DTU Management
Department of Technology, Management and Economics

Dwell time in public transport: a statistical model based on big data

Martine Grangé

Master Thesis

Dwell time in public transport: a statistical model based on big data
Thesis subtitle

Master Thesis
May, 2022

By
Martine Grangé

Cover photo: Vibeke Hempler, 2012
Published by: DTU, Department of Technology, Management and Economics, Brovej, Building 116, 2800 Kgs. Lyngby Denmark
www.man.dtu.dk
ISSN: [0000-0000] (electronic version)
ISBN: [000-00-0000-000-0] (electronic version)

ISSN: [0000-0000] (printed version)
ISBN: [000-00-0000-000-0] (printed version)

Approval

This thesis has been prepared over five months at the Transport Division, Department of Technology, Management and Economics at the Technical University of Denmark, DTU, in partial fulfilment for the degree Master of Science in Engineering, MSc Eng. It has been done in a collaboration with the French operator SNCF - Transilien, who has provided data and supervision.

Martine Grangé - s202255

Signature

03/06/2022

Date

Abstract

On the Parisian commuting network, train drivers have to respect carefully theoretical arrival and departure times. Due to this rule, time is lost between the end of the alighting and boarding and the door closure. Indeed, alighting and boarding time is different per door and is the longest at the critical door. But even at critical door, it is scarcely equal to dwell time. Thus, we propose a stochastic method to model the alighting and boarding time based on door used periods called clusters of passengers. We show that two behaviours of the alighting and boarding time exist in function of the number of passengers at the door. We also conclude on the time that could be save if the schedule were more adapted to passengers' flow. On line N of the Transilien network, 1min30 could be saved on a train run of 70 minutes. Finally, better spreading of passengers along the platform or specifying exclusively boarding doors are levers to reduce the alighting and boarding time.

Acknowledgements

At the first place, I am extremely grateful to Assistant Professor Jesper B. Ingvardson and to Professor Thomas K. Rasmussen for their regular advises and guidance during those five months. Besides geographic distance between three countries, they have generously offered their time to give constructive feedback and kind support. Through their wise questions, they have helped to enhance and enrich the whole analysis.

I would also like to express my deepest gratitude to SNCF-Transilien for providing the data and welcoming me. I am grateful to Agnès Grisoglio and Marc Deruelle for their kind inclusion in the MTA team and for their trust. I also could not have undertaken this journey without Rémi Coulaud. Many thanks for his support and encouragements, his clear explanations of the data, the Parisian context and the issues tackled by Transilien. His wise advises and guidance throughout the work were precious as well as his humour.

I would also thank Valentine Mazon, Sophie Marchal, Anis Ait Amira for their camaraderie during working hours. Lastly, I like to mention my dear friends at DTU Transport and especially Vincent Henrion for his help during the proofreading.

Professor Thomas Kjær Rasmussen, Professor in the Transport & Logistics department, DTU
DTU supervisor

Assistant Professor Jesper Bláfoss Ingvardson, Professor in the Transport & Logistics department, DTU
DTU supervisor

Rémi Coulaud, PhD student, SNCF - Transilien, Laboratoire de Mathématique d'Orsay (LMO)
SNCF supervisor

Contents

1 Introduction

With spreading and denser cities, the transportation network is growing as well. The transportation network is particularly impacted by the spreading of cities as city spreading often goes with a specification of the different places inside the urban area. Some business centers, malls or residential areas are created. Thus, people are more and more transferring to go from home to work and from leisure activities to home. Having efficient roads and public transport is a key element not to have a congested city.

Even if both roads and public transport are part of the transportation network, cities are more and more promoting public transports in the green transition context. However, promoting public transport goes hand in hand with a good reputation of performance and reliability of public transport. To evaluate their performance, commuting trains, metro and buses need to identify some indicators and to get some information about them. The indicators can be the number of delayed passengers, the number of cancelled trains, the evolution of the ridership...

In this master thesis, we will focus on commuting trains in the Paris area. Those trains are operated in mixed traffic and constrained by a schedule, so all kind of trains as fret trains, long-distance trains and commuting trains are using the same infrastructure. In such an environment, a schedule combined with signalling creates a secure and fluid train network. The schedule is the information shared with passengers. So it has a key role in the reputation of the commuter trains. For instance, delayed trains are defined from this schedule, as they are trains arriving after their theoretical departure time.

However, schedule and signalling constraint commuting trains and their speed. Indeed, it means that commuting trains cannot leave the station once passengers have alighted or boarded. Therefore, the schedule simultaneously defines delayed trains and has a key role to reduce their number. Thus, the schedule needs to be defined as close as possible to reality, taking into account the ridership and its evolution in function of the hour and of the day. The schedule defines some theoretical dwell times that the driver should respect.

1.1 Dwell time definition

To understand the dwell time process and how to improve it, let's first define its main components.

The dwell time is the time spent by the train at a station. It is noted DT and computed, following equation 1.1, as the difference between the observed train departure time (t_{dep}) and the observed train arrival time (t_{arr}).

$$DT = t_{dep} - t_{arr} \tag{1.1}$$

The dwell time can be separated into four sequences, as shown in Figure 1.1 : (1) a first technical time for door opening, (2) the alighting and boarding time, (3)

some margins and (4) a last technical time for door closing. Door opening and door closing depend on the rolling stock. In this analysis, door opening is fixed at 1 second and door closing at 14 seconds. Those two times summing to 15 seconds constitute technical times, called TT.

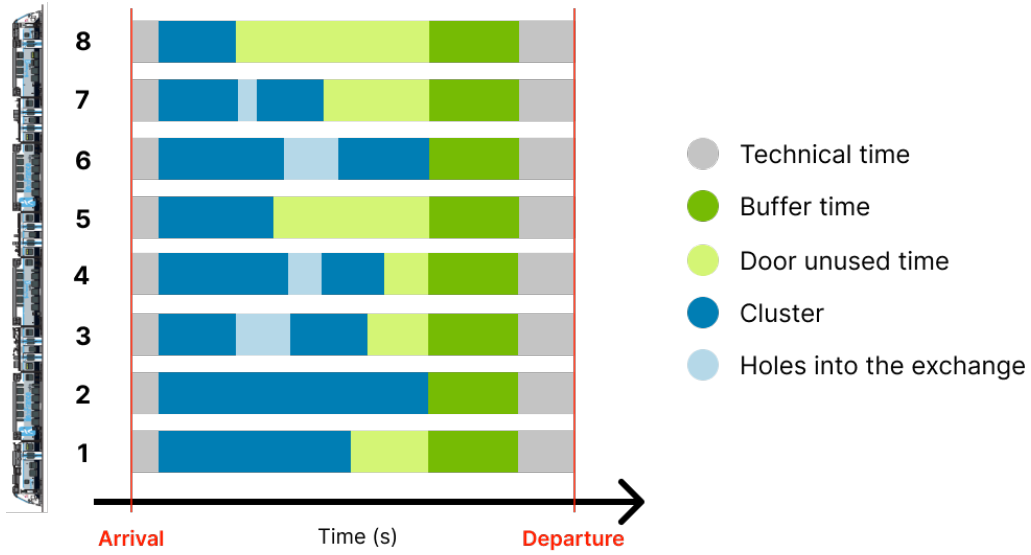


Figure 1.1: Example of the different sequences of the dwell time for a 8 doors carriage at one stop

The *alighting and boarding time* is hard to define and often depends on the available data. For instance, Harris [1], [2] studies alighting and boarding rate on data collected by a gathering of 90 lines around the world. This data is manually collected and therefore the alighting and boarding time is the time between the first passengers to cross the door and the last one. In his thesis Daamen [3] distinguishes the main group of boarders and the late runners. Late runners are opportunist passengers arriving on the platform when the train is already waiting. Thus, Daamen considers that the alighting and boarding time does not take into account late runners.

For this master thesis, a first and general definition can be drawn at the door level : alighting and boarding time is the time spent between the first passenger crossing door i during the stop (t^i_{first}) and the last one (t^i_{last}). The alighting and boarding time Y^i at door i is computed following equation (1.2).

$$Y^i = t^i_{last} - t^i_{first} \tag{1.2}$$

From the definition of alighting and boarding time, one can define the critical door. The *critical door* is the door whose alighting and boarding time is the longest of the train. Thus, the alighting and boarding time of the train is equal to the alighting and boarding time at the critical door : $Y^* = \max_i Y^i$.

Finally, margins can refer to two different realities : the buffer time and the door unused time. On the one hand, the *buffer time* is defined at the train level as the time between the last boarding passenger and the door closing. The buffer time will be called t^*_m and computed following equation (1.3). On the other hand, the

door unused time is defined at the door level as the time between the end of the alighting and boarding time at the given door and the beginning of the buffer time. The door unused time at door i will be called t_m^i and computed following equation (1.4).

$$t_m^* = DT - TT - Y^* \tag{1.3}$$

$$t_m^i = DT - TT - Y^i - t_m^* \tag{1.4}$$

Each of the components of the dwell time depends on the others. Thus, when modelling the dwell time, the role of each component will be interesting to understand.

1.2 Objective of the master thesis

Having a good picture of the time taken by passengers to alight and board, help to adapt the service through for example the timetable, and thus to improve the passengers' perception. The understanding of the dwell time is closely linked to the available data. The data set used in this master thesis gives the number of passengers alighting or boarding the train for each seconds of the stop. It is the first time that Transilien operator has such a precise data set. So, it creates the opportunity to compute margins and give a precise time for the alighting and boarding time. Then, this master thesis aims at answering the following questions:

How does a precise data set on the alighting and boarding of passengers during the dwell time give insight on the structure of the dwell time and particularly on the definition of the margins and their computations ? How can some behaviours of passengers be understood through detailed data ? Which model can be chosen to estimate the alighting and boarding time ? How can a model of the alighting and boarding time help in the timetable building process ?

It is important to note that the focus here was mainly on modelling the alighting and boarding time and on the knowledge it gives to Transilien in the timetable building process.

We first proceed a literature review on the modelling of the dwell time and the passengers behaviour during it. Then, the context of the study is presented including the Transilien network, the data set used and its preprocessing. From this data, some statistics are given to better understand it before the presentation of the method and models. The results are finally presented and discussed. Both the method and results parts present first the estimation of the alighting and boarding time before the margin study and adaptation of the timetable.

1.3 Bibliography review

The literature review tries to get an understanding of the modelling already done globally. Thus, definition of the dwell time and separation with the alighting and boarding time are investigated. Different modelling approaches are identified as

well as variables selected into the models. Furthermore, factors influencing passengers' speed during their alighting or boarding often conclude observations and modelling.

1.3.1 Dwell time and alighting and boarding time definition

First of all, the dwell time is categorised and split in several ways. Those categorisation often depends on the available data. So it leads to different definitions of the alighting and boarding time.

In their article from 1992, Lin & Wilson [4] distinguished different types of dwell times based on the public transport kind. Buses dwell times that are dense, frequent and led by the alighting and boarding time. So margins are not defined for headways based service. On the other hand, commuter train dwell times are much less frequent, on much longer trains and led by scheduled timetable. Margins appear in schedule based service.

Focusing specifically on trains networks and working on data close to the ones studied in this thesis, Cornet [5] introduced the concept of minimum dwell time. This minimum dwell time is the time needed for a given number of passengers to alight and board and for the train to manage the technical times at a given station. This minimum dwell time becomes a lower bound for the scheduled dwell time.

Buchmüller [6] is on of the authors trying to separate the boarding and alighting time from the dwell time and to give a representation of it. In Buchmüller's article, the network under study is similar to the Transilien network: constrained by a schedule and operated in mixed traffic. Mixed traffic networks are networks where long-distance trains, suburban trains, regional trains and freight trains all share the same infrastructure. Finally, the Swiss network studied by Buchmüller was equipped by 30% with automatic passenger counting system. Automatic passengers counting system collect really precise data but need heavy computationally storage. All those characteristics make this article very close to Transilien context. Buchmüller broke down the dwell time into subsets, tried to define each part and gave some statistics on each of them. However, he considered only delayed trains, so margins are not defined as part of the dwell time and he did not describe any model or prediction.

Finally, in 2001, Wiggenraad [7] is the first author to define some "clusters of passengers" on Dutch railway stations. He assumes that a passenger is part of a cluster if the time interval between his predecessor and himself is less than 3 seconds. He identified thanks to those clusters some passengers' behaviour during the alighting and boarding time. However, choosing 3 seconds between passengers of the same cluster is not justified and we will define a new threshold later on.

Therefore, several notions have been defined by the different researchers to better capture the dwell time and its complexity. The issue for the researchers have thus been to grasp a reality if the alighting and boarding time.

1.3.2 Alighting and boarding time modelling

A lot of models have already been set up to estimate the dwell time and the behaviour of passengers when boarding and alighting the train.

Regressions In 1992, Lin & Wilson [4] tried some linear and non-linear regression models whose variables were the number of boarding, the number of alighting and the load of the train. Those simple models are evaluated and compared through t-statistics. They do not care about multicollinearity between passengers' variables.

It is to avoid the multicollinearity that Cornet [5] introduces a new variable p depending on the station. One value of p at one station leads to a value of boarding passengers, a value of alighting passengers and a train load when leaving the station. In his three articles, Harris [8], [9], [1] uses the data from more than 90 operators around the world to understand the alighting and boarding time. CoMet and Nova metro benchmarking groups gather more than 90 lines around the world and are managed by the Railway Technology Strategy Centre at Imperial College London. Using this huge panel of public transport, a unified method of collecting data have been set up. Each operator part of these groups should send two observers in the stations to collect times and number of passengers at the critical door. Thus, characteristics of the rolling stock and of the layout of the platform are part of the set of variables used by Harris. He was able to give a precise regression of the alighting and boarding rates. His models are based on delayed trains and his research on the Weston formula. Weston formula is written by Weston in 1970 and is the first formula about dwell time in any scientific paper.

Finally, the normality have been questionned by Li in 2014 [10]. Li used track data registering the train arrival and departure on the Dutch network. So to approximate alighting and boarding time, his studies are using delayed trains. Li concluded that alighting and boarding time follows a log-normal distribution.

Queues model Palmqvist [11] in a literature review made in 2021 report analysis on formation of lanes and queues around the door and in the way between seats in the train during the exchange time. In his thesis Daamen [3], tries to model the passengers flows inside transport facilities. Therefore, a small part of the simulation is dedicated to the alighting and boarding processes. He first describes how the dwell time can be split, then he assumes that alighting and boarding processes happen as queuing systems. His simulation is limited since he assumes that boarding passengers wait for alighting passengers to alight before beginning to board. Secondly, he assumes that only one person can use the door at the same time.

Other models In 2016, Li [12] tried to create a model avoiding counting data and heavy computations. His goal was to estimate in real time the alighting and boarding time using characteristics of the day and dwell times at previous stations or on previous trains. Although this model is trained on the Dutch network, it should be appropriate for all transport facilities around the world. As he does not consider passengers behaviour, substitute variables are chosen such that they

are available in real-time. Ten parametric models and a non-parametric model using the **k-NN method** are tested.

Zhang [13] creates in 2008 a **cellular automata** model to simulate the behavior of alighting and boarding passengers in Beijing metro stations. He defined rules to moves to the next cells, to manage conflicts, and he introduces some desire and energy to move. This model translates in a really good way some characteristics of the exchange of passengers during the alighting and boarding time. The main drawback of this model is the assumption that passengers choose a door and cannot modify their choice during the alighting and boarding time.

Finally, Su [14] creates in 2019 a simulation tool estimating the alighting, boarding and settling time of passengers. Su creates a simulation based on **agent-based modelling** to estimate the alighting and boarding time of passengers and evaluate it on data from Santiago de Chile. They consider a set of actions and interactions between the different passengers. However, in their simulation, it is assumed that boarders wait for all alighters to get off the train. This assumption cannot be kept when it comes to the Parisian network.

1.3.3 Factors speeding up or slowing down the speed of passengers

Factors that influence dwell time are number of passengers, occupancy of the train [15], spreading of passengers along the platform, driver behavior [10], presence of cumbersome luggage [15], on the rolling stock feature such as the number and width of doorways [1], and the station configuration such as the curvature, the height, the step between the platform and the train [1].

Palmqvist [11] defines the “usable door width”, or the door width that is left when waiting boarding passengers are narrowing the space available in front of the door. Daamen [15] gave some figures for the alighting and boarding rates. He particularly studied the impact of a gap between the train and the platform and the presence of cumbersome luggage on those rates. The results from Su’s article [14] show that an interior layout with the minimum number of seats leads to quicker exchange time.

Li considers the driver behaviour into the computation of the alighting and boarding time and he interested himself in some correlations [10]. He looked at the correlation between the hour of the day and the alighting and boarding time, the one between the dwell time of the previous station or the delay at the previous station and the alighting and boarding time at the current station.

Conclusion To conclude this literature review, dwell time and alighting and boarding time have not been often separated. The data collection does not ease the differentiation between them as it is hard and heavy to get precise and abundant data on the alighting and boarding of passengers. However, many models try to understand and shorten the time spent by passengers to alight and board, concluding on the main factors influencing the speed of passengers.

2 Network and data context

2.1 Network presentation

Transilien is the operator of the commuting network connecting the outer suburbs to the center of Paris including eleven lines named with letters (C, D, E, H, J, K, L, N, P, R, U). Different rolling stocks are running on the network, but in this study, we are using only data coming from Regio2N trains. Regio2N trains are running on lines N and R and are able to record detailed data. As line N transports more passengers than line R, it was chosen to study alighting and boarding time on line N and to validate the model on line R.

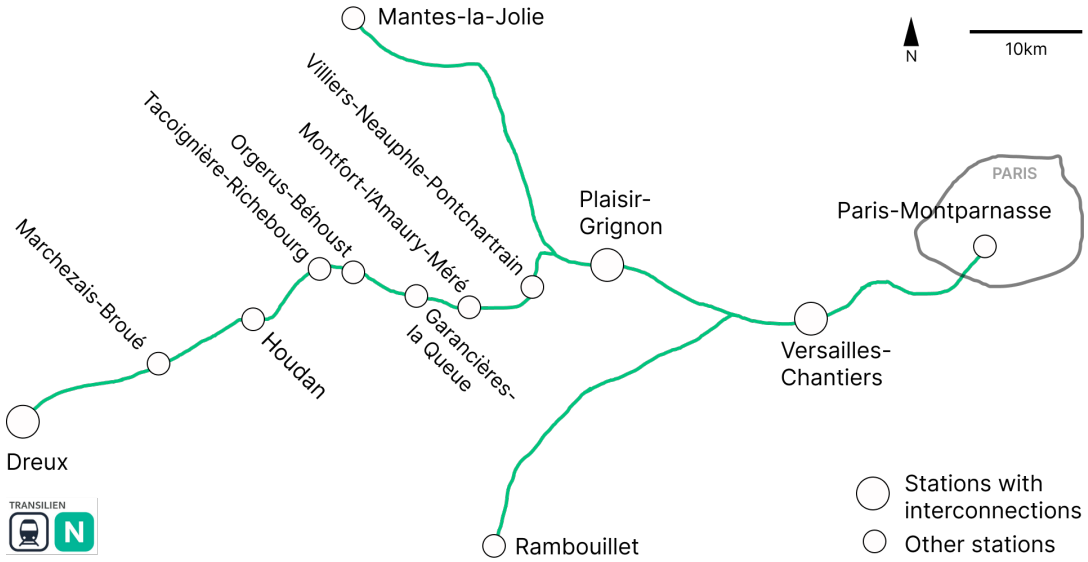


Figure 2.1: Schematic geographical situation of line N with studied stations and terminus

Figure 2.1 presents a map of line N with its terminus and the stations analysed here-below. Line N serves western suburbs leaving from Paris-Montparnasse train station. One of its terminus is Dreux situated outside Ile-de-France administrative region, 80km far from Paris. In this reprt, the alighting and boarding time is studied focusing on the Dreux branch since, in September and October, Regio2N trains were running only on that branch.

Trains serving from and to Dreux stop at nine stations represented on Figure 2.1 with a frequency depending in the hour of the day. On Dreux branch, there are three trains per hour between 6am and 8am towards Paris and two trains per hour between 16pm and 21pm in the suburbs direction. During off peak hours, only one train per hour is running in both directions. However, those trains are added to the trains coming from the two other branches of the line. Thus, during peak hours, one train stops every 4 minutes at Versailles-Chantiers station and every 10 minutes at Plaisir-Grignon. So, the frequency is high on the main branch

of the line. Therefore, line N is a busy line on the main branch but Dreux branch is less busy. The main nodes are thus Versailles-Chantiers and Plaisir-Grignon stations.

Coming to passenger numbers, at Versailles-Chantiers during peak hours, between 200 and 400 passengers alight or board. This number can go up to 700 passengers in really crowded days. As a comparison, during peak hours, between 100 and 300 passengers transfer at Plaisir Grignon, and between 25 and 125 at Houdan.

2.2 Data collection

As it has been said previously, data come from a specific rolling stock. The trains under study, called Regio2N, are used on lines N and R of the commuter train network. These double-deck trains are composed of either one or two train sets represented on Figure 2.2. Each train set is 110m long and has eight 1.6m-wide doors. The seated capacity of a train set is of 576 passengers while total capacity is of 1046 passengers. The seated capacity is the number of seats available in the train while the total capacity includes both seated and stood capacity. Stood capacity is considered to be 4 passengers per square meters. As one can see on Figure 2.2, the layout of Regio2N trains alternate between door-specific cars and double-deck cars with seats.

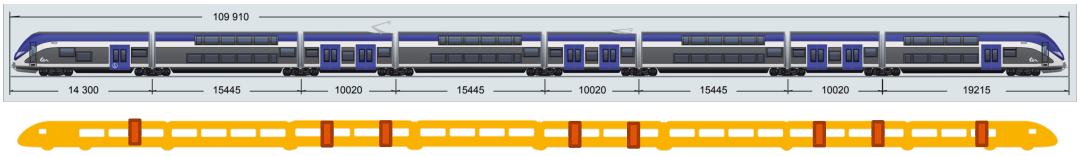


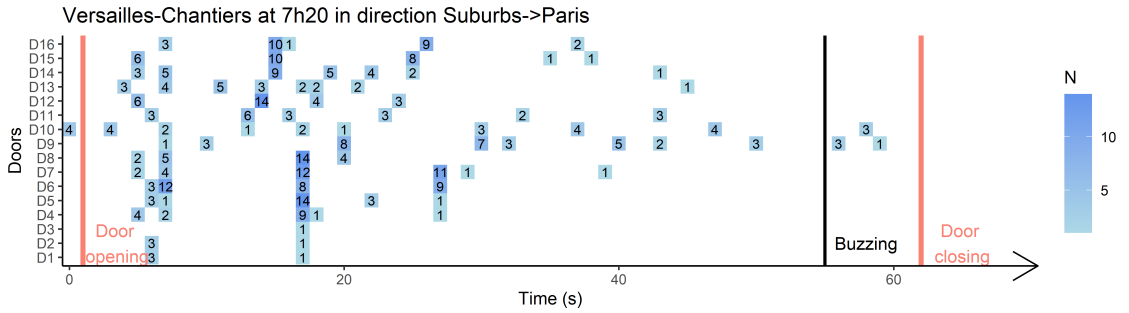
Figure 2.2: Layout of Regio2N trains

The trains are equipped with an automatic passenger counting (APC) system using infrared lights; captors are located above each doorway and are able to detect separately alighting and boarding movements. The captors are reliable at a 95% level. Captors are recording continuously both number of alighting and boarding at each door. Thus, data contains measure points described by a time, a number of boarders and a number of alighters. The period of observation goes from September to October 2021. Even if 2021 has been a year impacted by the coronavirus pandemic, the number of passengers was around 80% of the 2019 level in September and October.

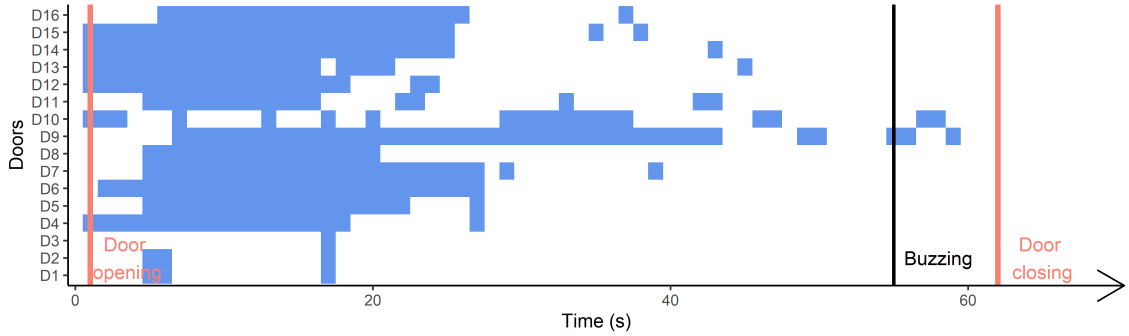
2.3 Preprocessing on the dataset

Data set gives us many information including the number of passengers alighting or boarding and the time of the observation. However, this information has to be transformed into alighting and boarding time. The way the alighting and boarding time is computed changes its definition and its accuracy. So we propose here a way to compute the alighting and boarding time and cleaning of the dataset. All computations are done with the coding software R.

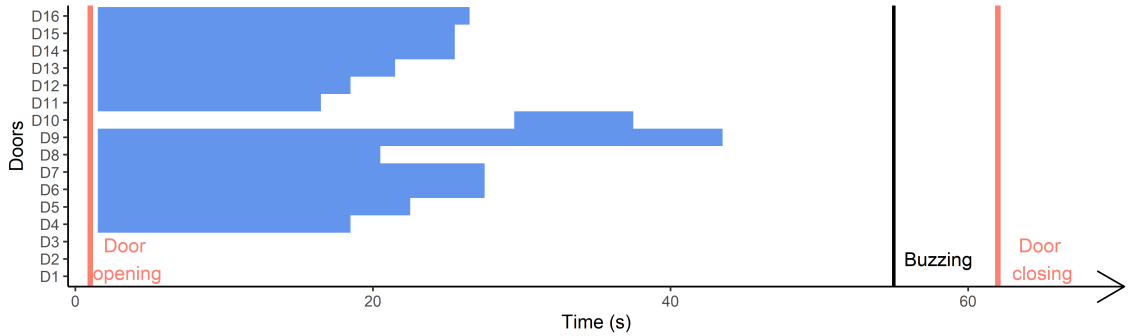
In the raw data set, the alighting and boarding of passengers are represented as measure points in time as shown in Figure 2.3a. Measures made outside the observed dwell time or when door are closed are considered as error from the captor and are disregarded. Finally, stops whose dwell time was higher than 180 seconds were discarded as this delay might come from factors outside the scope of this study.



(a) Raw data



(b) Clusters transformation



(c) Observed alighting and boarding times

Figure 2.3: Example of a stop at Versailles-Chantiers station, at 7:20am, in the direction Suburbs-Paris. In b) clusters are computed. Then, alighting and boarding time are deduced from the clusters in c)

Inspired by Wiggenraad [7], some clusters of passengers were defined. A cluster of passenger is defined as a lane of passengers such that two passengers part of this lane are separated by less than S seconds. *Late runners* are defined as passengers who are not part of any of the clusters. The alighting and boarding at

one specific door is composed of one or several clusters of passengers and late runners.

Wiggenraad [7] set the parameter S to 3s but his choice is not justified. Thus, in this study, different thresholds were tried and based on key indicators, 2 seconds has been chosen as the most realistic threshold to describe a cluster. The criteria used to make this decision are detailed in Table 2.1. A full Table of all considered indicators is provided in Appendix A.

It was essential to keep the more data as possible without transform reality. Thus S equal to 1s was discarded as it keeps too few data. Then the number of clusters per door was analysed. Indeed, as more than 2 clusters at one stop at one door is hardly observed.

Table 2.1: Key indicators computed on data from line N between September and October 2021 to choose the definition of a cluster

Time between 2 passengers of the same cluster (s)	Number of clusters	Number of different stops	Percentage of doors with less than 2 clusters (%)
<1	8,347	1,640	98.4
<1.5	13,761	2,420	98.7
<2	14,674	2,532	99.2
<2.5	16,039	2,746	99.5
<3	17,084	2,832	99.5

With the introduction of clusters, the alighting and boarding time at the door level is defined as the time spent between the beginning of the first cluster at this door and the end of the last cluster. This definition does not consider late runners as they are assumed to be opportunist passengers. The transformation of the data is illustrated for one stop in Figures 2.3.

Using this method, the alighting and boarding time can be computed for each door of each stop composing the data set. So during the computation of the clusters, three data sets are created. The first one, called ”Clusters” contains each clusters and its corresponding variables, the second contains all the stops where there are too few people to create any cluster. The stops part of ”Nobody” data set are considered to have an alighting and boarding time equal to 0s. The last one contains stops considered as outliers. Indeed, the flow of passengers using a given door was sometimes very high. The quantile 99% of this flow is chosen to separate outliers. It corresponds to a flow of 3 pass/s. Thus, the last data set, called ”Outliers”, contains all doors for which the flow of passengers was above 3 pass/s. This last dataset is kept to be estimated later on in the analysis.

The main data set containing the clusters and their variables is composed of 2,532 stops characterised by 13,253 doors and 14,674 clusters. Even if a train is composed of either 8 or 16 doors, not all doors are part of the data set because some of them are not open during the stop. The ”Clusters” set is then split into a training set and a testing set with proportions 75% and 25%. This split is based on train runs, so that full train runs are part of only one of the sets. This consideration is useful when coming to the computations of the time savings in Section 4.3.2. The

training set is used to, fit different models while the testing set enables to compute goodness-of-fit metric to rank such models.

2.4 Variables

For each measure of the dataset, are known the train number k, the station s, the date d, the theoretical arrival and departure times at the given stop, the observed arrival and departure times at the given stop, and the direction. The direction can be either Paris-Suburbs (PS) or Suburbs-Paris (SP). The direction is linked to the platform p on which train stops.

Table 2.2 summarizes all variables that have been considered for the model.

Table 2.2: List of variables used into regressions

Variables	**Type**	**Interval**	**Unit**	**Notation**
Alighting and boarding time at door i	Continuous	[0, 165]	Second	$Y^i_{k,s,d}$
Passengers-specific variables				
Number of boardings at door i	Continuous	[0, 86]	Passenger	$B^i_{k,s,d}$
Number of alightings at door i	Continuous	[0, 136]	Passenger	$A^i_{k,s,d}$
Total number of passengers at door i	Continuous	[0, 145]	Passenger	$N^i_{k,s,d}$
Occupancy of the train unit	Continuous	[0, 2058]	Passenger	$L_{k,s,d}$
Platform-specific variables				
Platform width	Categorical	{1, 2, 3}		$W_{s,p}$
Number of exits on the platform	Discrete	[1, 5]	Exit	$E_{s,p}$
Gap between the train and the platform:				
Vertical	Continuous	[0, 35]	cm	$H_{s,p}$
Horizontal	Continuous	[0, 26]	cm	$V_{s,p}$

The dependent variable is the alighting and boarding time designated as Y. The models proposed below will try to approximate Y.

Passengers-specific variables, including the number of alighting and boarding passengers (A, B and N) and the occupancy of the train set (L), are integers, so they are discrete variables. However, they can be considered as continuous variables as in Table 2.2. Additionally, one should underline that $N^i_{k,s,d} = B^i_{k,s,d} + A^i_{k,s,d}$.

The platform-specific variables include platform width. Platform width is a categorical variable taking values in {1,2,3}. 1 represents a narrow platform, 3 a wide platform and 2 a medium-size platform. Number of exits on the platform is a discrete variables. Versailles-Chantiers station has 5 exits on its platforms. All other platform have either one or two exits. Finally the horizontal and vertical gaps between the train and the platform are given in cm.

One should notice that, used data come from a unique kind of rolling stock. So, contrary to studies like Harris ones [1], the model does not include any variable describing the rolling stock.

In addition to variables presented in Table 2.2, transformations are applied to continuous variables. Square functions, logarithm function, root squared functions

are added to the set of variables. Finally, the interaction between variables are alsotaken into account.

To conclude, data from line N of the Transilien network is transformed into a ”Clusters” data set to compute alighting and boarding time. This ”Clusters” data set is split into a training set and a testing set. All data come from Regio2N trains, so passengers-specific variables and platform-specific variables are selected for the modelling of the alighting and boarding time. Some transformations and interactions of these variables are also included in the following analysis.

3 Statistical analysis

To begin with, some statistics are computed to give a global picture of the data set. In this chapter, we will first focus on the alighting and boarding times statistics and passengers statistics. Then, theoretical dwell times will be presented. Finally, door unused time and buffer time are statistically analysed.

3.1 Descriptives statistics

The data set will now and until Section 4.3.2 refer to the ”Clusters” data set containing 2,532 stops. As said previously, it contains 13,253 doors and 14,674 clusters.

Distribution of alighting and boarding time The data set is split into a training set and a testing set. The training set contains 9,484 doors while the testing set contains 2,881 doors. Thus, 9,484 alighting and boarding times are used for the modelling and the analysis and 2,881 times will enable to compute goodness-of-fit metrics. Alighting and boarding time distribution is presented in Figure 3.1.

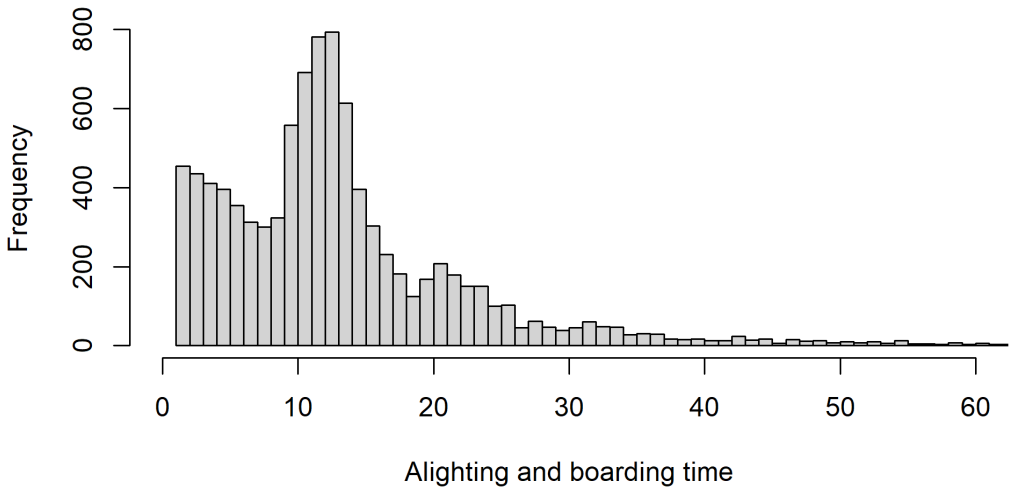
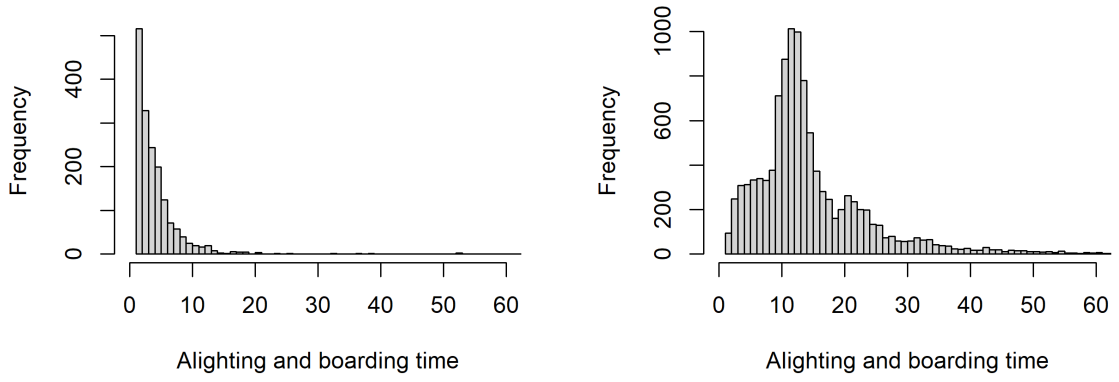


Figure 3.1: Distribution of the alighting and boarding time on the Cluster data set containing data on line N from September to October 2021

This distribution presents one main mode around 12s. However, close to 0 the distribution seems to increase as well. Figures 3.2 show two separated histograms if the data set is split according to the number of passengers N at the door during the given stop. These figures introduce the idea of two behaviours depending on the number of passengers at the door. The first one looks more exponential, while the second one presents one mode. The separation of those two behaviours will be studied in Section 4.2.2.



(a) Small N (b) High N

Figure 3.2: Separation of the two modes with a split of the data set according to the number of passengers. Data from line N between September and October 2021

Statistics on the number of passengers Let's give some statistics about the number of passengers. Table 3.1 compares the global statistics to statistic on the dataset reduced to doors with a small N and doors with a high N. In Section 5.1.3, we will show that the data set should be split around $N = 4$, this split is used in the following statistics. The doors used by less than 4 passengers represent 14% of the "Clusters" data set. So small transfers is a great part of the data set. It is interesting to observe that the load of the train is in general lower for a small number of passengers than for a high number of passengers. This means that when there is at least one door used by a small number of passengers, the train load is globally lower.

Table 3.1: Descriptive statistics of passengers variables at the door level of Clusters data set

	All observations			**Small N**			**High N**		
	A^i	B^i	**L**	A^i	B^i	**L**	A^i	B^i	**L**
Number of observations	12365			1690			10675		
Mean	6.9	6.5	194	1.3	1.2	106	7.7	7.3	207
Median	3	3	146	1	1	71	5	5	163
Standard deviation (Std)	9.3	8.1	187	1.0	1.0	136	9.7	8.4	190

Table 3.2: Statistics of the passengers variables for two stations at the train level

(a) Versailles-Chantiers
449 stops

	A	**B**	**L**	**Y**
Mean	108	63	350	31.9
Median	80	46	286	29.5
Std	98	61	255	17.0

(b) Plaisir-Grignon
420 stops

	A	**B**	**L**	**Y**
Mean	46	66	273	22.6
Median	26	33	203	19.5
Std	54	87	222	14.3

Table 3.2 focuses on two main stations of the line, namely Versailles-Chantiers and Plaisir-Grignon stations. One third of the observations occur at Versailles-Chantiers station (4045 doors) while one quarter occur at Plaisir-Grignon (3320 doors). Versailles-Chantiers station is a particularly dense station : the number of passengers alighting or boarding at Versailles are nearly twice higher than at Plaisir-Grignon station. Moreover, the load of the train is higher in Versailles than in Plaisir-Grignon, meaning that some of the passengers only transfer between Paris and Versailles.

Correlation matrix The correlation between all variables is computed to understand colinearity between variables. The correlation matrix is shown in Figure 3.3.

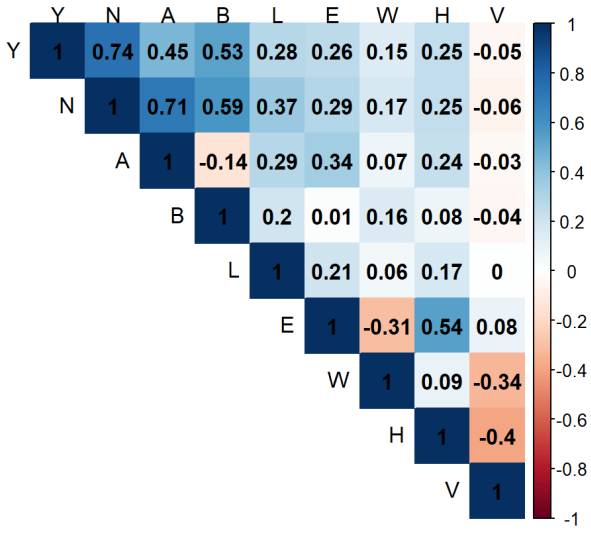


Figure 3.3: Correlation matrix between variables of the ”Clusters” data set

As we can see on Figure 3.3, the alighting and boarding time is highly correlated to the number of passengers and especially with the number of boarding passengers. The total number of passengers is highly correlated with A and B as N is the sum of A and B. To avoid multicollinearity, the model will include only N or a combination of A and B. Indeed, number of alighting passengers and number of boarding passengers are not correlated. Their correlation coefficient is -0.14. This coefficient seems reasonable as this study takes place at a given door during a given stop. So no effect of correlation between stops are studied.

All others correlation coefficients are less than 0.5 except for the correlation between the number of exits on the platform and the horizontal gap between the train and the platform.

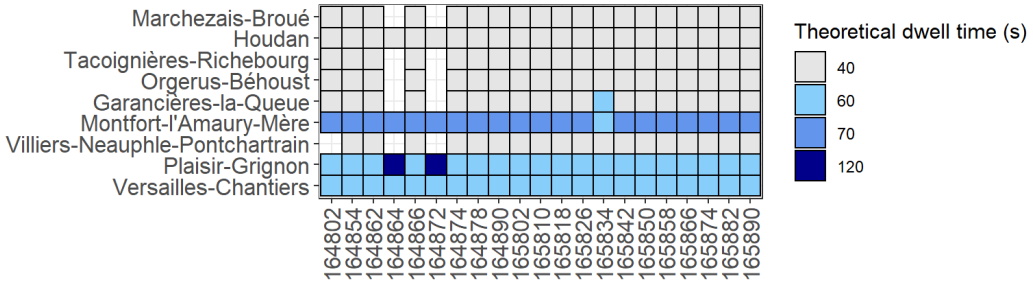
3.2 Theoretical dwell times

Theoretical dwell time definition The theoretical dwell time refers to the scheduled dwell time. As we have access to the scheduled arrival and departure time, the theoretical dwell time can be computed. Also, the theoretical alighting and

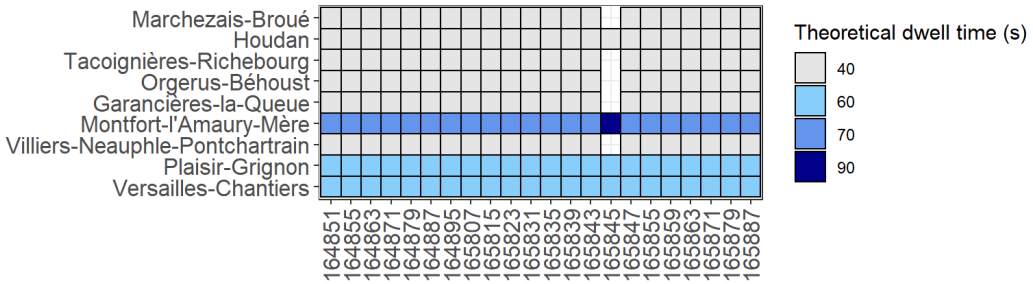
boarding time is the theoretical dwell time minus 15 seconds of technical time. Thus, looking at the schedule gives an idea of the planned alighting and boarding time.

The theoretical dwell times are set to fit the observed ones for nearly all stops. However, between Montparnasse and Dreux, a lot of time is lost at the station waiting for the theoretical departure time. One can wonder if theoretical dwell times are not too long.

Some stations have different dwell time along the day. Figures 3.4a and 3.4b shows the theoretical dwell times in function of the station and the train number.



(a) Even trains (direction Suburbs-Paris)



(b) Odd trains (direction Paris-Suburbs)

Figure 3.4: Theoretical dwell time in function of the station and the train number for line N between September and October 2021

Delayed trains because of a too long alighting and boarding time at the train level As we try to understand the alighting and boarding time, in this paragraph, we check the occurrence of situations where the theoretical alighting and boarding time is too short because of too many passengers. This study gives an overview of stops for which the scheduled dwell time should be widen. From this study, five main conclusions were drawn.

- Among 2532 stops, 189 stops have an observed alighting and boarding time longer than the scheduled one. In average, the delay is around 12 seconds.
- Among those 189 stops, only a third (62 stops) implies delayed trains. A delayed train is a train that arrives at the station after its theoretical departure time.

- Among the 637 delayed trains from the initial 2448 ones, 62 have an alighting and boarding time longer than the scheduled one. So 10% of delayed trains will have to manage crowded platform.
- Among the 189 long alighting and boarding times, 83 occurred at ”Versailles-Chantiers” station, so a half of them, and 52 at ”Houdan” station, so a third of them.
- Among 995 trains ahead of the schedule, 78 are delayed because the observed alighting and boarding time is higher than the theoretical one added to the additional seconds from the advance. It represents 8% of them.

To conclude, 7.7% of all stops are delayed because of a too long alighting and boarding time. Increasing the theoretical dwell time or adapting it in function of the hour of the day can solve those trains delays because of the ridership. However, other leverages exist to control the alighting and boarding time, including better understanding of the spreading of passengers along the platform.

3.3 Factors influencing on door unused time

The door unused time is defined by equation (1.4) as the time lost at the door level waiting for the passengers at the critical door to alight and board the train. This precise data set allows to deeper analyse the impact of the uneven spreading of passengers on the total alighting and boarding time. In this section, door unused time is analysed in function of the number of passengers (N) and of the number of exits on the platform (E). Also to keep only one measure per stop, the mean of all door unused time is kept. Indeed, the trains can be composed of either one or two train sets. So the mean is more robust than the sum of all door unused time.

3.3.1 Door unused time in function of N

One of the factors influencing the spreading of passengers on the platform is the number of passengers. Intuitively, the more passengers are waiting on the platform, the more spread they are. To understand the spreading of the passengers, the door unused time is expressed, following equation (3.1), as a fraction of the alighting and boarding time at the train level. This value will be called the *Spreading Indicator (SI)*. Thus if the spreading indicator represents 0% of the train alighting and boarding time, it means that passengers are perfectly spread. On the contrary, if the spreading indicator is equal to 100% of the train alighting and boarding time, it means that passengers are all grouped at the critical door.

$$SI = \frac{\sum_{i=1}^{I} t_m^i}{I \times Y^*} \tag{3.1}$$

Figure 3.5 presents the door unused time in function of the number of passengers. All passengers are considered including alighting and boarding passengers all along the train. As expected, the percentage decrease when the number of passengers increases, meaning that passengers spread more when they are more. However, even for high number of passengers, the door unused time represents 60% of the alighting and boarding time at the critical door. Thus, it seems

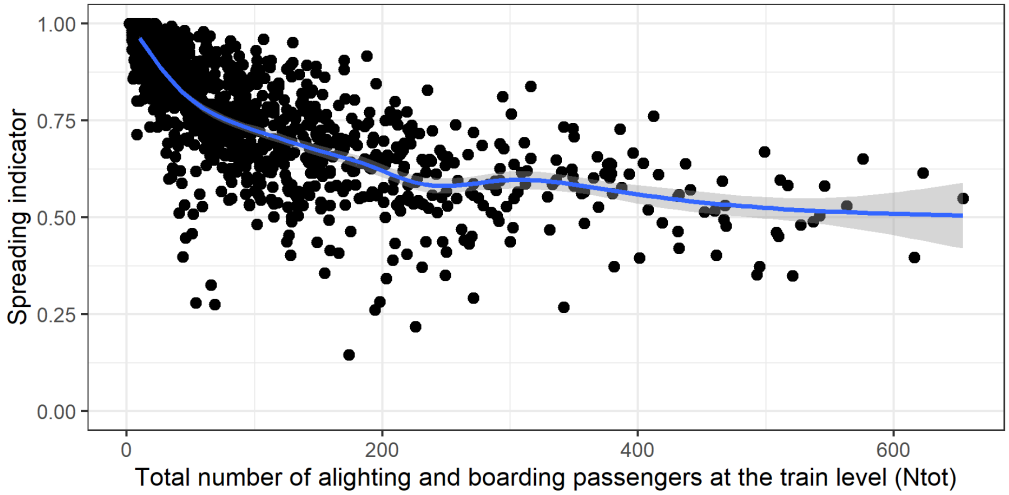


Figure 3.5: Spreading indicator in function of the total number of passengers alighting or boarding at the station on the training set. The blue line is the trend line and gray zone is the standard error of the trend line.

that even for a high number of passengers an uneven spreading still persists. To check this uneven spreading, Figure 3.6 represents the percentage of passengers at the critical door in function of the total number of passengers alighting and boarding at the station. Below 50 passengers, between 25% and 75% of passengers use the critical door. Moreover, 12% of passengers are using the critical door for number of passengers above 100. A perfect spreading of passengers would lead to a use of the critical door by 6% of the passengers. Therefore, even on a crowded platform, uneven spreading persists.

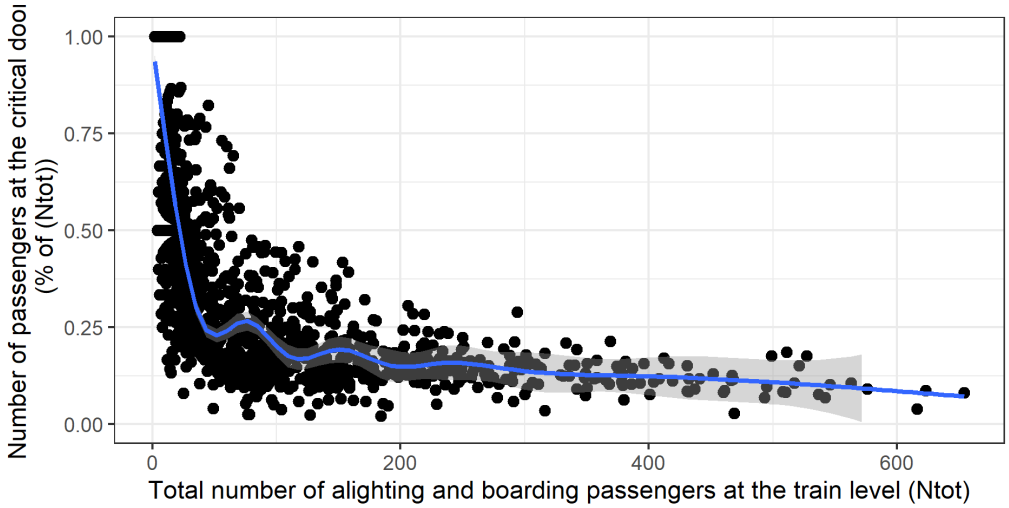


Figure 3.6: Percentage of passengers at the critical door in function of the total number of passengers alighting or boarding at the station on the training set. The blue line is the trend line and gray zone is the standard error of the trend line.

So, crowds better spread along the platform than sparse passengers but their spreading is not perfect. Thus, operators have tried to spread the exits along the

platform to encourage passengers to do so.

3.3.2 Door unused time in function of the layout

A second factor influencing the spreading of passengers on the platform is the number of exits and their location. Indeed, daily passengers know the position of the exit at their stop and might be located in the train in function of this. Therefore, if several exits exist, passengers might divide themselves between each door of the train. This is particularly important factor for crowded stations. Figure 3.7 shows the result of the door unused time in function of the layout. It shows that the passengers are effectively more spread when more exits are on the platform.

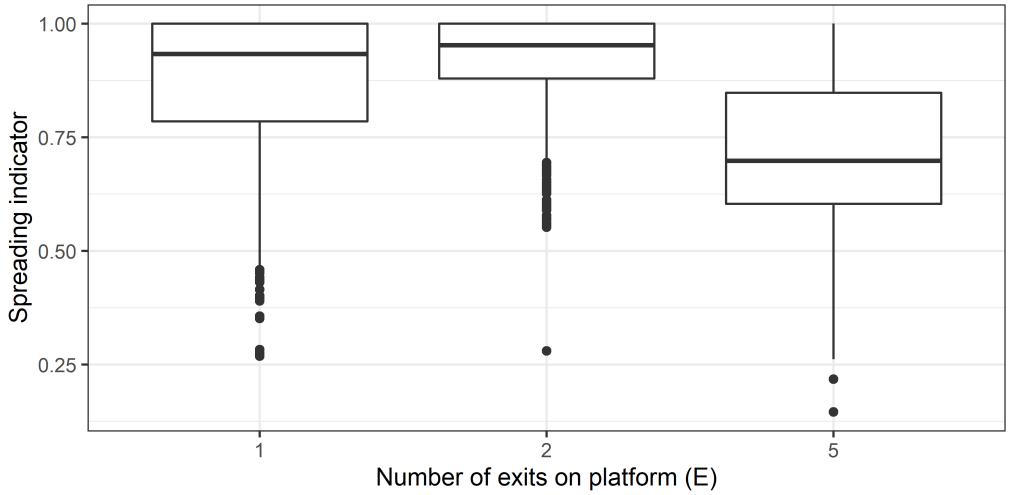


Figure 3.7: Spreading indicator in function of the number of exits on the platform, for stations of the training set

Figure 3.8: Door unused time in percentage of the alighting and boarding time at the train level, in function of the number of passengers and exits on the platform

However, when both factors are combined, it appears that the number of passengers has a greater effect than the number of exits. Figure 3.8 shows the door

unused time in function of the number of passengers on the platform and a trend line is drawn for each number of exits. The three lines have the same trend so it is hard to conclude on the effect of the number of exits on the spreading of passengers.

3.4 Buffer time and punctuality

The buffer time is defined by equation (1.3) as the time lost by the train at the station waiting for the signalling or its theoretical departure time. As buffer time has been rarely analysed in the Parisian network, an analysis of its characteristics was necessary. Firstly, statistical results will be given, then the method based on delayed trains will be compared to the one based on clusters.

3.4.1 Buffer time in function of punctuality

The buffer time is analysed in function of the punctuality of the train. A train is considered *ahead of its schedule* if it arrives at the station before its theoretical arrival time. A train is *delayed* if it arrives at the station after its theoretical departure time. Finally, a train is *on time* if it arrives at the station between its theoretical arrival and departure times.

Table 3.3 presents the mean and the median of the buffer time in function of the punctuality of the train.

Table 3.3: Buffer time in function of the punctuality of the train

	Ahead of schedule	**On time**	**Delayed**
Mean (s)	39.7	16	10.7
Median (s)	36.5	13.5	8.5

As expected, the later is the train, the shorter is the buffer time. More interestingly, for trains ahead of the schedule, the buffer time is really close to 40s, and according to Figures 3.4, 40s is also the theoretical dwell time for many of the stops. Another point of interest is that the buffer time for delayed trains is not so close to 0. So let's to evaluate the method based on delayed trains.

3.4.2 Evaluation of the method based on delayed trains

To compute alighting and boarding time, many articles [10] assume that delayed trains are leaving once all passengers have alighted or boarded. Indeed this method is used when detailed data is not available. Thus, it is assumed that there is no buffer time for delayed trains and therefore that the dwell time is only composed of the alighting and boarding time and the technical times. So, the method of delayed trains approximates the alighting and boarding time as the dwell time minus the technical time.

To evaluate this approximation, it was compared to two other methods. So we end up with three definitions of the alighting and boarding time:

- computation based only on delayed trains: $Y_{delay} = DT_{delay} - TT$;
- a method using all stops of all trains: $Y_{all} = DT - TT$;

- the method using clusters as presented in this report: $Y = \max_i Y^i$.

For all three definitions, only the alighting and boarding time at the critical door is computed. Additionally, all models include the same set of variables, namely A, B and the interaction between them ($A \times B$). Finally, the testing set is reduced to only delayed trains, so that the goodness of fit indicators are computed on the same number of data.

Then for each definition of the alighting and boarding time, the regression is trained on the training set and then tested on the testing set. The estimated alighting and boarding time, called $\hat{y}$, are compared to cluster definition y and some indicators of the goodness of fit are computed. The chosen indicators are the root mean squared error (RMSE) defined by equation 3.2, and the mean absolute error (MAE) defined by equation 3.3.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2} \tag{3.2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y| \tag{3.3}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y} - y|}{y} \tag{3.4}$$

Table 3.4: Evaluation of the definition of the alighting and boarding time based on delayed trains

Method	R^2	RMSE	MAE
Y_{all}	0.03	30	30
Y_{delay}	0.28	16	15
$Y_{cluster}$	0.40	8	7

The results are given in Table 3.4. One can observe that the values of the indicators are divided by two between each definition. Therefore selecting only delayed trains is far better than using all the trains but it is still an approximation. The clusters method is the most precise. Thus, when detailed data is available, it should be used.

4 Method

In this chapter, we focus on the stochastic model. As both buffer time and door unused time depend on the alighting and boarding time, it becomes necessary to characterize, model and estimate it.

Firstly, three different models will be compared and one of them will be selected. Then, the selected model is slightly modified to keep only influencing variables, and a split of the data set is proposed to better fit observations. Finally, margins are discussed and used to check the robustness of the timetable.

4.1 Alighting and Boarding time modelling

Regression is a simple way to describe the influence of some parameters on a phenomenon. Indeed, regression explains a dependent variable Y in function of some explanatory variables. To each variable v is associated a coefficient β_v, giving the variation observed in Y if variable v increases by 1 while all other variables are unchanged. Using regression allows then to interpret easily the effect of each variable on the variable of interest Y. In mathematical writing, with V the number of variables, β of length V, Y of length the number of observations n and X of dimensions (n,V), linear regression can be written following equation (4.1).

$$\hat{Y} = X\hat{\beta} + \varepsilon \tag{4.1}$$

4.1.1 Proposition of three different models

The most simple model is the Gaussian linear regression. Gaussian linear regressions can be interpreted easily through their coefficients. However they assumed a normally distributed error term. According to Figure 3.1, the distribution of alighting and boarding time is positive, continuous and right-skewed. So, a log-normal or gamma distribution would maybe better fit the observations.

At the end, three different models are tested and compared:

1. a linear Gaussian regression : $Y^i = \beta X^i + \epsilon$;
2. a linear log-normal regression : $\log(Y^i) = \beta X^i + \epsilon$;
3. a generalized linear model with the gamma distribution.

The two first models are Gaussian linear models. So they can be written $E(Y) = \beta X$. However, the log-normal distribution implies a transformation of the data. The estimation returned by the regression is $E(Y)$ but an error term exists and is normally distributed. The third model is a generalized linear model so it allow the error not to be Gaussian, it will be further explained in the next paragraph.

4.1.2 Insights on Generalized Linear Models (GLM)

The last proposed model is a generalised linear model used with the gamma distribution.

More generally, GLMs are composed of three parts [16] :

1. a random component: $E(Y) = \mu$
2. a systematic component: $\eta = \beta X$
3. and a link function: $\eta = g(\mu)$

Generalized linear models (GLMs) are more flexible than Gaussian linear models for two reasons. First, the distribution of the random component is not necessarily Gaussian. Secondly, the link function can be any differentiable function while in Gaussian linear models, the identity function is used. Thus, as the systematic component of GLMs is a regression, they keep high interpretability, and at the same time the random component and the link function ensure flexibility into the model. In the alighting and boarding time case, the distribution is continuous, positive and right-skewed. As those characteristics are also describing the gamma distribution, the gamma distribution is chosen for the random component. Gamma distribution can be written in function of a canonical parameter θ and a dispersion parameter ϕ following equation (4.2).

$$f(y, \theta, \phi) = \frac{1}{\Gamma(\frac{1}{\phi})} \frac{1}{y} \left(-\frac{\theta y}{\phi} \right)^{\frac{1}{\phi}} \exp \frac{\theta y}{\phi}, \quad \forall y > 0 \tag{4.2}$$

The Gamma function Γ is the function such that : $\forall x > 0, \quad \Gamma(x+1) = x\Gamma(x)$.

The canonical parameter is linked to the mean of the distribution though a function γ' such that $\mu = \gamma'(\theta) = -\frac{1}{\theta}$. The dispersion parameter is linked to the variance of the distribution, such that $Var = \phi\mu^2$ The principle of GLM can be summarized by the diagram represented on Figure 4.1 in the case of gamma distribution.

$$\eta = X\beta \underset{g^{-1}}{\overset{g}{\rightleftarrows}} \mu = E[Y] \underset{\gamma'^{-1}}{\overset{\gamma'}{\rightleftarrows}} \theta$$

Figure 4.1: Pattern of the interaction into GLMs

GLM are predicting μ using X, so the best β needs to be estimated. However, the maximum likelihood distribution is going through θ. Using the link function such that $\eta = \theta$ makes computations much easier. This particular link function is called the *canonical link*. To each distribution that might be chosen for the random component corresponds a canonical link.

The canonical link leads to a simplified derivation of the maximum likelihood estimator. Additionally, using the canonical link ensures that the sum of residual is 0.

The canonical link attributed to the gamma distribution is the reciprocal link, i.e. the function defined by : $\forall x > 0, \quad g(x) = 1/x$.

$$E[Y] = \frac{1}{\beta X} \tag{4.3}$$

So finally, for the estimation of the alighting boarding time, the Gamma distribution is chosen and equation 4.3 stands.

4.1.3 Criteria to choose among models

The best estimation of the alighting and boarding time is sought. All proposed models will then be trained on the training set and ranked on the testing set. So, some metric are selected to choose the best model among all. Four indicators have been selected.

- Root mean squared error (RMSE)
- Mean absolute error (MAE)
- Mean absolute percentage error (MAPE)
- Kullback-Leibler divergence (KL-distance)

The three first indicators are usual indicators for goodness of fit of linear regressions. However, they assume that the error of the distribution is Gaussian and that the estimation of interest is the mean of the distribution.

In the alighting and boarding time case, the distribution of the error might not be Gaussian. Secondly, as the study is turned around margins, the quantiles are the object of interest. Therefore, another goodness of fit metric is needed. The Kullback-Leibler divergence is chosen. Indeed, this measure was introduced in the probability theory context. It measures the information lost when a distribution Y is approximated by a distribution $\hat{Y}$. The general formula to compute this distance is the following:

$$D_{KL}(Y||\hat{Y}) = \sum_i Y^i \log\left(\frac{Y^i}{\hat{Y}^i}\right) \tag{4.4}$$

Such a criteria compares the distributions between them. We will be able to compare the distributions on the part that interest us the most, i.e. between their quantiles 0.9 and 1.

4.2 Improvements of the model

After having selected the best model, some refinements can be applied to make simpler and closer to the observations.

4.2.1 Variables selection

First of all, a selection of variables would keep only influencing variables. The model includes 6 continuous variables and their 3 transformations (log, square, squared root), so 18 different possibles variables, one categorical variable and

one discrete variables. Secondly, all interactions 2 by 2 are considered. At the end, all variables, their transformations and their interaction are at first included into the first model, so around 35 terms. Thus, hundreds of models can be drawn from all possible combinations of these terms. To select the best model among those, an algorithm of automatic selection is used. Such an algorithm considers all variables and return the best sub-set of variables for one Gaussian linear model.

Among algorithms of automatic selection, some begin from the null regression, in other words a regression containing only a constant, and then add only variables improving the model according to a chosen criterion. Those algorithms are said to run forward. Other algorithms begin from the full regression, i.e. the regression containing all variables, then remove variables to improve the model according to a chosen criterion. Those algorithms run backward. Finally, stepwise algorithms run both backward and forward.

In this report, we choose a stepwise algorithm to select the best combination of variables based on the Bayesian information criterion (BIC). The BIC definition is given in equation 4.5 (with V the number of β parameters into the regression, n the number of observations, and $L(\beta|Y)$ the likelihood). It has been chosen as it is more influenced by the number of parameters than the Aikake criterion. Moreover, the R^2 increases when adding variables, so it becomes better with the complexity of the model and favours more complex models. On the contrary, the BIC criterion takes into account the number of parameters, so favour simple models.

$$BIC = V \log(n) - 2 \log(L(\beta|Y)) \tag{4.5}$$

The stepwise algorithm is a systematic search algorithm running in both directions, forward and backward. A pseudo-code written in Algorithm 1 explains the different steps of the selection. Starting from a first full regression including all variables, any variable that does not provide an improvement in the model fit is removed. So we obtain a set of variables. Then, at each step, all variables are considered. If a variable is already selected, the algorithm computes the BIC of the regression without this variable. On the contrary, if a variable is not part of the regression, the algorithm computes the BIC of the regression including this variable. Finally, all BICs are compared and the minimum one is chosen. Thus at each step, one variable among all is either added or removed from the selection.

As a result, this algorithm returns around ten variables that are the most contributing to the model.

4.2.2 Split of the dataset

The second improvement of the model concerns the number of modes. Indeed, according to Figure 3.1, the distribution of the alighting and boarding time has two different modes: one around 2 seconds and one around 12 seconds. So it seems that two behaviours can be observed when it comes to alight or board a train. To translate this particularity, mixed distribution can be used, training two different models and adding them at the end.

Data: Training set
Initialization: Full regression composed of all variables (set X_0);
Delete variables that don't change the BIC of the regression;
X = set of variables included in the regression;
$\tilde{X}$ = set of variables not included in R;
$X_{selected}$ = an empty set representing the null regression;
while X *!=* $X_{selected}$ **do**
    X = $X_{selected}$ #except during the first iteration;
    Update $\tilde{X}$;
    B = {BIC of the regression including the variables in X} #set of all BIC;
    A = {X} #set of combinations of variables;
    forall *variables* v *in* X **do**
        $X' = X \setminus v$;
        Append (BIC of regression including variables of the set X') in B;
        Append (X') in A;
    end
    forall *variables* $\tilde{v}$ *in* $\tilde{X}$ **do**
        $X' = X \cup \{\tilde{v}\}$;
        Append (BIC of regression including variables of the set X') in B;
        Append (X') in A;
    end
    $X_{selected}$ = A[argmin(B)] #selection of the sub-set corresponding to the minimum BIC;
end
Return X

Algorithm 1: Pseudo-code of the Stepwise algorithm

So the data set is split into two smaller sets according to the number of passengers N. It appears that the distribution of the alighting and boarding time for small number of passengers differs from the distribution of alighting and boarding time for a higher number of passengers. This split is represented on Figures 3.2.

To determine the value of N around which splitting the data, the following method is used.

1. N_{split} varies between 2 and 15.
2. The testing set is filtered such that N > 15. Indeed, it is the condition that constraints the most the data. So to compare values on the same number of observations, the smaller date set is chosen, corresponding to the more constrained.
3. For each N_{split}, the training set is filtered such that N > N_{split}; then the selected model is trained on this filtered training set; next estimations are computed on the testing set; finally, Kullback-Leibler distance is computed to compare the estimations to the observations on the testing set.
4. The values of the Kullback-Leibler distance are plotted and compared. The chosen value corresponds to the elbow of the curve.

At the end of this process, we obtain a number N_{split} of passengers around which splitting the data set into two sets. Those two different sets have different parameters of the same distribution. So, two regressions will be trained : one for door with a small number of passengers and one for doors with a high number of passengers. At the end, we obtain an addition of two models.

Using mixed distributions in the model translates both behaviours. So it makes the model more precise and accurate.

4.3 From observed margins to estimated margins

In order to build a scheduled transport planning, the estimation of the margins are important as they ensure its robustness. Observed margins, including buffer time and door unused time, are computed in function of the alighting and boarding time. So estimation of margins is done through the estimation of the alighting and boarding time.

However, contrary to the analysis on observed margins, the estimation of margins will mainly focus on the buffer time. Indeed, buffer time can be planned to avoid to many delayed trains. This corresponds to the chosen margin level : the higher the margin level, the higher the buffer time might appear in reality, the less delayed are trains.

4.3.1 Margins computation

To compute margins, we define a time Y_α for which N passengers have exchanged at an α confidence level. So, time Y_α is defined as :

$$P(Y < Y_\alpha | N) = \alpha \tag{4.6}$$

Different α correspond to different scenarios, they are chosen from $\{50, 90, 95, 99\}$. A higher α leads to a careful timetable leaving few risk for passengers not to have the time to alight or board. So a higher α leads to a longer planned alighting and boarding time, so to less time savings when computing margins.

From those margins, two analysis can be conducted: time saved compared to the current schedule, testing the robustness of the current time table in some particular cases.

4.3.2 Computation of time saving using margins

Time savings are the comparison between the estimated alighting and boarding time through the model and the current theoretical alighting and boarding time. The theoretical alighting and boarding time is the theoretical dwell time to which technical times are subtracted. For this section, as we study full train runs, the alighting and boarding time of stops part of the ”Outliers” data set are estimated, and stops part of the ”Nobody” data set are estimated to have 0 passengers. All those stops are added to the testing data set to have as many complete train runs as possible.

Two analysis are conducted and for both of them all variables of the model are fixed, only margin levels are varying. On the first hand, time savings at the train

level are computed. To do so, a train run is pointed as a train number k and a day d. So we can check if all runs have all their stations. Then, time savings at each station of the run are computed for different level of margins and then summed up. On the other hand, time savings at a specific station are also analysed. Similarly as for the analysis on trains trips, time savings are computed in seconds for different margin levels. This analysis however focuses more on the variation of time savings in function of the time of the day and in function of the chosen margin level.

4.3.3 Robustness of the timetable

To reduce the number of delayed trains, it is important to check that the theoretical dwell times holds out to some variation of the demand. To evaluate the robustness of the timetable, the margin level 90% is selected. Then variables from the model are varying to test some experimental cases. In the model, the variables are linked either to the number of passengers or to the gap between the train and the platform. As the gap between the train and the platform is a fixed number, only the number of passengers, influences on the alighting and boarding speed.

1. First, to compute the robustness of the timetable, the maximum number of passengers that can alight or board is estimated according to the current theoretical dwell time.
2. Secondly, to evaluate the effect of specifying one alighting door and one boarding door, we compute the maximum number of passengers that can alight or board if we designate one door only for alighting, another door only for boarding.
3. Thirdly, the spreading of passengers can also speed up the alighting and boarding time. So, the maximum number of passengers that can alight or board if passengers are perfectly spread along the platform is estimated.

To compute the maximum number of passengers able to alight or board, in the first analysis, the following method is used.

1. The mean R of the ratio alighters over the total number of passengers is computed at the station level according to equation (4.7).
2. The number of passengers alighting at a specific door is varying from 0 to 50.
3. So the number of boarders is computed as $\frac{1-R}{R} \times A$.
4. The alighting and boarding time is estimated for all number of passengers
5. They are compared to the theoretical alighting and boarding time
6. The minimum number of passengers for which the estimated alighting and boarding time is above the theoretical one is noted m. m is also the maximum number of passengers that can alight or board given a theoretical alighting and boarding time.

$$R = \frac{A}{A+B} \tag{4.7}$$

The second analysis follows the same method as the first one. The ration R is then set to 1 and both A and B are varying from 0 to 50. The obtained percentages are compared to ones computed in the first analysis. Moreover, the number of additional passengers being able to alight or board can be observed. Similarly, this can be translated into time savings.

Coming to the last analysis, the uneven spreading of passengers along the platform has been observed in Section 3.3. The conclusion was that when there are between 100 and 400 passengers alighting or boarding at the station, the door unused time represents between 50 and 60% of the train alighting and boarding time and that 12% of passengers are using the critical door. So let's compare the alighting and boarding time (1) when passengers are perfectly spread along the platform and (2) when the critical door is used by 12% of passengers. To do so, we assume that the number of alighters A is equal to the number of boarders B. Then let's make N varying from 100 to 400, so A and B will both vary from 50 to 200. When passengers are perfectly spread along the platform, each door is used by $N/16$ passengers. Otherwise, the critical door is used by $0.12N$ and the other doors by $0.88N/15 = 0.06N$. Then, the alighting and boarding time in both cases are estimated and compared.

5 Results

In this chapter, all the results of the computations described in Chapter 4 are presented and discussed. Beginning from the estimation of the alighting and boarding time, we will then present the results on the margins. To conclude, the model is performed on line R, the estimation is then compared to the observations so we can validate the model on another line.

5.1 Estimation of the alighting and boarding time

The estimation of the alighting and boarding time is the main point of this study. First the selection of variables is presented, then the three models are compared, finally the results on the split of the data set are given.

5.1.1 Selection of the variables

Considering all variables, their transformations and interactions, around 40 variables were available. So, a selection of variables have been performed on the training set through a stepwise algorithm assuming a linear Gaussian regression. Table 5.1 presents the final selection of variables for the model. It also gives their coefficient and significancy.

Table 5.1: Selection of variables made through a stepwise algorithm assuming a linear Gaussian regression

Variable	Estimated coefficient	p-value (t-test)
Intercept	4.63	-
A	0.73	<1e-06
B	-0.42	<1e-06
H	-0.16	<1e-06
log(A)	0.53	<1e-06
log(B)	1.58	<1e-06
A²	-0.0040	<1e-06
B²	-0.0010	0.25
AxB	-0.0088	<1e-06
BxH	0.075	<1e-06

From Table 5.1, only the coefficient of B² is significant at a 90% level, all other coefficients are different from 0 with a degree of confidence of 99%. Secondly, one can analyse the sign of the coefficients.

- A, $\log(A)$ and $\log(B)$ have positive coefficients as the higher they are, the longer is the alighting and boarding time.
- A^2, B^2 and $A \times B$ have negative coefficients. Indeed above a threshold, if the number of passengers increases, it will slow down passengers as less space will be available for them. Through this result, we find again the fundamental diagram of the speed in function of density.

- $B \times H$ positive coefficient leads to the conclusion that horizontal gaps slow down boarders (compared to alighters).
- B has a not intuitive negative coefficient. However, it might be due to the presence of $B \times H$. As H is around 10 cm, the addition of the coefficients of B and $B \times H$ results being positive.
- H has weirdly negative coefficient.

The selection of variable considerably reduces the number of variables in the final model. However, different variables could have been selected for each of the three proposed models. Indeed, variables would have been specific to the model and better translate its behaviour. To conclude, the variables used in the following analysis are A, B, their transformations $\log(A)$, $\log(B)$, A^2, B^2, their interaction $A \times B$, and H with its interaction with B, $H \times B$.

5.1.2 Comparison between the three distributions

Based on these variables, three different statistical models were trained on the training set and tested on the testing set : a Gaussian model, a log-normal model and a Gamma model. As presented in section 4.1, different goodness-of-fit indicators are chosen, not only RMSE, MAE and MAPE but also the Kullback-Leibler distance. Indeed the Kullback-Leibler distance indicates how similar two distributions are.

Table 5.2: Comparison between the three statistical models through mean indicators on the testing set

Model	RMSE (s)	MAE (s)	MAPE
Gaussian	5.93	3.68	0.47
Log-normal	6.54	4.04	0.43
Gamma	6.67	4.19	0.59

From Table 5.2, the log-normal model is the one giving the closest mean estimations to observations. However, when computing margins, quantiles between 0.9 and 0.99 are used in the estimation to be sure that in more than 90% of the stops, the estimated alighting and boarding is long enough for the number of passengers. So to choose between models, mean indicators must be completed by an indicator on the shape of distributions. The Kullback-Leibler divergence from observations is thus computed on quantile intervals of length 0.1 for each of the estimation distribution. Table 5.3 presents this distance for all three models and each interval.

Therefore, the most interesting model is the closest one to observations on the [0.9 - 1] interval. This model is the Gamma model.

This result can be visualised on two graphs. Figure 5.1 is a quantile-quantile plot to understand how far are the estimated quantile from the observed one. Estimations on the testing set with all three models are represented and the Gamma model is the closest from the observations for high quantiles. Indeed on the quantile-quantile plot, the last quantiles of the gamma distribution are closer to the identity line.

Table 5.3: Kullback-Leibler distance for estimations of all three models compared to observations for different intervals of quantiles

Interval of α	Observed quantile q_α (s)	D_{KL} Gaussian	Log-normal	Gamma
0 - 0.1	1.0	8.28	1.78	6.71
0.1 - 0.2	3.0	0.50	0.21	0.037
0.2 - 0.3	5.7	1.88e-02	1.07e-02	1.02
0.3 - 0.4	9.0	1.27e-03	1.28e-03	2.00e-03
0.4 - 0.5	10.5	1.15e-03	4.87e-04	1.09e-03
0.5 - 0.6	11.8	1.04e-03	2.74e-04	1.09e-03
0.6 - 0.7	12.9	5.49e-04	8.82e-04	6.90e-04
0.7 - 0.8	14.1	1.71e-04	9.15e-05	7.79e-05
0.8 - 0.9	17.5	8.12e-05	3.80e-05	9.86e-05
0.9 - 1	22.5	2.52e-05	5.03e-05	1.64e-05
0 - 1	-	3.06e-02	1.00e-02	5.76e-02

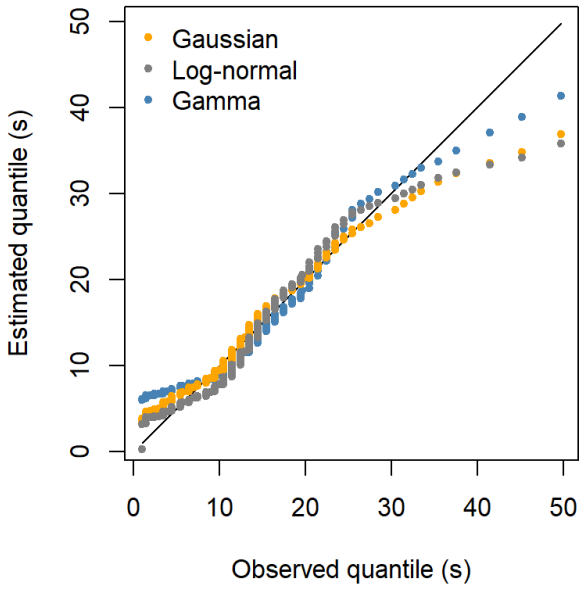


Figure 5.1: Quantile-quantile plot with estimations made on the testing set

Figure 5.2 represents the cumulative distributions of the observations and of estimations from all three models. For the observed cumulative distribution, the value of a time T is the number of observations lower than T, divided by the total number of observations. Again the Gamma model is the closest one from the observation in the interval [0.9, 1].

As a result, even if the Gamma distribution does not perform well on the mean indicators, it estimates well high quantiles. As high quantiles will be used for computation of margins, the Gamma distribution is chosen. A last improvement of the

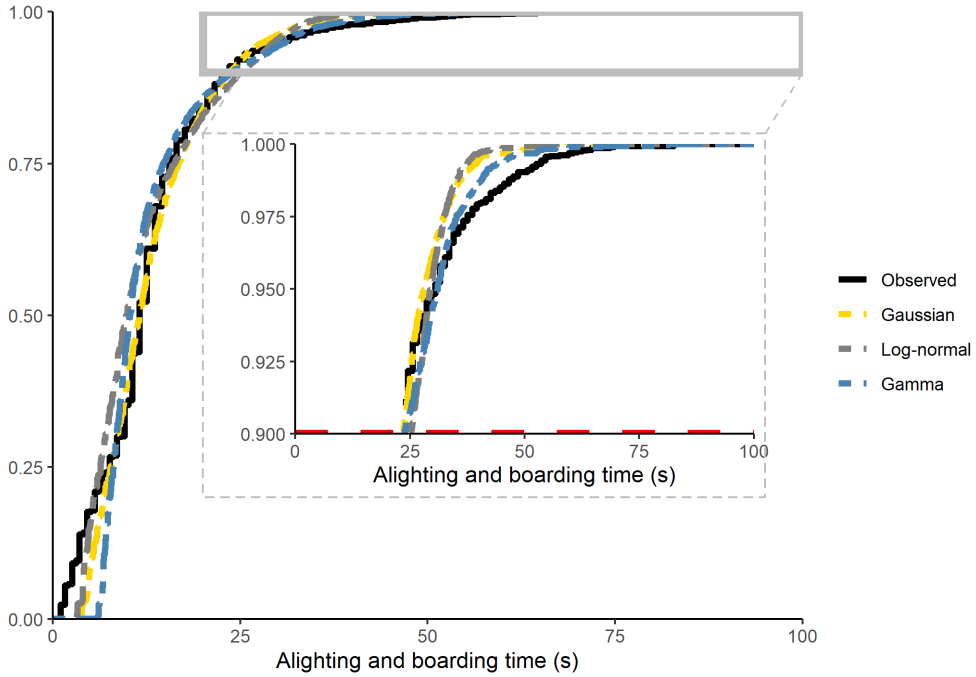


Figure 5.2: Comparison of cumulative distributions of the observations with the one of the models estimations

model is tried to better fit observations.

5.1.3 Two behaviours in function of the number of passengers

Figure 3.2 suggests two behaviours in the alighting and boarding time. It seems that doors used by a high number of passengers differ from doors used by few passengers. To study those two behaviours, we first look for the number of passengers around which the behaviour changes. The selected number is the one minimizing the KL-distance between estimated and observed data. According to Figure 5.3, $N = 4$ is chosen.

The minimum is $N = 4$. Choosing $N = 4$ means that a model is fit for all doors used by less than 4 passengers and that another model is fit for doors used by more than 4 passengers. The combination of those two models is the mixed model.

Then, Mixed Gamma model is compared to the Gamma model through different indicators. Table 5.4 gives RMSE, MAE, MAPE and Kullback distance for both cases. All three indicators are smaller for the mixed model. It seems to perform better : the mixed model is closer to the observations than the simple model.

The Mixed Gamma model better predicts the alighting and boarding time. It will thus be preferred to the Gamma distribution. This result confirms that when there are less than 4 passengers at a door, then the alighting and boarding time is close to an exponential distribution, so this process is close to a queuing process. When passengers are more than 4 at a door, then the alighting and boarding time is close to a gamma distribution.

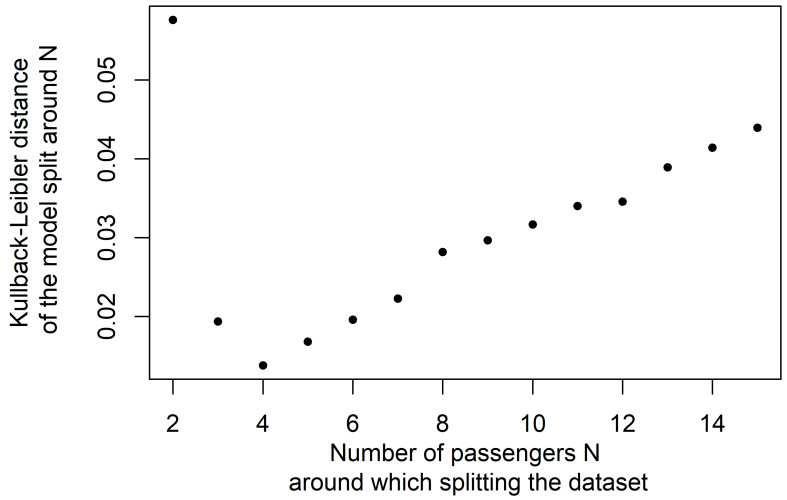


Figure 5.3: KL-distance between estimated and observed distributions for different splitting of the data set. The estimation is done through the Gamma distribution

Table 5.4: Comparison between the Mixed Gamma model and Gamma model through indicators computed on estimations made on the testing set

	RMSE (s)	MAE (s)	MAPE	KL-distance
Gamma model	6.67	4.19	0.59	5.76e-02
Mixed Gamma model	6.23	3.77	0.47	1.38e-02

Conclusion of all models To conclude, the performance of all models are summed up in Table 5.5. Even if Gaussian and Log-normal models perform better globally, the Gamma model is preferred for its distribution close to observations for high quantiles. Moreover, splitting the data set improves the estimation. So, the mixed gamma model is chosen for estimations in next Sections because it performs as well as Gaussian model globally and has quantiles close to observed ones.

Table 5.5: Conclusion on all tested models

Model	RMSE (s)	MAE (s)	MAPE	KL-distance
Gaussian	5.93	3.68	0.47	3.06e-02
Log-normal	6.54	4.04	0.43	1.00e-02
Gamma	6.67	4.19	0.59	5.76e-02
Mixed Gamma	6.23	3.77	0.47	1.38e-02

5.2 Estimation of margins

Once the alighting and boarding time correctly estimated, we are interested in the advantages of taking into account the ridership into the theoretical timetable. Also, the robustness of the timetable to special situations can be evaluate.

5.2.1 Possible time savings

The Mixed Gamma model returns an estimated alighting and boarding time. These estimations is compared to the theoretical alighting and boarding time and time

savings can be deduced as the difference between them. Time savings are linked with the alighting and boarding time at the train level, and depends on the quantile chosen. Indeed, the alighting and boarding time is estimated through quantiles based on the Mixed Gamma model. The quantile chosen corresponds to a margin level, equivalent of the buffer time. The higher the margin level chosen, the more cautious it is about the time needed for people to alight or board. Then the estimated alighting and boarding time will be higher. So time savings will be less important. If time savings are negative, it means that the theoretical alighting and boarding time is not long enough for this stop. Time savings are computed at the train level at each stop and then aggregated either for a train run or for a station.

In this Section, the station and the number of passengers are fixed and the margin is varying.

Time savings on train runs A full train run lasts 70 minutes. For computations on train runs, the more complete are train runs, the better. So data sets ”Outliers” and ”Nobody” are aggregated to the testing set for computations of time savings. Alighting and boarding time of stops from the ”Outliers” set are estimated through the Mixed Gamma model. The stops from the ”Nobody” set are considered to have a null alighting and boarding time. Table 5.6 shows the mean time savings for train runs for different margin levels. It also gives the percentage of train runs with a positive cumulative time savings. Indeed, for some train runs, theoretical alighting and boarding time is too short compared to the estimated one. So no time is saved but the timetable should be adapted to take into account the ridership. For instance, with a margin level at 90%, in average 155 seconds are saved on a train run, but for 2.1% of the rides, estimated alighting and boarding time is longer than the theoretical one.

Table 5.6: Time savings in second for different margin levels for train runs

Margin level (α)	**Mean (s)**	**Trains with positive time savings (%)**
50%	217	100%
90%	155	97.9%
95%	132	96.8%
99%	84	76.6%

Surprisingly even for the 99% margin level, the average time savings on a train run is of 84 seconds. Reported to the 277 trains running on line N per day, this time saving represents 6h30 of train runs. This time could be saved on the number of drivers or train sets needed to operate the line. However, this results hides some variations through stations and times of the day.

Time savings at stations To better understand station specificities, the same study is done at the station level and through the time of the day. Then, critical hours are identified at one specific station for one direction. As said in Section 3.2, alighting and boarding time is longer that the theoretical one mainly in two stations : Versailles-Chantiers and Houdan. So analysis have been done on

Versailles-Chantiers and Houdan to understand particularly rush hours. Graphs 5.4 represent the time savings in function of the chosen margin level and the hour.

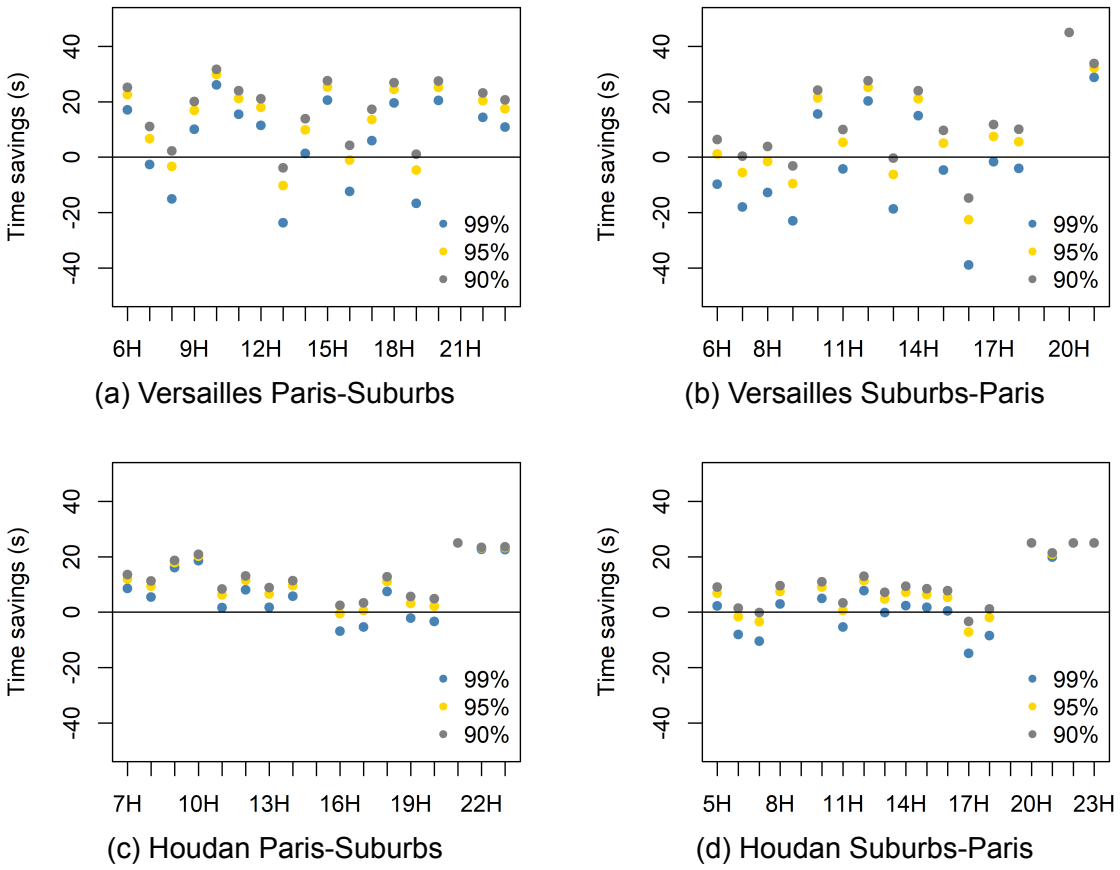


Figure 5.4: Time savings at Houdan and Versailles in function of the hour of the day and direction

From Figures 5.4, rush hours are particularly critical at Versailles-Chantiers station. For a 99% margin level, 20 seconds are missing to the theoretical alighting and boarding time. At Houdan station, the difference between the scheduled and the estimated alighting and boarding time is lower. The timetable is better constructed.

The main deficiency of this analysis is the number of available data for each hour at each station. For instance, the result observed for Versailles in direction Suburbs-Paris (SP) at 1pm is based on only one estimation. Conclusions from this analysis would be more reliable with more data.

Such an analysis could lead to a timetable building adapted to passengers flows However, when reducing the theoretical dwell time, a risk of losing passengers appears.

5.2.2 Simulation to evaluate the robustness of the current timetable

To check whether the timetable holds out specific situations, alighting and boarding time is estimated in different scenarios. To test the robustness of the schedule,

A and B are varying and the alighting and boarding time is estimated. This way the maximum number of passengers able to alight and board during a stop can be estimated and compared for different scenarios.

In this Section, the margin is fixed to 95% and the number of passengers is varying.

Specifying doors for exclusively boarding or alighting To see the effect of specifying doors exclusively for boarding or alighting, different scenarios are computed. Scenario 1 (S1) approximates observations as the observed mean of the proportion between alighters and boarders at a given station in a given direction. This scenario is close to observations. Then, three other scenarios are chosen: if there are only alighters using the door (S2), if only boarders are using the door (S3) and if the number of alighters is the same as the number of boarders, in other word, if the ration is equal to 0.5 (S4).

Table 5.7: Maximum number of passengers that can alight or board for different scenarios with a 95% margin level

Station	**Direction**	N_{max}			
		S1	**S2**	**S3**	**S4**
Garancières-la-Queue	PS	14	18	14	14
	SP	16	19	15	14
Houdan	PS	16	19	16	14
	SP	16	19	15	14
Marchezais-Broué	PS	16	20	20	16
	SP	16	21	23	16
Orgerus-Béhoust	PS	15	19	16	14
	SP	15	19	16	14
Plaisir-Grignon	PS	27	37	31	28
	SP	29	37	32	28
Tacoignières-Richebourg	PS	16	19	16	14
	SP	15	19	15	14
Versailles-Chantiers	PS	28	37	32	28
	SP	25	36	27	24
Villiers-Neauphle-Pontchartrain	PS	16	19	15	14
	SP	17	19	15	14

In Table 5.7, it seems that between four and ten more passengers can alight or board if a door is specified for alighting. The effect is less pronounced for boarding doors. Moreover, the number of passengers able to alight or board through a door seems quite small. The 95% margin level is really cautious but the model might overestimate the time taken by one passenger.

To conclude, the gain depends on the station but globally, specifying a boarding door and an alighting door speed up the transfer. This is quite easy in Regio2N trains, as doors are grouped by two along the train.

Spreading of passengers The final way to reduce the dwell time at station is to spread passengers on the platform. Two scenarios are compared. As it is observed on crowded days, Scenario 1 (S1) supposes that 12% of passengers are using the critical door. Scenario 2 (S2) assumes that passengers are evenly spread along the platform, so that 6% of passengers are situated at the critical door. The study is done computing the alighting and boarding time for a number of passengers going from 100 to 400 at the train level, so between 6 and 48 passengers at the critical door. Figure 5.5 presents boxplots of the estimated alighting and boarding time.

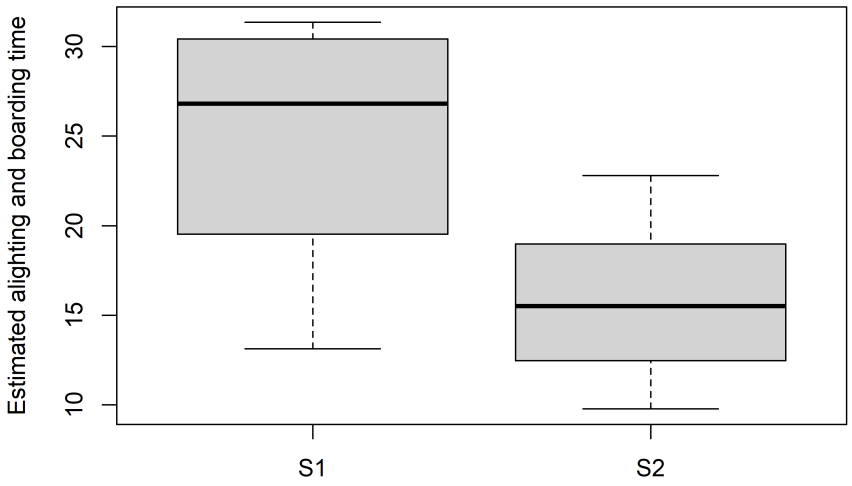


Figure 5.5: Boxplots of the estimated alighting and boarding time when 12% of passengers are using the critical door and when passengers are evenly spread along the platform

The median of the scenario with an uneven spreading is around 27s while the median of the alighting and boarding time with an even spreading is around 15s. Spreading passengers really reduce the alighting and boarding time. Moreover, one can compute the time saved in percentage of the alighting and boarding time with an uneven spreading. Results say that spreading passengers leads to time savings between 22% and 42% of the alighting and boarding time. This corresponds to a time saving between 3s and 12s at a given stop.

The spreading of

Spreading passengers along the platform is therefore one efficient method to reduce the alighting and boarding time on the train run.

In conclusion, the alighting and boarding time modelling enables to computes some time savings for different scenarios and to propose some clue to reduce the duration of a train run.

5.3 Validation of the model

Finally, Transilien is interested to use this model for other lines on the network. Therefore, it is important to evaluate how well the model can be transferred to another line. Thus, data from line R are collected and processed. This new data is preprocessed as data from line N, a clusters data set is built and will be considered as a new testing set. However, data of horizontal gaps at station is not available on line R. So the Mixed Gamma model is trained on the line N training set with passengers related variables. In other words, the Mixed Gamma model is fitting on line N without horizontal gap variable to be transferred to other lines. We checked that this slight modification has not a too great impact on the results. The results are really close and the model can be modified to widen its use.

Line R characteristics The main characteristic of line R is that the same rolling stock, Regio2N trains, as on line N is running. Regio2N trains have been running on line R since the beginning of 2021. So, data from January to July 2021 are available on all branches of the line.

Line R serves the southern suburbs of Paris. Its first station after having left Paris is Melun, a town 42km far from Paris, and the last one is Montargis situated outside the Ile-de-France administrative region. A route map of the line is provided on Figure 5.6. Trains leave Paris-Gare de Lyon every 30min in the morning peak hours and every 20 min in the evening peak hours (16h – 20h). During non-peak hours, only one train per hour is running in both directions.

Preprocessing on the dataset The same preprocessing is applied to the new data set. Therefore, the clusters are computed and we obtain a data set containing 7,361 stops, 23,735 doors and 26,185 clusters of passengers. 10,479 or one third of those clusters occur at Melun station. This station is thus the main node of the line. The second main station is Fontainebleau-Avon : 5,803 clusters happen at that station.

Few statistics of the number of alighting passengers, boarding passengers and the alighting and boarding time are given in Tables 5.8 for those two specific stations.

Table 5.8: Descriptive statistics of the number of passengers and the alighting and boarding time on the two main stations of line R

(a) Melun
1703 stops

	A	B	Y (s)
Mean	49	48	27
Median	27	21	23
Standard deviation	62	56	18

(b) Fontainebleau-Avon
1481 stops

	A	B	Y (s)
Mean	18	17	15
Median	8	7	12
Standard deviation	25	25	9

From Tables 5.8, it appears that even if Melun and Fontainebleau-Avon stations are the two main stations of the line, not many passengers are transferring at those stations. The number of passengers transferring at Melun is twice as the one at

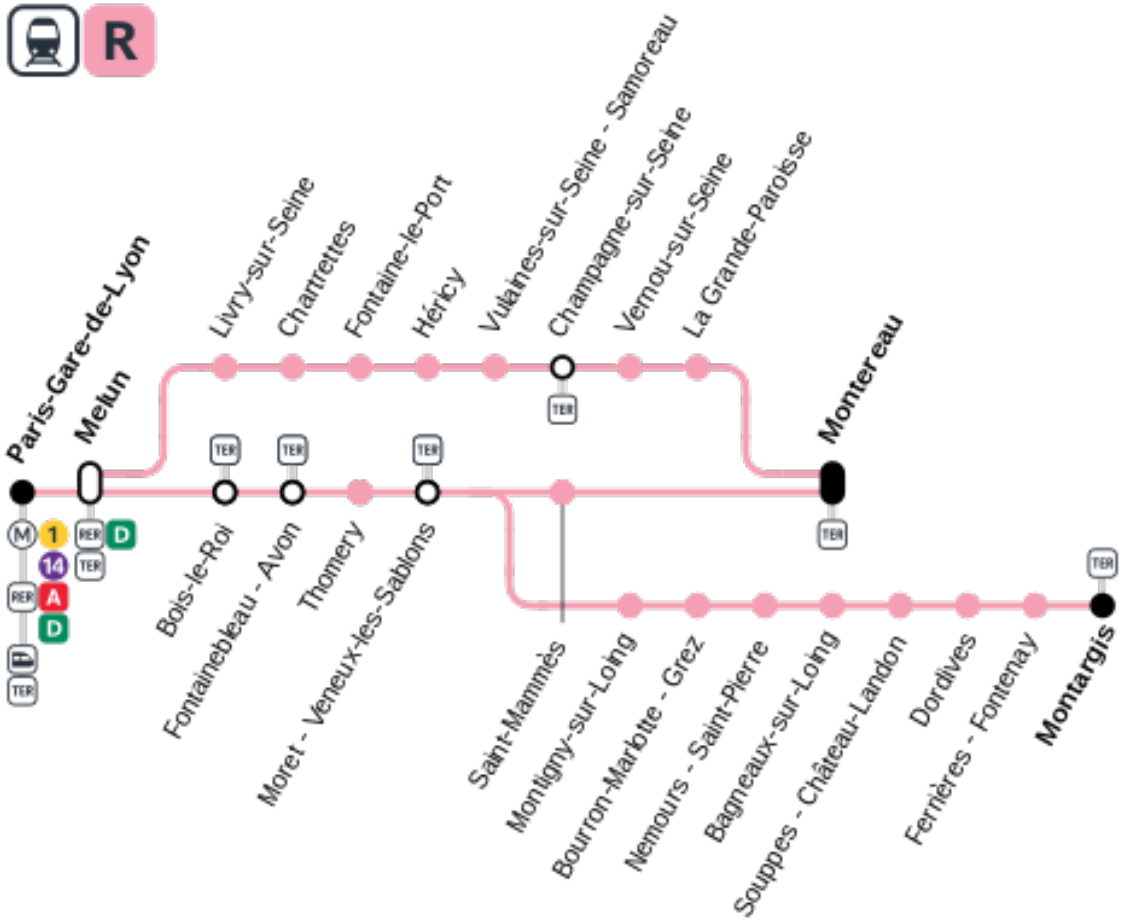


Figure 5.6: Route map of line R. Transilien ®

Fontainebleau-Avon, leading to around 20 passengers boarding at Fontainebleau-Avon along the whole platform. The alighting and boarding time is also twice lower at Fontainebleau-Avon than at Melun station.

Modelling and conclusion on the validation Then, the mixed gamma model is used to estimate the alighting and boarding time for all the doors part of the data set. This estimated alighting and boarding time is then compared to the observed one.

Table 5.9: Results of the modelling of the alighting and boarding time on all data from line R compared to the results on the testing set

Line	**RMSE** (s)	**MAE** (s)	**MAPE**
N	6.50	3.86	0.47
R	7.27	3.87	0.62

Table 5.9 gives the results of the estimation on line R and compares them to the results of the model on the testing set of line N. The RMSE is longer of one second for line R than for line N and the MAPE of line R is 0.15 above the MAPE of line

N. So results for line R are slightly worse than those for line N but the difference is acceptable. Overall, the model can be transferred and used on line R.

Time savings are thus computed for line R. Full train runs last between 1h and 1h40 in function of the served branch. Table 5.10 gives the time savings in function of the margin. Compared to line N, time savings are smaller showing that transport timetable is better built on line R. However for a 95% margin level, 88 seconds can be saved on a full run and nearly all trains save time.

Table 5.10: Time savings in second for different margin levels for R train runs

Margin level (α)	**Mean (s)**	**Trains with positive time savings (%)**
90%	108	99.8%
95%	88	98.9%
99%	43	89.1%

The study on data from line R is made on the same data than those from line N. It permits to validate the model and to compute some results on another line.

6 Conclusion

The greater density of Paris in number of inhabitants, employment and cultural activities implies more passengers flows across the city and its suburbs. The transport network is therefore adapted and spread, especially the public transport network. A dense network has to have a high frequency and short headways with no delays. Obviously those goals are ambitious and research is helping the operators to achieve their objectives.

This master thesis aims to give some clues about modelling alighting and boarding time. As it was the first time similar data are recorded and studied, the descriptive statistics already gave results about the door unused time and the buffer time.

Global conclusions Firstly, it has been confirmed that delayed trains have small buffer time. However, buffer times of delayed trains are not null but around 10 seconds. This means that contrary to the assumption of some studies, the delayed trains do not leave the station right after the end of the alighting and boarding time. So the method computing the alighting and boarding time based on delayed trains have been evaluated. It appears that the method based on clusters is much more precise that the method based on delayed trains. Thus when the data are available, it is preferred to use the clusters method. When the data are not available, the method based on delayed trains is still an acceptable method.

Secondly, the alighting and boarding time has been estimated through four different models on a fixed data set. We concluded that the Mixed Gamma distribution is the most similar to the observation distribution. The Gamma model is a generalised linear model using the gamma distribution and the reciprocal link function. The Mixed Gamma model is composed of two Gamma models translating two behaviours in the alighting and boarding time. Therefore, if less than 4 passengers are using the door at a stop, the alighting and boarding time follows an exponential distribution, close to the queuing theory. If more than 4 passengers are using the door, then the distribution is continuous, positive and right-skewed. The model has been finally transferred to another line of the network. This validation leads to comparable results for lines N and R on the estimation of alighting and boarding time.

Thirdly, some levers where given to reduce the dwell time of trains. A study of the estimated buffer time shows that around one minute and a half can be saved on a 70 minutes train run. On a whole day on line N, this represents 6h30 of train runs. A similar analysis was performed at the station level to identify critical and empty hours and give some clues to adapt the timetable. Moreover, we showed that if doors are specified as alighting or boarding doors, between 0.4 and 0.6 times more passengers can alight or board. Finally, the analysis on the spreading of passengers along the platform leads to an alighting and boarding time divided by two. The statistics on door unused time confirm that crowds spread naturally

along the platform. It seems that the number of passengers impact more their spreading than the number of exits.

Recommendations Therefore, few recommendations are given to Transilien operator. First, when developing models estimating alighting and boarding time and when data is available, the cluster method is more accurate than the method based on delayed trains. Secondly, in the current time table, the theoretical dwell times depends mainly on the station. We propose to adapt these theoretical dwell time to the hour as well, especially in crowded stations. This adaptation can lead to a gain of 6h30 of train run on the whole line per day. Finally, as Regio2N train layout groups doors by two on specific car, proposing to specified one door for boarding and one door for alighting will speed up the passengers flow. However such a proposition will need to manage passengers' flow inside the train.

Future work The novelty of the data can feed many new studies. The definition of the alighting and boarding time can especially be further analysed, understanding clusters and gaps between them. The particularity of this data could also lead to a deeper analysis of the passengers' behaviour during the alighting and boarding time. For instance, one could ask if alighting always occur before boarding or if sometimes passengers begin to board before the end of the alighting. Also we confirmed that the spreading of passengers along the platform is a key change to reduce the alighting and boarding time. A further analysis on the factors influencing on the passengers' position could lead to give some clues to encourage a better dispersal.

Coming to the modelling, dependency between the stations during a train run could be included into the model. It might better estimate the alighting and boarding time. Finally, the modelled developed here can be useful to Transilien operator, it would be extremely interesting to generalize it to more lines than the one equipped with the adapted rolling stock. Such a tools would give to the lines more clues for the timetable building process to adapt it to the passengers' flow and optimize the resources use.

Bibliography

[1] Nigel G. Harris et al. “The Impact of Urban Rail Boarding and Alighting Factors”. In: *TRB Annual Meeting* (2014).

[2] Nigel G Harris, Flavia de Simone, and Ben Condry. “A Comprehensive Analysis of Passenger Alighting and Boarding Rates”. In: *Urban Rail Transit* 8 (2022), pp. 67–98.

[3] Winnie Daamen. “Modelling Passenger Flows in Public Transport Facilities”. In: *TRAIL Research School* (2004).

[4] Tyh-Ming Lin and Nigel H. M. Wilson. “Dwell Time Relationship for Light Rail Systems”. In: *Transportation research record* 1361 (1992), pp. 287–295.

[5] Sélim Cornet et al. “Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas”. In: *HAL* (2019).

[6] S. Buchmueller, U. Weidmann, and A. Nash. “Development of a dwell time calculation model for timetable planning”. In: *WIT Press* 103.XI (2008), pp. 525–534.

[7] Ir. Paul. B.L. Wiggenraad. “Alighting and boarding times of passengers at Dutch railway stations”. In: *TRAIL Research School* (2001).

[8] Nigel G. Harris. “Train Boarding and Alighting Rates at High Passenger Loads”. In: *Journal of Advanced Transportation* 40.3 (2005), pp. 249–263.

[9] N G Harris and R J Anderson. “An international comparison of urban rail boarding and alighting rates”. In: *Proc. IMechE* 221.F (1997), pp. 521–526.

[10] D. Li et al. “Train Dwell Time Distributions at Short Stop Stations”. In: *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)* (2014).

[11] Ruben A. Kuipers et al. “The passenger’s influence on dwell times at station platforms: a literature review”. In: *Transport Reviews* (2021), DOI: 10.1080/01441647.2021

[12] Dewei Li, Winnie Daamen, and Rob M. P. Goverde. “Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station”. In: *JOURNAL OF ADVANCED TRANSPORTATION* 50 (2016), pp. 877–896.

[13] Zhang Qi, Han Baoming, and Li Dewei. “Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations”. In: *Transportation Research* 16.5 (2008), pp. 635–649.

[14] Boyi Su et al. “An Agent-Based Model for Evaluating the Boarding and Alighting Efficiency of Autonomous Public Transport Vehicles”. In: *LNCS* 11536 (2019), pp. 534–547.

[15] Winnie Daamen, Yu-chen Lee, and Paul Wiggenraad. “Boarding and Alighting Experiments, Overview of Setup and Performance and Some Preliminary Results”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2042 (2008), pp. 71–81.

[16] P. McCullagh and J.A. Nelder. *Generalized Linear Models. Second Edition*. Chapman and Hall, 1989.

List of Figures

List of Tables

A Cluster definition

Table A.1: Key indicators to choose the definition of a cluster computed on the whole dataset from line N between September and October 2021

Time between 2 passengers of the same cluster	Number of clusters	Number of different stops	Number of stops with clusters for 16 doors	Number of kept measures
<1s	8347	1640	3	19937
<1.5s	13761	2420	29	37916
<2s	14674	2532	35	42536
<2.5s	16039	2746	43	48476
<3s	17084	2832	46	51777

Time between 2 passengers of the same cluster	Quantiles of number of passengers into the cluster						Number of different doors	Percentage of doors with C clusters			
	25%	50%	75%	95%	99%	max		C = 1	C = 2	C = 3	C = 4
<1s	6	11	16	31	49	125	7119	84,53%	13,84%	1,47%	0,15%
<1.5s	4	9	15	33	51	125	12086	87,72%	10,94%	1,17%	0,14%
<2s	4	9	15	34	52	125	13253	90,27%	8,88%	0,72%	0,09%
<2.5s	4	8	15	34	53	138	14826	92,42%	7,03%	0,51%	0,05%
<3s	4	8	15	34	53	138	15902	93,08%	6,46%	0,42%	0,03%

Technical
University of
Denmark

Brovej, Building 116
2800 Kgs. Lyngby
Tlf. 4525 1700

www.man.dtu.dk