p. It will be used as the only representation of passenger numbers 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 3. 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.

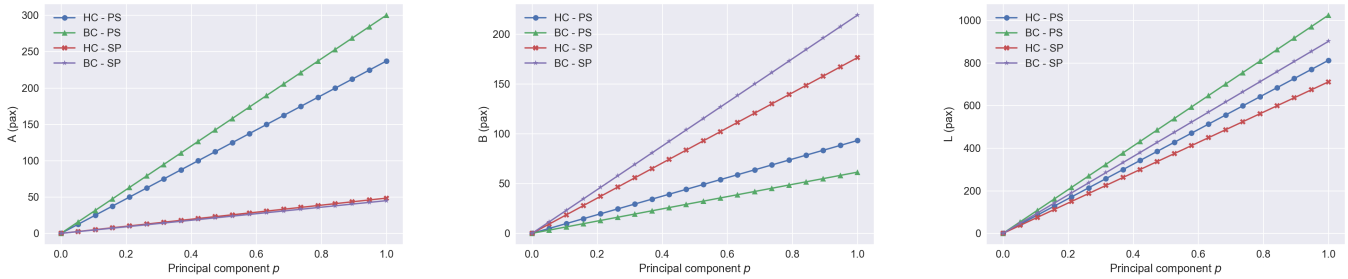

Figure 3: Correspondence between values of p and values of A, B, L

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 4, where the reduced 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 numbers 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), seen as a function of the reduced passenger flow p. The next section is dedicated to a definition and an algorithm for computing it.

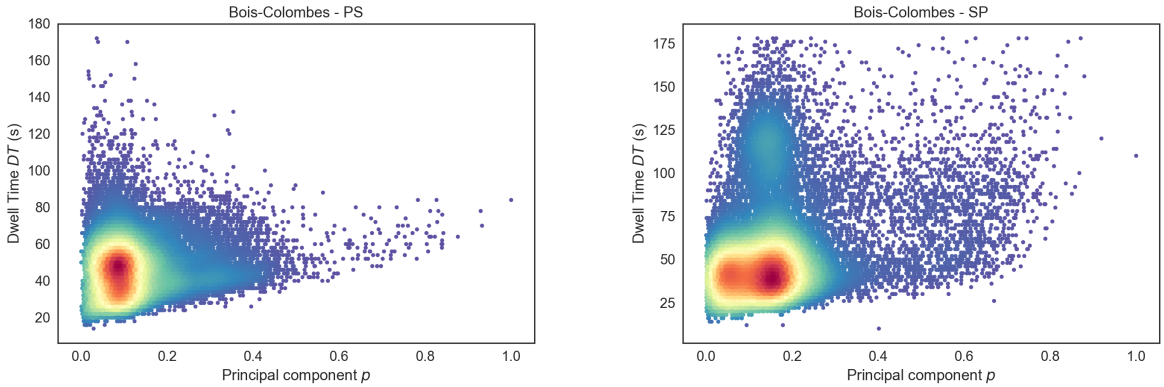

Figure 4: Dwell times for different values of p. Observations are coloured according to density.

4 Estimating the minimum dwell time for a given passenger flow

4.1 Definition

We refer to minimum dwell time (MDT) for a given number 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 (including door opening and closing). It depends on the reduced passenger flow p, but also on the rolling stock features (time required for door opening and closing, 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 the concept of minimum dwell time 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 allowing the exchange of passengers to complete. Therefore they conduct analyses by considering only delayed trains. In our opinion, this approach has several shortcomings and can be improved. First, we believe that not all delayed trains close doors and 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 numbers is the actual cause of the observed variation of the MDT according to temporal factors.

To model the phenomenon observed on figure 4, let us introduce the following random variables:

- DT that represents the actual dwell time
- DTT that represents the minimum time the train is required to dwell at the station in order to satisfy the no-early-departure constraint. It is equal to the scheduled dwell time minus the delay on arrival (or 0 if this difference is negative).
- DTP that represents the time required for door opening, passenger exchange and door closing.

These three random variables are linked by the relation

$$DT \geq \max(DTT, DTP) \tag{1}$$

We also assume that random variable DTP can be decomposed the following way:

$$DTP = MDT(p) + Dev \tag{2}$$

where the minimum dwell time $MDT(p)$ is a deterministic function of the reduced passenger flow p, and the deviation Dev is a non-negative random variable that accounts for stochastic behaviors such as inhomogeneous passenger distribution on the platform or passenger carrying luggage. All these random variables can be visualised on figure 1 in the introduction. 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. A similar modeling approach was used in [Medeossi, 2008] for performing stochastic simulation of traffic. In that thesis however, the minimal dwell time is a constant value determined by the rolling stock features. The dependence on the passenger flow is therefore not evaluated.

Note that although MDT is chosen as a function of a single variable p, it is however influenced e.g. by temporal factors through p. Indeed, during peak hours in the week, passenger flows will be important and so will be the value of p, whereas during the week off-peak hours and the week-end, the values of p will be lower.

4.2 Computing the minimum dwell time

4.2.1 Algorithm

Our datasets are compound of observations of the random variable DT. On figure 4, we can observe that the minimum dwell time is an increasing and sometimes approximately linear function of the reduced passenger flow p. We propose to estimate it using a two-step process, whose principle can be presented the following way. We first seek to select observations that correspond to situations where $DT = DTP$ and $DTP \simeq MDT(p)$, meaning $Dev \simeq 0$ (step 1). Those observations are then used for estimating the function $MDT(p)$ (step 2).

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. Select the one with the lowest value of DT among all those satisfying $DTT < DT$. Discard all the other observations in $[0, \Delta p)$.
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 ([Altman, 1992]) to the selected data. This method consists, to obtain the value of $MDT(p)$ for a given value of p, to select the k nearest observations along the p axis and take the average of the corresponding values of MDT. We subsequently applied a slight modification to the obtained function. 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 Results on real-world data

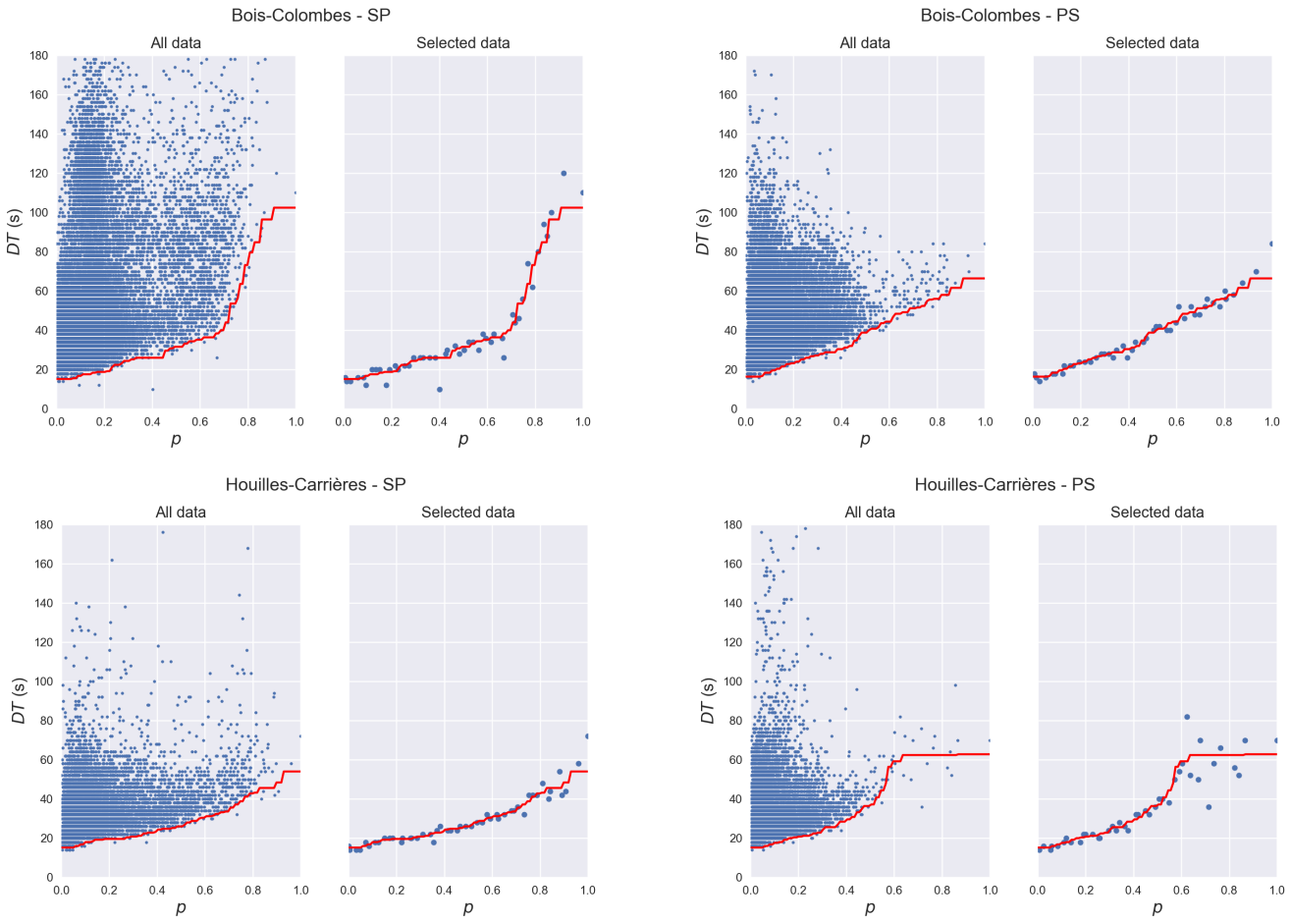

Figure 5: Selected data of each dataset and kNN regression

We first applied the method to fictitious datasets where the function MDT was known, for purposes of calibration and validation. This lead to the choice of parameters $k = 5$ and $\Delta p = 0.005$. Details about this step are provided in appendix. The same method was subsequently applied to the BC-SP,

BC-PS, HC-SP and HC-PS datasets. The selected data and obtained regression curve are given on Figure 5. 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 flows 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 MDT for these passenger flows and compared the obtained results. Figure 6 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 flows, 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).

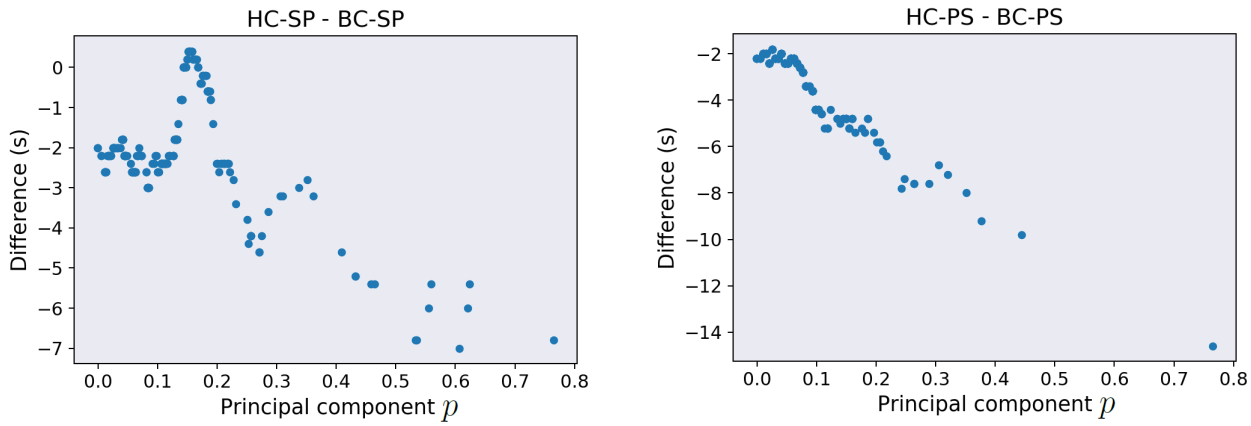

Figure 6: Difference between predictions of MDT for HC and BC datasets

5 Identifying determinants of dwell time

Let us recall that dwell time of trains in stations is not only determined by the passenger exchange process 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 DTT, by $DTT = \max(\widetilde{DT} - Delay, 0)$ where $\widetilde{DT}$ 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 MDT with the method described in the previous section.

Observations are then classified in two sets: those where $MDT > DTT$ correspond to events were passengers numbers justified a longer dwell time in station than planned in the timetable. Conversely, those where $MDT \leq DTT$ are events where the dwell time was likely determined by the no-early-departure rule. Classified observations on BC-SP dataset are plotted on Figure 7. There are 16306 observations in the $MDT \leq DTT$ case, and 7558 observations in the $MDT > DTT$ (similar proportions are observed on the other datasets).

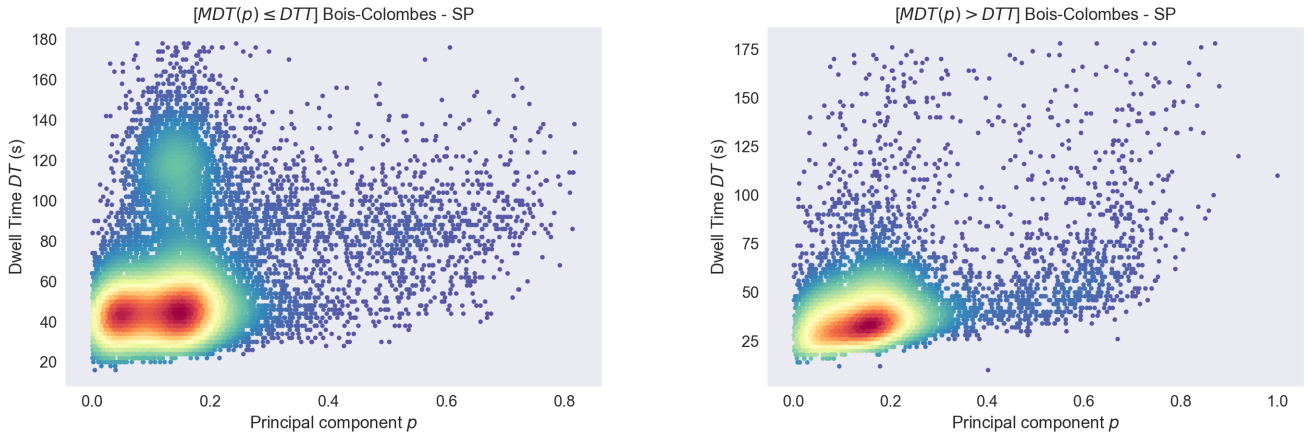

Figure 7: Classification according to the main determinant of dwell time.

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 in the case of Bois-Colombes SP to the right part of Figure 7) 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 (reduced) passenger flow. From now on, we assume passenger flow 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 flow p, the conditional distribution of DT given $P = p$ 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 8; similar curves were obtained on other datasets and are provided for illustrative purposes in appendix, on Figure 14. 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. This is particularly true in the SP datasets. This is due to the fact that in this direction, there are more passengers boarding than alighting. The boarding movement being slower than the alighting movement (due to the fact that a step has to be climbed), it is also more prone to congestion.

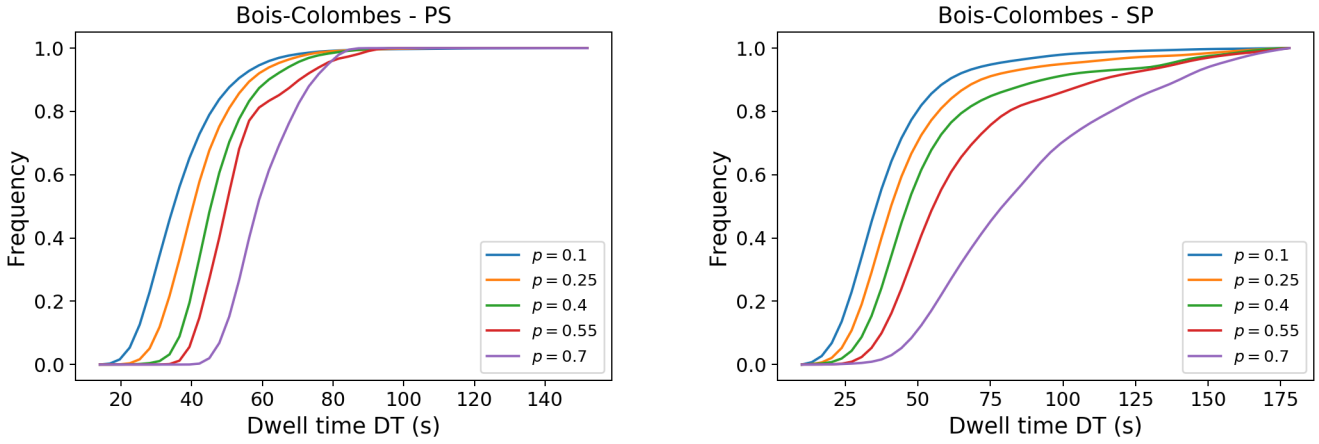

Figure 8: Conditional cdf of DT for different values of p

In order to verify the accuracy of these distribution estimations, we then randomly split each data set into a training set and a test set (with equal sizes). The observations from the training set were used for estimating the conditional distribution of DT given p as described previously. For different values of $\overline{p} \in [0,1]$, we selected the observations in the test set satisfying $p \in [\overline{p} - 0.01, \overline{p} + 0.01]$. We

then computed the deciles of observations contained in the test set as well as the deciles of distributions estimated using the training set. The distributions were then compared by plotting those deciles on a Quantile-Quantile plot [Chambers et al., 2018], see Figure 9.

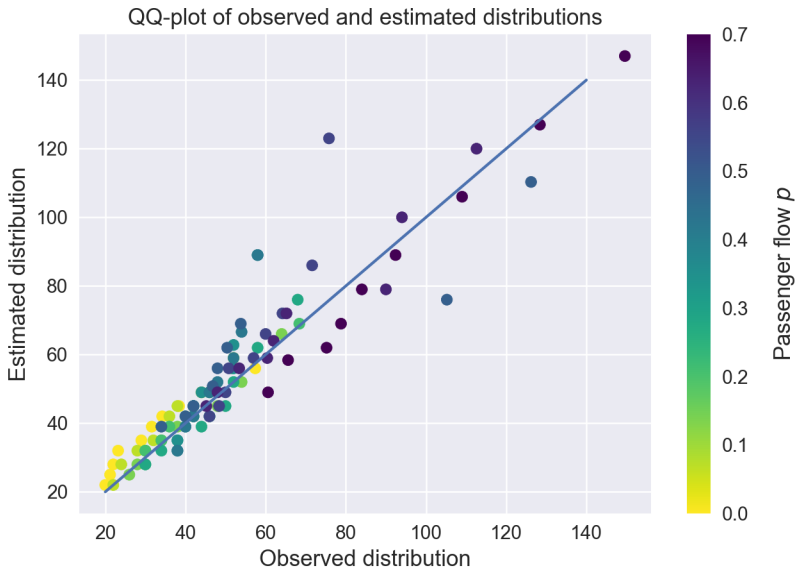

Figure 9: Q-Q plot of observed and estimated conditional distributions of dwell times

The accuracy of the estimation is measured by the distance between the points and the straight line where the points should be if the observed and estimated distributions were perfectly equal. We observe that the conditional distributions of DT given the passenger flow are estimated with reasonable accuracy for the lower values of p. The estimates, in particular the higher quantiles, are less accurate for extreme passenger flows. This is mainly due, as already mentioned in previous parts of the paper, to the relatively small number of observations corresponding to this situation.

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 flow 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 and dimensioning of such systems. 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. These distributions can be used coupled with a passenger flow prediction model. This would then allow to perform quantitative assessments of timetable robustness, forecast delay propagation in real time and provide reliable information to travelers, as well as to give 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. Authors want to thank the anonymous reviewers for their useful comments, as well as W. Daamen, A. Furno and C. Grodecoeur for inspiring discussions during the writing process.

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.

[Altman, 1992] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175–185.

[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.

[Ceder, 2007] Ceder, A. (2007). Passenger demand. In *Public Transit Planning and Operation: Modeling, Practice and Behavior*, pages 309 – 332. CRC Press.

[Chambers et al., 2018] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (2018). *Graphical Methods for Data Analysis*. CRC 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.

[Jolliffe, 2011] Jolliffe, I. (2011). Principal Component Analysis. In Lovric, M., editor, *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg.

[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, Á., 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.

[Wei and Chen, 2012] Wei, Y. and Chen, M.-C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, 21(1):148–162.

[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

Calibration and validation on fictitious test datasets

We verified the ability of our 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

$$MDT_1(p) = 20 + 60p \quad (3.1) \qquad \text{and} \qquad MDT_2(p) = 20 + 60p^2 \quad (3.2)$$

We built these datasets by selecting the values of p in the BC-SP dataset, and randomly generated the values of DT based on equations (1) and (2). We assumed Dev to follow an exponential distribution of parameter 20 and added a random noise following a normal distribution $\mathcal{N}(0, 1)$. Some outliers (representing 0.1% of the total number of observations) were added following a uniform distribution in the space (p, DT). The corresponding points are plotted on Figure 10.

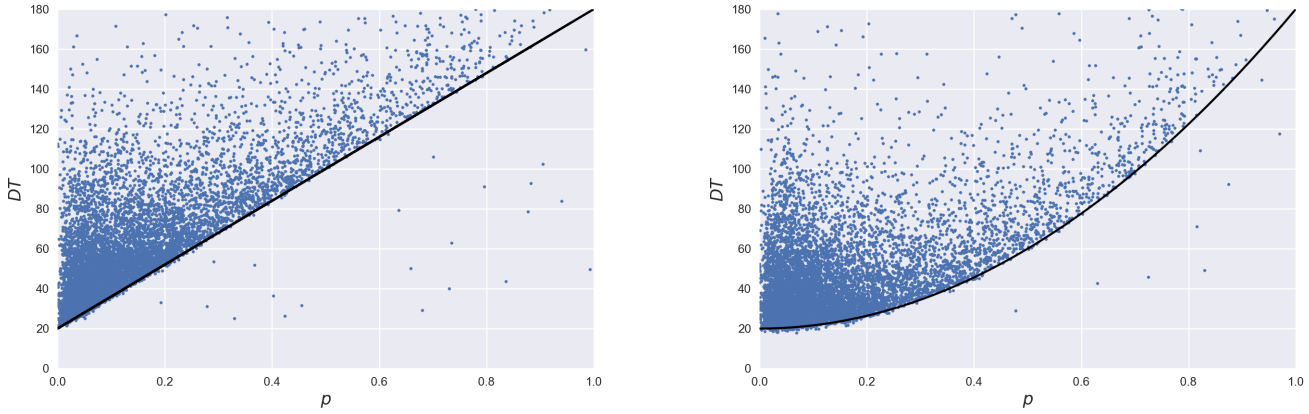

Figure 10: Fictitious test datasets. Left: eq (3.1), right: eq. (3.2)

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 11. 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 12.

We note that in both cases, the value of MDT is estimated with good accuracy when the observations are dense enough, which is the case 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

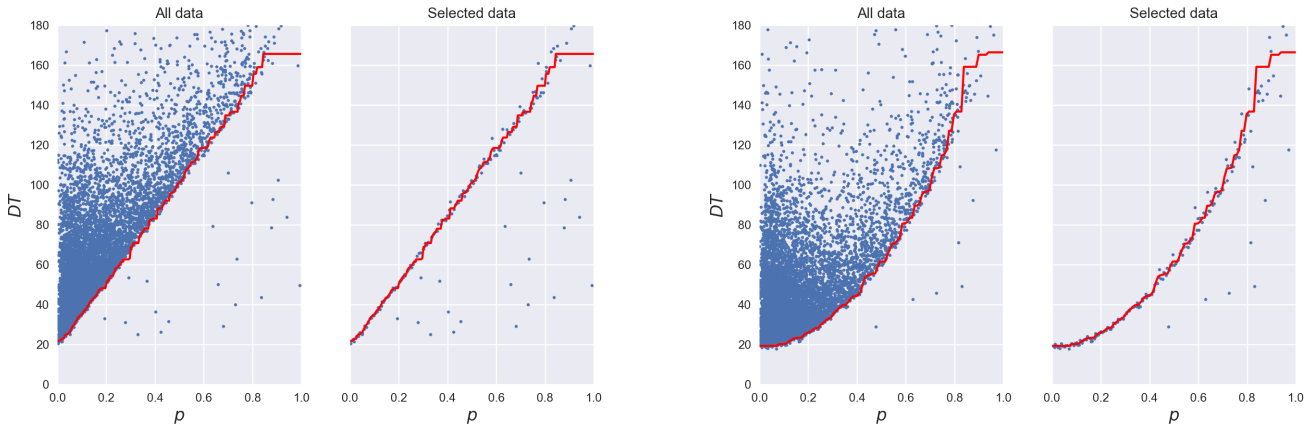

Figure 11: Selected data and estimated MDT function on test datasets. Left: eq (2.1), right: eq. (2.2)

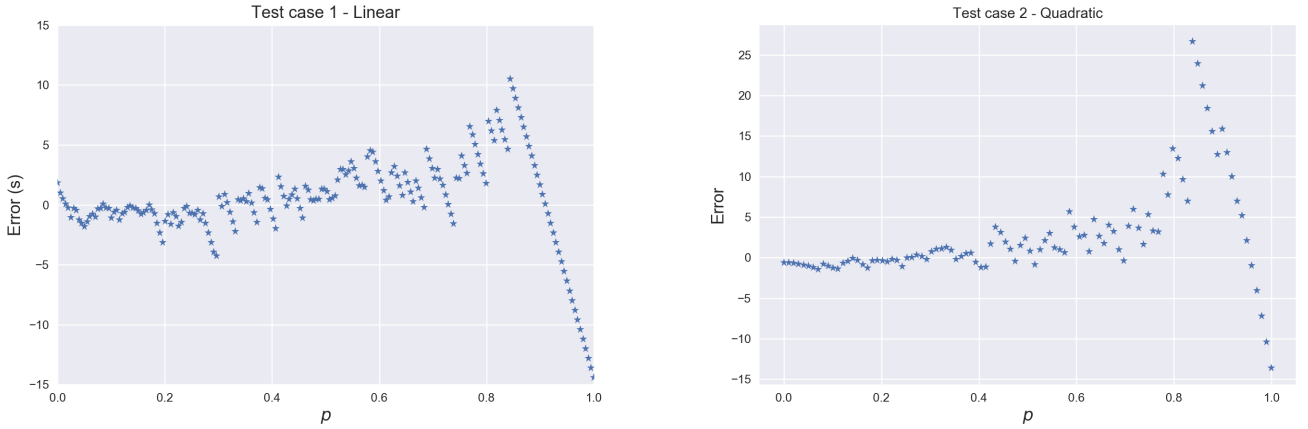

Figure 12: Error on estimated MDT function on test datasets. Left: eq (2.1), right: eq. (2.2). Note that for $p > 0.8$, the lack of points make the prediction inaccurate.

value of MDT_2. The Mean Absolute Error (MAE) was used as criterion:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |MDT_2(p_i) - \widehat{MDT}(p_i)| \tag{7}$$

The values of MAE for different sets of parameters $(\Delta p, k)$ are provided on the colormap of Figure 13. Note that a variety of parameter sets provide satisfying results. In the sequel we used $\Delta p = 0.005$ in order to select a sufficient number of observations, and $k = 5$ as a compromise between a small value (that would yield a curve with possibly many steep variations) and a large one (that would result in a probably too smooth curve). However, some other choices of parameters would have suited as well.

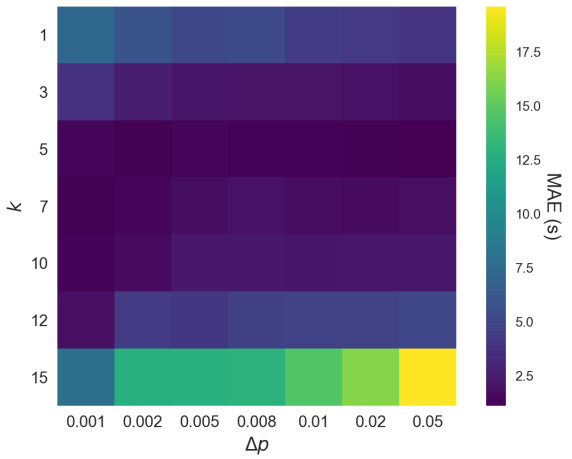

Figure 13: Colormap of MAE for various values of k and Δp

Results on other real-world datasets

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

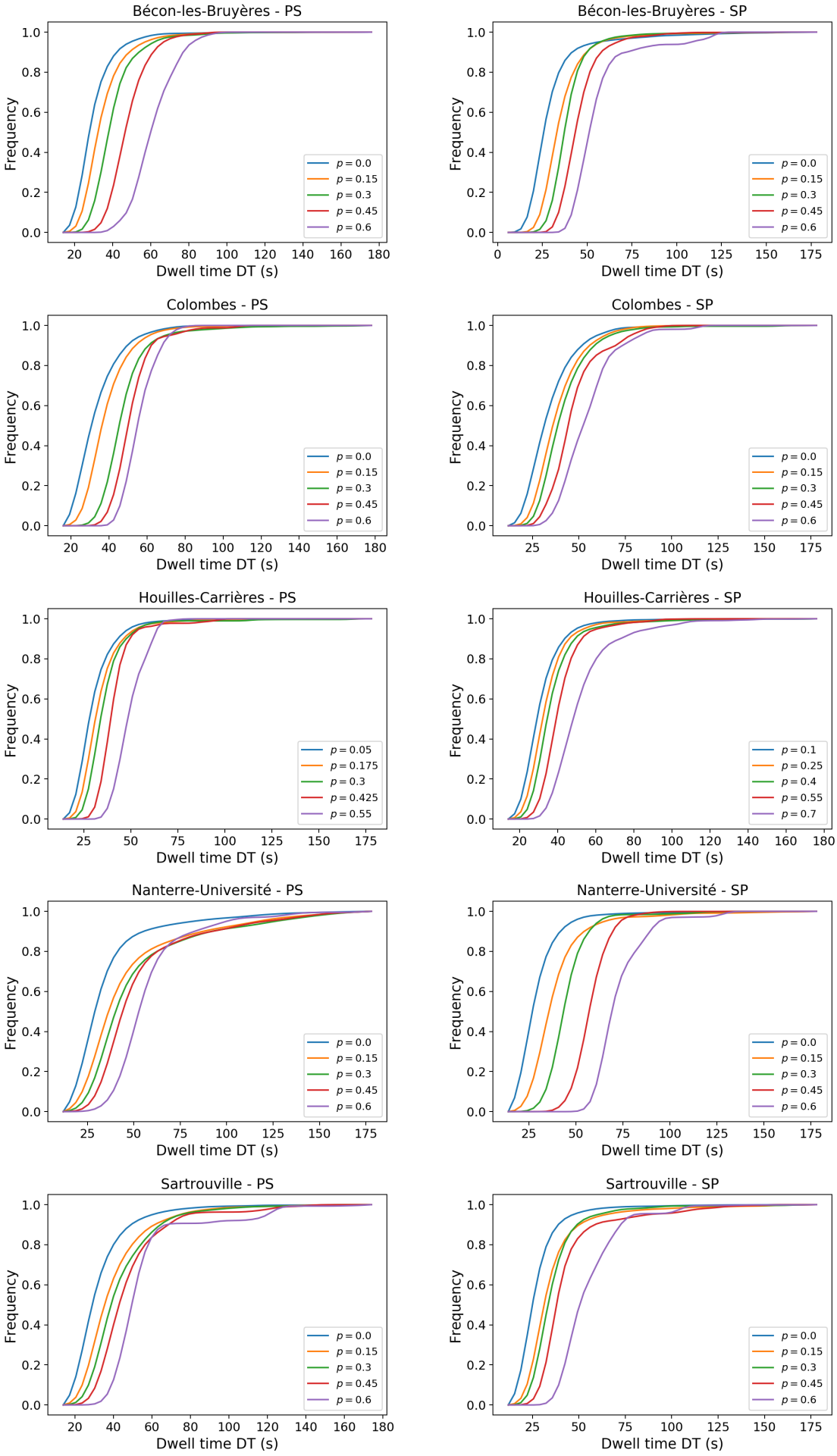

Figure 14: Conditional cdf of dwell time given passenger flow